6-1-2017

OPTIMAL, TWO STAGE, ADAPTIVE ENRICHMENT DESIGNS FOR RANDOMIZED TRIALS USING SPARSE LINEAR PROGRAMMING

Michael Rosenblum
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, mrosen@jhu.eduMi

Xingyuan Fang
Department of Operations Research and Financial Engineering, Princeton University

Han Liu
Department of Operations Research and Financial Engineering, Princeton University

Optimal, Two Stage, Adaptive Enrichment Designs for Randomized Trials, using Sparse Linear Programming

Michael Rosenblum[*], Ethan X. Fang[†], and Han Liu[‡]

June 1, 2017

Abstract

Adaptive enrichment designs involve preplanned rules for modifying enrollment criteria based on accruing data in a randomized trial. We focus on designs where the overall population is partitioned into two predefined subpopulations, e.g., based on a biomarker or risk score measured at baseline. The goal is to learn which populations benefit from an experimental treatment. Two critical components of adaptive enrichment designs are the decision rule for modifying enrollment, and the multiple testing procedure. We provide a general method for simultaneously optimizing these components for two stage, adaptive enrichment designs. We minimize the expected sample size under constraints on power and the familywise Type I error rate. It is computationally infeasible to directly solve this optimization problem due to its nonconvexity. The key to our approach is a novel, discrete representation of this optimization problem as a sparse linear program, which is large but computationally feasible to solve using modern optimization techniques. Applications of our approach produce new, approximately optimal designs.

[*]mrosen@jhu.edu, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

[†]xxf13@psu.edu, Department of Statistics and Department of Industrial and Manufacturing Engineering, Pennsylvania State University

[‡]hanliu@princeton.edu, Department of Operations Research and Financial Engineering, Princeton University

1 Introduction

Consider the problem of planning a randomized trial of a new treatment versus control, when the population of interest is partitioned into two subpopulations. The subpopulations could be defined in terms of a biomarker or risk score measured at baseline. Our goal is to test the null hypotheses of no average treatment benefit for each subpopulation and for the combined population. Standard randomized trial designs may have low power to detect a treatment effect if the treatment only benefits one subpopulation.

Adaptive enrichment designs have been proposed for this problem, e.g., Follmann (1997), Russek-Cohen and Simon (1997), Jennison and Turnbull (2007), Wang et al. (2007), Wang et al. (2009), Brannath et al. (2009), Rosenblum and van der Laan (2011), Jenkins et al. (2011), Friede et al. (2012), Boessen et al. (2013), Stallard et al. (2014), Graf et al. (2015), Krisam and Kieser (2015), Götte et al. (2015). This related work either does not involve optimization, or optimizes over designs that depend on a few, real-valued parameters. In contrast, we simultaneously optimize over a very large class of designs and multiple testing procedures, described below. Wason and Jaki (2012) and Hampson and Jennison (2015) consider the related problem of optimizing adaptive designs involving multiple treatments for a single population. Their approaches do not apply to our problem, as we show in Section 6.

A two-stage, adaptive enrichment design consists of a decision rule for potentially modifying enrollment at the end of stage 1, and a multiple testing procedure at the end of stage 2. The decision rule is a function from the stage 1 data to a finite set of possible enrollment choices for stage 2. The multiple testing procedure is a function from the stage 1 and 2 data to the set of null hypotheses that are rejected. We put no restrictions on these functions except that they are measurable and discretized, as described below. The resulting class of possible designs is quite large. Our goal is to construct new adaptive enrichment designs that minimize expected sample size under constraints on power and Type I error, over this class of possible designs. This is a nonconvex optimization problem that is computationally infeasible to solve directly.

Our approach is to approximate the original optimization problem by a sparse linear program. This idea was applied to standard designs, which do not have an enrollment modification rule, by Rosenblum et al. (2014); they optimized power over different multiple testing procedures. We tackle the substantially more challenging problem of simultaneously

optimizing the decision rule and multiple testing procedure in two stage, adaptive enrichment designs. The added difficulty of the latter problem is twofold: it is harder to construct a representation as a sparse, linear program, and the resulting linear program is harder to solve computationally. Another difference between Rosenblum et al. (2014) and our problem is that we consider not only power, but also expected sample size. In practice, both are important in trial planning.

We show that our designs control the familywise Type I error rate in the strong sense defined by Hochberg and Tamhane (1987, pg. 3). Control of the familywise Type I error rate in confirmatory trials is generally required by regulators such as the U.S. Food and Drug Administration and the European Medicines Agency (FDA and EMEA, 1998).

As in all of the above related work, we require the subpopulations to be defined before the trial starts. Such a definition could be based on prior trial data and scientific understanding of the disease being treated. Designs exist that try to solve the more challenging problem of defining a subpopulation based on accruing data and then testing for a treatment effect in that subpopulation, e.g., Freidlin and Simon (2005); Lai et al. (2014).

In our examples, the optimized designs substantially improve power compared to standard designs and some existing adaptive designs. A limitation of our approach is that it becomes computationally difficult or infeasible for more than 2 stages or subpopulations, as described in Section 8. Also, our approach requires that each participant's outcome is measured relatively soon after her/his enrollment. We focus on designs where the only allowed adaptation is to modify enrollment for stage 2. We do not consider other types of adaptation such as modifying randomization probabilities, or modifying the treatment for each individual in response to his/her outcomes over time.

In Section 2, we define our optimization problem. We prove that it suffices to consider decision rules and multiple testing procedures that depend on the data only through minimal sufficient statistics, in Section 3. In Section 4, we discretize the optimization problem and then tranform it into a sparse, linear program. The discretization involves partitioning the sample space of the sufficient statistics into small rectangles; we then restrict to decision rules and multiple testing procedures that depend on the data only through the rectangle that contains the sufficient statistics. The sparse linear program in solved in two examples, in Section 5. In Section 6, we describe the structure of the sparse linear program. We present our method for verifying strong control of the familywise Type I error rate, in Section 7. Limitations and open problems are discussed in Section 8.

2 Problem Definition

2.1 Null Hypotheses

We assume that the population is partitioned into two subpopulations, defined in terms of variables measured before randomization. Let p_s denote the proportion of the population in subpopulation $s \in \{1, 2\}$, which we assume to be known; $p_1 + p_2 = 1$. Each enrolled participant is assigned to treatment ($a = 1$) or control ($a = 0$) with probability 1/2.

For each subpopulation $s \in \{1, 2\}$ and stage $k \in \{1, 2\}$, we assume exactly half the participants are assigned to each study arm $a \in \{0, 1\}$. This can be approximately achieved by using block randomization stratified by subpopulation. For each participant i from subpopulation $s \in \{1, 2\}$ enrolled in stage $k \in \{1, 2\}$, denote her/his study arm assignment by $A_{s,i}^{(k)} \in \{0, 1\}$ and outcome by $Y_{s,i}^{(k)} \in \mathbb{R}$. Throughout, the *sub*population indicator s is in the *sub*script, and the stage number k is in the superscript.

For clarity, we focus on normally distributed outcomes with known variances. Under regularity conditions, our results can be extended to different outcome distributions as long as one uses asymptotically linear statistics, e.g., the difference between sample means or the estimated coefficient in a proportional hazards model. We assume that conditioned on study arm $A_{s,i}^{(k)} = a$, the outcome $Y_{s,i}^{(k)} \sim N(\mu_{sa}, \sigma_{sa}^2)$ and is independent of the data from all previously enrolled participants. Let $\boldsymbol{\sigma}^2 = (\sigma_{10}^2, \sigma_{11}^2, \sigma_{20}^2, \sigma_{21}^2)$, which we assume is known; we also assume $\sigma_{s0}^2 = \sigma_{s1}^2$ for each $s \in \{1, 2\}$. Let $X^{(k)}$ denote all the data from stage k, and let $X = X^{(1)} \cup X^{(2)}$ denote the cumulative data at the end of stage 2. Let $\mathcal{X}^{(k)}$ and $\mathcal{X}$ denote the sample spaces of $X^{(k)}$ and X, respectively. We assume that each participant's outcome is observed relatively soon after enrollment, so that all stage 1 outcome data are available at the interim analysis.

Denote the average treatment effect for each subpopulation $s \in \{1, 2\}$ by $\Delta_s = \mu_{s1} - \mu_{s0}$, and for the combined population by $\Delta_C = p_1\Delta_1 + p_2\Delta_2$. Let $\boldsymbol{\Delta} = (\Delta_1, \Delta_2)$. We do not assume any relationships among the subpopulation-specific treatment effects Δ_1, Δ_2; their magnitudes and signs can differ arbitrarily. For simplicity, we assume $\mu_{s1} = \Delta_s/2$ and $\mu_{s0} = -\Delta_s/2$ for each subpopulation $s \in \{1, 2\}$, so that the only unknown parameters in our problem are the subpopulation-specific treatment effects (Δ_1, Δ_2).

Define H_{01}, H_{02}, H_{0C}, to be the null hypotheses of no average treatment benefit in

subpopulation 1, subpopulation 2, and the combined population, respectively, i.e.,

$$H_{01} : \Delta_1 \leq 0; \qquad H_{02} : \Delta_2 \leq 0; \qquad H_{0C} : \Delta_C \leq 0.$$

Let $\mathcal{H} = \{H_{01}, H_{02}, H_{0C}\}$, and let $\mathcal{S}$ denote the power set of $\mathcal{H}$. For any $\boldsymbol{\Delta} \in \mathbb{R}^2$, define $\mathcal{H}_{\text{TRUE}}(\boldsymbol{\Delta})$ to be the set of true null hypotheses at $\boldsymbol{\Delta}$. For each $s \in \{1, 2\}$, this set contains H_{0s} if $\Delta_s \leq 0$; it contains H_{0C} if $p_1\Delta_1 + p_2\Delta_2 \leq 0$.

2.2 Two Stage, Adaptive Enrichment Designs

In stage 1, $n_s^{(1)}$ participants are enrolled from each subpopulation s. At the interim analysis following stage 1, a decision rule D determines the number of participants to enroll from each subpopulation in stage 2, based on the stage 1 data. At the end of stage 2, a multiple testing procedure M determines which subset (if any) of the null hypotheses to reject, based on the data from stages 1 and 2. A two stage, adaptive enrichment design is defined by the following quantities, which must be specified before the trial starts:

i. The stage 1 sample sizes $n_1^{(1)}, n_2^{(1)}$ for subpopulations 1 and 2, respectively.

ii. The number $K < \infty$ of possible stage 2 decisions, and for each decision $d \in \mathcal{E} = \{1, \ldots, K\}$ the stage 2 sample sizes $n_1^{(2),d}, n_2^{(2),d}$ for subpopulations 1 and 2, respectively.

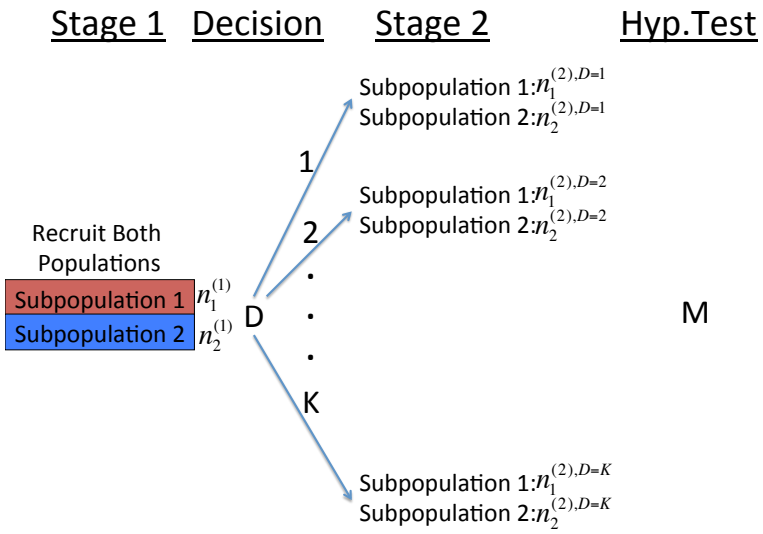
iii. A decision rule D mapping the stage 1 data $X^{(1)}$ to an enrollment decision in $\mathcal{E}$.

iv. A multiple testing procedure M mapping the stage 1 and 2 data X to a (possibly empty) subset of null hypotheses $H \subseteq \mathcal{H}$ to reject.

In our examples, we set the stage 1 sample sizes $n_1^{(1)}, n_2^{(1)}$ proportional to the subpopulation sizes p_1, p_2; however, our general method does not require this.

Define an adaptive design template, denoted by $\mathbf{n}$, to be the quantities in (i)-(ii), i.e., the set of possible decisions and corresponding sample sizes $\mathbf{n} = (\mathcal{E}, n_1^{(1)}, n_2^{(1)}, \{n_1^{(2),d}, n_2^{(2),d}\}_{d\in\mathcal{E}})$. A generic adaptive design template is depicted in Figure 1a. A specific example for $p_1 = 1/2$ is given in Figure 1b, where for a given $n > 0$, the stage 1 sample sizes satisfy $n_1^{(1)} = n_2^{(1)} = n/4$, and there are four choices for stage 2 enrollment: $D = 1$: stop the trial, i.e., $n_1^{(2),1} = n_2^{(2),1} = 0$; $D = 2$: enroll exactly as in stage 1, i.e., $n_1^{(2),2} = n_2^{(2),2} = n/4$; $D = 3$: only enroll from subpopulation 1, i.e., $n_1^{(2),3} = 3n/4, n_2^{(2),3} = 0$; $D = 4$: only enroll from subpopulation 2, i.e., $n_1^{(2),4} = 0, n_2^{(2),4} = 3n/4$. This adaptive design template, denoted $\mathbf{n}^{(1b)}$, is used in Section 5.

a.



b.

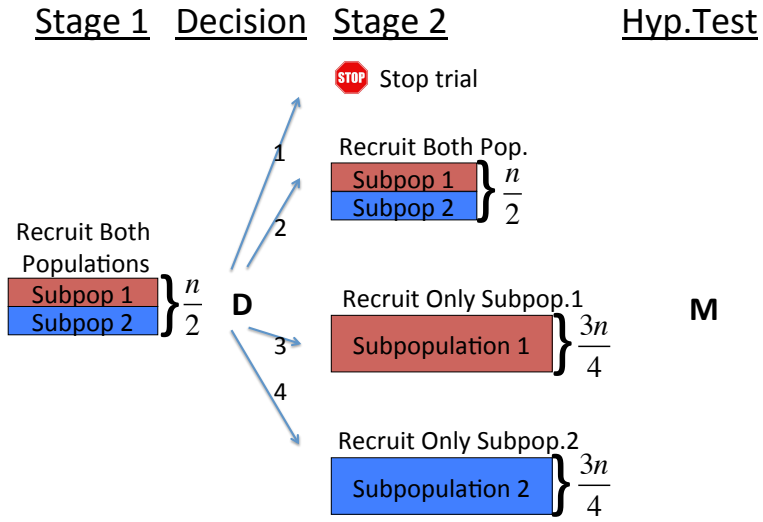


Figure 1: (a) Generic adaptive enrichment design template; (b) Example of an adaptive enrichment design template, denoted $\mathbf{n}^{(1b)}$, with four possible stage 2 decisions, parametrized by n.

It allows enrichment of subpopulation 1 ($D = 3$) or subpopulation 2 ($D = 4$), in which case the total enrolled from the enriched subpopulation is n ($n/4$ from stage 1 plus $3n/4$ from stage 2). This choice of sample sizes was motivated by the problems in Section 5.

For a given adaptive design template $\mathbf{n}$ (which we consider fixed), we aim to simultaneously optimize the decision rule D and multiple testing procedure M, as described in Section 2.3. For reasons given in Section 4.4, we consider randomized decision rules and multiple testing procedures, i.e., we allow D and M to additionally take as input a uniformly distributed random variable U on $[0, 1]$ that is independent of the data. For conciseness, we sometimes suppress dependence on U in our notation, and omit the word "randomized" when referring to "randomized decision rules" and "randomized multiple testing procedures".

Define the class of decision rules $\mathcal{D}^*$ to be all measurable functions from the stage 1 sample space $\mathcal{X}^{(1)}$ and the support $[0, 1]$ of U to $\mathcal{E}$. Define the class of multiple testing procedures $\mathcal{M}^*$ to be all measurable functions from $\mathcal{X} \times \mathcal{E}$ and the support $[0, 1]$ of U to the power set $\mathcal{S}$ of null hypotheses. The importance of measurability in adaptive designs is discussed by Liu et al. (2002, Section 5).

For a given adaptive design template $\mathbf{n}$, an adaptive enrichment design is defined as a pair $(D, M) \in \mathcal{D}^* \times \mathcal{M}^*$. For a given $(p_1, \mathbf{n}, D, M, \boldsymbol{\Delta}, \boldsymbol{\sigma})$, let $P_{\boldsymbol{\Delta}}$ denote the corresponding distribution of the data X and let $E_{\boldsymbol{\Delta}}$ denote expectation with respect to this distribution (where we suppress dependence on the other parameters $p_1, \mathbf{n}, D, M, \boldsymbol{\sigma}$ for clarity).

2.3 General Optimization Problem

Our optimization problem is represented using the decision theory framework. We define the quantity to be minimized, called the objective function, in terms of a loss function L and a distribution Λ on the subpopulation treatment effects $\boldsymbol{\Delta} \in \mathbb{R}^2$. This allows a variety of choices for what to optimize, including power, expected sample size, and other possibilities as described below. The loss function L can be any bounded, integrable function of the treatment effect $\boldsymbol{\Delta}$, the enrollment decision D, and the set of hypotheses rejected M. For a given loss function L, the risk at treatment effect vector $\boldsymbol{\Delta} \in \mathbb{R}^2$ is defined as $R_L(\boldsymbol{\Delta}) = E_{\boldsymbol{\Delta}} L[M\{X, D(X^{(1)})\}, D(X^{(1)}), \boldsymbol{\Delta}]$. The objective function is the Bayes risk $\int R_L(\boldsymbol{\Delta}) d\Lambda(\boldsymbol{\Delta})$.

By selecting an appropriate loss function L, the objective function can be made to represent, e.g., power, expected sample size, expected number assigned to an ineffective treatment

(or weighted combinations of these). For example, the loss function could be set as the total number of enrolled participants (sample size) $L^{\mathrm{SS}} = n_1^{(1)} + n_2^{(1)} + n_1^{(2),D(X^{(1)})} + n_2^{(2),D(X^{(1)})}$; the corresponding risk at $\boldsymbol{\Delta} \in \mathbb{R}^2$ is the expected sample size under treatment effect vector $\boldsymbol{\Delta}$.

Alternatively, one can encode one minus power for different null hypotheses using the following loss functions (where $1[B]$ is the indicator function taking value 1 if B is true and 0 otherwise):

$$L^{(s)} = 1[H_{0s} \notin M\{X, D(X^{(1)})\}, \Delta_s \geq \Delta^{\min}], \text{ for each } s \in \{1, 2\}; \tag{1}$$

$$L^{(C)} = 1[H_{0C} \notin M\{X, D(X^{(1)})\}, \Delta_1 \geq \Delta^{\min}, \Delta_2 \geq \Delta^{\min}]; \tag{2}$$

where $\Delta^{\min}$ represents the minimum, clinically meaningful treatment effect, which is user-specified. The loss function $L^{(s)}$ penalizes 1 unit for failing to reject H_{0s} when the true treatment effect for subpopulation s is at least the clinically meaningful level $\Delta^{\min}$; the loss function $L^{(C)}$ penalizes 1 unit for failing to reject H_{0C} when both subpopulation treatment effects are at least $\Delta^{\min}$. For each subpopulation $s \in \{1, 2\}$, if its treatment effect Δ_s is at least $\Delta^{\min}$, then the risk $R_{L^{(s)}}(\boldsymbol{\Delta})$ equals one minus the power to reject H_{0s}. Similarly, if both treatment effects Δ_1, Δ_2 are at least $\Delta^{\min}$, the risk $R_{L^{(C)}}(\boldsymbol{\Delta})$ equals one minus the power to reject H_{0C}. In both cases, minimizing risk corresponds to maximizing power.

We aim to minimize the Bayes risk, i.e., the risk integrated with respect to a distribution Λ on the treatment effect vector $\boldsymbol{\Delta} \in \mathbb{R}^2$. For example, we could let Λ denote a weighted sum of the four point masses in the set $Q = \{(0, 0), (\Delta^{\min}, 0), (0, \Delta^{\min}), (\Delta^{\min}, \Delta^{\min})\}$, which correspond to no treatment effect, only subpopulation 1 benefiting at the minimum level, only subpopulation 2 benefiting at the minimum level, and both subpopulations benefiting at the minimum level, respectively. Let Λ^{pm} denote this distribution with weight 1/4 on each point mass. Then the Bayes risk corresponding to the pair $(L, \Lambda) = (L^{\mathrm{SS}}, \Lambda^{\mathrm{pm}})$ is the expected sample size under $\boldsymbol{\Delta}$, averaged over the four scenarios $\boldsymbol{\Delta} \in Q$. As another example, let Λ^{mix} denote the mixture of four bivariate normal distributions with one centered at each point in Q, and each having covariance matrix $\sigma_\Lambda^2 \mathbf{I_2}$ for $\mathbf{I_2}$ the 2×2 identity matrix and $\sigma_\Lambda^2 > 0$.

Our optimization problem has two types of constraints. The first are familywise Type I error constraints. The second are encoded similarly as the objective function; these constraints are represented by triples $(L_j, \Lambda_j, \beta_j)$, for $j = 1, \ldots, J$, of loss function L_j, distribution Λ_j on $\boldsymbol{\Delta} \in \mathbb{R}^2$, and threshold $\beta_j \in \mathbb{R}$ defined below.

Constrained Bayes Optimization Problem: For a given vector of inputs:

$$\left(p_1, \mathbf{n}, \boldsymbol{\sigma}^2, \alpha, \{(L_j, \Lambda_j, \beta_j) : j = 0, \ldots, J\}\right), \tag{3}$$

find the adaptive enrichment design $(D, M) \in (\mathcal{D}^* \times \mathcal{M}^*)$ minimizing Bayes risk:

$$\int E_{\boldsymbol{\Delta}} L_0[M\{X, D(X^{(1)})\}, D(X^{(1)}), \boldsymbol{\Delta}] d\Lambda_0(\boldsymbol{\Delta}), \tag{4}$$

under the familywise Type I error constraints:

$$P_{\boldsymbol{\Delta}} \{M \text{ rejects any null hypotheses in } \mathcal{H}_{\text{TRUE}}(\boldsymbol{\Delta})\} \leq \alpha, \text{ for any } \boldsymbol{\Delta} \in \mathbb{R}^2, \tag{5}$$

and additional constraints: for each $j \in \{1, \ldots, J\}$:

$$\int E_{\boldsymbol{\Delta}} L_j[M\{X, D(X^{(1)})\}, D(X^{(1)}), \boldsymbol{\Delta}] d\Lambda_j(\boldsymbol{\Delta}) \leq \beta_j. \tag{6}$$

The constraints (5) represent strong control of the familywise Type I error rate, i.e., for any pair of treatment effects Δ_1, Δ_2, the probability of rejecting one or more true null hypotheses is at most α. This is more stringent than Type I error control at the global null hypothesis $(\Delta_1, \Delta_2) = (0, 0)$. We say an adaptive enrichment design $(D, M) \in (\mathcal{D}^* \times \mathcal{M}^*)$ is feasible if it satisfies all of the constraints (5) and (6).

Consider the special case where $J = 0$, i.e., there are no additional constraints (6). Then the constrained Bayes optimization problem is to minimize the Bayes risk subject to strong control on the familywise Type I error rate at level α. The additional constraints (6) allow one to define a broader set of problems, such as optimizing expected sample size subject to power and Type I error constraints, as described next.

The only role of the distributions $\{\Lambda_j, j = 0, 1, \ldots, J\}$ on $\boldsymbol{\Delta}$ is in defining the objective function (4) and constraints (6). The familywise Type I error constraints (5) are over all $\boldsymbol{\Delta} \in \mathbb{R}^2$ and do not involve these distributions. We refer to the distributions $\{\Lambda_j, j = 0, 1, \ldots, J\}$ as priors, with the understanding that our optimization problem uses the decision theory framework. Our general approach can also be used to solve a minimax version of the above optimization problem where the outer integral in the objective function (4) is replaced by the maximum over $\boldsymbol{\Delta}$ in a finite set $\mathcal{P} \subseteq \mathbb{R}^2$.

2.4 Example Optimization Problems

We solve the following two example optimization problems in Section 5, for values of $p_1, \mathbf{n}$, $\boldsymbol{\sigma}^2$, α, β, $\Delta^{\min}$ defined there:

Example 2.1. *Consider the problem of minimizing expected sample size averaged over the four point masses in Q, under the Type I error constraints (5) and the following power constraints for given Type II error $\beta > 0$:*

P1. At $(\Delta_1, \Delta_2) = (\Delta^{\min}, 0)$, the power to reject H_{01} is at least $1 - \beta$.

P2. At $(\Delta_1, \Delta_2) = (0, \Delta^{\min})$, the power to reject H_{02} is at least $1 - \beta$.

P3. At $(\Delta_1, \Delta_2) = (\Delta^{\min}, \Delta^{\min})$, the power to reject H_{0C} is at least $1 - \beta$.

This problem can be represented by setting $(L_0, \Lambda_0) = (L^{SS}, \Lambda^{pm})$ and the following $J = 3$ additional constraints of the form $(L_j, \Lambda_j, \beta_j)$:

$$(L^{(1)}, \boldsymbol{\delta}_{(\Delta^{\min},0)}, \beta); \qquad (L^{(2)}, \boldsymbol{\delta}_{(0,\Delta^{\min})}, \beta); \qquad (L^{(C)}, \boldsymbol{\delta}_{(\Delta^{\min},\Delta^{\min})}, \beta),$$

where $\boldsymbol{\delta}_{(x,y)}$ denotes a point mass at $\boldsymbol{\Delta} = (x, y)$.

Example 2.2. *We modify the above example by replacing the four point mass prior Λ^{pm} by the the mixture of Gaussian prior Λ^{mix} (defined above). We set the variance of each component of the Gaussian mixture prior to be $\sigma^2_\Lambda = \Delta^2_{\min}$. This modified problem is encoded as above, except setting $\Lambda_0 = \Lambda^{mix}$.*

3 Reducing Problem Complexity through Sufficient Statistics

We show that it suffices to consider decision rules D and multiple testing procedures M that depend only on sufficient statistics defined below. This dramatically reduces the problem complexity from having to search over arbitrarily complex functions of the data X, to the easier (but still very challenging) problem of searching over functions of the 2-dimensional sufficient statistics at each stage. Let $N_s^{(k)}$ denote the number enrolled from subpopulation $s \in \{1, 2\}$ during stage $k \in \{1, 2\}$. The stage 1 sample sizes are set in advance, while the stage 2 sample sizes are functions of the stage 1 data; specifically, $N_s^{(1)} = n_s^{(1)}$ and $N_s^{(2)} = n_s^{(2),D(X^{(1)})}$ for each $s \in \{1, 2\}$.

For each subpopulation $s \in \{1, 2\}$ and stage $k \in \{1, 2\}$, define the z-statistic

$$Z_s^{(k)} = \left\{ \frac{\sum_{i=1}^{N_s^{(k)}} Y_{s,i}^{(k)} A_{s,i}^{(k)}}{\sum_{i=1}^{N_s^{(k)}} A_{s,i}^{(k)}} - \frac{\sum_{i=1}^{N_s^{(k)}} Y_{s,i}^{(k)} (1 - A_{s,i}^{(k)})}{\sum_{i=1}^{N_s^{(k)}} (1 - A_{s,i}^{(k)})} \right\} \left\{ \frac{\sigma_{s1}^2 + \sigma_{s0}^2}{N_s^{(k)}/2} \right\}^{-1/2}, \tag{7}$$

where the quantity inside curly braces on the right is the variance of the difference between sample means on the left. Define the final (cumulative) z-statistic based on pooling all stage 1 and 2 data for subpopulation s by

$$Z_s^{(F)} = \left\{ \frac{\sum_{k=1}^{2} \sum_{i=1}^{N_s^{(k)}} Y_{s,i}^{(k)} A_{s,i}^{(k)}}{\sum_{k=1}^{2} \sum_{i=1}^{N_s^{(k)}} A_{s,i}^{(k)}} - \frac{\sum_{k=1}^{2} \sum_{i=1}^{N_s^{(k)}} Y_{s,i}^{(k)} (1 - A_{s,i}^{(k)})}{\sum_{k=1}^{2} \sum_{i=1}^{N_s^{(k)}} (1 - A_{s,i}^{(k)})} \right\} \left\{ \frac{\sigma_{s1}^2 + \sigma_{s0}^2}{(N_s^{(1)} + N_s^{(2)})/2} \right\}^{-1/2}. \tag{8}$$

Let $\mathbf{Z}^{(k)} = (Z_1^{(k)}, Z_2^{(k)})$ for each stage $k \in \{1, 2\}$, and $\mathbf{Z}^{(F)} = (Z_1^{(F)}, Z_2^{(F)})$. The joint distribution of these random vectors is given in Section A of the Supplementary Materials. The first stage z-statistics $\mathbf{Z}^{(1)}$ are bivariate normal, as are the second stage statistics $\mathbf{Z}^{(2)}$ conditional on the decision D for stage 2 enrollment; the final z-statistic $Z_s^{(F)}$ for subpopulation s is a weighted combination of the corresponding first and second stage statistics, with each subpopulation s participant contributing equal information. We prove the following in Section E of the Supplementary Materials:

Theorem 3.1. *If the constrained Bayes optimization problem in Section 2.3 over $\mathcal{D}^* \times \mathcal{M}^*$ is feasible, then there exists an optimal solution (D, M) such that D depends on the data only through $\mathbf{Z}^{(1)}$, and M depends on the data only through $\mathbf{Z}^{(F)}$ and the decision D.*

The proof involves showing that $\mathbf{Z}^{(1)}$ is a minimal sufficient statistic at the end of stage 1, and $(\mathbf{Z}^{(F)}, D)$ is a minimal sufficient statistic at the end of stage 2.

By the above theorem, it suffices to consider decision rules D that depend on the data only through $\mathbf{Z}^{(1)}$, and multiple testing procedures M that depend on the data only through $\mathbf{Z}^{(F)}$ and D. Let $\mathcal{D}$ denote the class of all measurable functions D from $\mathbb{R}^2 \times [0, 1]$ $\big($representing all possible values of $(\mathbf{Z}^{(1)}, U)\big)$ to the set of stage 2 enrollment decisions $\mathcal{E}$. Let $\mathcal{M}$ denote the class of all measurable functions from $\mathbb{R}^2 \times \mathcal{E} \times [0, 1]$ to $\mathcal{S}$; the domain represents all possible values of $(\mathbf{Z}^{(F)}, D(\mathbf{Z}^{(1)}, U), U)$ and the range represents all possible subsets of null hypotheses. For conciseness, we let $D = D(\mathbf{Z}^{(1)}, U)$ and $M = M\{\mathbf{Z}^{(F)}, D(\mathbf{Z}^{(1)}, U), U\}$ for the rest of the paper. Let $\mathcal{A}^{SUFF} = \{(D, M) : D \in \mathcal{D}, M \in \mathcal{M}\}$, i.e., the class of adaptive enrichment designs that only use the data through $\mathbf{Z}^{(1)}$ at the end of stage 1 and $(\mathbf{Z}^{(F)}, D)$ at the end of stage 2. This is a subclass of the adaptive enrichment designs $\mathcal{D}^* \times \mathcal{M}^*$ defined in Section 2.2.

4 Discretization of Constrained Bayes Optimization Problem and Transformation into Sparse Linear Program

4.1 Overview

Even after simplifying the constrained Bayes optimization problem by using sufficient statistics as in the previous section, the problem is still extremely difficult or impossible to solve directly. This is because the optimization is over the very large class of decision rules $\mathcal{D}$ and multiple testing procedures $\mathcal{M}$, and involves infinitely many familywise Type I error constraints (5).

We propose a novel approach to solving a discretized version of the above problem, involving four steps. We first discretize the decision rule, multiple testing procedure, and familywise Type I error constraints in Section 4.2. The resulting discretized problem can be naturally represented in terms of a finite set of $[0, 1]$-valued variables, as shown in Section 4.3. However, this representation is nonconvex and so is still extremely difficult to solve. Step two, handled in Section 4.4, involves reparametrizing this problem so that it can be represented as a sparse, linear program, a class of problems that is much easier to solve than nonconvex problems. The third step, handled in Section 6, is to apply large-scale optimization methods to solve the sparse, linear program. Lastly, we verify strong control of the familywise Type I error rate, i.e., that (5) holds for all $\mathbf{\Delta} \in \mathbb{R}^2$, using the method in Section 7.

4.2 Definition of Discretized Problem and Class of Designs $\mathcal{A}^{DISC}$

The first of the above steps is to discretize the constrained Bayes optimization problem. This involves partitioning $\mathbb{R}^2$ into a finite set of rectangles. One way to construct such a partition is to start with a box $B = [-b, b] \times [-b, b]$ for a given integer $b > 0$. Let $\tau = (\tau_1, \tau_2)$ be such that b/τ_s is an integer for each $s \in \{1, 2\}$. For each $j, j' \in \mathbb{Z}$, define the rectangle $R_{j,j'} = [j\tau_1, (j + 1)\tau_1) \times [j'\tau_2, (j' + 1)\tau_2)$. Let $\mathcal{R}_{\mathcal{B}}$ denote the set of such rectangles in the bounded region B, i.e., $\{R_{j,j'} : j, j' \in \mathbb{Z}, R_{j,j'} \subset B\}$. Define the following partition of $\mathbb{R}^2$: $\mathcal{R} = \mathcal{R}_{\mathcal{B}} \cup \{\mathbb{R}^2 \setminus B\}$. Though $\mathbb{R}^2 \setminus B$ is not a rectangle, we still refer to $\mathcal{R}$ as a partition of rectangles, with a slight abuse of notation.

Let $\mathcal{R}_{\text{dec}}$ denote a partition of $\mathbb{R}^2$ into rectangles. We restrict to the subclass of decision rules $D \in \mathcal{D}$ that only depend on the data through the rectangle that contains the first stage

z-statistics, i.e., decision rules $D \in \mathcal{D}$ such that for any rectangle $r \in \mathcal{R}_{\text{dec}}$ and $u \in [0,1]$,

$$D(\mathbf{z}^{(1)}, u) = D(\mathbf{z}^{(1)'}, u) \text{ for any } \mathbf{z}^{(1)} \in r, \mathbf{z}^{(1)'} \in r. \tag{9}$$

For each $d \in \mathcal{E}$, let $\mathcal{R}_{\text{mtp},d}$ denote a partition of $\mathbb{R}^2$ into rectangles. We restrict to multiple testing procedures $M \in \mathcal{M}$ that only depend on the data through the enrollment decision D and the rectangle that contains the cumulative statistics $\mathbf{Z}^{(F)}$ at the end of stage 2; that is, we restrict to $M \in \mathcal{M}$ such that for any $d \in \mathcal{E}$, $r \in \mathcal{R}_{\text{mtp},d}$, and $u \in [0,1]$,

$$M(\mathbf{z}^{(F)}, d, u) = M(\mathbf{z}^{(F)'}, d, u) \text{ for any } \mathbf{z}^{(F)} \in r, \mathbf{z}^{(F)'} \in r. \tag{10}$$

It remains to discretize the set $\mathbf{\Delta} \in \mathbb{R}^2$ in the Type I error constraints (5) by selecting a finite subset $G \subseteq \mathbb{R}^2$. Define the boundaries of the null spaces for H_{01}, H_{02}, H_{0C} to be $\{(0, \Delta_2) : \Delta_2 \in \mathbb{R}\}$, $\{(\Delta_1, 0) : \Delta_1 \in \mathbb{R}\}$, $\{(\Delta_1, \Delta_2) \in \mathbb{R}^2 : p_1\Delta_1 + p_2\Delta_2 = 0\}$, respectively. Let G denote a grid of points on the union of these boundaries; an example is given in Section 6. The motivation for this choice of G is based on the conjecture that the active constraints among (5) will be on the boundaries of the null spaces. We will demonstrate that by a careful selection of G, the solutions to the discretized problem in our examples satisfy (5) at all $\mathbf{\Delta} \in \mathbb{R}^2$, if we solve the discretized problem using a value of α slightly lower than the required value in (5).

Define the class $\mathcal{A}^{DISC}$ of discretized adaptive enrichment designs to be all pairs $(D, M) \in \mathcal{A}^{SUFF}$ that satisfy (9) and (10). The discretized version of the constrained Bayes optimization problem from Section 2.3 is defined as that problem optimized over the class of discretized adaptive enrichment designs $\mathcal{A}^{DISC}$ instead of the larger class of adaptive enrichment designs $\mathcal{D}^* \times \mathcal{M}^*$, and involving only Type I error constraints (5) for $\mathbf{\Delta} \in G$. Throughout the remainder of the paper we fix the discretization defined by $\mathcal{R}_{\text{dec}}, \{\mathcal{R}_{\text{mtp},d} : d \in \mathcal{E}\}$ and focus on solving the discretized problem.

4.3 (Nonconvex) Representation of Discretized Problem Using Finitely Many $[0,1]$-Valued Variables

We show that the discretized problem can be equivalently represented in terms of a finite set of variables $x_{rd}, y_{rdr's}$, called the design variables, each taking values in $[0,1]$. This involves first showing that each design in $(D, M) \in \mathcal{A}^{DISC}$ can be represented in terms of these variables. Next, we show that the objective function and constraints (4)-(6) can each

be represented in terms of $x_{rd}, y_{rdr's}$ and probabilities that can be computed based on the problem inputs (3). The benefit of this representation is that optimizing over such designs is equivalent to optimizing over a finite set of variables, which is a key step in making the problem computationally feasible.

Consider an arbitrary design $(D, M) \in \mathcal{A}^{DISC}$. We show how to equivalently represent it in terms of variables $x_{rd}, y_{rdr's}$. For each $r \in \mathcal{R}_{\text{dec}}$ and $d \in \mathcal{E}$, define x_{rd} to be the probability that decision d is made conditioned on $\mathbf{Z}^{(1)} \in r$, i.e.,

$$x_{rd} = P\left\{D(\mathbf{Z}^{(1)}, U) = d | \mathbf{Z}^{(1)} \in r\right\}. \tag{11}$$

For each $r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{E}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S}$, define $y_{rdr's}$ to be the probability that precisely the subset s is rejected conditioned on $\mathbf{Z}^{(1)} \in r,\ D(\mathbf{Z}^{(1)}, U) = d, \mathbf{Z}^{(F)} \in r'$, i.e.,

$$y_{rdr's} = P\left\{M(\mathbf{Z}^{(F)}, d, U) = s | \mathbf{Z}^{(1)} \in r, D(\mathbf{Z}^{(1)}, U) = d, \mathbf{Z}^{(F)} \in r'\right\}. \tag{12}$$

By (9) and (10), the variables $x_{rd}, y_{rdr's}$ do not depend on the unknown population parameter $\boldsymbol{\Delta}$; that is why we omit the subscript $\boldsymbol{\Delta}$ on P in (11) and (12).

Next, we represent (4)-(6), which define the constrained Bayes optimization problem, in terms of $x_{rd}, y_{rdr's}$. Summations over indices r, d, r', s are with respect to the sets $\mathcal{R}_{\text{dec}}, \mathcal{E}, \mathcal{R}_{\text{mtp},d}, \mathcal{S}$, respectively, unless otherwise stated.

The initial step toward representing (4) in terms of $x_{rd}, y_{rdr's}$ is to do this for the probability $P_{\boldsymbol{\Delta}}$ of making enrollment decision $d \in \mathcal{E}$ and then rejecting precisely the subset $s \in \mathcal{S}$:

$$\begin{aligned}
& P_{\boldsymbol{\Delta}}\left[D(\mathbf{Z}^{(1)}, U) = d, M\{\mathbf{Z}^{(F)}, D(\mathbf{Z}^{(1)}, U), U\} = s\right] && (13)\\
= & \sum_{r,r'} P_{\boldsymbol{\Delta}}\left\{M(\mathbf{Z}^{(F)}, d, U) = s, \mathbf{Z}^{(F)} \in r', D(\mathbf{Z}^{(1)}, U) = d, \mathbf{Z}^{(1)} \in r\right\} \\
= & \sum_{r,r'} \left[P_{\boldsymbol{\Delta}}\left\{M(\mathbf{Z}^{(F)}, d, U) = s | \mathbf{Z}^{(F)} \in r', D(\mathbf{Z}^{(1)}, U) = d, \mathbf{Z}^{(1)} \in r\right\} \times \right. \\
& \left. P_{\boldsymbol{\Delta}}\left\{\mathbf{Z}^{(F)} \in r' | D(\mathbf{Z}^{(1)}, U) = d, \mathbf{Z}^{(1)} \in r\right\} P_{\boldsymbol{\Delta}}\left\{D(\mathbf{Z}^{(1)}, U) = d | \mathbf{Z}^{(1)} \in r\right\} P_{\boldsymbol{\Delta}}\left\{\mathbf{Z}^{(1)} \in r\right\}\right] \\
= & \sum_{r,r'} x_{rd} y_{rdr's} p(\boldsymbol{\Delta}, r, d, r'), && (14)
\end{aligned}$$

where (14) follows from (11) and (12), and we let

$$p(\boldsymbol{\Delta}, r, d, r') = P_{\boldsymbol{\Delta}}\left\{\mathbf{Z}^{(F)} \in r' | D(\mathbf{Z}^{(1)}, U) = d, \mathbf{Z}^{(1)} \in r\right\} P_{\boldsymbol{\Delta}}\left\{\mathbf{Z}^{(1)} \in r\right\}. \tag{15}$$

The value of $p(\boldsymbol{\Delta}, r, d, r')$ depends on neither D nor M, which follows from (9). It equals the probability of $(\mathbf{Z}^{(1)}, \mathbf{Z}^{(F)}) \in (r \times r')$ under $P_{\boldsymbol{\Delta}}$ and the (non-adaptive) decision rule that always makes enrollment choice d at the end of stage 1. This probability can be computed using the multivariate normal distribution function with mean and covariance that depend only on $\boldsymbol{\Delta}, d, \mathbf{n}, \boldsymbol{\sigma}^2$ as shown in Section B of the Supplementary Material.

We can express the objective function (4) of the constrained Bayes optimization problem in terms of the variables $x_{rd}, y_{rdr's}$, since for any $\boldsymbol{\Delta} \in \mathbb{R}^2$ the expectation inside the integral in (4) satisfies

$$\begin{aligned} E_{\boldsymbol{\Delta}}\left\{L_0(M, D, \boldsymbol{\Delta})\right\} &= \sum_{s,d} L_0(s, d, \boldsymbol{\Delta}) P_{\boldsymbol{\Delta}}\left\{D(\mathbf{Z}^{(1)}, U) = d, M(\mathbf{Z}^{(F)}, D(\mathbf{Z}^{(1)}, U), U) = s\right\} \\ &= \sum_{s,d} L_0(s, d, \boldsymbol{\Delta}) \sum_{r,r'} x_{rd} y_{rdr's} p(\boldsymbol{\Delta}, r, d, r') \\ &= \sum_{r,d,r',s} x_{rd} y_{rdr's} \left\{L_0(s, d, \boldsymbol{\Delta}) p(\boldsymbol{\Delta}, r, d, r')\right\}, \end{aligned}$$

where the second line follows from the equality of (13) and (14). Since the variables $x_{rd}, y_{rdr's}$ do not depend on $\boldsymbol{\Delta}$, the above display implies that the objective function (4) equals

$$\int E_{\boldsymbol{\Delta}} L_0(M, D, \boldsymbol{\Delta}) d\Lambda_0(\boldsymbol{\Delta}) = \sum_{r,d,r',s} x_{rd} y_{rdr's} \left\{ \int L_0(s, d, \boldsymbol{\Delta}) p(\boldsymbol{\Delta}, r, d, r') d\Lambda_0(\boldsymbol{\Delta}) \right\}.$$

The quantity in curly braces in the above display does not depend on (D, M), and can be computed from the problem inputs (3). The familywise Type I error constraints (5) and additional constraints (6) can similarly be expressed as a function of $x_{rd}, y_{rdr's}$, as shown in Section D of the Supplementary Materials.

Recall that the discretized version of the constrained Bayes optimization problem from Section 2.3 is defined as that problem optimized over the class of discretized adaptive enrichment designs $\mathcal{A}^{DISC}$ instead of the larger class of adaptive enrichment designs $\mathcal{D}^* \times \mathcal{M}^*$, and involving only Type I error constraints (5) for $\boldsymbol{\Delta} \in G$. The above arguments show that this discretized problem can be equivalently represented in terms of an optimization problem over the finite set of variables $\{x_{rd}, y_{rdr's}\}$ as follows:

Discretized Problem:

$$\min \sum_{r,d,r',s} x_{rd} y_{rdr's} \int L_0(s,d,\boldsymbol{\Delta}) p(\boldsymbol{\Delta},r,d,r') d\Lambda_0(\boldsymbol{\Delta}) \tag{16}$$

over the variables $x_{rd}, y_{rdr's}$, under the following constraints:

$$\text{for each } \boldsymbol{\Delta} \in G, \sum_{r,d,r'} \sum_{\{s \in \mathcal{S}: s \cap \mathcal{H}_{TRUE}(\boldsymbol{\Delta}) \neq \emptyset\}} x_{rd} y_{rdr's} p(\boldsymbol{\Delta},r,d,r') \leq \alpha; \tag{17}$$

$$\text{for each } j \in \{1,\ldots,J\}, \sum_{r,d,r',s} x_{rd} y_{rdr's} \int L_j(s,d,\boldsymbol{\Delta}) p(\boldsymbol{\Delta},r,d,r') d\Lambda_j(\boldsymbol{\Delta}) \leq \beta_j; \tag{18}$$

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, \sum_{d \in \mathcal{E}} x_{rd} = 1; \tag{19}$$

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{E}, r' \in \mathcal{R}_{\text{mtp},d}, \sum_{s \in \mathcal{S}} y_{rdr's} = 1; \tag{20}$$

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{E}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S} : x_{rd} \geq 0, y_{rdr's} \geq 0, \tag{21}$$

where the sum $\sum_{r,d,r',s}$ in (16) and (18) is taken over $r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{E}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S}$. The objective function (16) represents (4). The constraints (17) and (18) represent the familywise Type I error constraints (5) and additional constraints (6), respectively. The remaining constraints encode properties of x_{rd} and $y_{rdr's}$ that follow from their definitions (11)-(12) as conditional probabilities; the constraints (19) and (20) follow from the law of total probability, and the constraints (21) follow from probabilities being nonnegative. The values of $p(\boldsymbol{\Delta},r,d,r')$ and the integrals in (16) and (18), can be computed using numerical integration based on the inputs (3) to the constrained Bayes optimization problem from Section 2.3.

4.4 Transformation of (Nonconvex) Discretized Problem into Sparse Linear Program

The discretized problem from Section 4.3 is not linear (and not convex) in the variables $\{x_{rd}, y_{rdr's}\}$. Therefore, this problem is generally computationally intractable to solve, since only ad hoc methods exist for solving nonconvex optimization problems and even if a local minimum is found there is no general way to determine if it is the global minimum. We transform this problem into a sparse, linear program by defining the new variables:

$$v_{rdr's} = x_{rd} y_{rdr's}, \text{ for all } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{E}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S}. \tag{22}$$

For each $d \in \mathcal{E}$, let r'_d denote an arbitrary element in $\mathcal{R}_{\text{mtp},d}$, e.g., the first element under a fixed ordering of $\mathcal{R}_{\text{mtp},d}$. The discretized problem (16)-(21) can be equivalently represented in terms of the variables $v_{rdr's}$ as follows:

Sparse linear program:

$$\min \sum_{r,d,r',s} v_{rdr's} \int L_0(s,d,\boldsymbol{\Delta}) p(\boldsymbol{\Delta},r,d,r') d\Lambda_0(\boldsymbol{\Delta}) \tag{23}$$

under the constraints:

$$\text{for each } \boldsymbol{\Delta} \in G, \sum_{r,d,r'} \sum_{s \in \mathcal{S}: s \cap \mathcal{H}_{TRUE}(\boldsymbol{\Delta}) \neq \emptyset} v_{rdr's} p(\boldsymbol{\Delta},r,d,r') \leq \alpha; \tag{24}$$

$$\text{for each } j \in \{1,\ldots,J\}, \sum_{r,d,r',s} v_{rdr's} \int L_j(s,d,\boldsymbol{\Delta}) p(\boldsymbol{\Delta},r,d,r') d\Lambda_j(\boldsymbol{\Delta}) \leq \beta_j; \tag{25}$$

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, \sum_{d \in \mathcal{E}} \sum_{s \in \mathcal{S}} v_{rdr'_d s} = 1; \tag{26}$$

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{E}, \widetilde{r}' \in \mathcal{R}_{\text{mtp},d}, \sum_{s \in \mathcal{S}} v_{rdr'_d s} = \sum_{s \in \mathcal{S}} v_{rd\widetilde{r}'s}; \tag{27}$$

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{E}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S} : v_{rdr's} \geq 0. \tag{28}$$

We call the above linear program "sparse" since the vast majority of elements of the corresponding constraint matrix are 0, as described in Section 6. The discretized problem's objective function (16) and constraints (17)-(18) are equivalent to (23)-(25), respectively, of the sparse linear program. The discretized problem's constraints (19)-(21) are equivalent to the constraints (26)-(28) of the sparse linear program. This claim and the following theorem are proved in Section F of the Supplementary Material:

Theorem 4.1. *i. (Equivalence of discretized problem and sparse linear program) The optimum value of the sparse linear program equals the optimum value of the discretized problem. ii. (Map from solution of sparse linear program to solution of discretized problem) For any optimal solution* $\{v_{rdr's} : r \in \mathcal{R}_{dec}, d \in \mathcal{E}, r' \in \mathcal{R}_{mtp,d}, s \in \mathcal{S}\}$ *to the sparse linear program, define the variables* $\{x_{rd}, y_{rdr's}\}$ *by the transformation:*

$$x_{rd} = \sum_{s \in \mathcal{S}} v_{rdr's}; \qquad y_{rdr's} = \begin{cases} v_{rdr's}/x_{rd}, & \text{if } x_{rd} > 0 \\ 1/|\mathcal{S}|, & \text{otherwise.} \end{cases} \tag{29}$$

Then $\{x_{rd}, y_{rdr's}\}$ *is a well-defined, feasible, optimal solution to the discretized problem.*

The importance of Theorem 4.1 is that we have derived a computationally feasible approximation (23)-(28) of the original constrained Bayes optimization problem (4)-(6). This

relies on the fact that even very large, sparse linear programs are computationally feasible to solve; we defer presentation of the method used to solve the above sparse linear program to Section 6.

The reason we consider randomized decision rules and multiple testing procedures is that these correspond to the variables $x_{rd}, y_{rdr's}$ (and therefore $v_{rdr's}$) being in the interval $[0, 1]$, rather than being integer-valued. Solving the above sparse linear program (where the variables are $[0, 1]$ valued) is much easier, computationally, than the corresponding problem in which variables are required to be integer-valued. Fortunately, the solutions to the linear programs for our example problems turn out to be mostly integer-valued, as shown below.

5 Solutions to Examples 2.1 and 2.2: Minimizing Expected Sample Size under Power and Type I Error Constraints

5.1 Problem Definition

We solve the optimization problems in Examples 2.1 and 2.2 from Section 2.4 over the class of discretized adaptive enrichment designs $\mathcal{A}^{DISC}$. These problems involve the power constraints (P1)-(P3) defined in Section 2.4. The problem inputs depend on $p_1, \mathbf{n}, \boldsymbol{\sigma}^2, \alpha, \beta, \Delta^{\min}$, which we specify next. Let $p_1 = 1/2$, $\alpha = 0.05$, and let each σ^2_{sa} equal a common value $\sigma^2 > 0$.

We use the adaptive design template $\mathbf{n}^{(1b)}$ defined in Section 2.2 and depicted in Figure 1b; the corresponding sample sizes are functions of n, i.e., the total sample size under $D = 2$ (where both subpopulations are enrolled during stage 2). This adaptive design template allows enrichment of subpopulation 1 ($D = 3$) or subpopulation 2 ($D = 4$), in which case the total enrolled from the enriched subpopulation is n. We next describe the intuition for this choice of sample sizes. The constraints (P1)-(P3) require the same power $(1 - \beta)$ to reject H_{0C} when $\Delta_1 = \Delta_2 = \Delta_{\min}$ as to reject H_{0s} when $\Delta_s = \Delta_{\min}, \Delta_{s'} = 0$, for $s, s' \in \{1, 2\}$, $s \neq s'$. We chose the stage 2 sample sizes in $\mathbf{n}^{(1b)}$ so that the information at the end of stage 2 for Δ_C under $D = 2$ equals the information for Δ_s under $D = 2 + s$, for each $s \in \{1, 2\}$; that is, it's possible to generate the same information for the parameter of interest in each of (P1)-(P3) by a corresponding choice for stage 2 enrollment. Our choice of sample sizes in $\mathbf{n}^{(1b)}$ is not necessarily optimal in any sense; it is an area of future research to solve the

above optimization problems using different stage 2 enrollment choices.

For each of Examples 2.1 and 2.2, the optimal solution to the constrained Bayes optimization problem depends on the inputs $(\sigma^2, \Delta^{\min}, n)$ only through the non-centrality parameter $(n/8)^{1/2}\Delta^{\min}/\sigma$, as proved in Section G of the Supplementary Material. We set $(n/8)^{1/2}\Delta^{\min}/\sigma = 2^{1/2}\Phi^{-1}(1-0.05)$, where Φ is the cumulative distribution function of the standard normal; this is equivalent to setting, for any given $\sigma^2 > 0, \Delta_{\min} > 0$,

$$n = 16\sigma^2\{\Phi^{-1}(1-0.05)\}^2(\Delta^{\min})^{-2}, \tag{30}$$

We use n in (30) as a benchmark sample size, since it is the smallest n such that in a standard (non-adaptive) design enrolling $n/2$ from each subpopulation, the uniformly most powerful test of H_{0C} at level $\alpha = 0.05$ has power 0.95 at the alternative $\boldsymbol{\Delta} = (\Delta^{\min}, \Delta^{\min})$; this power constraint is identical to (P3) at $1-\beta = 0.95$. In contrast, our optimization problem has the more stringent set of power constraints (P1)-(P3), which involve null hypotheses for subpopulations as well as the combined population. We therefore expect our optimization problems to be solvable only if we set the required power in constraints (P1)-(P3) to be lower than $1-\beta = 0.95$. Below, we determine the greatest value of $1-\beta$ for which our optimization problems can be solved; for this and smaller values of $1-\beta$, we determine the minimum expected sample size averaged over the distributions $\Lambda_0 = \Lambda^{\text{pm}}$ and $\Lambda_0 = \Lambda^{\text{mix}}$ for Examples 2.1 and 2.2, respectively.

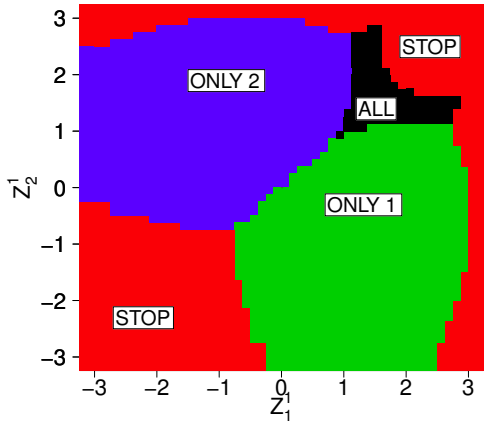
5.2 Optimal Solutions to Examples 2.1 and 2.2 over Adaptive Enrichment Design Class $\mathcal{A}^{DISC}$

We applied the method from Section 4.4 to construct the sparse linear program corresponding to each of the two problems in Section 5.1, for the class of discretized adaptive enrichment designs $\mathcal{A}^{DISC}$. Details of the sparse linear programs, including the fineness of the discretization, are given in Section 6. We separately solved each sparse linear program at every power constraint threshold $\beta \in \{0.01, \ldots, 0.99\}$. The value of $1-\beta$ represents the required power in each constraint (P1)-(P3). Larger values of $1-\beta$ correspond to stricter constraints. Our results show the problems are feasible, i.e., the Type I error and power constraints (P1)-(P3) can be simultaneously satisfied, if and only if $1-\beta < 0.83$.

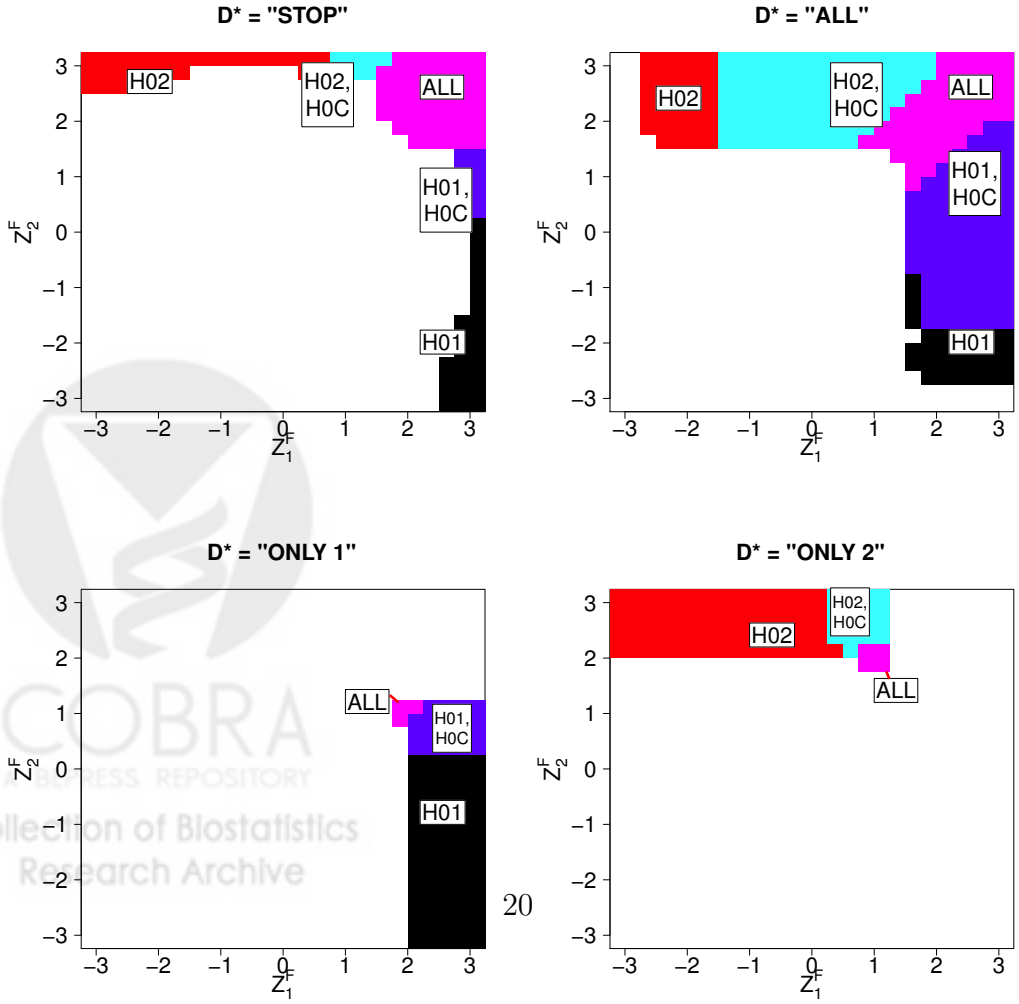
For the case of $1-\beta = 0.82$, Figure 2 depicts the optimal solution $(D^*, M^*) \in \mathcal{A}^{DISC}$ to Example 2.2. We first focus on the top plot (Figure 2a), which represents the decision

Figure 2: Optimal design (D^*, M^*) for discretized problem in Example 2.2 at $1-\beta = 0.82$. Stage 2 enrollment choices STOP, ALL, ONLY 1, ONLY 2 correspond to $D^* = 1, 2, 3, 4$.

a. Decision Rule D^* for Stage 2 Enrollment (z-statistics correspond to $\mathbf{Z}^{(1)}$):



b. Rejection Regions of M^* after each decision D^* (z-stats. correspond to $\mathbf{Z}^{(F)}$):



rule D^*. The different regions correspond to the four possible stage 2 enrollment choices from the adaptive design template $\mathbf{n}^{(1b)}$. The top right and bottom left regions (in red) of this plot correspond to stopping the trial after stage 1 (i.e., $D^* = 1$, marked STOP in the plot). Intuitively, the top right region represents stopping early for efficacy (since, as described below, at least one null hypothesis is rejected whenever the first stage statistic $\mathbf{Z}^{(1)}$ is in this region), while the bottom left region represents stopping early for futility (since no null hypothesis is rejected if $\mathbf{Z}^{(1)}$ is in this region). This pattern, just as those below, naturally emerged as the solution to the optimization problem, and was not imposed a priori. The black region marked ALL represents the choice $D^* = 2$ to continue enrollment from both subpopulations in stage 2. Intuitively, this occurs when the stage 1 z-statistics for each subpopulation both indicate a non-neglible, positive signal that is not strong enough to allow outright rejection of any null hypothesis; this motivates the investment of stage 2 enrollment from both subpopulations, in order to determine which (if any) null hypotheses to reject. The green and blue regions marked ONLY 1 (representing $D^* = 3$), ONLY 2 (representing $D^* = 4$), respectively, represent choosing stage 2 enrollment to be only from the corresponding subpopulation.

The four plots in Figure 2b represent the multiple testing procedure M^* that is used after each of the four enrollment choices, respectively. For each possible value of the enrollment decision D^*, the corresponding plot shows the mapping from the final z-statistics $\mathbf{Z}^{(F)}$ to the set of null hypotheses that are rejected. The top left, top right, bottom left, bottom right plots of M^* correspond to $D^* = 1, 2, 3, 4$, respectively; equivalently, these correspond to $D^* =$ STOP, ALL, ONLY 1, ONLY 2, respectively. Each plot has a white region (corresponding to not rejecting any null hypothesis) and colored regions where the specified null hypotheses are rejected.

The plot of M^* for $D^* =$ STOP has colored regions whose union is approximately identical to the red STOP region in the upper right of Figure 2a. This means that when the first stage z-statistics are in the red STOP region in the upper right of Figure 2a, at least one null hypothesis will be rejected by M^* (since when $D^* =$ STOP, the first stage z-statistics $\mathbf{Z}^{(1)}$ are identical to the final z-statistics $\mathbf{Z}^{(F)}$). Intuitively, this corresponds to stopping early for efficacy. (The match between the aforementioned regions is only approximate since a coarser level of discretization was used for M^* compared to D^*, a choice we made in order to reduce the computational requirements for solving the optimization problem.)

Next, consider the plot of M^* for $D^* =$ ALL. This is qualitatively similar to the plot of

M^* for D^* = STOP, except for two important differences. First, the rejection thresholds are generally lower (i.e., the rejection regions are larger), which makes sense since the final z-statistics after the enrollment decision D^* = ALL incorporate twice as much data as under D^* = STOP and therefore more information is available. This property is analogous to what occurs in standard group sequential designs, e.g., using efficacy stopping boundaries of O'Brien and Fleming (1979), which decrease (on the z-statistic scale) at each stage due to more information being available. The second difference is that there are white areas to the left of the H_{02} region and under the H_{01} region in the plot of M^* for D^* = ALL where one might have expected red and black (i.e., extensions of the these regions), respectively. We conjecture that this is due to the very small probability of $D^*(\mathbf{Z}^{(1)})$ = ALL and M^* being in these white areas; these very small probabilities would not contribute enough to the objective function or constraints to lead to added value in rejecting null hypotheses in these areas, up to the precision used in solving the sparse linear program.

Consider the plot of M^* for D^* = ONLY 1. An interesting feature is that no null hypothesis is rejected when $Z_2^{(F)} > 1.25$. In fact, it is not possible to have both $D^*(\mathbf{Z}^{(1)}) =$ ONLY 1 and $Z_2^{(F)} > 1.25$. This is a consequence of the green ONLY 1 region in Figure 2a being below the horizontal line $Z_2^{(1)} = 1.25$, as explained next. Since the enrollment decision D^* = ONLY 1 occurs precisely when $\mathbf{Z}^{(1)}$ is in the green ONLY 1 region in Figure 2a, and since $Z_2^{(1)} = Z_2^{(F)}$ whenever D^* = ONLY 1 (due to no new subpopulation 2 data being collected in stage 2), it is not possible to have $D^*(\mathbf{Z}^{(1)})$ = ONLY 1 and $Z_2^{(F)} > 1.25$. The plot of M^* for D^* = ONLY 2 is (approximately) a symmetric version of the plot for D^* = ONLY 1.

We did the same as above for Example 2.1, whose optimal solution looks qualitatively similar to Example 2.2.

5.3 Comparison to Standard (Non-adaptive) Designs and Adaptive Designs using p-value Combination Approach

The comparator design classes defined below are based on simpler decision rules and multiple testing procedures than the class $\mathcal{A}^{DISC}$. We compare the performance of the optimal design from each of the simpler classes to the optimal design (computed by sparse linear programming) over $\mathcal{A}^{DISC}$, for each problem in Section 5.1. All of the designs below use the adaptive design template defined in Section 5.1, i.e., $\mathbf{n}^{(1b)}$ with n defined in (30).

Let $D^{STD} \in \mathcal{D}$ denote the decision rule corresponding to the standard (non-adaptive) design that always enrolls from both subpopulations in stage 2, i.e., $D^{STD} = 2$ for all values of the stage 1 statistics. This is equivalent to a design with no interim analysis that enrolls n participants, with $p_s n$ from each subpopulation (where each $p_s = 1/2$ in our case). Define the class of standard (non-adaptive) designs to be $\mathcal{A}^{STD} = \{(D^{STD}, M) : M \in \mathcal{M}\}$.

We next define a class $\mathcal{A}^{COMB}$ of adaptive enrichment designs based on the p-value combination approach of Bauer (1989), Bauer and Köhne (1994), Lehmacher and Wassmer (1999), with the closed testing principle of Marcus et al. (1976); this approach has been used to construct adaptive enrichment designs by, e.g., Bretz et al. (2006); Schmidli et al. (2006); Jennison and Turnbull (2007); Brannath et al. (2009); Jenkins et al. (2011); Boessen et al. (2013). Since it is an open problem how to optimize over the class of all possible designs that can be constructed using this approach, we instead define a low-dimensional, simple class $\mathcal{A}^{COMB}$ of such designs. The full description of $\mathcal{A}^{COMB}$ is given in Section H of the Supplementary Material, but we summarize the key features. The multiple testing procedure, denoted M^{pv}, uses the Dunnett intersection test (Dunnett, 1955; Jennison and Turnbull, 2007), with p-values combined across stages using the weighted inverse normal rule with equal weights for each stage. We slightly modified this approach to incorporate early stopping for efficacy after stage 1 as in, e.g., Jennison and Turnbull (2007), using the equivalent of the boundaries of O'Brien and Fleming (1979) for the stage 1 p-values. We consider a class of decision rules that involve two thresholds t_c and t_i. Define the decision rule $D^{(t_c,t_i)}(\mathbf{Z}^{(1)})$ as follows: If the multiple testing procedure M^{pv} rejects any null hypothesis at the end of stage 1, stop the entire trial; else, if the combined population statistic $(Z_1^{(1)} + Z_2^{(1)})/\sqrt{2} > t_c$, enroll both subpopulations in stage 2; else, enroll in stage 2 from each subpopulation s for which $Z_s^{(1)} > t_i$. Let $\mathcal{A}^{COMB} = \{(D^{(t_c,t_i)}, M^{\text{pv}}) : (t_c, t_i) \in (-3, -2.9, \ldots, 3) \times (-3, -2.9, \ldots, 3)\}$. Each design in $\mathcal{A}^{COMB}$ strongly controls the familywise Type I error rate at level 0.05. An example of the decision rule $D^{(t_c,t_i)}$ is depicted in Figure 3.

We next compare the expected sample size of the optimal design in each of the three classes $\mathcal{A}^{DISC}$, $\mathcal{A}^{COMB}$, $\mathcal{A}^{STD}$, as we vary the power constraint $1 - \beta$. Let ESS denote the value of the objective function (4), which represents the expected sample size with respect to the corresponding prior. For each of Examples 2.1 and 2.2 and each value of $1 - \beta$ in the top row of Table 1, we solved the constrained Bayes optimization problem from Section 5.1 over each class of designs $\mathcal{A}^{DISC}$, $\mathcal{A}^{COMB}$, $\mathcal{A}^{STD}$. For the first and third classes, we used the sparse linear programming method from Section 4.4. For $\mathcal{A}^{COMB}$, we did an exhaustive

Figure 3: Enrollment decision rule $D^{(t_c,t_i)}$ for $(t_c, t_i) = (1.6, 0.6)$, which corresponds to the optimal design over $\mathcal{A}^{COMB}$ for the problem in Example 2.2 under the power constraints (P1)-(P3) at $1 - \beta = 0.74$. The z-statistics in the plot correspond to first stage statistics $\mathbf{Z}^{(1)}$. Stage 2 enrollment choices "STOP", "ALL", "ONLY 1", "ONLY 2" correspond to decisions 1, 2, 3, 4, respectively, from the adaptive design template $\mathbf{n}^{(1b)}$. The red areas in the lower left and upper right corners correspond to stopping the trial at the end of stage 1.

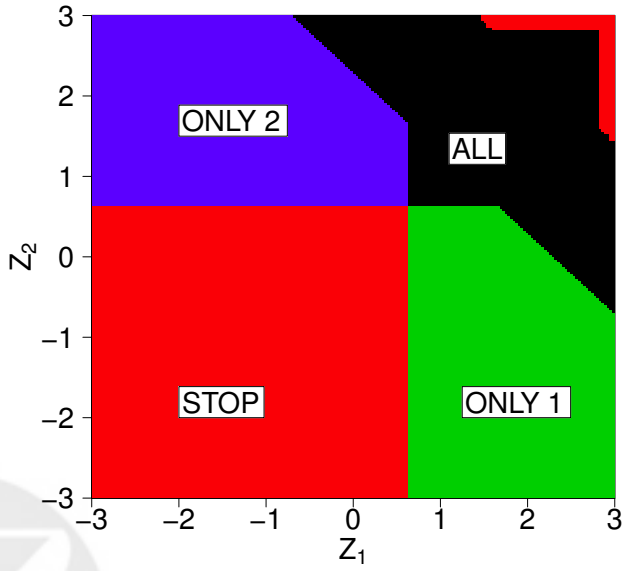


Table 1: Optimal solution values (representing ESS) for the constrained Bayes optimization problems in Examples 2.1 and 2.2, comparing the two classes of adaptive designs $\mathcal{A}^{DISC}$ and $\mathcal{A}^{COMB}$. The symbol $\times$ indicates that no design in the class satisfies the Type I error constraints and power constraints (P1)-(P3) at the required power threshold $1-\beta$.

Power Constraint $(1-\beta)$	58%	62%	66%	70%	74%	78%	82%
Example 2.1							
Min. ESS over $\mathcal{A}^{DISC}$	$0.65n$	$0.69n$	$0.73n$	$0.79n$	$0.84n$	$0.92n$	$1.03n$
Min. ESS over $\mathcal{A}^{COMB}$	$0.86n$	$0.89n$	$0.92n$	$0.97n$	$1.01n$	$\times$	$\times$
Example 2.2							
Min. ESS over $\mathcal{A}^{DISC}$	$0.75n$	$0.75n$	$0.76n$	$0.80n$	$0.84n$	$0.90n$	$0.99n$
Min. ESS over $\mathcal{A}^{COMB}$	$0.89n$	$0.92n$	$0.95n$	$0.98n$	$1.01n$	$\times$	$\times$

search over the set of thresholds (t_c, t_i) given above.

Table 1 gives the optimal solution values (representing ESS) for the constrained Bayes optimization problems in Examples 2.1 and 2.2, comparing the two classes of adaptive designs $\mathcal{A}^{DISC}$ and $\mathcal{A}^{COMB}$. At all values of $1-\beta$ we considered, the minimum value of ESS was substantially lower for the optimal design over $\mathcal{A}^{DISC}$ (computed based on our sparse linear programming approach) compared to the optimal design over $\mathcal{A}^{COMB}$ (which use the p-value combination approach). E.g., in Example 2.1 at power constraint $1-\beta = 0.74$, the value of ESS for the latter is 20% larger than for the former. The optimization problems are infeasible for the p-value combination designs $\mathcal{A}^{COMB}$ at $1-\beta \geq 0.78$, i.e., it is not possible to simultaneously satisfy the Type I error constraints and power constraints (P1)-(P3); in contrast, the problem is feasible for the class $\mathcal{A}^{DISC}$ up to power threshold $1-\beta = 0.82$. Our sparse linear programming method made it possible to compute the optimal design over $\mathcal{A}^{DISC}$. The results in Table 1 should not be taken to mean that our approach outperforms any possible design using the p-value combination approach; we only showed that substantial improvements are possible when comparing to the simple class of adaptive designs $\mathcal{A}^{COMB}$.

We next compare the optimal designs over $\mathcal{A}^{DISC}$ versus the class of standard designs $\mathcal{A}^{STD}$, which have fixed sample size n. The problems in Examples 2.1 and 2.2 are infeasible

for the class $\mathcal{A}^{STD}$ whenever the power constraint $1-\beta > 0.65$, i.e., it is not possible to simultaneously satisfy the Type I error constraints and power constraints (P1)-(P3); in contrast, the problem is feasible for the class $\mathcal{A}^{DISC}$ up to power threshold $1-\beta = 0.82$. We similarly considered the above optimization problems over the class of standard designs with total sample size $5n/4$, i.e., the maximum total sample size that can occur in any adaptive enrichment design in $\mathcal{A}^{DISC}$ (which uses adaptive design template $\mathbf{n}^{(1b)}$); these problems are infeasible for any such standard design when $1-\beta > 0.73$. This shows that there is a substantial advantage in using adaptive enrichment designs versus the standard designs for the problems in Examples 2.1 and 2.2.

6 General Form of Sparse Linear Program

We describe the general form of the sparse linear program from Section 4.4. Let $\mathbf{v}$ denote the column vector consisting of all variables $v_{rdr's}$ for $r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{E}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S}$ in lexicographic order. Define $\psi = \sum_{d\in\mathcal{E}} |\mathcal{R}_{\text{mtp},d}|$ and $w = |\mathcal{R}_{\text{dec}}| \times \psi \times |\mathcal{S}|$, where $|Q|$ denotes the number of elements in the set Q. Then $\mathbf{v}$ has w components. Let $\mathbb{R}_+$ denote the nonnegative reals. The general form of the sparse linear program from Section 4.4 is

$$\min_{\mathbf{v}\in\mathbb{R}_+^w} \quad \mathbf{c^T v} \qquad \text{s.t.} \quad \mathbf{A^{(1)} v} \le \mathbf{a}^{(1)}, \mathbf{A^{(2)} v} = \mathbf{a}^{(2)}; \tag{31}$$

for matrices $\mathbf{A^{(1)}}, \mathbf{A^{(2)}}$ and vectors $\mathbf{c}, \mathbf{a^{(1)}}, \mathbf{a^{(2)}}$. The matrix $\mathbf{A^{(1)}}$ has dimensions $(|G|+J)\times w$ and encodes the Type I error constraints (24) and additional constraints (25), which in Examples 2.1 and 2.2 are power constraints. The matrix $\mathbf{A^{(2)}}$ has dimensions $(1+\psi)|\mathcal{R}_{\text{dec}}|\times w$ and encodes the equality constraints (26) and (27). The matrix $\mathbf{A^{(1)}}$ is dense (most entries are non-zero), while the matrix $\mathbf{A^{(2)}}$ is sparse (most entries are 0) and has the form:

$$\mathbf{A^{(2)}} = \left[\begin{array}{c} |\mathcal{R}_{\text{dec}}| \text{ rows, each with } |\mathcal{E}| \times |\mathcal{S}| \text{ entries with 1 and the rest 0's.} \\ \hline \psi|\mathcal{R}_{\text{dec}}| \text{ rows, each with } |\mathcal{S}| \text{ entries = 1, } |\mathcal{S}| \text{ entries = -1, and the rest 0's.} \end{array}\right].$$

Though the matrix $\mathbf{A^{(2)}}$ is typically much larger than $\mathbf{A^{(1)}}$, the former does not dramatically impact the computational difficulty since it is sparse.

The vector $\mathbf{c}$ represents the objective function (23) and is dense, and the vectors $\mathbf{a^{(1)}}, \mathbf{a^{(2)}}$

have the following forms:

$$\mathbf{a}^{(\mathbf{1})^T} = \begin{pmatrix} |G| \text{ entries} & J \text{ entries} \\ \alpha, \ldots, \alpha, & \beta_1, \ldots, \beta_J \end{pmatrix}, \ \mathbf{a}^{(\mathbf{2})^T} = \begin{pmatrix} |\mathcal{R}_{\text{dec}}| \text{ entries} & \psi|\mathcal{R}_{\text{dec}}| \text{ entries} \\ 1, \ldots 1, & 0, \ldots, 0 \end{pmatrix}.$$

We next describe the discretization and two step approach to solving the optimization problem in Example 2.2 of Section 5; the problem in Example 2.1 had a similar structure and was solved analogously. In step one, a sparse linear program was constructed using the following discretization: the decision region partition $\mathcal{R}_{\text{dec}}$ consisted of length 0.5 squares covering the region $[-3,3] \times [-3,3]$ and unit squares covering $[-6,6] \times [-6,6] \setminus ([-3,3] \times [-3,3])$; for each possible decision d, the multiple testing procedure partition $\mathcal{R}_{\text{mtp},d}$ consisted of unit squares covering $[-6,7] \times [-6,7]$ except that for $d \neq$ STOP we replaced all squares in the lower left quadrant $[-6,0] \times [-6,0]$ by a single large square. (Recall that the units in $\mathcal{R}_{\text{dec}}$ and $\mathcal{R}_{\text{mtp},d}$ are on the z-scale.) We defined G to be the 541 Type I error constraints corresponding to the pairs of non-centrality parameters $(\Delta_1\{n/8\}^{1/2}/\sigma, \Delta_2\{n/8\}^{1/2}/\sigma)$ in the set $\{(x,y) : [x \in \{-9, -8.9, \ldots, 9\}, y = 0]$ or $[x = 0, y \in \{-9, -8.9, \ldots, 9\}]$ or $[x \in \{-9, -8.9, \ldots, 9\}, y = -x]\}$, which are grids along the boundaries of the null spaces for the null hypotheses in $\mathcal{H}$. This resulted in $w \approx 10^6$ variables in $\mathbf{v}$ and $\approx 10^5$ equality constraints in $\mathbf{A}^{(\mathbf{2})}$. We call the solution to the above sparse linear program the "step one" solution.

In step two, we used features of the step one solution to refine the choice of G and the discretization in $\mathcal{R}_{\text{dec}}$ and $\mathcal{R}_{\text{mtp},d}$; we then solved the resulting discretized problem, and iterated this refinement process. The refinement of G involved using the dual of the step one solution to approximately identify the active Type I error constraints; we then augmented G by points $\boldsymbol{\Delta}$ concentrated in small neighborhoods of these active constraints. Further augmentation of G was done as described in Section 7. A finer discretization was obtained by iteratively breaking some rectangles in $\mathcal{R}_{\text{dec}}$ into smaller rectangles; this was done for rectangles near the decision region boundary of the current solution, i.e., rectangles for which an adjacent rectangle made a different decision for stage 2 enrollment. To offset the computational cost of adding such rectangles, we merged rectangles that were far from the boundary. A similar process was applied to refine each $\mathcal{R}_{\text{mtp},d}$. We incorporated additional constraints as described in Section I of the Supplementary Materials to produce an easier to visualize solution, as long as this did not affect the value of the optimization problem.

The resulting solution after several iterations of step two is the design in Figure 2. The solution's active Type I error constraints correspond to the following pairs of non-centrality pa-

rameters $(\Delta_1\{n/8\}^{1/2}/\sigma, \Delta_2\{n/8\}^{1/2}/\sigma) : (0,0), (0,1.91), (0,1.96), (1.91,0),\ (1.96,0)$. That is, the familywise Type I error equals α for each such pair of non-centrality parameters.

Hampson and Jennison (2015) solve a two stage optimization problem related to ours, but involving multiple treatments instead of multiple populations. If applied to our example problems and class of designs, their method would not work since it requires the solution to the optimization problem that only constrains Type I error at the global null hypothesis $\boldsymbol{\Delta} = (0,0)$ to also control the familywise Type I error constraints at all other values of $\boldsymbol{\Delta}$; the approach of Wason and Jaki (2012) has a similar requirement. This requirement does not hold for our example problems, whose optimal solutions have more active Type I error constraints than just the global null hypothesis (e.g., as listed in the previous paragraph).

To solve each sparse linear programming problem, we used the IBM CPLEX solver, version 12.4. To take advantage of the extreme sparse structure of the problem, we used an interior point algorithm. To achieve high precision, we set the tolerance of the relative duality gap to be 10^{-10}. The solution $\mathbf{v}$ to Example 2.2 (Figure 2) had 97% of its components equal to 0 or 1, with the remaining components in $(0,1)$. This means that the corresponding adaptive enrichment design (D^*, M^*) is deterministic (non-randomized) except on a small fraction of rectangles; for the few such rectangles in Figure 2, we set their colors based on rounding the corresponding probabilities.

7 Augmenting G and Verifying Strong Control of the Familywise Type I Error Rate

The sparse linear program in Section 4.4 is an approximation to the original, constrained Bayes optimization problem in Section 2.3. The familywise Type I error constraints (5) in the latter are approximated by the finite set G of constraints (24) in the former. By construction, the solution to the sparse linear program controls the familywise Type I error rate at each $\boldsymbol{\Delta} \in G$.

Below, we describe an iterative procedure that we used to construct the set G for the optimization problem in Section 5.2 corresponding to Example 2.2 and $1 - \beta = 0.82$, i.e., the problem whose optimal solution is the design in Figure 2. We were able to verify the familywise Type I error constraints (5) hold for all $\boldsymbol{\Delta} \in \mathbb{R}^2$ for this design, as described below. If the procedure below terminates, as it did for this example problem, the result is a

verification of strong control of the familywise Type I error rate. The procedure should be viewed as heuristic, since it is unknown whether it will terminate for arbitrary problems.

The procedure involves iteratively augmenting an initial set G and solving the corresponding sparse linear program until we can verify strong control of the familywise Type I error rate. The verification is based on a combination of analytic bounds and numerical checking at values $\boldsymbol{\Delta}$ in a grid $\widetilde{G}$ (defined below) that covers a larger area than G.

Step 1 of the procedure is to solve the sparse linear program with the Type I error threshold α slightly lower than required, i.e., with $\alpha = 0.05 - 10^{-4}$ in the familywise Type I error constraints (24) over $\boldsymbol{\Delta} \in G$. Second, we conduct a search over a grid $\widetilde{G}$ of points $\boldsymbol{\Delta}$ in the square $\widetilde{B} = ([-\widetilde{b}, \widetilde{b}] \times [-\widetilde{b}, \widetilde{b}])$ for $\widetilde{b} > b$, where we use numerical integration to compute the familywise Type I error (5) at each such point; let $\bar{\alpha}$ denote the maximum such value over the grid points $\widetilde{G}$. Let w denote the maximum distance between adjacent grid points in $\widetilde{G}$. In Section C of the Supplementary Material, we use a second-order Taylor expansion of (5) to prove an analytic bound $\epsilon(w)$ on the maximum difference between the familywise Type I error at any point in $\widetilde{B}$ and its closest point in the grid $\widetilde{G}$; this bound is a decreasing function of the distance w between adjacent grid points. If $\bar{\alpha} + \epsilon(w) \leq 0.05$, this implies the familywise Type I error rate is at most 0.05 for all points $\boldsymbol{\Delta} \in \widetilde{B}$. If $\bar{\alpha} + \epsilon(w) > 0.05$, then we either augment the grid $\widetilde{G}$ so as to decrease w, or else we augment the set G of familywise Type I error constraints; in either case, we iterate the above procedure. Augmenting the grid $\widetilde{G}$ decreases the bound $\epsilon(w)$, while adding more points to G forces the sparse linear program to control Type I error at more points $\boldsymbol{\Delta}$. (One strategy is to augment G by all the points in the grid $\widetilde{G}$ where the familywise Type I error exceeds $0.05 - 10^{-4}$.) The above procedure is iterated until $\bar{\alpha} + \epsilon(w) \leq 0.05$. Lastly, we show analytically that the familywise Type I error rate is at most 0.05 for all points $\boldsymbol{\Delta} \notin \widetilde{B}$; the value of $\widetilde{b}$ was initially set large enough to allow such a proof as described in Section C of the Supplementary Material.

8 Discussion

Our approach optimizes the decision rule and multiple testing procedure for a given adaptive design template $\mathbf{n}$ (which includes the first stage sample size and the set of possible stage 2 decisions). An area of future work is to incorporate larger numbers of stage 2 decisions. The size of the search space for the discretized optimization problem is linear in the number K of stage 2 decisions, and we conjecture that it is computationally feasible to set $K = 20$.

The output of such a procedure would be challenging to visualize, unless the optimal rule concentrates on a small subset of the stage 2 decisions.

Another area of future work is to investigate the tradeoff between expected sample size and maximum sample size of two stage, adaptive enrichment designs. This could be done by solving optimization problems such as in Section 5 but using different sets of stage 2 enrollment choices, and plotting the expected and maximum sample sizes achieved by each design. It also may be of interest to explore the impact of the interim analysis timing, e.g., to solve an optimization separately using different stage 1 sample sizes, to determine the best time to make the decision regarding an enrollment change.

A limitation of our approach is that it becomes computationally difficult or infeasible for more than two subpopulations or stages. This is because the discretized search space grows exponentially in the number of subpopulations and the number of stages.

We conjecture that for finer discretizations, the optimal value for the discretized problem will be closer to that of the original constrained Bayes optimization problem from Section 2.3. It is an open problem to bound the difference between these optima as a function of the level of discretization.

The adaptive enrichment designs generated by our approach are probably too complex to be directly used in practice. However, these optimal designs could be used as a benchmark to determine how much can be gained, in principle, from adaptive enrichment for a given adaptive design template $\mathbf{n}$. When the added value is substantial, the designs generated by our approach could later be approximated by simpler designs, e.g., by replacing the discretized regions in Figure 2 by simpler curves.

Acknowledgments

This work was funded by the Patient-Centered Outcomes Research Institute (ME-1306-03198) and the US Food and Drug Administration (HHSF223201400113C); we used IBM CPLEX software that was generously made available through the IBM Academic Initiative. This publication's contents are solely the responsibility of the authors and do not represent the views of the above organizations.

References

Bauer, P. (1989). Multistage testing with adaptive designs (with discussion). *Biometrie und Informatik in Medizin und Biologie 20*, 130–148.

Bauer, P. and K. Köhne (1994). Evaluations of experiments with adaptive interim analyses. *Biometrics 50*, 1029–1041.

Boessen, R., F. van der Baan, R. Groenwold, A. Egberts, O. Klungel, D. Grobbee, M. Knol, and K. Roes (2013). Optimizing trial design in pharmacogenetics research: comparing a fixed parallel group, group sequential, and adaptive selection design on sample size requirements. *Pharmaceutical Statistics 12*(6), 366–374.

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

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

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

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

Follmann, D. (1997). Adaptively changing subgroup proportions in clinical trials. *Statistica Sinica 7*, 1085–1102.

Freidlin, B. and R. Simon (2005). Adaptive signature design: An adaptive clinicaltrial design for generating and prospectivelytesting agene expression signature for sensitive patients. *Clin Cancer Res 11*, 7872–7878.

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

Götte, H., M. Donica, and G. Mordenti (2015). Improving probabilities of correct interim decision in population enrichment designs. *Journal of biopharmaceutical statistics 25*(5), 1020–1038.

Graf, A. C., M. Posch, and F. Koenig (2015). Adaptive designs for subpopulation analysis optimizing utility functions. *Biometrical Journal 57*(1), 76–89.

Hampson, L. V. and C. Jennison (2015). Optimizing the data combination rule for seamless phase II/III clinical trials. *Statistics in Medicine 34*(1), 39–58.

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

Jenkins, M., A. Stone, and C. Jennison (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 B. W. Turnbull (2007). Adaptive seamless designs: Selection and prospective testing of hypotheses. *J. Biopharmaceutical Statistics 17*(6), 1135–1161, doi: 10.1080/10543400701645215.

Krisam, J. and M. Kieser (2015). Optimal decision rules for biomarker-based subgroup selection for a targeted therapy in oncology. *International Journal of Molecular Sciences 16*(5), 10354–10375.

Lai, T. L., P. W. Lavori, and O. Y.-W. Liao (2014). Adaptive choice of patient subgroup for comparing two treatments. *Contemporary clinical trials 39*(2), 191–200.

Lehmacher, W. and G. Wassmer (1999). Adaptive sample size calculations in group sequential trials. *Biometrics 55*(4), 1286–1290.

Lehmann, E. L. and H. Scheffé (1950). Completeness, similar regions, and unbiased estimation. i. *Sankhyā 10*(4), 305–340.

Liu, Q., M. A. Proschan, and G. W. Pledger (2002). A unified theory of two-stage adaptive designs. *JASA 97*(460), 1034–1041.

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

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

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

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

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

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

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

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

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

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

Supplementary Material for Optimal, Two Stage, Adaptive Enrichment Designs for Randomized Trials, using Sparse Linear Programming, by Michael Rosenblum, Ethan X. Fang, and Han Liu

A Distribution of Sufficient Statistics

The distribution of $\mathbf{Z}$ is characterized as follows:

a. $\mathbf{Z}^{(1)}$ is bivariate normal with mean vector $\left(\Delta_1 \left\{\frac{n_1^{(1)}}{2\left(\sigma_{11}^2+\sigma_{10}^2\right)}\right\}^{1/2}, \Delta_2 \left\{\frac{n_2^{(1)}}{2\left(\sigma_{21}^2+\sigma_{20}^2\right)}\right\}^{1/2}\right)$ and covariance matrix $\mathbf{I}_2$, i.e., the 2×2 identity matrix.

b. $\mathbf{Z}^{(2)}$, which uses only stage 2 data, is conditionally independent of $\mathbf{Z}^{(1)}$ given the decision $D(\mathbf{Z}^{(1)}, U)$. The conditional distribution of $\mathbf{Z}^{(2)}$ given $D(\mathbf{Z}^{(1)}, U) = d$ is bivariate normal with mean vector $\left(\Delta_1 \left\{\frac{n_1^{(2),d}}{2\left(\sigma_{11}^2+\sigma_{10}^2\right)}\right\}^{1/2}, \Delta_2 \left\{\frac{n_2^{(2),d}}{2\left(\sigma_{21}^2+\sigma_{20}^2\right)}\right\}^{1/2}\right)$ and covariance matrix $\mathbf{I}_2$.

c. For each subpopulation $s \in \{1, 2\}$, for $D = D(\mathbf{Z}^{(1)}, U)$, we have the following relationship between the final (cumulative) z-statistic and the stagewise z-statistics:

$$Z_s^{(F)} = \left\{\frac{n_s^{(1)}}{n_s^{(1)} + n_s^{(2),D}}\right\}^{1/2} Z_s^{(1)} + \left\{\frac{n_s^{(2),D}}{n_s^{(1)} + n_s^{(2),D}}\right\}^{1/2} Z_s^{(2)}. \tag{S-1}$$

B Evaluating $p(\boldsymbol{\Delta}, r, d, r')$

We show that the quantity $p(\boldsymbol{\Delta}, r, d, r')$ defined in (15) equals the probability that a multivariate normal distribution (with mean vector and covariance matrix given below, which depend on $\boldsymbol{\Delta}, d, \mathbf{n}, \boldsymbol{\sigma}^2$) is in the rectangle $r \times r'$.

Consider any $\boldsymbol{\Delta} \in \mathbb{R}^2$, $r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{E}, r' \in \mathcal{R}_{\text{mtp},d}$. Let $D^{(d)}$ denote the decision rule defined to equal d regardless of the stage 1 data. It follows from (9) and U being independent of the data that the quantity $p(\boldsymbol{\Delta}, r, d, r')$ equals the probability under $P_{\boldsymbol{\Delta}}$ and decision rule $D^{(d)}$ that $\mathbf{Z} = (Z_1^{(1)}, Z_2^{(1)}, Z_1^{(F)}, Z_2^{(F)}) \in r \times r'$. It follows from Section A of the Supplementary Material that under decision rule $D^{(d)}$ and $P_{\boldsymbol{\Delta}}$, the vector $\mathbf{Z}$ has a multivariate Gaussian distribution with covariance matrix

$$\boldsymbol{\Sigma}_d = \begin{pmatrix} 1 & 0 & \gamma_1 & 0 \\ 0 & 1 & 0 & \gamma_2 \\ \gamma_1 & 0 & 1 & 0 \\ 0 & \gamma_2 & 0 & 1 \end{pmatrix},$$

where for each $s \in \{1, 2\}$, we have $\gamma_s = \sqrt{n_s^{(1)} / \left(n_s^{(1)} + n_s^{(2),d}\right)}$, i.e., the correlation between $Z_s^{(1)}, Z_s^{(F)}$. The mean of $\mathbf{Z}$ under $P_{\boldsymbol{\Delta}}$ and decision rule $D^{(d)}$ is $\boldsymbol{\nu}_d(\boldsymbol{\Delta}) = 2^{-1/2} \times$

$$\left(\Delta_1 \left\{\frac{n_1^{(1)}}{(\sigma_{11}^2+\sigma_{10}^2)}\right\}^{1/2}, \Delta_2 \left\{\frac{n_2^{(1)}}{(\sigma_{21}^2+\sigma_{20}^2)}\right\}^{1/2}, \Delta_1 \left\{\frac{n_1^{(1)}+n_1^{(2),d}}{(\sigma_{11}^2+\sigma_{10}^2)}\right\}^{1/2}, \Delta_2 \left\{\frac{n_2^{(1)}+n_2^{(2),d}}{(\sigma_{21}^2+\sigma_{20}^2)}\right\}^{1/2}\right).$$

In computing the coefficients for the sparse linear program (which are functions of $p(\boldsymbol{\Delta}, r, d, r')$), we computed the above probability using the multivariate normal distribution function implemented in the R package mvtnorm.

The vector $\boldsymbol{\nu}_d(\boldsymbol{\Delta})$ can equivalently be expressed as the row vector $\boldsymbol{\Delta} = (\Delta_1, \Delta_2)$ multiplied by the matrix:

$$\mathbf{N}_d = 2^{-1/2} \begin{pmatrix} \left\{ \frac{n_1^{(1)}}{(\sigma_{11}^2+\sigma_{10}^2)} \right\}^{1/2} & 0 & \left\{ \frac{n_1^{(1)}+n_1^{(2),d}}{(\sigma_{11}^2+\sigma_{10}^2)} \right\}^{1/2} & 0 \\ 0 & \left\{ \frac{n_2^{(1)}}{(\sigma_{21}^2+\sigma_{20}^2)} \right\}^{1/2} & 0 & \left\{ \frac{n_2^{(1)}+n_2^{(2),d}}{(\sigma_{21}^2+\sigma_{20}^2)} \right\}^{1/2} \end{pmatrix}. \quad \text{(S-2)}$$

We will use in Section C of the Supplementary Material that

$$\boldsymbol{\Sigma}_d^{-1} = \begin{pmatrix} (1-\gamma_1^2)^{-1} & 0 & -\gamma_1(1-\gamma_1^2)^{-1} & 0 \\ 0 & (1-\gamma_2^2)^{-1} & 0 & -\gamma_2(1-\gamma_2^2)^{-1} \\ -\gamma_1(1-\gamma_1^2)^{-1} & 0 & (1-\gamma_1^2)^{-1} & 0 \\ 0 & -\gamma_2(1-\gamma_2^2)^{-1} & 0 & (1-\gamma_2^2)^{-1} \end{pmatrix},$$

and

$$\mathbf{N}_d \boldsymbol{\Sigma}_d^{-1} \mathbf{N}_d^T = \begin{pmatrix} \frac{n_1^{(1)}+n_1^{(2),d}}{2(\sigma_{11}^2+\sigma_{10}^2)} & 0 \\ 0 & \frac{n_2^{(1)}+n_2^{(2),d}}{2(\sigma_{21}^2+\sigma_{20}^2)} \end{pmatrix}. \quad \text{(S-3)}$$

C Analytic Bounds Used in Verification of Strong Control of the Familywise Type I Error Rate

Denote the familywise Type I error rate of an adaptive design $(D, M) \in \mathcal{A}^{DISC}$ at a given $\boldsymbol{\Delta} \in \mathbb{R}^2$ by

$$F_{D,M}(\boldsymbol{\Delta}) = P_{\boldsymbol{\Delta}}[M\{\mathbf{Z}^{(F)}, D(\mathbf{Z}^{(1)}, U), U\} \cap \mathcal{H}_{\text{TRUE}}(\boldsymbol{\Delta}) \neq \emptyset]. \quad \text{(S-4)}$$

Let $\boldsymbol{\nu}_d(\boldsymbol{\Delta})$ and $\boldsymbol{\Sigma}_d$ denote the mean vector and covariance matrix, respectively, of the z-statistics $\mathbf{Z}$ under the decision rule $D^{(d)}$ and treatment effects $\boldsymbol{\Delta}$ as defined in Section B of the Supplementary Material. Note that $\boldsymbol{\Sigma}_d$ does not depend on $\boldsymbol{\Delta}$. Recall that $\boldsymbol{\nu}_d(\boldsymbol{\Delta})$ equals the matrix product $\boldsymbol{\Delta}\mathbf{N}_d$ for $\mathbf{N}_d$ the matrix defined in (S-2).

For any $\boldsymbol{\Delta}, \boldsymbol{\Delta}' \in \mathbb{R}^2$ and $\lambda \in [0, 1]$, define the convex combination of $\boldsymbol{\Delta}$ and $\boldsymbol{\Delta}'$ as $\widetilde{\boldsymbol{\Delta}}(\lambda) = \lambda\boldsymbol{\Delta} + (1-\lambda)\boldsymbol{\Delta}'$. The next lemma bounds the difference between the familywise Type I error at $\boldsymbol{\Delta}$ and $\boldsymbol{\Delta}'$.

Lemma C.1. *Assume there exists a set $H \subset \mathcal{H}$ such that for all $\lambda \in [0,1]$, $\mathcal{H}_{TRUE}(\widetilde{\boldsymbol{\Delta}}(\lambda)) = H$. Then for any $(D,M) \in \mathcal{A}^{DISC}$, we have*

$$F_{D,M}(\boldsymbol{\Delta}) \leq F_{D,M}(\boldsymbol{\Delta}') + \frac{d}{d\lambda} F_{D,M}(\widetilde{\boldsymbol{\Delta}}(\lambda))|_{\lambda=0} + \sum_{s=1}^{2} \frac{(\Delta_s - \Delta_s')^2}{\sigma_{s1}^2 + \sigma_{s0}^2} \sum_{d \in \mathcal{E}} \left(n_s^{(1)} + n_s^{(2),d} \right),$$

where $\frac{d}{d\lambda} F_{D,M}(\widetilde{\boldsymbol{\Delta}}(\lambda))|_{\lambda=0}$ is given by the formula (S-5) below.

Proof. A second order Taylor expansion of $F_{D,M}(\widetilde{\boldsymbol{\Delta}}(\lambda))$ gives:

$$F_{D,M}(\boldsymbol{\Delta}) - F_{D,M}(\boldsymbol{\Delta}') = \frac{d}{d\lambda} F_{D,M}(\widetilde{\boldsymbol{\Delta}}(\lambda))|_{\lambda=0} + \frac{d^2}{d\lambda^2} F_{D,M}(\widetilde{\boldsymbol{\Delta}}(\lambda))|_{\lambda=\bar{\lambda}},$$

for some $\bar{\lambda} \in [0,1]$. We next bound the terms on the right side of the above display.

Let $\mathbf{z} = (z_1^{(1)}, z_2^{(1)}, z_1^{(F)}, z_2^{(F)})$. Let η_d denote the multivariate normal density with covariance matrix $\boldsymbol{\Sigma}_d$ and mean vector $\mathbf{0}$. Unless indicated otherwise, the integrals below are each

over $u \in [0,1], \mathbf{z} \in \mathbb{R}^4$. We have

$$
\begin{aligned}
&\frac{d}{d\lambda} F_{D,M}(\widetilde{\boldsymbol{\Delta}}(\lambda)) \\
=&\frac{d}{d\lambda} P_{\boldsymbol{\Delta}}[M\{\mathbf{Z}^{(F)}, D(\mathbf{Z}^{(1)}, U), U\} \cap \mathcal{H}_{\text{TRUE}}(\boldsymbol{\Delta}) \neq \emptyset] \\
=&\frac{d}{d\lambda} \sum_{d\in\mathcal{E}} P_{\boldsymbol{\Delta}}[D(\mathbf{Z}^{(1)}, U) = d, M\{\mathbf{Z}^{(F)}, d, U), U\} \cap \mathcal{H}_{\text{TRUE}}(\boldsymbol{\Delta}) \neq \emptyset] \\
=&\frac{d}{d\lambda} \sum_{d\in\mathcal{E}} \int \mathbf{1}\left[D(\mathbf{z}^{(1)}, u) = d, M\{\mathbf{z}^{(F)}, d, u\} \cap H \neq \emptyset\right] \eta_d[\mathbf{z} - \boldsymbol{\nu}_d\{\widetilde{\boldsymbol{\Delta}}(\lambda)\}] d\mathbf{z} du \\
=&\sum_{d\in\mathcal{E}} \int \mathbf{1}\left[D(\mathbf{z}^{(1)}, u) = d, M\{\mathbf{z}^{(F)}, d, u\} \cap H \neq \emptyset\right] \frac{d}{d\lambda}\left(\eta_d[\mathbf{z} - \boldsymbol{\nu}_d\{\widetilde{\boldsymbol{\Delta}}(\lambda)\}]\right) d\mathbf{z} du \\
=&\sum_{d\in\mathcal{E}} \int \mathbf{1}\left[D(\mathbf{z}^{(1)}, u) = d, M\{\mathbf{z}^{(F)}, d, u\} \cap H \neq \emptyset\right] (\mathbf{z} - \boldsymbol{\nu}_d\{\widetilde{\boldsymbol{\Delta}}(\lambda)\}) \boldsymbol{\Sigma}_d^{-1} \frac{d}{d\lambda} \boldsymbol{\nu}_d\{\widetilde{\boldsymbol{\Delta}}(\lambda)\} \\
&\qquad \eta_d[\mathbf{z} - \boldsymbol{\nu}_d\{\widetilde{\boldsymbol{\Delta}}(\lambda)\}] d\mathbf{z} du \\
=&\sum_{d\in\mathcal{E}} \int \mathbf{1}\left[D(\mathbf{z}^{(1)}, u) = d, M\{\mathbf{z}^{(F)}, d, u\} \cap H \neq \emptyset\right] (\mathbf{z} - \boldsymbol{\nu}_d\{\widetilde{\boldsymbol{\Delta}}(\lambda)\}) \boldsymbol{\Sigma}_d^{-1} \boldsymbol{\nu}_d (\boldsymbol{\Delta} - \boldsymbol{\Delta}')^T \\
&\qquad \eta_d[\mathbf{z} - \boldsymbol{\nu}_d\{\widetilde{\boldsymbol{\Delta}}(\lambda)\}] d\mathbf{z} du \\
=&\sum_{d\in\mathcal{E}} \boldsymbol{\nu}_d (\boldsymbol{\Delta} - \boldsymbol{\Delta}') \boldsymbol{\Sigma}_d^{-1} \\
&\int (\mathbf{z} - \boldsymbol{\nu}_d\{\widetilde{\boldsymbol{\Delta}}(\lambda)\})^T \mathbf{1}\left[D(\mathbf{z}^{(1)}, u) = d, M\{\mathbf{z}^{(F)}, d, u\} \cap H \neq \emptyset\right] \eta_d[\mathbf{z} - \boldsymbol{\nu}_d\{\widetilde{\boldsymbol{\Delta}}(\lambda)\}] d\mathbf{z} du \\
=&\sum_{d\in\mathcal{E}} (\boldsymbol{\Delta} - \boldsymbol{\Delta}') \mathbf{N}_d \boldsymbol{\Sigma}_d^{-1} \\
&\int (\mathbf{z} - \boldsymbol{\nu}_d\{\widetilde{\boldsymbol{\Delta}}(\lambda)\})^T \mathbf{1}\left[D(\mathbf{z}^{(1)}, u) = d, M\{\mathbf{z}^{(F)}, d, u\} \cap H \neq \emptyset\right] \eta_d[\mathbf{z} - \boldsymbol{\nu}_d\{\widetilde{\boldsymbol{\Delta}}(\lambda)\}] d\mathbf{z} du.
\end{aligned}
\tag{S-5}
$$

where the last line follows from (S-2).

We next consider the second derivative of the familywise Type I error with respect to λ:

$$
\begin{aligned}
&\Big|\frac{d^2}{d\lambda^2}F_{D,M}(\widetilde{\boldsymbol{\Delta}}(\lambda))\Big| \\
=&\Big|\frac{d^2}{d\lambda^2}\sum_{d\in\mathcal{E}}\int \mathbf{1}\left[D(\mathbf{z}^{(1)},u)=d, M\{\mathbf{z}^{(F)},d,u\}\cap H\neq\emptyset\right]\eta_d[\mathbf{z}-\boldsymbol{\nu}_d\{\widetilde{\boldsymbol{\Delta}}(\lambda)\}]d\mathbf{z}du\Big| \\
=&\Big|\sum_{d\in\mathcal{E}}\int \mathbf{1}\left[D(\mathbf{z}^{(1)},u)=d, M\{\mathbf{z}^{(F)},d,u\}\cap H\neq\emptyset\right]\frac{d^2}{d\lambda^2}\left(\eta_d[\mathbf{z}-\boldsymbol{\nu}_d\{\widetilde{\boldsymbol{\Delta}}(\lambda)\}]\right)d\mathbf{z}du\Big| \\
\leq&\sum_{d\in\mathcal{E}}\int \Big|\mathbf{1}\left[D(\mathbf{z}^{(1)},u)=d, M\{\mathbf{z}^{(F)},d,u\}\cap H\neq\emptyset\right]\frac{d^2}{d\lambda^2}\left(\eta_d[\mathbf{z}-\boldsymbol{\nu}_d\{\widetilde{\boldsymbol{\Delta}}(\lambda)\}]\right)\Big|d\mathbf{z}du \\
\leq&\sum_{d\in\mathcal{E}}\int \Big|\frac{d^2}{d\lambda^2}\eta_d[\mathbf{z}-\boldsymbol{\nu}_d\{\widetilde{\boldsymbol{\Delta}}(\lambda)\}]\Big|d\mathbf{z}du \\
=&\sum_{d\in\mathcal{E}}\int \Big|\frac{d}{d\lambda}\left((\mathbf{z}-\widetilde{\boldsymbol{\Delta}}(\lambda))\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\nu}_d(\boldsymbol{\Delta}-\boldsymbol{\Delta}')^T\eta_d[\mathbf{z}-\boldsymbol{\nu}_d\{\widetilde{\boldsymbol{\Delta}}(\lambda)\}]\right)\Big|d\mathbf{z}du \\
\leq&\sum_{d\in\mathcal{E}}\int \Big|\boldsymbol{\nu}_d(\boldsymbol{\Delta}-\boldsymbol{\Delta}')\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\nu}_d(\boldsymbol{\Delta}-\boldsymbol{\Delta}')^T\Big|\eta_d[\mathbf{z}-\boldsymbol{\nu}_d\{\widetilde{\boldsymbol{\Delta}}(\lambda)\}]d\mathbf{z}du \\
&+\sum_{d\in\mathcal{E}}\int |(\mathbf{z}-\boldsymbol{\nu}_d\{\widetilde{\boldsymbol{\Delta}}(\lambda)\})\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\nu}_d(\boldsymbol{\Delta}-\boldsymbol{\Delta}')^T|^2\eta_d[\mathbf{z}-\boldsymbol{\nu}_d\{\widetilde{\boldsymbol{\Delta}}(\lambda)\}]d\mathbf{z}du \qquad \text{(S-6)} \\
=&\sum_{d\in\mathcal{E}}\boldsymbol{\nu}_d(\boldsymbol{\Delta}-\boldsymbol{\Delta}')\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\nu}_d(\boldsymbol{\Delta}-\boldsymbol{\Delta}')^T+\sum_{d\in\mathcal{E}}\boldsymbol{\nu}_d(\boldsymbol{\Delta}-\boldsymbol{\Delta}')\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\nu}_d(\boldsymbol{\Delta}-\boldsymbol{\Delta}')^T \qquad \text{(S-7)} \\
=&2(\boldsymbol{\Delta}-\boldsymbol{\Delta}')\left(\sum_{d\in\mathcal{E}}\mathbf{N}_d\boldsymbol{\Sigma}_d^{-1}\mathbf{N}_d^T\right)(\boldsymbol{\Delta}-\boldsymbol{\Delta}')^T, \qquad \text{(S-8)} \\
=&2(\boldsymbol{\Delta}-\boldsymbol{\Delta}')\begin{pmatrix} \frac{\sum_{d\in\mathcal{E}}n_1^{(1)}+n_1^{(2),d}}{2\left(\sigma_{11}^2+\sigma_{10}^2\right)} & 0 \\ 0 & \frac{\sum_{d\in\mathcal{E}}n_2^{(1)}+n_2^{(2),d}}{2\left(\sigma_{21}^2+\sigma_{20}^2\right)}\end{pmatrix}(\boldsymbol{\Delta}-\boldsymbol{\Delta}')^T, \qquad \text{(S-9)} \\
=&\sum_{s=1}^{2}\frac{(\Delta_s-\Delta_s')^2}{\sigma_{s1}^2+\sigma_{s0}^2}\sum_{d\in\mathcal{E}}\left(n_s^{(1)}+n_s^{(2),d}\right),
\end{aligned}
$$

where (S-7) follows from the integral in (S-6) being equal to the expected value of the squared magnitude of a multivariate normal distribution with mean 0 and variance $\boldsymbol{\nu}_d(\boldsymbol{\Delta}-\boldsymbol{\Delta}')\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\Sigma}_d\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\nu}_d(\boldsymbol{\Delta}-\boldsymbol{\Delta}')^T=\boldsymbol{\nu}_d(\boldsymbol{\Delta}-\boldsymbol{\Delta}')\boldsymbol{\Sigma}_d^{-1}\boldsymbol{\nu}_d(\boldsymbol{\Delta}-\boldsymbol{\Delta}')^T$; (S-8) follows from the definitions of $\boldsymbol{\nu}_d$ and $\mathbf{N}_d$ in Section B of the Supplementary Material; (S-9) follows from (S-3). This completes the proof of the above lemma. □

Since by the constraint (9), the first stage decision only depends on the rectangle r where the first stage z-statistics $\mathbf{Z}^{(1)}$ falls in. Given Δ and d, the quantity $p(\boldsymbol{\Delta},r,d,r')$ in (15)

is the probability that a multivariate Gaussian distribution with covariance matrix Σ and mean ν falls in the rectangle $r \times r'$, which can be computed efficiently using the function pmvnorm in the R package mvtnorm.

We claimed in Section 7 that $\widetilde{b}$ can be selected such that the familywise Type I error rate of each solution to our example problems from Section 5 is at most 0.05 for all $\boldsymbol{\Delta} \notin [-\widetilde{b}, \widetilde{b}] \times [-\widetilde{b}, \widetilde{b}]$. To show this, it is simpler to use the non-centrality parameter scale, where for a given $\boldsymbol{\Delta} \in \mathbb{R}^2$ the corresponding non-centrality parameter vector is $(\upsilon_1, \upsilon_2) = (\Delta_1\{n/8\}^{1/2}/\sigma, \Delta_2\{n/8\}^{1/2}/\sigma)$. This vector equals the mean of $\boldsymbol{Z}^{(F)}$ under $P_{\boldsymbol{\Delta}}$ for the decision rule $D^{(2)}$ that always makes enrollment decision 2 (using adaptive design template $\mathbf{n}$), which follows from Section A of the Supplementary Material. By construction of the partition of rectangles for the multiple testing procedure described in Section 6, no null hypothesis is rejected when $\boldsymbol{Z}^{(F)} \notin ([-6, 7] \times [-6, 7])$. Therefore, the probability of $D = 2$ and $M \neq \emptyset$ is at most $\Phi(-3) < 0.002$ when any component of the non-centrality parameter vector is outside $[-9, 10] \times [-9, 10]$. By a similar argument (and using that the mean of $\boldsymbol{Z}^{(F)}$ under $P_{\boldsymbol{\Delta}}$ for the decision rule $D^{(d)}$ that always makes enrollment decision d is $(\upsilon_1/\sqrt{2}, \upsilon_2/\sqrt{2})$ for $d = 1$, (υ_1, υ_2) for $d = 2$, $(\upsilon_1\sqrt{2}, \upsilon_2/\sqrt{2})$ for $d = 3$, $(\upsilon_1/\sqrt{2}, \upsilon_2\sqrt{2})$ for $d = 4$), it follows that for any $d \in \mathcal{E}$ the probability of $D = d$ and $M \neq \emptyset$ is at most $\Phi(-3) < 0.002$ when any component of the non-centrality parameter vector is outside $[-9\sqrt{2}, 10\sqrt{2}] \times [-9\sqrt{2}, 10\sqrt{2}]$. Therefore, the familywise Type I error rate is at most $4(0.002) = 0.008 < 0.05$ for any $\boldsymbol{\Delta}$ for which the corresponding non-centrality parameter $(\upsilon_1, \upsilon_2) \notin [-9\sqrt{2}, 10\sqrt{2}] \times [-9\sqrt{2}, 10\sqrt{2}]$.

D Representation of Familywise Type I Error Constraints and Additional Constraints Using $x_{rd}, y_{rdr's}$

Consider any vector $\boldsymbol{\Delta} \in \mathbb{R}^2$ and any discretized adaptive enrichment design $(D, M) \in \mathcal{A}^{DISC}$. We show how to equivalently express the corresponding familywise Type I error constraint (5) from the constrained Bayes optimization problem as the constraint (17) in the discretized problem. The quantity on the left side of the familywise Type I error constraint

(5) is

$$\begin{aligned} & P_{\boldsymbol{\Delta}}\left\{M \text{ rejects any null hypotheses in } \mathcal{H}_{\mathrm{TRUE}}(\boldsymbol{\Delta})\right\} \\ = & \sum_{s \in \mathcal{S}: s \cap \mathcal{H}_{TRUE}(\boldsymbol{\Delta}) \neq \emptyset} P_{\boldsymbol{\Delta}}\left\{M \text{ rejects precisely the set of null hypotheses } s\right\} \\ = & \sum_{s \in \mathcal{S}: s \cap \mathcal{H}_{TRUE}(\boldsymbol{\Delta}) \neq \emptyset} \sum_{r,d,r'} x_{rd} y_{rdr's} p(\boldsymbol{\Delta}, r, d, r'), \end{aligned} \tag{S-10}$$

where the last line, which is identical to (17), follows from (13)-(14).

The additional constraints (6) are equivalently represented in terms of the variables $x_{rd}, y_{rdr's}$ by (18), which follows from analogous argument as in Section 4.3 for the equivalence of the objective function (4) and its discretized counterpart (16).

E Proof of Theorem 3.1

We first prove the following lemma:

Lemma E.1. *At the interim analysis, $\boldsymbol{Z}^{(1)}$ is a minimal sufficient statistic. At the end of the trial, $(D(\mathbf{Z}^{(1)}), \boldsymbol{Z}^{(F)})$ is a minimal sufficient statistic.*

Proof. Let $n_{sa}^{(k)}$ denote the number of participants from subpopulation s with arm assignment a enrolled during stage k. Define $\sigma_s = \sigma_{s0} = \sigma_{s1}$. Let $\widehat{\mu}_{sa}^{(k)}$ denote the sample mean of the outcomes in participants from subpopulation s, arm a, and stage k. At the interim analysis, the likelihood function is (for a constant C that does not depend on the parameter $\boldsymbol{\Delta}$):

$$\begin{aligned} \mathcal{L}(\boldsymbol{\Delta}; X^{(1)}) &= C \prod_{s=1}^{2} \left(2\pi\sigma_s^2\right)^{-n_s^{(1)}/2} \exp\left\{ -\sum_{i=1}^{n_s^{(1)}} \frac{\left(Y_{s,i}^{(1)} - \mu_{s,A_{s,i}^{(1)}}\right)^2}{2\sigma^2_{s,A_{s,i}^{(1)}}} \right\} \\ &= C \prod_{s=1}^{2} \left(2\pi\sigma_s^2\right)^{-n_s^{(1)}/2} \prod_{a=0}^{1} \exp\left\{ -\frac{n_{sa}^{(1)}\left(\mu_{sa} - \widehat{\mu}_{sa}^{(1)}\right)^2 + \sum_{i:A_{s,i}^{(1)}=a}\left(Y_{s,i}^{(1)} - \widehat{\mu}_{sa}^{(1)}\right)^2}{2\sigma_s^2} \right\} \\ &= C \prod_{s=1}^{2} \left(2\pi\sigma_s^2\right)^{-n_s^{(1)}/2} \exp\left[-\frac{n_s^{(1)}\left\{\Delta_s - (\widehat{\mu}_{s1}^{(1)} - \widehat{\mu}_{s0}^{(1)})\right\}^2 + n_s^{(1)}\left(\widehat{\mu}_{s1}^{(1)} + \widehat{\mu}_{s0}^{(1)}\right)^2}{8\sigma_s^2} \right] \times \\ &\quad \prod_{a=0}^{1} \exp\left\{ -\frac{\sum_{i:A_{s,i}^{(1)}=a}\left(Y_{s,i}^{(1)} - \widehat{\mu}_{sa}^{(1)}\right)^2}{2\sigma_s^2} \right\}, \end{aligned}$$

where the last equality holds by the assumption from Section 2.1 that $\mu_{s1} = \Delta_s/2$ and $\mu_{s0} = -\Delta_s/2$ for each subpopulation $s \in \{1, 2\}$. The Fisher-Neyman factorization theorem implies that $\left(\widehat{\mu}_{11}^{(1)} - \widehat{\mu}_{10}^{(1)}, \widehat{\mu}_{21}^{(1)} - \widehat{\mu}_{20}^{(1)}\right)$ is a sufficient statistic at the end of stage 1. By (7), we have

$$\boldsymbol{Z}^{(1)} = \left(\frac{\widehat{\mu}_{11}^{(1)} - \widehat{\mu}_{10}^{(1)}}{(4\sigma_1^2/n_1^{(1)})^{1/2}}, \frac{\widehat{\mu}_{21}^{(1)} - \widehat{\mu}_{20}^{(1)}}{(4\sigma_2^2/n_2^{(1)})^{1/2}} \right). \tag{S-11}$$

This implies $\boldsymbol{Z}^{(1)}$ is a sufficient statistic at the end of stage 1, since we assumed $\sigma_1^2, \sigma_2^2, n_1^{(1)}, n_2^{(1)}$ are known.

To prove the minimal sufficiency, let $\widetilde{X}^{(1)}$ be an independent set of stage 1 outcomes, i.e., the data collected in stage 1 of a separate, independent trial. Let $\widetilde{\mu}_{sa}^{(1)}$ denote the corresponding sample mean for participants from subpopulation s and arm a enrolled during stage 1, which is a function of $\widetilde{X}^{(1)}$. It follows from the above derivation of $\mathcal{L}(\boldsymbol{\Delta}; X^{(1)})$ that the likelihood ratio is (where C_1 is a function of the data and not of the parameter $\boldsymbol{\Delta}$):

$$\frac{\mathcal{L}(\boldsymbol{\Delta}; X^{(1)})}{\mathcal{L}(\boldsymbol{\Delta}; \widetilde{X}^{(1)})} = C_1\left(X^{(1)}, \widetilde{X}^{(1)}\right) \prod_{s=1}^{2} \exp\left[-\frac{n_s^{(1)}\left\{\Delta_s - (\widehat{\mu}_{s1}^{(1)} - \widehat{\mu}_{s0}^{(1)})\right\}^2 - n_s^{(1)}\left\{\Delta_s - (\widetilde{\mu}_{s1}^{(1)} - \widetilde{\mu}_{s0}^{(1)})\right\}^2}{8\sigma_s^2} \right].$$

It follows that the likelihood ratio does not depend on (Δ_1, Δ_2) if and only if for each $s \in \{1, 2\}$, $\widehat{\mu}_{s1} - \widehat{\mu}_{s0} = \widetilde{\mu}_{s1} - \widetilde{\mu}_{s0}$, by Lehmann-Scheffé Theorem (Lehmann and Scheffé, 1950). By (S-11), this implies the z-statistic $\mathbf{Z}^{(1)}$ is minimal sufficient at the end of stage 1.

We next prove that at the end of the trial, the statistic $\left(D(\mathbf{Z}^{(1)}), \mathbf{Z}^{(F)}\right)$ is minimal sufficent. Let $\widehat{\mu}_{sa}$ denote the sample mean from subpopulation s and arm a pooling across participants in both stages. By a similar derivation as for the case of stage 1 data only, at the end of stage 2 the likelihood function is (for C_2 a function of the data and not of the parameter $\boldsymbol{\Delta}$):

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\Delta}; X) \\
&= C_2(X) \prod_{s=1}^{2} \left(2\pi\sigma_s^2\right)^{-(n_s^{(1)} + N_s^{(2)})/2} \exp\left[-\frac{(n_s^{(1)} + N_s^{(2)})\left\{\Delta_s - (\widehat{\mu}_{s1} - \widehat{\mu}_{s0})\right\}^2}{8\sigma_s^2} \right] \times \\
&\qquad \exp\left[-\frac{(n_s^{(1)} + N_s^{(2)})\left(\widehat{\mu}_{s1} + \widehat{\mu}_{s0}\right)^2}{8\sigma_s^2} \right] \prod_{a=0}^{1} \prod_{k=1}^{2} \exp\left\{ -\frac{\sum_{i:A_{s,i}^{(k)}=a}\left(Y_{s,i}^{(k)} - \widehat{\mu}_{sa}^{(k)}\right)^2}{2\sigma_s^2} \right\}.
\end{aligned}$$

By the Fisher-Neyman factorization theorem and (S-11), the enrollment decision D (which is equivalent to knowing $n_s^{(1)} + N_s^{(2)}$ for each $s \in \{1, 2\}$) together with the cumulative z-statistics $\mathbf{Z}^{(F)}$ are sufficient statistics.

To prove minimal sufficiency, let $\widetilde{X}$ be an independent set of outcomes, i.e., the data collected in a separate, independent trial, with corresponding quantities denoted by $\widetilde{\cdot}$. Let $\widetilde{\mu}_{sa}$ denote the corresponding sample mean for participants from subpopulation s and arm a (from both stages), which is a function of $\widetilde{X}$. By similar arguments as above, the likelihood ratio is (for a function C_3 that depends on the data but not on the parameter $\boldsymbol{\Delta}$):

$$\frac{\mathcal{L}(\boldsymbol{\Delta}; X)}{\mathcal{L}(\boldsymbol{\Delta}; \widetilde{X})} = C_3\left(X, \widetilde{X}\right) \times \prod_{s=1}^{2} \exp\left[-\frac{(n_s^{(1)} + N_s^{(2)})\left\{\Delta_s - (\widehat{\mu}_{s1} - \widehat{\mu}_{s0})\right\}^2 - (n_s^{(1)} + N_s^{(2)})\left\{\Delta_s - (\widetilde{\mu}_{s1} - \widetilde{\mu}_{s0})\right\}^2}{8\sigma_s^2}\right].$$

The likelihood ratio does not depend on (Δ_1, Δ_2) if and only if for each $s \in \{1, 2\}$, $\widehat{\mu}_{s1} - \widehat{\mu}_{s0} = \widetilde{\mu}_{s1} - \widetilde{\mu}_{s0}$ and $n_s^{(1)} + N_s^{(2)} = \widetilde{n}_s^{(1)} + \widetilde{N}_s^{(2)}$. This implies the minimal sufficiency of the statistic $\left(\mathbf{Z}^{(F)}, D(\mathbf{Z}^{(1)}, U)\right)$, by the Lehmann-Scheffé Theorem (Lehmann and Scheffé, 1950). □

Next, we prove Theorem 3.1.

Proof. Let $P_{\boldsymbol{\Delta},D}$ denote the distribution on X under treatment effect $\boldsymbol{\Delta}$ and using the adaptive enrichment design with decision rule D. For any $d \in \mathcal{E}$, let $P_{\boldsymbol{\Delta},d}$ denote the distribution on X under treatment effect $\boldsymbol{\Delta}$ and using the adaptive enrichment design that always makes the enrollment choice d for stage 2 enrollment.

Consider any adaptive enrichment design $(D, M) \in \mathcal{D}^* \times \mathcal{M}^*$. We will prove that there exists an adaptive enrichment design $(D', M') \in \mathcal{A}^{SUFF}$ with identical risk for any bounded loss function L and any $\boldsymbol{\Delta} \in \mathbb{R}^2$, i.e., we will prove

$$E_{\boldsymbol{\Delta},D} L(M, D, \Delta) = E_{\boldsymbol{\Delta},D'} L(M', D', \Delta). \tag{S-12}$$

Let U, U' be independent, uniformly distributed random variables on $[0, 1]$ that are exogenous, i.e., independent of the data X, and define $\widetilde{U} = (U, U')$. It follows from Lemma E.1 that the conditional distribution of X given $(\mathbf{Z}^{(F)}, D = d, U)$ does not depend on Δ. Therefore, there exists a function $f_2 : \mathbb{R}^2 \times \mathcal{E} \times [0, 1]^2 \to \mathcal{X}$ such that $f_2(\mathbf{Z}^{(F)}, d, U, U')$ and X have the same conditional distribution given $(\mathbf{Z}^{(F)}, D = d, U)$. It follows that $M(f_2(\mathbf{Z}^{(F)}, d, U, U'), d, U)$ and $M(X, d, U)$ have the same conditional distribution given $(\mathbf{Z}^{(F)}, D = d, U)$. Define

$$M'(\mathbf{Z}^{(F)}, d, U, U') = M\{f_2(\mathbf{Z}^{(F)}, d, U, U'), d, U\},$$

for any $d \in \mathcal{E}$. By construction, $M'(\mathbf{Z}^{(F)}, D, \widetilde{U})$ and $M(X, D, U)$ have the same conditional distribution given $(\mathbf{Z}^{(F)}, D = d, U)$.

For any bounded loss function L and $\mathbf{\Delta} \in \mathbb{R}^2$, we have (where $D = D(X^{(1)}, U)$) that

$$
\begin{aligned}
& E_{\mathbf{\Delta},D} L(M, D, \Delta) && \text{(S-13)} \\
= \; & E_{\mathbf{\Delta},D} L\left[M\{X, D, U\}, D, \mathbf{\Delta}\right] \\
= \; & \sum_{d \in \mathcal{E}} E_{\mathbf{\Delta},d} \mathbf{1}(D = d) L\left[M\{X, d, U\}, d, \mathbf{\Delta}\right] \\
= \; & \sum_{d \in \mathcal{E}} E_{\mathbf{\Delta},d} \mathbf{1}(D = d) E_{\mathbf{\Delta},d}\left(L\left[M\{X, d, U\}, d, \mathbf{\Delta}\right] \big| \mathbf{Z}^{(F)}, D = d, U\right) \\
= \; & \sum_{d \in \mathcal{E}} E_{\mathbf{\Delta},d} \mathbf{1}(D = d) E_{\mathbf{\Delta},d}\left(L\left[M'\{\mathbf{Z}^{(F)}, d, \widetilde{U}\}, d, \mathbf{\Delta}\right] \big| \mathbf{Z}^{(F)}, D = d, U\right) \\
= \; & \sum_{d \in \mathcal{E}} E_{\mathbf{\Delta},d} \mathbf{1}(D = d) L\left[M'\{\mathbf{Z}^{(F)}, d, \widetilde{U}\}, d, \mathbf{\Delta}\right] \\
= \; & \sum_{d \in \mathcal{E}} E_{\mathbf{\Delta},D} \mathbf{1}(D = d) L\left[M'\{\mathbf{Z}^{(F)}, D, \widetilde{U}\}, D, \mathbf{\Delta}\right] \\
= \; & E_{\mathbf{\Delta},D} L\left[M'\{\mathbf{Z}^{(F)}, D, \widetilde{U}\}, D, \mathbf{\Delta}\right] \\
= \; & E_{\mathbf{\Delta},D} L(M', D, \mathbf{\Delta}). && \text{(S-14)}
\end{aligned}
$$

It remains to show that (S-14) equals the right side of (S-12). It follows from the arguments in Lemma E.1 that the first stage data $X^{(1)}$ can be expressed as a function (denoted by f_1) of the minimal sufficient statistic $\mathbf{Z}^{(1)}$ and an ancillary part R with the following property: R is independent of $(\mathbf{Z}^{(1)}, \mathbf{Z}^{(F)})$ under $P_{\mathbf{\Delta},d}$ for any $d \in \mathcal{E}$. Define $\widetilde{R}$ to be an exogenous random vector with the same distribution as R and independent of U, U'. Define the decision rule $D'(\mathbf{Z}^{(1)}, \widetilde{R}, U) = D(f_1(\mathbf{Z}^{(1)}, \widetilde{R}), U)$. It follows that for any $d \in \mathcal{E}$, we have

$$
\begin{aligned}
& E_{\mathbf{\Delta},d} \mathbf{1}\{D(X^{(1)}, U) = d\} L(M'\{\mathbf{Z}^{(F)}, d, \widetilde{U}\}, d, \Delta) \\
= \; & E_{\mathbf{\Delta},d} \mathbf{1}\{D(f_1(\mathbf{Z}^{(1)}, R), U) = d\} L(M'\{\mathbf{Z}^{(F)}, d, \widetilde{U}\}, d, \Delta) \\
= \; & E_{\mathbf{\Delta},d} \mathbf{1}\{D(f_1(\mathbf{Z}^{(1)}, \widetilde{R}), U) = d\} L(M'\{\mathbf{Z}^{(F)}, d, \widetilde{U}\}, d, \Delta) \\
= \; & E_{\mathbf{\Delta},d} \mathbf{1}\{D'(\mathbf{Z}^{(1)}, \widetilde{R}, U) = d\} L(M'\{\mathbf{Z}^{(F)}, d, \widetilde{U}\}, d, \Delta).
\end{aligned}
$$

Therefore,

$$\begin{aligned}
& E_{\mathbf{\Delta},D} L(M', D, \mathbf{\Delta}) \\
= \; & E_{\mathbf{\Delta},D} L\left[M'\{\mathbf{Z}^{(F)}, D(X^{(1)}, U), \widetilde{U}\}, D(X^{(1)}, U), \mathbf{\Delta}\right] \\
= \; & \sum_{d \in \mathcal{E}} E_{\mathbf{\Delta},d} \mathbf{1}\{D(X^{(1)}, U) = d\} L\left[M'\{\mathbf{Z}^{(F)}, d, \widetilde{U}\}, d, \mathbf{\Delta}\right] \\
= \; & \sum_{d \in \mathcal{E}} E_{\mathbf{\Delta},d} \mathbf{1}\{D'(\mathbf{Z}^{(1)}, \widetilde{R}, U) = d\} L\left[M'\{\mathbf{Z}^{(F)}, d, \widetilde{U}\}, d, \mathbf{\Delta}\right] \\
= \; & E_{\mathbf{\Delta},D'} L\left[M'\{\mathbf{Z}^{(F)}, D', \widetilde{U}\}, D', \mathbf{\Delta}\right] \\
= \; & E_{\mathbf{\Delta},D'} L(M', D', \mathbf{\Delta}).
\end{aligned}$$

The above display shows that (S-14) equals the right side of (S-12), which by the equality of (S-13) and (S-14) proves (S-12). The exogenous part $(\widetilde{R}, \widetilde{U})$ can be expressed as a function of a single, exogenous, uniformly distributed random variable V on $[0, 1]$, which implies that the exogenous parts in each of M' and D' can be written as functions of V. Since the definitions of D' and M' do not depend on $\mathbf{\Delta}$ or L, this completes the proof of Theorem 3.1 □

F Proof of Theorem 4.1

For any $\{v_{rdr's}\}$, define the corresponding variables for the discretized problem $\{x_{rd}, y_{rdr's}\}$ by (29). This mapping is well defined, since by (27) we have $\sum_{s \in \mathcal{S}} v_{rdr's}$ does not depend on r'. Similarly, for any $\{x_{rd}, y_{rdr's}\}$, define the corresponding variables for the sparse linear program $\{v_{rdr's}\}$ by (22). We prove the following lemma:

Lemma F.1. *The discretized problem's constraints (19)-(21) on the variables $\{x_{rd}, y_{rdr's}\}$ are equivalent to the sparse linear program's constraints (26)-(28) on the variables $\{v_{rdr's}\}$.*

Proof. First, consider any $\{v_{rdr's}\}$ that satisfy (26)-(28). Define the corresponding $\{x_{rd}, y_{rdr's}\}$ using the mapping (29) in part (ii) of Theorem 4.1. The nonnegativity constraint (28) for the sparse linear program implies the analogous constraint (21) from the discretized problem. The formula (29) for x_{rd} and the constraint (26) imply (19). The formula (29) for $y_{rdr's}$ implies (20). We have shown that the constraints (26)-(28) from the sparse linear program imply the constraints (19)-(21) from the discretized problem. We next show the reverse implication.

Consider any $\{x_{rd}, y_{rdr's}\}$ that satisfies the constraints (19)-(21). Define the corresponding $\{v_{rdr's}\}$ using the mapping (22). The nonnegativity constraint (21) from the discretized problem implies the analogous constraint (28) for the sparse linear program. For any $r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{E}, r' \in \mathcal{R}_{\text{mtp},d}$, we have by (22) and (20) that

$$\sum_{s\in\mathcal{S}} v_{rdr's} = \sum_{s\in\mathcal{S}} x_{rd} y_{rdr's} = x_{rd} \sum_{s\in\mathcal{S}} y_{rdr's} = x_{rd}. \tag{S-15}$$

For each $r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{E}$, and any two rectangles $r', \widetilde{r}' \in \mathcal{R}_{\text{mtp},d}$, we have

$$\sum_{s\in\mathcal{S}} v_{rdr's} = \sum_{s\in\mathcal{S}} v_{rd\widetilde{r}'s}, \tag{S-16}$$

since the left and right sides of the above display both equal x_{rd} by (S-15). The above display implies (27). For each $r \in \mathcal{R}_{\text{dec}}$,

$$\sum_{d\in\mathcal{E}} \sum_{s\in\mathcal{S}} v_{rdr'_d s} = \sum_{d\in\mathcal{E}} x_{rd} = 1, \tag{S-17}$$

where the first equality follows from (S-15) and the second follows from (19). The above display implies (26). We have shown that the constraints (19)-(21) from the discretized problem imply the constraints (26)-(28) for the sparse linear program. This completes the proof of Lemma F.1. □

Proof of Theorem 4.1. Consider any feasible solution $\mathbf{v}$ (the vector representation of $\{v_{rdr's}\}$) to the sparse linear program. Define the corresponding solution $(\mathbf{x}, \mathbf{y})$ (the vector representation of $\{x_{rd}, y_{rdr's}\}$) to the discretized problem from Section 4 through the mapping (29). The equations (29) imply (22) holds. The solution $\mathbf{x}, \mathbf{y}$ is feasible for the discretized problem since (24) and (22) imply (17); (25) and (22) imply (18); Lemma F.1 implies (19)-(21). The value of the objective function (16) of the discretized problem from Section 4 equals the value of the objective function (23) of the sparse linear program, which follows by (22). This shows $\mathbf{x}, \mathbf{y}$ is a feasible solution to the discretized problem with the same value (of the objective function) as the corresponding solution to the sparse linear program evaluated at $\mathbf{v}$. Therefore, the value of the optimal solution to the sparse linear program is an upper bound on the value of the optimal solution to the discretized problem.

Next, consider any feasible solution $\mathbf{x}, \mathbf{y}$ to the discretized problem from Section 4, and define the corresponding solution $\mathbf{v}$ to the sparse linear program using the transformation (22). We next show that $\mathbf{v}$ is a feasible solution to the sparse linear program with the same

value (of the objective function) as the corresponding discretized problem evaluated at $\mathbf{x}, \mathbf{y}$. The solution $\mathbf{v}$ is feasible for the discretized problem since (17) and (22) imply (24); (18) and (22) imply (25); Lemma F.1 implies (26)-(28). The value of the objective function (23) equals (16). We have shown $\mathbf{v}$ is a feasible solution to the sparse linear program with the same value (of the objective function) as the corresponding discretized problem evaluated at $\mathbf{x}, \mathbf{y}$. Therefore, the value of the optimal solution to the discretized problem is an upper bound on the value of the optimal solution to the sparse linear program.

The results of the above two paragraphs prove the claim (i) in the theorem. Claim (ii) then follows from the result in the first paragraph. This completes the proof of Theorem 4.1. □

As a side note, the set of constraints (27) is equivalent to:

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{E}, \text{ and each pair } r', \widetilde{r}' \in \mathcal{R}_{\text{mtp},d}, \sum_{s \in \mathcal{S}} v_{rdr's} = \sum_{s \in \mathcal{S}} v_{rd\widetilde{r}'s}. \tag{S-18}$$

The reason we use (27) instead of the constraints in the above display is that the former has fewer constraints, which makes the corresponding linear program smaller.

G Proof that for each of Examples 2.1 and 2.2, the optimal solution depends on $(\sigma^2, \Delta^{\min}, n)$ only through the non-centrality parameter

Proof. Consider the constrained Bayes optimization problems corresponding to Examples 2.1 and 2.2 with inputs as specified in Section 5.1, which uses adaptive design template $\mathbf{n}^{(1b)}$. By Theorem 3.1, it suffices to consider decision rules and multiple testing procedures that depend on the data only through the sufficient statistics given in that theorem. Denote the non-centrality parameter by $\upsilon = (n/8)^{1/2}\Delta^{\min}/\sigma$. We will show that the solutions to the optimization problems depend on $(n, \Delta^{\min}, \sigma^2)$ only through υ.

We reparametrize the optimization problems in terms of $(\Delta_1/\Delta_{\min}, \Delta_2/\Delta_{\min})$ in place of (Δ_1, Δ_2). Denote $\widetilde{\mathbf{\Delta}} = (\widetilde{\Delta}_1, \widetilde{\Delta}_2) = (\Delta_1/\Delta_{\min}, \Delta_2/\Delta_{\min})$. Let $P_{\widetilde{\mathbf{\Delta}}} = P_{\mathbf{\Delta}/\Delta_{\min}}$ and $E_{\widetilde{\mathbf{\Delta}}} = E_{\mathbf{\Delta}/\Delta_{\min}}$. It follows from Section A of the Supplementary Material that the distribution of $\mathbf{Z}$ under $P_{\widetilde{\mathbf{\Delta}}}$ and adaptive design template $\mathbf{n}^{(1b)}$ is characterized in terms of υ as follows:

a. $\mathbf{Z}^{(1)}$ is bivariate normal with mean vector $(\upsilon/2^{1/2})\widetilde{\mathbf{\Delta}}$ and covariance matrix $\mathbf{I}_2$.

b. $\mathbf{Z}^{(2)}$ is conditionally independent of $\mathbf{Z}^{(1)}$ given D. We next give the conditional distribution of $\mathbf{Z}^{(2)}$ given $D = d$ for each $d \in \mathcal{E}$. For $d = 1$, we have $\mathbf{Z}^{(2)} = \mathbf{0}$. Given $D = 2$, we have $\mathbf{Z}^{(2)}$ is bivariate normal with mean vector $(\upsilon/2^{1/2})\widetilde{\boldsymbol{\Delta}}$ and covariance matrix $\mathbf{I}_2$. Given $D = 3$, the distribution of $Z_1^{(2)}$ is normal with mean $\upsilon(3/2)^{1/2}\widetilde{\Delta}_1$ and variance 1, and $Z_2^{(2)} = 0$. Given $D = 4$, the distribution of $Z_2^{(2)}$ is normal with mean $\upsilon(3/2)^{1/2}\widetilde{\Delta}_2$ and variance 1, and $Z_1^{(2)} = 0$.

c. For each subpopulation $s \in \{1, 2\}$, for $D = D(\mathbf{Z}^{(1)}, U)$, we have the following relationship between the final (cumulative) z-statistic and the stagewise z-statistics:

$$Z_s^{(F)} = \{f(D, s)\}^{1/2} Z_s^{(1)} + \{1 - f(D, s)\}^{1/2} Z_s^{(2)}, \tag{S-19}$$

where we define

$$\begin{aligned} f(d, s) \;\; = \;\; & 1(d = 1) + 1(d = 2)(1/2) + 1(d = 3, s = 1)(1/4) + 1(d = 3, s = 2) \\ & +1(d = 4, s = 1) + 1(d = 4, s = 2)(1/4). \end{aligned}$$

It follows that the distribution of the statistics $(\mathbf{Z}^{(1)}, D(\mathbf{Z}^{(1)}, U), \mathbf{Z}^{(F)})$ under $P_{\widetilde{\boldsymbol{\Delta}}}$ does not depend on the problem inputs $(n, \Delta^{\min}, \sigma^2)$ except through υ. We next show that the constraints (5)-(6) and objective function (4) can be represented in terms of $P_{\widetilde{\boldsymbol{\Delta}}}$, $E_{\widetilde{\boldsymbol{\Delta}}}$, and that the resulting optimization problem does not depend on $(n, \Delta^{\min}, \sigma^2)$ except through υ.

The set of familywise Type I error constraints (5), which is over $\boldsymbol{\Delta} \in \mathbb{R}^2$, is equivalent to the reparametrized set of constraints replacing $P_{\boldsymbol{\Delta}}$ by $P_{\widetilde{\boldsymbol{\Delta}}}$ over the set $\widetilde{\boldsymbol{\Delta}} \in \mathbb{R}^2$, i.e.,

$$P_{\widetilde{\boldsymbol{\Delta}}}\left\{M \text{ rejects any null hypotheses in } \mathcal{H}_{\text{TRUE}}(\widetilde{\boldsymbol{\Delta}})\right\} \leq \alpha, \text{ for any } \widetilde{\boldsymbol{\Delta}} \in \mathbb{R}^2, \tag{S-20}$$

The loss functions (1)-(2) in the additional constraints (6) can be equivalently represented in terms of $\widetilde{\boldsymbol{\Delta}}$ as

$$\begin{aligned} \widetilde{L}^{(s)}(M, D, \widetilde{\boldsymbol{\Delta}}) &= 1\big(H_{0s} \notin M; \widetilde{\Delta}_s \geq 1\big), \text{ for each } s \in \{1, 2\}; \\ \widetilde{L}^{(C)}(M, D, \widetilde{\boldsymbol{\Delta}}) &= 1\big(H_{0C} \notin M; \widetilde{\Delta}_1 \geq 1, \widetilde{\Delta_2} \geq 1\big). \end{aligned}$$

The additional constraints (6) in the context of the problems from Examples 2.1 and 2.2 can be expressed as follows:

$$\begin{aligned} &\text{For } \widetilde{\boldsymbol{\Delta}} = (1, 0), E_{\widetilde{\boldsymbol{\Delta}}}\widetilde{L}^{(1)}(M, D, \widetilde{\boldsymbol{\Delta}}) \leq \beta; \\ &\text{For } \widetilde{\boldsymbol{\Delta}} = (0, 1), E_{\widetilde{\boldsymbol{\Delta}}}\widetilde{L}^{(2)}(M, D, \widetilde{\boldsymbol{\Delta}}) \leq \beta; \\ &\text{For } \widetilde{\boldsymbol{\Delta}} = (1, 1), E_{\widetilde{\boldsymbol{\Delta}}}\widetilde{L}^{(C)}(M, D, \widetilde{\boldsymbol{\Delta}}) \leq \beta, \end{aligned}$$

where the above lines are equivalent to (P1)-(P3), respectively, from Section 2.4.

We next reparametrize the objective function (4) in terms of $\widetilde{\boldsymbol{\Delta}}$. The loss function $L_0 = L^{\mathrm{SS}}$ depends on $M, D, \boldsymbol{\Delta}$ only through D (which depends on the data only through $\mathbf{Z}^{(1)}$) and equals

$$L^{\mathrm{SS}} = n_1^{(1)} + n_2^{(1)} + n_1^{(2),D} + n_2^{(2),D} = (n/2) + 1(D = 2)(n/2) + 1(D \in \{3, 4\})(3n/4).$$

The optimal solution is unchanged if the objective function is multiplied by a positive constant; therefore, we can replace L^{SS} by $\widetilde{L}^{\mathrm{SS}} = (1/2) + 1(D = 2)(1/2) + 1(D \in \{3, 4\})(3/4)$ and the optimal solution will not change. (Though the value attained by the optimal solution will be changed, this value as a fraction of n–which is reported in Table 1–is not changed). In Example 2.1, the objective function (4) is equivalent to:

$$\frac{1}{4} \sum_{\widetilde{\boldsymbol{\Delta}} \in Q'} E_{\widetilde{\boldsymbol{\Delta}}} \widetilde{L}^{\mathrm{SS}}(D), \tag{S-21}$$

for $Q' = \{(0, 0), (1, 0), (0, 1), (1, 1)\}$. In Example 2.2, the objective function (4) is equivalent to:

$$\int_{\widetilde{\boldsymbol{\Delta}} \in \mathbb{R}^2} E_{\widetilde{\boldsymbol{\Delta}}} \widetilde{L}^{\mathrm{SS}}(D) d\Lambda^{\mathrm{mix}}(\widetilde{\boldsymbol{\Delta}})$$

for Λ^{mix} the distribution on $\widetilde{\Delta}$ that is a mixture of four bivariate normal distributions with mean vectors in Q' and each having covariance matrix $\mathbf{I}_2$ (from our assumption in Section 2.4 that $\sigma_\Lambda^2 = \Delta_{\min}^2$).

The above reparametrization removed all dependence on the problem inputs $(\sigma^2, \Delta^{\min}, n)$ except through υ. We have shown that the constrained Bayes optimization problems in Examples 2.1 and 2.2 with inputs as specified in Section 5.1 are functions of the inputs $(\sigma^2, \Delta^{\min}, n)$ only through υ.

□

H Multiple Testing Procedure Based on P-Value Combination Approach Used in $\mathcal{A}^{COMB}$

We define the class of adaptive enrichment designs $\mathcal{A}^{COMB}$ used in Section 5.3, which are based on the p-value combination approach and the closed testing principle. The set of elementary null hypotheses is $\mathcal{H} = \{H_{01}, H_{02}, H_{0C}\}$. We consider subsets of null hypotheses

indexed by $I \subseteq \{1, 2, C\}$. For each subset I of null hypotheses, we define a first stage p-value $p_{1,I}$, a second stage p-value $p_{2,I}$ (under every possible decision for stage 2 enrollment), and a combination test $C_I(p_{1,I}, p_{2,I})$. In order to apply the closure principle, the requirements that must be satisfied are, for each $I \subseteq \{1, 2, C\}$: [Note: the phrase "under the intersection null hypothesis $\bigcap_{i \in I} H_{0i}$" means "assuming H_{0i} is true for each $i \in I$".]

i. The distribution of first stage p-value $p_{1,I}$ must stochastically dominate the distribution $U[0, 1]$ under the intersection null hypothesis $\bigcap_{i \in I} H_{0i}$.

ii. Conditioned on the first stage statistics and enrollment decision, the second stage p-value $p_{2,I}$ must stochastically dominate the distribution $U[0, 1]$ under the intersection null hypothesis $\bigcap_{i \in I} H_{0i}$.

iii. The combination test $C_I(p_{1,I}, p_{2,I})$ is a mapping from $[0, 1] \times [0, 1]$ to $\{0, 1\}$ where 1 indicates rejection of the intersection $\bigcap_{i \in I} H_{0i}$, and 0 indicates failure to reject $\bigcap_{i \in I} H_{0i}$. We require $\int_{(x,y) \in [0,1] \times [0,1]} C_I(x, y) dx dy \leq 0.05$, and that for any $x' \leq x, y' \leq y, C_I(x', y') \geq C_I(x, y)$ (called monotonicity).

The above conditions imply for each $I \subseteq \{1, 2, C\}$ that the probability of $C_I(p_{1,I}, p_{2,I}) = 1$ is at most 0.05 under the intersection null hypothesis $\bigcap_{i \in I} H_{0i}$. (Note that the combination test is the same regardless of the decision as to stage 2 enrollment, e.g., if the inverse-normal combination test is used, the weights must be set in advance and cannot depend on the decision for stage 2 enrollment.)

For each null hypothesis $i \in \{1, 2, C\}$, let $Z_i^{(k)}$ denote the z-statistic based on stage k data for population i defined as $Z_i^{(k)}$ if $i \in \{1, 2\}$ and otherwise $Z_C^{(k)} = p_1^{1/2} Z_1^{(k)} + p_2^{1/2} Z_2^{(k)}$.

For each $I \subseteq \{1, 2, C\}$, we describe our choices for $p_{1,I}, p_{2,I}$, and C_I, each of which satisfies the above requirements. For stage 1, we set $p_{1,I}$ to be the p-value corresponding to the Dunnett test based on the maximum of $\{Z_i^{(1)}, i \in I\}$. We set C_I to be the inverse-weighted normal combination function, using weights $w_1 = w_2 = 1/\sqrt{2}$, i.e., $C_I(x, y) = 1$ if $1 - \Phi\{w_1 \Phi^{-1}(1 - x) + w_2 \Phi^{-1}(1 - y)\} < 0.05$, and equals 0 otherwise. For stage 2, for each $I \subseteq \{1, 2, C\}$, we define $p_{2,I}$ under each possible enrollment decision:

- If both subpopulations are enrolled in stage 2, we set $p_{2,I}$ to be the p-value corresponding to the Dunnett test based on the maximum of stage 2 z-statistics $\{Z_i^{(2)}, i \in I\}$.
- If the trial is stopped completely after stage 1, we set $p_{2,I} = 1$.

- If only subpopulation $s \in \{1, 2\}$ is enrolled in stage 2, we set $p_{2,I}$ to be $1 - \Phi(Z_s^{(2)})$ if $s \in I$, else we set $p_{2,I} = 1$.

For each $I \subseteq \{1, 2, C\}$, the corresponding intersection null hypothesis $\bigcap_{i \in I} H_{0i}$ is rejected at the end of the trial if and only if $C_I(p_{1,I}, p_{2,I}) = 1$. By the closure principle, for each elementary null hypothesis $H \in \mathcal{H}$, it is rejected at the end of the trial if and only if for each $I \subseteq \{1, 2, C\}$ containing the index of H, the intersection null hypothesis $\bigcap_{i \in I} H_{0i}$ is rejected. The closure principle ensures the familywise Type I error rate is at most 0.05.

For example, if the decision was to enroll only subpopulation 1 in stage 2, then the stage 2 p-values are computed as follows: $p_{2,I} = 1 - \Phi(Z_1^{(2)})$ for each $I \subseteq \{1, 2, C\}$ for which $1 \in I$, and $p_{2,I} = 1$ otherwise. I.e., $p_{2,\{H_{01}\}} = p_{2,\{H_{01},H_{0C}\}} = p_{2,\{H_{01},H_{02}\}} = p_{2,\{H_{01},H_{02},H_{0C}\}} = 1 - \Phi(Z_1^{(2)})$, and all other stage 2 p-values equal 1. The elementary null hypothesis H_{01} is rejected if $C_I(p_{1,I}, p_{2,I}) = 1$ for each $I \subseteq \{1, 2, C\}$ for which $1 \in I$; i.e., we need $C_I(p_{1,I}, p_{2,I}) = 1$ for the case of I being the indices corresponding to $\{H_{01}\}, \{H_{01}, H_{0C}\}, \{H_{01}, H_{02}\}$, and $\{H_{01}, H_{02}, H_{0C}\}$. (Note: each such I has identical stage 2 p-values in this case, but the stage 1 p-values may differ, so it is necessary to compute all four values and check $C_I(p_{1,I}, p_{2,I}) = 1$ in each case.)

To allow early stopping for efficacy, we also consider a modified combination test motivated by the O'Brien-Fleming group sequential test, and a modified decision rule for stage 2 enrollment. Let $C'_I(x, y) = 1$ if $x < 1 - \Phi(2.37)$ or $1 - \Phi\{w_1\Phi^{-1}(1 - x) + w_2\Phi^{-1}(1 - y)\} < 1 - \Phi(1.68)$, and equal 0 otherwise. This modified combination test satisfies assumption (iii) above, and allows rejection of null hypotheses for efficacy after stage 1. This occurs for a hypothesis $H \in \mathcal{H}$ if for all intersection tests involving H, the corresponding first stage p-value is at most $1 - \Phi(2.37)$ (since this guarantees that this intersection test would be rejected at the end of stage 2, regardless of the stage 2 statistics). We modify the decision rule to stop the entire trial early if any elementary null hypothesis is rejected in stage 1.

More generally, for any $u_1 \geq \Phi^{-1}(1 - \alpha)$, define u_2 to be the unique solution to

$$P\{(V_1 > u_1) \text{ or } (w_1V_1 + w_2V_2 > u_2)\} = \alpha,$$

for V_1, V_2 independent, standard normal random variables. Define the modified combination test $C_I^{(u_1)}(x, y) = 1$ if $x < 1 - \Phi(u_1)$ or $1 - \Phi\{w_1\Phi^{-1}(1 - x) + w_2\Phi^{-1}(1 - y)\} < 1 - \Phi(u_2)$, and equal 0 otherwise. For the special case that $u_1 = 2.37$ and $w_1 = w_2 = 1/\sqrt{2}$, we have $u_2 = 1.68$ and therefore the combination test $C_I^{(u_1)}(x, y)$ is identical to $C'_I(x, y)$. Throughout, we used $u_1 = 1.68$ and $u_2 = 2.37$.

I Additional Constraints

I.1 Constraints that Aim to Remove the Dependence of $y_{rdr's}$ on r

The set of constraints below was conjectured to be satisfied at the optimum solution to our example problems. This was verified by solving the optimization problems without these constraints, and then with them, and checking that the value of the optimal solutions are equal to high precision. The motivation for these constraints was to produce a solution that can be more easily visualized.

We added new, sparse linear constraints to the sparse linear program that, in some cases, force $y_{rdr's}$ to depend only on $dr's$ and not on r. To give the intuition behind these constraints, consider any solution $\{x_{rd}, y_{rdr's}\}$ to the discretized problem such that the variables $y_{rdr's}$ do not depend on r, i.e., for which $y_{r_1dr's} = y_{r_2dr's}$ for all $r_1, r_2 \in \mathcal{R}_{\text{dec}}, d \in \mathcal{E}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S}$. Let $\{v_{rdr's}\}$ denote the corresponding solution to the sparse linear program defined by the mapping (22). By the arguments in the proof of Theorem 4.1, $\{v_{rdr's}\}$ is a solution to the corresponding sparse linear program. Consider any $r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{E}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S}$. Let $w_{dr's}$ denote the value of $y_{rdr's}$, which by our choice above does not depend on r; it follows from (20) that $\sum_{s\in\mathcal{S}} w_{dr's} = 1$. We have

$$
\begin{aligned}
w_{dr's} &= y_{rdr's} \\
&= x_{rd} y_{rdr's} + (1 - x_{rd}) y_{rdr's} \\
&= v_{rdr's} + (1 - x_{rd}) y_{rdr's} && \text{(S-22)} \\
&\leq v_{rdr's} + (1 - x_{rd}) && \text{(S-23)} \\
&= v_{rdr's} + \sum_{\widetilde{d}\in\mathcal{E}\setminus\{d\}} x_{r\widetilde{d}} && \text{(S-24)} \\
&= v_{rdr's} + \sum_{\widetilde{d}\in\mathcal{E}\setminus\{d\}} \sum_{s'\in\mathcal{S}} v_{r\widetilde{d}r'_{\widetilde{d}}s'},
\end{aligned}
$$

where (S-22) follows from (22); (S-23) follows from $x_{rd}, y_{rdr's} \in [0, 1]$; (S-24) follows from (19); and, the last line follows from (20) and (22). The above arguments imply that if there is an optimal solution $\{x_{rd}, y_{rdr's}\}$ to the discretized problem such that the variables $y_{rdr's}$ do not depend on r, then the corresponding optimal solution $\{v_{rdr's}\}$ to the sparse linear program satisfies the following additional set of constraints:

$$\text{for each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{E}, r' \in \mathcal{R}_{\text{mtp},d}, s \in \mathcal{S},\ w_{dr's} \leq v_{rdr's} + \sum_{\tilde{d} \in \mathcal{E} \setminus \{d\}} \sum_{s' \in \mathcal{S}} v_{r\tilde{d}r'_{\tilde{d}}s'};$$

$$\text{for each } d \in \mathcal{E}, r' \in \mathcal{R}_{\text{mtp},d}, \qquad \sum_{s \in \mathcal{S}} w_{dr's} = 1.$$

We introduced the new variables $w_{dr's}$ and added the above, new constraints to our sparse linear programs.

I.2 Monotonicity Constraints

The set of monotonicity constraints below was conjectured to be satisfied at the optimum solution to our example problems. This was verified by solving the optimization problems without these constraints, and then with them, and checking that the value of the optimal solutions are equal to high precision. For each $r' \in \mathcal{R}_{\text{mtp},d},$, let r'_R, r'_A denote the rectangle immediately to its right, and the rectangle immediately above it, respectively (if one exists).

The set of constraints below encodes that the probability of rejecting H_{01} is a non-decreasing function of $Z_1^{(F)}$ in the following sense: the conditional probability of rejecting H_{01} given $\mathbf{Z}^{(1)} \in r, D(\mathbf{Z}^{(1)}, U) = d, \mathbf{Z}^{(F)} \in r'$ is at most the the conditional probability of rejecting H_{01} given $\mathbf{Z}^{(1)} \in r, D(\mathbf{Z}^{(1)}, U) = d, \mathbf{Z}^{(F)} \in r'_R$. This is encoded as follows:

$$\text{For each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{E}, r' \in \mathcal{R}_{\text{mtp},d}, \sum_{s \in \mathcal{S}: H_{01} \in s} v_{rdr's} \leq \sum_{s \in \mathcal{S}: H_{01} \in s} v_{rdr'_R s}. \tag{S-25}$$

The analogous set of monotonicity constraints with respect to H_{02} is the following:

$$\text{For each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{E}, r' \in \mathcal{R}_{\text{mtp},d}, \sum_{s \in \mathcal{S}: H_{02} \in s} v_{rdr's} \leq \sum_{s \in \mathcal{S}: H_{02} \in s} v_{rdr'_A s}. \tag{S-26}$$

The following set of constraints encodes that the conditional probability of rejecting H_{0C} given $\mathbf{Z}^{(1)} \in r, D(\mathbf{Z}^{(1)}, U) = d, \mathbf{Z}^{(F)} \in r'$ is at most the the conditional probability of rejecting H_{0C} given $\mathbf{Z}^{(1)} \in r, D(\mathbf{Z}^{(1)}, U) = d, \mathbf{Z}^{(F)} \in \overline{r}'$ for $\overline{r}' = r'_R$ and similarly for $\overline{r}' = r'_A$:

$$\text{For each } r \in \mathcal{R}_{\text{dec}}, d \in \mathcal{E}, r' \in \mathcal{R}_{\text{mtp},d}:$$
$$\sum_{s \in \mathcal{S}: H_{0C} \in s} v_{rdr's} \leq \sum_{s \in \mathcal{S}: H_{0C} \in s} v_{rdr'_R s} \text{ and } \sum_{s \in \mathcal{S}: H_{0C} \in s} v_{rdr's} \leq \sum_{s \in \mathcal{S}: H_{0C} \in s} v_{rdr'_A s}. \tag{S-27}$$

The design in Figure 2 corresponds to the solution where we included the above sets of monotonicity constraints.