Published in final edited form as:
J Am Stat Assoc. 2014 January 1; 109(507): 1216–1228. doi:10.1080/01621459.2013.879063.

Optimal Tests of Treatment Effects for the Overall Population and Two Subpopulations in Randomized Trials, using Sparse Linear Programming

Michael Rosenblum[*], **Han Liu**[†], and **En-Hsu Yen**[‡]

Michael Rosenblum: mrosenbl@jhsph.edu

[*]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA, 21205

[†]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, USA, 08544

[‡]Department of Computer Science, University of Texas at Austin, TX, USA, 78712

Abstract

We propose new, optimal methods for analyzing randomized trials, when it is suspected that treatment effects may differ in two predefined subpopulations. Such subpopulations could be defined by a biomarker or risk factor measured at baseline. The goal is to simultaneously learn which subpopulations benefit from an experimental treatment, while providing strong control of the familywise Type I error rate. We formalize this as a multiple testing problem and show it is computationally infeasible to solve using existing techniques. Our solution involves a novel approach, in which we first transform the original multiple testing problem into a large, sparse linear program. We then solve this problem using advanced optimization techniques. This general method can solve a variety of multiple testing problems and decision theory problems related to optimal trial design, for which no solution was previously available. In particular, we construct new multiple testing procedures that satisfy minimax and Bayes optimality criteria. For a given optimality criterion, our new approach yields the optimal tradeoff between power to detect an effect in the overall population versus power to detect effects in subpopulations. We demonstrate our approach in examples motivated by two randomized trials of new treatments for HIV.

1 Introduction

An important goal of health research is determining which populations, if any, benefit from new treatments. Randomized trials are generally considered the gold standard for producing evidence of treatment effects. Most randomized trials aim to determine how a treatment compares to control, on average, for a given population. This results in trials that may fail to detect important differences in benefits and harms for subpopulations, such as those with a certain biomarker or risk factor. This problem affects trials in virtually all disease areas.

Consider planning a randomized trial of an experimental treatment versus control, where there is prior evidence that treatment effects may differ for two, predefined subpopulations. Such evidence could be from past trials or observational studies, or from medical knowledge of how the treatment is conjectured to work. Our goal is to construct a multiple testing

procedure with optimal power to detect treatment effects for the overall population and for each subpopulation. We consider both Bayes and minimax optimality criteria. Existing multiple testing procedures in general do not satisfy either of these criteria.

It is a challenging problem to construct optimal multiple testing procedures. According to Romano et al. (2011), “there are very few results on optimality in the multiple testing literature. ” The problems we consider are especially challenging since we require strong 2 control of the familywise Type I error rate, also called the studywide Type I error rate, as defined by Hochberg and Tamhane (1987). That is, we require that under any data generating distribution, the probability of rejecting one or more true null hypotheses is at most a given level α. We incorporate these constraints because control of the studywide Type I error rate is generally required by regulatory agencies such as the U.S. Food and Drug Administration and the European Medicines Agency for confirmatory randomized trials involving multiple hypotheses (FDA and EMEA, 1998).

Strong control of the familywise Type I error rate implies infinitely many constraints, i.e., one for every possible data generating distribution. The crux of our problem is constructing multiple testing procedures satisfying all these constraints and optimizing power at a given set of alternatives. In the simpler problem of testing only the null hypothesis for the overall population, the issue of infinitely many constraints can be sidestepped; this is because for most reasonable tests, strong control of the Type I error is implied by control of the Type I error at the global null hypothesis of zero average treatment effect. In contrast, when dealing with multiple populations, procedures that control the familywise Type I error at the global null hypothesis can have quite large Type I error at distributions corresponding to a positive effect for one subpopulation and a nonpositive effect for another. For this reason, optimization methods designed for a single null hypothesis, such as those of Jennison (1987); Eales and Jennison (1992); Banerjee and Tsiatis (2006); and Hampson and Jennison (2013), do not directly apply to our problem. Though in principle these methods could be extended to handle more Type I error constraints, such extensions are computationally infeasible in our problems, as we discuss in Section 7.

Our solution hinges on a novel method for transforming a fine discretization of the original multiple testing problem into a large, sparse linear program. The resulting linear program typically has over a million variables and constraints. We tailor advanced optimization tools 3 to solve the linear program. To the best of our knowledge, this is the first computationally feasible method for constructing Bayes or minimax optimal tests of treatment effects for subpopulations and the overall population, while maintaining strong control of the familywise Type I error rate.

We apply our approach to answer the following open questions: What is the maximum power that can be gained to detect treatment effects in subpopulations if one is willing to sacrifice *x*% power for detecting an effect in the overall population? What is the minimum additional sample size required to increase power for detecting treatment effects in subpopulations by *x*%, while maintaining a desired power for the overall population?

A motivating data example is given in Section 2. We define our problem in Section 3, present our method for solving it in Section 4, and demonstrate this method in Section 5. We explain how we overcome computational challenges in our problem in Sections 6 and 7. Section 8 extends our method to decision theory problems. The sparse linear programming algorithm we use is given in Section 9. We conclude with a discussion of limitations of our approach and future directions for research in Section 10.

2 Example: Randomized Trials of New Antiretroviral Treatments for HIV

We demonstrate our approach in scenarios motivated by two recently completed randomized trials of maraviroc, an antiretroviral medication for treatment-experienced, HIV positive individuals (Fätkenheuer et al., 2008). There is suggestive evidence from these trials that the treatment benefit may differ depending on the suppressive effect of an individual's background therapy, as measured by the phenotypic sensitivity score (PSS) at baseline. The estimated average treatment benefit of maraviroc among individuals with PSS less than 3 was larger than that among individuals with PSS 3 or more. This pattern has been observed for other antiretroviral medications, e.g., in randomized trials of etravirine (Katlama et al., 2009). We refer to those with PSS less than 3 as subpopulation 1, and those with PSS 3 or more as subpopulation 2. In the combined maraviroc trials, 63% of participants are in subpopulation 1.

In planning a trial of a new antiretroviral medication, it may be of interest to determine the average treatment effect for the overall population and for each of these subpopulations. We construct multiple testing procedures that maximize power for detecting treatment benefits in each subpopulation, subject to constraints on the familywise Type I error rate and on power for the overall population.

3 Multiple Testing Problem

3.1 Null Hypotheses and Test Statistics

Consider a randomized trial comparing a new treatment (a=1) to control (a=0), in which there are two prespecified subpopulations that partition the overall population. Denote the fraction of the overall population in subpopulation $k \in \{1, 2\}$ by p_k. We assume each patient is randomized to the new treatment or control with probability 1/2, independent of the patient's subpopulation. Below, for clarity of presentation, we focus on normally distributed outcomes with known variances. In Section A of the Supplementary Materials, we describe asymptotic extensions allowing a variety of outcome types, and where the variances are unknown and must be estimated.

For each subpopulation $k \in \{1, 2\}$ and study arm $a \in \{0, 1\}$, assume the corresponding patient outcomes are independent and distributed as $Y_{ka,i} \sim N(\mu_{ka}, \sigma^2_{ka})$, for each patient $i = 1, 2, \ldots n_{ka}$. For each subpopulation $k \in \{1, 2\}$, define the population average treatment effect as $\Delta_k = \mu_{k1} - \mu_{k0}$. For each $k \in \{1, 2\}$, define H_{0k} to be the null hypothesis $\Delta_k \leq 0$, i.e., that treatment is no more effective than control, on average, for subpopulation k; define H_{0C} to be the null hypothesis $p_1\Delta_1 + p_2\Delta_2 \leq 0$, i.e., that treatment is no more effective than 5 control, on average, for the combined population.

Let n denote the total sample size in the trial. For each $k \in \{1, 2\}$ and $a \in \{0, 1\}$, we assume the corresponding sample size $n_{ka} = p_k n/2$; that is, the proportion of the sample in each subpopulation equals the corresponding population proportion p_k, and exactly half of the participants in each subpopulation are assigned to each study arm. This latter property can be approximately achieved by block randomization within each subpopulation.

We assume the subpopulation fractions p_k and the variances σ^2_{ka} are known. This implies the following z-statistics are sufficient statistics for (Δ_1,Δ_2):

$$\text{for each subpopulation} k \in \{1,2\}, Z_k = \left(\frac{1}{n_{k1}} \sum_{i=1}^{n_{k1}} Y_{k1,i} - \frac{1}{n_{k0}} \sum_{i=1}^{n_{k0}} \right) \upsilon_k^{-1/2},$$

for $\upsilon_k = \sigma^2_{k1}/n_{k1} + \sigma^2_{k0}/n_{k0}$.

We also consider the pooled z-statistic for the combined population,

$$Z_C = \sum_{k=1}^{2} p_k \left(\frac{1}{n_{k1}} \sum_{i=1}^{n_{k1}} Y_{k1,i} - \frac{1}{n_{k0}} \sum_{i=1}^{n_{k0}} Y_{k0,i} \right) \left(p_1^2 \upsilon_1 + p_2^2 \upsilon_2 \right)^{-1/2}.$$

We then have $Z_C = \rho_1 Z_1 + \rho_2 Z_2$, for $\rho_k = \left[\rho_k^2 \upsilon_k / (p_1^2 \upsilon_1 + \rho_2^2 \upsilon_2) \right]^{1/2}$, which is the covariance of Z_k and Z_C. The vector of sufficient statistics (Z_1,Z_2) is bivariate normal with mean $(\delta_1, \delta_2) = (\Delta_1/\sqrt{\upsilon_1}, \Delta_2/\sqrt{\upsilon_2})$ and covariance matrix the identity matrix. We call (δ_1, δ_2) the non-centrality parameters of (Z_1,Z_2). For $\Delta^{\min} > 0$ the minimum, clinically meaningful treatment effect, let $\delta_1^{\min}$ and $\delta_2^{\min}$ be the non-centrality parameters that correspond to $\Delta_1 = \Delta^{\min}$ and $\Delta_2 = \Delta^{\min}$, respectively.

Define $\delta_C = EZ_C = \rho_1\delta_1 + \rho_2\delta_2$. We use the following equivalent representation of the null hypotheses above:

$$H_{01}:\delta_1 \le 0; H_{02}:\delta_2 \le 0; H_{0C}:\rho_1\delta_1 + \rho_2\delta_2 \le 0. \quad (1)$$

For any (δ_1, δ_2), denote the corresponding set of true null hypotheses in the family $\mathcal{H} = \{H_{01},H_{02},H_{0C}\}$ by $\mathcal{H}_{\text{TRUE}}(\delta_1,\delta_2)$; for each $k \in \{1, 2\}$, this set contains H_{0k} if and only if $\delta_k \le 0$, and contains H_{0C} if and only if $\rho_1\delta_1 + \rho_2\delta_2 \le 0$.

3.2 Multiple Testing Procedures and Optimization Problem

The multiple testing problem is to determine which subset of $\mathcal{H}$ to reject, on observing a single realization of (Z_1,Z_2). The pair (Z_1,Z_2) is drawn from the distribution P_{δ_1,δ_2}, defined to be the bivariate normal distribution with mean vector (δ_1,δ_2) and covariance matrix the 2×2 identity matrix.

Let $\mathcal{S}$ denote an ordered list of all subsets of the null hypotheses $\mathcal{H}$. Consider multiple testing procedures for the family $\mathcal{H}$, i.e., maps from each possible realization of (Z_1,Z_2) to an element of $\mathcal{S}$, representing the null hypotheses rejected upon observing (Z_1,Z_2). It will be

useful to consider the class $\mathscr{M}$ of randomized multiple testing procedures, defined as the maps M from each possible realization of (Z_1,Z_2) to a random variable taking values in $\mathscr{S}$. Formally, a randomized multiple testing procedure is a measurable map $M = M(Z_1,Z_2, U)$ that depends on (Z_1,Z_2) but also may depend on an independent random variable U that has a uniform distribution on [0, 1]. Define the class of deterministic multiple testing procedures $\mathscr{M}_{det}$ to be all $M \in \mathscr{M}$ such that for any $(z_1, z_2) \in \mathbb{R}^2$ and $u, u' \in [0, 1]$, we have $M(z_1, z_2, u) = M(z_1, z_2, u')$; for such procedures, we let $M(z_1, z_2)$ denote the value of $M(z_1, z_2, u)$, which does not depend on u.

The reason we use randomized procedures, rather than restricting to deterministic procedures, is computational. We show in Section 4 that the discretized version of our optimization problem reduces to a linear program, when we optimize over a class of randomized procedures. In contrast, if we restrict to deterministic procedures, the optimization problem reduces to an integer program. Linear programs are generally much easier to solve than integer programs. This computational advantage is especially important in our context where we have a large number of variables and constraints. Though we optimize over randomized procedures, it turns out that each optimal solution in the examples in Section 5.1 is a deterministic procedure, as we discuss in Section 10. For conciseness, we write "multiple testing procedure" instead of "randomized multiple testing procedure," with the understanding that unless otherwise stated, we deal with the latter throughout.

Let L denote a bounded loss function, where $L(s; \delta_1,\delta_2)$ represents the loss if precisely the subset $s \subseteq \mathscr{H}$ is rejected when the true non-centrality parameters are (δ_1, δ_2). An example is the loss function that imposes a penalty of 1 unit for failing to reject the null hypothesis for each subpopulation when the average treatment effect is at least the minimum, clinically meaningful level in that subpopulation. This loss function can be written as $\tilde{L}(s;\delta_1,\delta_2)=\sum_{k=1}^{2} 1[\delta_k \geq \delta_k^{\min}, H_{0k} \notin s]$, where $1[C]$ is the indicator function taking value 1 if C is true and 0 otherwise. In Section D of the Supplementary Materials, we consider modifications of $\tilde{L}$ where the penalty is proportional to the corresponding treatment benefit, up to a given maximum penalty. Our general method can be applied to any bounded loss function that can be numerically integrated with respect to δ_1, δ_2 by standard software with high precision. In particular, we allow L to be non-convex in (δ_1, δ_2), which is the case in all our examples.

We next state the Bayes version of our general optimization problem. Let Λ denote a prior distribution on the set of possible pairs of non-centrality parameters (δ_1, δ_2). We assume Λ is a distribution with compact support on $(\mathbb{R}^2, \mathscr{B})$, for $\mathscr{B}$ a σ-algebra over $\mathbb{R}^2$.

Constrained Bayes Optimization Problem—For given $\alpha > 0$, $\beta > 0$, $\delta_1^{\min}$, $\delta_2^{\min}$, L, and Λ, find the multiple testing procedure $M \in \mathscr{M}$ minimizing

$$\int E_{\delta_1,\delta_2}\{L(M(Z_1, Z_2, U);\delta_1, \delta_2)\}d\Lambda(\delta_1, \delta_2), \quad (2)$$

under the familywise Type I error constraints: for any $(\delta_1, \delta_2) \in \mathbb{R}^2$,

$$P_{\delta_1,\delta_2}\{M \text{rejects any null hypotheses in} \mathscr{H}_{\text{TRUE}}(\delta_1,\delta_2)\} \leq \alpha, \quad (3)$$

and the power constraint for the combined population:

$$P_{\delta_1^{\min},\delta_2^{\min}}\{M \text{rejects} H_{0C}\} \geq 1-\beta. \quad (4)$$

The objective function (2) encodes the expected loss incurred by the testing procedure *M*, averaged over the prior distribution Λ. The constraints (3) enforce strong control of the familywise Type I error rate.

The corresponding minimax optimization problem replaces the objective function (2) by

$$\sup_{(\delta_1,\delta_2)\in\wp} E_{\delta_1,\delta_2} L(M(Z_1,Z_2,U);\delta_1,\delta_2), \quad (5)$$

for $\wp$ a subset of $\mathbb{R}^2$ representing the alternatives of interest.

4 Solution to Constrained Bayes Optimization Problem

The above constrained Bayes optimization problem is either very difficult or impossible to solve analytically, due to the continuum of Type I error constraints that must be satisfied. Our approach involves discretizing the constrained Bayes optimization problem. We approximate the infinite set of constraints (3) by a finite set of constraints, and restrict to multiple testing procedures that are constant over small rectangles. This transforms the constrained Bayes optimization problem, which is non-convex, into a large, sparse linear program that we solve using advanced optimization tools. In Section 6, we bound the approximation error in the discretization using the dual linear program; we apply this to show the approximation error is very small in our examples.

We first restrict to the class of multiple testing procedures $\mathscr{M}_B \subset \mathscr{M}$ that reject no hypotheses outside the region $B = [-b, b] \times [-b, b]$ for a fixed integer $b > 0$. Intuitively, if we select *b* large enough that $(Z_1,Z_2) \in B$ with high probability under the prior Λ, we may expect the Bayes risk of the optimal solution among procedures in $\mathscr{M}_B$ to be within a small value $\in$ of the optimal solution over $\mathscr{M}$. For the examples in Section 5.1, we verify that it is sufficient to set $b = 5$ to achieve this at $\varepsilon = 0.005$, as shown in Section 6. In Section B of the Supplementary Materials, we show how to augment the structure of an approximately optimal procedure among $\mathscr{M}_B$ to allow rejection of null hypotheses outside of *B*.

We next restrict to a finite subset of the familywise Type I error constraints (3). These will be selected from points in $G = \{(\delta_1, \delta_2) : \delta_1 = 0 \text{ or } \delta_2 = 0 \text{ or } \rho_1\delta_1 + \rho_2\delta_2 = 0\}$, which represents the pairs of non-centrality parameters at which the first subpopulation has zero average benefit, the second subpopulation has zero average benefit, or the combined population has zero average benefit. Our restricting to *G* is motivated by the conjecture that the worst-case, familywise Type I error occurs on the union of the boundaries of the null spaces for H_{01},H_{02},H_{0C}. We verified this holds for each example in Section 5.1. We also prove in Section 5.2 that for a class of multiple testing procedures with certain intuitively

appealing properties, the worst-case, familywise Type I error always occurs at some $(\delta_1, \delta_2) \in G$. Let G' denote a finite subset of G; e.g., for some $\tau_1, \tau_2 > 0$, we could set G' to be

$$G_{\tau,b}=[\{(k\tau_1,0):k\in\mathbb{Z}\}\cup\{(0,k\tau_2):k\in\mathbb{Z}\}\cup\{(\rho_2 k\tau_1,-\rho_1 k\tau_1):k\in\mathbb{Z}\}]\cap B.$$

In Section 6, we discuss why certain carefully selected, finite subsets G' of G lead to solutions that are very close to optimal for the original problem, and that satisfy all constraints of the original problem.

The next step is to define a subclass of multiple testing procedures that are constant over small rectangles. For fixed $\tau = (\tau_1, \tau_2)$, for each $k, k' \in \mathbb{Z}$, define the rectangle $R_{k,k'} = [k\tau_1, (k+1)\tau_1)\times[k'\tau_2, (k'+1)\tau_2)$. Let $\mathscr{R}$ denote the set of such rectangles in the bounded region B, i.e., $\mathscr{R} = \{R_{k,k'}: k, k' \in \mathbb{Z}, R_{k,k'} \subset B\}$. Define $\mathscr{M}_{\mathscr{R}}$ to be the subclass of multiple testing procedures $M \in \mathscr{M}_B$ that, for any $u \in [0, 1]$ and rectangle $r \in \mathscr{R}$, satisfy

$M(z_1, z_2, u)=M(z_1', z_2', u)$ whenever (z_1, z_2) and (z_1', z_2') are both in r. For any procedure $M \in \mathscr{M}_{\mathscr{R}}$, its behavior is completely characterized by the finite set of values $\mathbf{m} = \{m_{rs}\}_{r\in\mathscr{R},s\in\mathscr{S}}$, where

$$m_{rs}=P[M(Z_1,Z_2,U)\text{rejects precisely the subset of null hypothesess}|(Z_1,Z_2)\in r]. \quad (6)$$

For any $r \in \mathscr{R}$, it follows that

$$\sum_{s\in\mathscr{S}} m_{rs}=1,\text{and}m_{rs}\geq 0\text{for any}s\in\mathscr{S}. \quad (7)$$

Also, for any set of real values $\{m_{rs}\}_{r\in\mathscr{R},s\in\mathscr{S}}$ satisfying (7), there is a multiple testing procedure $M \in \mathscr{M}_{\mathscr{R}}$ satisfying (6), i.e., the procedure M that rejects precisely the subset of null hypotheses s with probability m_{rs} when $(Z_1,Z_2) \in r$.

The advantage of the above discretization is that if we restrict to procedures in $\mathscr{M}_{\mathscr{R}}$, the objective function (2) and constraints (3)–(4) in the constrained Bayes optimization problem are each linear functions of the variables $\mathbf{m}$. This holds even when the loss function L is non-convex. To show (2) is linear in $\mathbf{m}$, first consider the term inside the integral in (2):

$$E_{\delta_1,\delta_2}L(M(Z_1,Z_2,U);\delta_1,\delta_2) \quad (8)$$

$$\begin{aligned}&=\sum_{r\in\Re,s\in\mathscr{S}} E_{\delta_1,\delta_2}[L(M(Z_1,Z_2,U);\delta_1,\delta_2)|M(Z_1,Z_2,U)\\&\quad=s,(Z_1,Z_2)\in r]\\&\quad\times P_{\delta_1,\delta_2}[M(Z_1,Z_2,U)\\&\quad=s|(Z_1,Z_2)\in r]P_{\delta_1\delta_2}[(Z_1,Z_2)\in r]\\&=\sum_{r\in\Re,s\in\mathscr{S}} L(s;\delta_1,\delta_2)P_{\delta_1,\delta_2}[(Z_1,Z_2)\in r]m_{rs}.\end{aligned} \quad (9)$$

The objective function (2) is the integral over $\Lambda$of (8), which by the above argument equals

$$\sum_{r\in\Re,s\in\mathscr{S}} \{\int L(s;\delta_1,\delta_2)P_{\delta_1,\delta_2}[(Z_1,Z_2)\in r]d\Lambda(\delta_1,\delta_2)\}\, m_{rs}. \quad (10)$$

The constraints (3) and (4) can be similarly represented as linear functions of **m**, as we show in Section C of the Supplementary Materials.

Define the discretized problem to be the constrained Bayes optimization problem restricted to procedures in $\mathscr{M}_{\mathscr{R}}$, and replacing the familywise Type I error constraints (3) by those corresponding to $(\delta_1, \delta_2) \in G'$. The discretized problem can be expressed as:

Sparse Linear Program Representing Discretization of Original Problem (2)–(4)

For given $\alpha > 0$, $\beta > 0$, $\delta_1^{\min}$, $\delta_2^{\min}$, τ, b, G', L, and Λ, find the set of real values $\mathbf{m} = \{m_{rs}\}_{r\in\mathscr{R},s\in\mathscr{S}}$ minimizing (10) under the constraints:

$$\text{for all}(\delta_1,\delta_2)\in G', \sum_{r\in\Re}\sum_{s\in\mathscr{S}:s\cap\mathscr{H}_{\text{TRUE}}(\delta_1,\delta_2)\neq\varnothing} P_{\delta_1,\delta_2}[(Z_1,Z_2)\in r]m_{rs}\le\alpha; \quad (11)$$

$$\sum_{r\in\Re}\sum_{s\in\mathscr{S}:H_{0C}\in s} P_{\delta_1^{\min},\delta_2^{\min}}[(Z_1,Z_2)\in r]m_{rs}\ge 1-\beta; \quad (12)$$

$$\text{for all} r\in\Re, \sum_{s\in\mathscr{S}} m_{rs}=1; \quad (13)$$

$$\text{for all} r\in\Re, s\in\mathscr{S}, m_{rs}\ge 0. \quad (14)$$

The constraints (11) represent the familywise Type I error constraints (3) restricted to $(\delta_1, \delta_2) \in G'$ and $M \in \mathscr{M}_{\mathscr{R}}$; (12) represents the power constraint (4) restricted to $\mathscr{M}_{\mathscr{R}}$. We refer to the value of the Bayes objective function (10) evaluated at **m** as the Bayes risk of m.

Denote the optimal solution to the above problem as $\mathbf{m}^*=\{m_{rs}^*\}_{r\in\mathscr{R},s\in\mathscr{S}}$, which through (6) characterizes the corresponding multiple testing procedure which we denote by $M^* \in \mathscr{M}_{\mathscr{R}}$.

The constraint matrix for the above linear program is quite sparse, that is, a large fraction of its elements are 0. This is because for any $r \in \mathscr{R}$ the constraint (13) has only $|\mathscr{S}|$ nonzero elements, and for any $r \in \mathscr{R}$, $s \in \mathscr{S}$, the constraint (14) has only 1 nonzero element. The power constraint (12) and the familywise Type I error rate constraints (11) generally have many nonzero elements, but there are relatively few of these constraints compared to (13) and (14).

The coefficients in (11) and (12) can be computed by evaluating the bivariate normal probabilities $P_{\delta_1,\delta_2}[(Z_1,Z_2) \in r]$. This can be done with high precision, essentially instantaneously, by standard statistical software such as the pmvnorm function in the R package `mvtnorm`. For each $r \in \mathscr{R}$, $s \in \mathscr{S}$, the term in curly braces in the objective function (10) can be computed by numerical integration over $(\delta_1, \delta_2) \in \mathbb{R}^2$ with respect to the prior

distribution Λ. We give R code implementing this in the Supplementary Materials. The minimax version (5) of the optimization problem from Section 3.2 can be similarly represented as a large, sparse linear program, as described in Section J of the Supplementary Materials. We show in Section 9 how to efficiently solve the resulting discretized problems using advanced optimization tools.

5 Application to HIV Example in Section 2

5.1 Solution to Optimization Problem in Four Special Cases

We illustrate our method by solving special cases of the constrained Bayes optimization problem. We use the loss function $\tilde{L}$ defined in Section 3.2. The risk corresponding to $\tilde{L}$ has an interpretation in terms of power to reject subpopulation null hypotheses. We define the power of a procedure to reject a null hypothesis $H \in \mathscr{H}$ as the probability it rejects at least H (and possibly other null hypotheses). For any non-centrality parameters $\delta_1 \geq \delta_1^{\min}, \delta_2 < \delta_2^{\min}$ and any $M \in \mathscr{M}_{\mathscr{R}}$, the risk $E_{\delta_1,\delta_2} \tilde{L}(M(Z_1,Z_2, U); \delta_1, \delta_2)$ equals one minus the power of M to reject H_{01} under (δ_1, δ_2); an analogous statement holds for subpopulation 2. For $\delta_1 \geq \delta_1^{\min}, \delta_2 < \delta_2^{\min}$, the risk equals the sum of one minus the power to reject each subpopulation null hypothesis.

We specify the following prior on the non-centrality parameters $(\delta_1, \delta_2) : \Lambda = \sum_{j=1}^{4} w_j \lambda_j$, where $\mathbf{w} = (w_1, w_2, w_3, w_4)$ is a vector of weights. Let λ_1, λ_2, λ_3, λ_4 be point masses at $(0,0), (\delta_1^{\min}, 0), (0, \delta_2^{\min})$, and $(\delta_1^{\min}, \delta_2^{\min})$ respectively. We consider two cases below. In the first, called the symmetric case, we set the subpopulation proportions $p_1 = p_2 = 1/2$ and use the symmetric prior Λ_1 defined by weights $\mathbf{w}^{(1)} = (0.25, 0.25, 0.25, 0.25)$. In the second, called the asymmetric case, we set $p_1 = 0.63$ and use the prior Λ_2 defined by weights $\mathbf{w}^{(2)} = (0.2, 0.35, 0.1, 0.35)$; this case is motivated by the example in Section 2, where subpopulation 1 is 63% of the total population and is believed to have a greater likelihood of benefiting from treatment than subpopulation 2. In Section D of the Supplementary Materials, we give examples using a continuous prior distribution on $\mathbb{R}^2$.

For each case, we solved the corresponding linear program using the algorithm in Section 9. The dimensions of the rectangles in the discretization are set at $\tau = (0.02, 0.02)$, and we set $b = 5$. We describe how G' is determined in Section 6.2. Each discretized linear program has approximately 1.5 million variables and 1.8 million constraints; all but a couple hundred constraints are sparse. We give the precise structure of this linear program in Section 9.

We set $\alpha = 0.05$ and set each variance σ_{ka}^2 to be a common value σ^2. Let $M_{H_{0C}}^{\mathrm{UMP}}$ denote the uniformly most powerful test of the single null hypothesis H_{0C} at level α, which rejects H_{0C} if (Z_1,Z_2) is in the region $R_{\mathrm{UMP}} = \{(z_1, z_2) : \rho_1 z_1 + \rho_2 z_2 > \Phi^{-1}(1 - \alpha)\}$, for Φ the standard normal cumulative distribution function. To allow a direct comparison with $M_{H_{0C}}^{\mathrm{UMP}}$, we set the total sample size n equal to $n_{\min}$, defined to be the minimum sample size such that $M_{H_{0C}}^{\mathrm{UMP}}$ has 90% power to reject H_{0C} when the treatment benefit in both populations equals $\Delta^{\min}$. We round all results to two decimal places.

Consider the symmetric case. Let $\mathbf{m}^*_{\text{sym}}(1-\beta)$ denote the solution to the discretized problem at H_{0C} power constraint 1 − β. For 1 − β = 0.9, any multiple testing procedure that satisfies the power constraint (4) and the familywise Type I error constraint (3) at the global null hypothesis $(\delta_1, \delta_2) = (0, 0)$ must reject H_{0C} whenever $(Z_1,Z_2) \in R_{\text{UMP}}$ and cannot reject any null hypothesis when $(Z_1,Z_2) \notin R_{\text{UMP}}$, except possibly on a set of Lebesgue measure zero; this follows from Theorem 3.2.1 of Lehmann and Romano (2005). Since this must hold for the optimal procedure $\mathbf{m}^*_{\text{sym}}(0,9)$ what remains to be determined is what regions in R_{UMP} correspond to $\mathbf{m}^*_{\text{sym}}(0,9)$ rejecting H_{01},H_{02}, both, or neither. The rejection regions for $\mathbf{m}^*_{\text{sym}}(0,9)$, computed using our method, are depicted in Figure 1a. For each $s \in \mathscr{S}$, the region where $\mathbf{m}^*_{\text{sym}}(0,9)$ rejects precisely s is shown in a different color.

Consider weakening the H_{0C} power constraint from 1 − β = 0.9 to 0.88. The optimal solution $\mathbf{m}^*_{\text{sym}}(0.88)$ is shown in Figure 1b. Unlike $\mathbf{m}^*_{\text{sym}}(0.9)$, the procedure $\mathbf{m}^*_{\text{sym}}(0.88)$ has substantial regions outside R_{UMP} where it rejects a single subpopulation null hypothesis. However, there is a small region in R_{UMP} where $\mathbf{m}^*_{\text{sym}}(0.88)$ does not reject any null hypothesis. Also, in some parts of R_{UMP} corresponding to one z-statistic being large and positive while the other is negative, $\mathbf{m}^*_{\text{sym}}(0.88)$ only rejects the null hypothesis corresponding to the large z-statistic, while $\mathbf{m}^*_{\text{sym}}(0.9)$ rejects both this and H_{0C}.

The optimal solutions $\mathbf{m}^*_{\text{sym}}(0.9)$ and $\mathbf{m}^*_{\text{sym}}(0.88)$ illustrate a tradeoff between power for H_{0C} and for H_{01},H_{02}, as shown in the first two columns of Table 1. For each procedure, the first row gives one minus the Bayes risk, which is a weighted sum of power under the three alternatives $(\delta_1^{\min}, 0), (0, \delta_2^{\min})$, and $(\delta_1^{\min}, \delta_2^{\min})$; these alternatives correspond to the treatment only benefiting subpopulation 1, only benefiting subpopulation 2, and benefiting both subpopulations, respectively, at the minimum, clinically meaningful level. The contributions from each of these are given in rows 2–4 of Table 1. There is no contribution from the alternative (0, 0) since the loss function $\tilde{L}$ is identically zero there.

The upshot is that using the procedure $\mathbf{m}^*_{\text{sym}}(0.88)$ in place of $\mathbf{m}^*_{\text{sym}}(0.9)$ involves sacrificing 2% power for H_{0C} at $(\delta_1^{\min}, \delta_2^{\min})$, but gaining 11% power to reject H_{01} at $(\delta_1^{\min}, 0)$ plus an identical increase in power to reject H_{02} at $(0, \delta_2^{\min})$. We further discuss this tradeoff over a range of β values in Section 5.3.

Next consider the asymmetric case, corresponding to $p_1 = 0.63$ and prior Λ_2. Let $\mathbf{m}^*_{\text{asym}}(1-\beta)$ denote the solution to the corresponding discretized problem at 1 − β. Figures 1c and 1d show the optimal solutions $\mathbf{m}^*_{\text{asym}}(0.9)$ and $\mathbf{m}^*_{\text{asym}}(0.88)$. The main difference between these and the solutions for the symmetric case is that $\mathbf{m}^*_{\text{asym}}(0.9)$ and $\mathbf{m}^*_{\text{asym}}(0.88)$ have larger rejection regions for H_{01} and smaller rejection regions for H_{02}. The power tradeoff between $\mathbf{m}^*_{\text{asym}}(0.9)$ and $\mathbf{m}^*_{\text{asym}}(0.88)$ is given in the last two columns of Table 1. Sacrificing 2% power for H_{0C} at $(\delta_1^{\min}, \delta_2^{\min})$ leads to an increase in 12% power to reject H_{01} at $(\delta_1^{\min}, 0)$, and an increase in 5% power to reject H_{02} at $(0, \delta_2^{\min})$.

5.2 Monotonicity Properties and Approach for Verifying (3)

Let $\mathscr{M}^*$ denote the set $\{\mathbf{m}^*_{\text{sym}}(0.9), \mathbf{m}^*_{\text{asym}}(0.9), \mathbf{m}^*_{\text{sym}}(0.88), \mathbf{m}^*_{\text{asym}}(0.88)\}$. Each procesdure $M^* \in \mathscr{M}^*$ satisfies properties that we define next. For any $M \in \mathscr{M}_{\text{det}}$ and any $R \subseteq \mathbb{R}^2$, define the following monotonicity properties with respect to R: for any $(z_1, z_2) \in R$,

- **a.** if $H_{01} \in M(z_1, z_2)$, then $H_{01} \in M(z_1', z_2)$ for any $(z_1', z_2) \in R$ for which $z_1' \geq z_1$;
- **b.** if $H_{02} \in M(z_1, z_2)$, then $H_{01} \in M(z_1, z_2')$ for any $(z_1, z_2') \in R$ for which $z_1' \geq z_1, z_2' \geq z_2$;
- **c.** if $H_{0C} \in M(z_1, z_2)$, then $H_{0C} \in M(z_1', z_2')$ for any (z_1', z_2') for which $z_1' \geq z_1, z_2' \geq z_2$;
- **d.** if $M(z_1, z_2) \neq \emptyset$, then $M(z_1+x, z_2+x) \neq \emptyset$ for any $x > 0$ such that $(z_1+x, z_2+x) \in R$.

We verified that each procedure $M^* \in \mathscr{M}^*$ satisfies all of these monotonicity properties with respect to the region $R = B$. These properties are intuitively appealing. Also, they simplify the process of verifying all the familywise Type I error constraints (3) of the original problem; below, we give an overview of the main steps involved in verifying this for the procedures $\mathscr{M}^*$. The full argument is given in Section H of the Supplementary Materials, including the proof of the following theorem:

Theorem 1: (a.) For any $M \in \mathscr{M}_{det} \cap \mathscr{M}_B$ that satisfies (a)–(d) with respect to $R = B$,

$$\sup_{(\delta_1,\delta_2)\in\mathbb{R}^2} P_{\delta_1,\delta_2}[M(Z_1, Z_2) \cap \mathscr{H}_{TRUE}(\delta_1, \delta_2) \neq \emptyset] = \sup_{(\delta_1,\delta_2)\in\mathbb{R}^2} P_{\delta_1,\delta_2}[M(Z_1, Z_2) \cap \mathscr{H}_{TRUE}(\delta_1, \delta_2) \neq \emptyset]. \quad (15)$$

(b.) For any $M \in \mathscr{M}_{det}$ that satisfies (a)–(d) with respect to $R = \mathbb{R}^2$, (15) holds.

By part (a) of Theorem 1, to verify the familywise Type I error constraints (3) of the original problem for all $(\delta_1, \delta_2) \in \mathbb{R}^2$, it suffices to check the constraints for all $(\delta_1, \delta_2) \in G$. We check these latter constraints by first partitioning G into multiple regions. For each region that is sufficiently far from B, we directly prove (3) holds over that region, using that $\mathscr{M}^* \subseteq \mathscr{M}_B$. Each of the remaining regions is discretized, and we compute the familywise Type I error at each point in the discretization; we combine this with an analytic bound on the maximum possible discrepancy between the familywise Type I error rate at any point in that region, and the familywise Type I error rate at the nearest point in the discretization.

To upper bound (15) by 0.05 using this approach, it was necessary to solve the discretized problems at $\alpha = 0.05 - 10^{-4}$. This reduction from 0.05 had a negligible effect on the Bayes risk of the resulting procedures, as described in Section H of the Supplementary Materials.

5.3 Optimal Power Tradeoff for Combined Population versus Subpopulations

We explore the tradeoffs in power for rejecting a subpopulation null hypothesis when the treatment only benefits one subpopulation, versus power for rejecting the combined population null hypothesis when the treatment benefits both subpopulations. Figure 2 shows the Bayes risk and its components for the optimal procedure $\mathbf{m}^*_{\text{sym}}(1-\beta)$, for each value of

$1 - \beta$ in a grid of points on the interval [0.8, 0.9], for the symmetric case. The solid curve in Figure 2a gives the optimal tradeoff between the Bayes risk and the constraint $1 - \beta$ on the power to reject H_{0C} at $(\delta_1^{\min}, \delta_2^{\min})$. Figures 2b–d show the contribution to the Bayes risk from power under the three alternatives $(\delta_1^{\min}, 0)$, $(0, \delta_2^{\min})$, and $(\delta_1^{\min}, \delta_2^{\min})$.

In each plot, we included points corresponding to $\mathbf{m}^*_{\text{sym}}(0.9)$ and $\mathbf{m}^*_{\text{sym}}(0.88)$, as well as three existing multiple testing procedures. The first is a procedure of Rosenbaum (2008) that rejects H_{0C} when $M^{\text{UMP}}_{H_{0C}}$ does, and if so, additionally rejects each subpopulation null hypothesis H_{0k} for which $Z_k > \Phi^{-1}(1 - \alpha)$. The second existing method is an improvement on the Bonferroni and Holm procedures by Bergmann and Hommel (1988) for families of hypotheses that are logically related, as is the case here. The third is a special case of the method of Song and Chi (2007) that trades off power for H_{0C} to increase power for H_{01}, we augmented their procedure to additionally reject H_{02} in some cases. The details of the latter two procedures are given in Section E of the Supplementary Materials. Each of the three existing procedures strongly controls the familywise Type I error rate at level α.

The procedure of Rosenbaum (2008) is quite close to the optimal threshold at $1 - \beta = 0.9$, being suboptimal compared to $\mathbf{m}^*_{\text{sym}}(0.9)$ by only 0.4% in terms of the Bayes risk; the corresponding rejection regions are very similar to those of $\mathbf{m}^*_{\text{sym}}(0.9)$. The procedure of Bergmann and Hommel (1988) is suboptimal by 5% in power for rejecting H_{01} at $(\delta_1^{\min}, 0)$ and for rejecting H_{02} at $(0, \delta_2^{\min})$. The procedure of Song and Chi (2007) is close to optimal for rejecting H_{01} at $(\delta_1^{\min}, 0)$, but is 9% suboptimal for H_{02} at $(0, \delta_2^{\min})$. This is not surprising since their procedure was designed with a focus on the null hypothesis for a single subpopulation, rather than for both a subpopulation and its complement.

The tradeoff curves are steep near $1 - \beta = 0.9$, indicating that a small sacrifice in power to reject H_{0C} at $(\delta_1^{\min}, \delta_2^{\min})$ leads to a relatively large gain in power to detect subpopulation treatment effects when the treatment benefits only one subpopulation. The first two columns of Table 1, which compare $\mathbf{m}^*_{\text{sym}}(0.9)$ versus $\mathbf{m}^*_{\text{sym}}(0.88)$, are an example of this tradeoff. Diminishing returns set in for $1 - \beta$ less than 0.84, in that there is negligible improvement in the Bayes risk or any of its components if one further relaxes the power constraint for H_{0C}.

Consider the impact of increasing the total sample size n above $n_{\min}$, holding $\Delta^{\min}$ and the variances σ^2_{sa} fixed. Define the multiple testing procedure $\mathbf{m}^*_{SS}(n)$ to be the solution to the discretized optimization problem in the symmetric case at $1 - \beta = 0.9$ and sample size n, for $n \geq n_{\min}$. As n increases from $n_{\min}$, the rejection regions of $\mathbf{m}^*_{SS}(n)$ progress from $\mathbf{m}^*_{\text{sym}}(0.9)$ as in Figure 1a to rejection regions qualitatively similar to $\mathbf{m}^*_{\text{sym}}(0.88)$ as in Figure 1b; these regions are given in Section F of the Supplementary Materials. Increasing sample size from $n = n_{\min}$ to $n = 1.06 n_{\min}$, the power of $\mathbf{m}^*_{SS}(n)$ to reject H_{01} at $(\delta_1^{\min}, 0)$ increases from 42% to 52%; there is an identical increase in power to reject H_{02} at $(0, \delta_2^{\min})$.

To give a sense of the value of increasing power from 42% to 52%, consider testing the single null hypothesis H_{01} based on Z_1, using the uniformly most powerful test of H_{01} at

level α'. Consider the sample size for which the power of this test is 42% at a fixed alternative $\Delta_1 = \Delta_1^{\min} > 0 \text{ and } \sigma_{10}^2 > 0, \sigma_{11}^2 > 0$. To increase power to 52%, one needs to increase the sample size by 38%, 31%, or 28%, for α' equal to 0.05, 0.05/2 or 0.05/3, respectively. In light of this, the above 10% gains in power for detecting subpopulation treatment effects at the cost of only a 6% increase in sample size (and while maintaining 90% power for H_{0C}), as $\mathbf{m}_{SS}^*(n)$ does, is a relatively good bargain.

The tradeoff curve in Figure 2a is optimal, i.e., no multiple testing procedure satisfying the familywise Type I error constraints (3) can have Bayes risk and power for H_{0C} corresponding to a point that exceeds this curve. The Bayes risk is a weighted combination of power at the three alternatives given above, as shown in Figures 2b–d. It follows that no multiple testing procedure satisfying (3) can simultaneously exceed all three power curves in Figures 2b–d. However, there do exist procedures that have power greater than one or two of these curves but that fall short on the other(s). By solving the constrained Bayes optimization problem using different priors Λ, one can produce examples of such procedures.

A similar pattern as in Figure 2 holds for the asymmetric case. The main difference is that power to reject H_{01} at $(\delta_1^{\min}, 0)$ is larger than power to reject H_{02} at $(0, \delta_2^{\min})$. In Section F of the Supplementary Materials, we answer the question posed in Section 1 of what minimum additional sample size is required to achieve a given power for detecting treatment effects in each subpopulation, while maintaining 90% power for H_{0C} and strongly controlling the familywise Type I error rate. We do this for $p_1 = p_2$, but the general method can be applied to any subpopulation proportions.

6 Using the Dual of the Discretized Problem to Bound the Bayes Risk of the Original Problem

6.1 Active Constraints in the Dual Solution of the Discretized Problem

For each optimal procedure from Section 5.1, Figure 3 shows the constraints among (11) and (12) that are active, i.e., for which the corresponding inequalities hold with equality. In all cases, the global null hypothesis $(\delta_1, \delta_2) = (0, 0)$, the power constraint (12), and one constraint on the boundary of the null space for each of H_{01} and H_{02}, are active. In addition, each of the optimal procedures at $1 - \beta = 0.88$ has two active constraints on the boundary of the null space for H_{0C}. The active familywise Type I error constraints correspond to the least-favorable distributions for a given procedure.

To illustrate the importance of all these constraints, consider what would happen if we only imposed the familywise Type I error constraint (3) at the global null hypothesis and the power constraint (4) at $1 - \beta = 0.88$. The optimal solution to the corresponding constrained Bayes optimization problem in the symmetric case has familywise Type I error 0.54 at non-centrality parameters $(\delta_1^{\min}, 0)$ and $(0, \delta_2^{\min})$; in the asymmetric case, the familywise Type I error at each of these alternatives is 0.39 and 0.69, respectively. The rejection regions are given in Section G of the Supplementary Materials. This demonstrates the importance of the additional familywise Type I error constraints.

As described in Section 5.2, the optimal solutions to the discretized problems in Section 5.1 satisfy all constraints (3); this holds despite our not having imposed the constraints (3) for $(\delta_1, \delta_2) \in G \setminus B$. Intuitively, this is because the optimal solutions are driven by the active constraints, all of which are contained in B for the value $b = 5$ used in defining our discretized problem.

6.2 Improving Accuracy by More Closely Approximating the Active Constraints

For each example in Section 5.1, we first solved the discretized problem at an initial, relatively coarse discretization, where we set $b = 5$, $\tau_1 = \tau_2 = 0.1$, and $G' = G_{\tau,b}$. The locations of active constraints in the resulting solution were then used to construct a new, more focused set $G'_{new} \subset G$ of constraints (3). Specifically, for each active familywise Type I error constraint (δ_1, δ_2) from the solution at the initial discretization, we included in G'_{new} a high concentration of points along a small line segment in G containing (δ_1, δ_2); we did not include any other points. The motivation was to simultaneously obtain closer approximations to the active constraints of the original problem, and to reduce the total number of constraints. We then solved the discretized problem at the finer discretization $b = 5$, $\tau_1 = \tau_2 = 0.02$, using familywise Type I error constraints $G' = G'_{new}$. The set G'_{new} for each example from Section 5.1 is given in Section K of the Supplementary Materials. As one example, the number of constraints in G'_{new} is 106 for the symmetric case at $1 - \beta = 0.88$.

6.3 Bounding the Bayes Risk of the Optimal Solution to the Original Problem

The optimal multiple testing procedures shown in Figure 1 are the solutions to versions of the discretized problem (10)–(14), which is an approximation to the constrained Bayes optimization problem (2)–(4). We refer to the latter as the original problem. A natural question is how the optimal Bayes risk for the discretized problem compares to the optimal Bayes risk achievable in the original problem.

We use the optimal solution ν^* to the dual of the discretized problem to obtain a lower bound on optimal Bayes risk of the original problem. For a given discretized problem and optimal dual solution ν^*, let C_{FWER} denote the set of indices of active familywise Type I error constraints among (11); these are the indices j of the pairs $(\delta_{1,j}, \delta_{2,j}) \in G'$ corresponding to the nonzero components ν_j^* of ν^*. Let ν_p^* denote the value of the dual variable corresponding to the power constraint (12). Let $\mathscr{M}_c$ denote the subclass of multiple testing procedures in $\mathscr{M}$ that satisfy all the constraints (3) and (4) of the original problem. Then we have the following lower bound on the objective function (2) of the original problem:

$$\inf_{M \in \mathscr{M}_c} \int E_{\delta_1,\delta_2} L(M(Z_1, Z_2, U); \delta_1, \delta_2) d\Lambda(\delta_1, \delta_2) \quad (16)$$

$$\geq \inf_{M \in \mathscr{M}} \left[\int E_{\delta_1,\delta_2} L(M(Z_1, Z_2, U); \delta_1, \delta_2) d\Lambda(\delta_1, \delta_2) + \nu_p^* \{1 - \beta - P_{\delta_1^{\min}, \delta_2^{\min}}(M \text{rejects} H_{0C})\} + \sum_{j \in C_{\text{FWER}}} \nu_j^* \{P_{\delta_{1,j},\delta_{2,j}}(M \text{rejects any n}$$

which follows since all components of ν* are nonnegative, by definition. The minimization problem (17) is straightforward to solve since it is unconstrained. We give the solution in Section I of the Supplementary Materials, which is computed by numerical integration. We then computed the absolute value of the difference between this lower bound and the Bayes risk of the optimal solution to the discretized problems in Section 5.1, which is at most 0.005 in each case. This shows the Bayes risk for the optimal solution to each discretized problem is within 0.005 of the optimum achievable in the original problem, so little is lost by restricting to the discretized procedures at the level of discretization we used.

7 Computational Challenge and Our Approach to Solving It

Previous methods, such as those of Jennison (1987); Eales and Jennison (1992); and Banerjee and Tsiatis (2006) are designed to test a null hypothesis for a single population. These methods require specifying one or two constraints that include the active constraints for a given problem. This can be done for a single population since often the global null hypothesis of zero treatment effect and a single power constraint suffice. However, as shown in the previous section, in our problem there can be 6 active constraints in cases of interest. Especially in the asymmetric case shown in Figure 3d, it would be difficult to a priori guess this set of constraints or to do an exhaustive search over all subsets of 6 constraints in G'. Even if the set of active constraints C_{FWER} for our problems from Section 5.1 were somehow known or correctly guessed, the problems could still be challenging to solve using standard optimization methods such as Lagrange multipliers. We discuss this in Section L of the Supplementary Materials.

Our approach overcomes the above computational obstacles by transforming a fine discretization of the original problem to a sparse linear program that contains many constraints; we then leverage the machinery of linear program solvers, which are expressly designed to optimize under many constraints simultaneously. The sparsity of the constraint matrix of the discretized linear program is crucial to the computational feasibility of our approach. This sparsity results from being able to a priori specify a subset G' of the familywise Type I error constraints that contains close approximations to the active constraints, where G' is not so large as to make the resulting linear program computationally intractable. The size of G' in the examples from Section 5.1 and in the examples in the Supplementary Materials was never more than 344. More generally our method is computationally feasible with G' having up to a thousand constraints.

8 Application to Decision Theory Framework

A drawback of the hypothesis testing framework when considering subpopulations is that it does not directly translate into clear treatment recommendations. For example, if the null hypotheses H_{0C} and H_{01} are rejected, it is not clear whether to recommend the treatment to subpopulation 2. We propose a decision theory framework that formalizes the goal of recommending treatments to precisely the subpopulations who benefit at a clinically meaningful level. The framework allows one to explore tradeoffs in prioritizing different types of errors in treatment recommendations to different subpopulations. The resulting

optimization problems, which were not solvable previously, are solved using our general approach.

We use the definitions in Section 3.1. Our goal is to construct a decision procedure D, i.e., a measurable map from any possible realization of (Z_1,Z_2) to a set of subpopulations ($\emptyset$, {1}, {2}, or {1, 2}) to recommend the new treatment to. We consider randomized decision procedures, i.e., we allow D to additionally depend on a random variable U that is independent of Z_1,Z_2 and that has uniform distribution on [0, 1].

We next define a class of loss functions. For each subpopulation $k \in \{1, 2\}$, let $l_{k,FP}$ be a user-defined penalty for recommending the treatment to subpopulation k when $\delta_k < \delta_k^{\min}$ (a False Positive); let $l_{k,FN}$ be the penalty for failing to recommend the treatment to subpopulation k when $\delta_k \geq \delta_k^{\min}$ (a False Negative). Define the loss function $L_D(d; \delta_1, \delta_2) = L_{D,1}(d; \delta_1, \delta_2) + L_{D,2}(d; \delta_1, \delta_2)$, where for each $d \subseteq \{1, 2\}$ and $k \in \{1, 2\}$, $L_{D,k}(d;\delta_1, \delta_2) = l_{k,FP} 1[\delta_k < \delta_k^{\min}, k \in d] + l_{k,FN} 1[\delta_k \geq \delta_k^{\min}, k \notin d]$. For illustration, we consider two loss functions. The first, $L_D^{(1)}$, is defined by $l_{k,FN} = 1$ and $l_{k,FP} = 2$ for each k; the second, $L_D^{(2)}$, is defined by $l_{\mathrm{k,FN}} = 2$ and $l_{k,FP} = 1$ for each k.

We minimize the Bayes criterion $\int E_{\delta_1,\delta_2} \{L(D(Z_1,Z_2, U); \delta_1, \delta_2)\}\, d\Lambda(\delta_1, \delta_2)$, over all decision procedures D as defined above, under the constraints that for any $(\delta_1, \delta_2) \in \mathbb{R}^2$, $P_{\delta_1,\delta_2} \{\sum_{k \in D(Z_1,Z_2,U)} p_k \Delta_k \leq 0\} \leq \alpha\}$ These constraints impose a bound of α on the probability of recommending the new treatment to an aggregate population (defined as the corresponding single subpopulation if $D = \{1\}$ or $\{2\}$, or the combined population if $D = \{1, 2\}$) having no average treatment benefit.

We consider the symmetric case from Section 5.1. The optimal decision regions are given in Figure 4. The optimal decision rule under $L_D^{(1)}$, denoted by $D^{(1)*}$, is more conservative in recommending the treatment than the optimal rule under $L_D^{(2)}$, denoted by $D^{(2)*}$. This is because the former loss function penalizes more for false positive recommendations. Table 2 contrasts $D^{(1)*}$ and $D^{(2)*}$. When $(\delta_1, \delta_2) = (\delta_1^{\min}, \delta_2^{\min})$, the conservative rule $D^{(1)*}$ recommends treatment to both subpopulations 21% less often compared to $D^{(2)*}$. However, when the treatment only benefits one subpopulation, the conservative rule $D^{(1)*}$ has 11% greater accuracy in recommending it to just that subpopulation.

One may prefer to strengthen the constraints above to require for any $(\delta_1, \delta_2) \in \mathbb{R}^2$, $P_{\delta_1,\delta_2} [D(Z_1,Z_2, U) \cap \{k : \delta_k \leq 0\} \neq \emptyset] \leq \alpha$, that is, to require probability at most α of recommending the new treatment to any subpopulation having no average treatment benefit. Our framework allows computation of the tradeoff between optimal procedures under these different sets of constraints, which is an area of future research.

9 Algorithm to Solve Our Large, Sparse Linear Programs

The discretized problem from Section 4 can be represented as a large-scale linear programming problem. To show this, define the following ordering of subsets of $\mathscr{H}$:

$$\mathscr{S}'=(s_0,\ldots,s_6)=(\varnothing,\{H_{01}\},\{H_{02}\},\{H_{0C}\},\{H_{01},H_{0C}\},\{H_{02},H_{0C}\},\{H_{01},H_{02},H_{0C}\}).$$

We leave out the subset $\{H_{01},H_{02}\}$, since by the results of Sonnemann and Finner (1988) it suffices to consider only coherent multiple testing procedures, which in our context are those that reject H_{0C} whenever $\{H_{01},H_{02}\}$ is rejected. For a given ordering $r_1, r_2, \ldots$ of the rectangles $\mathscr{R}$, define $\mathbf{x} = (m_{r_1s_1}, \ldots, m_{r_1s_6}, m_{r_2s_1}, \ldots, m_{r_2s_6}, m_{r_3s_1} \ldots)$, which has $n_\upsilon = |\mathscr{R}|(|\mathscr{S}'| - 1)$ components. We do not include the variables $m_{r_is_0}$ in $\mathbf{x}$, since by (13) these variables are functions of variables already in $\mathbf{x}$; in particular, $m_{r_is_0}=1-\sum_{j=1}^{6} m_{r_is_j}$.

The discretized problem from Section 4 can be expressed in the canonical form:

$$\max_{x\in\mathbb{R}^{n_\upsilon}} \mathbf{c}^{\mathbf{T}}\mathbf{x}\text{s.t.}\mathbf{A}\mathbf{x}\le\mathbf{b}. \quad (18)$$

The objective function $\mathbf{c}^{\mathbf{T}}\mathbf{x}$ represents the Bayes objective function (10). We set the first $n_d = |G\curlywedge + 1$ rows of $\mathbf{A}$ to comprise the dense constraints, which include the familywise Type I error constraints (11) and the H_{0C} power constraint (12). The remaining ns rows of $\mathbf{A}$ comprise the sparse constraints (13) and (14). Since $|\mathscr{R}| = (2b/\tau + 1)^2$, for the symmetric case at $1-\beta = 0.88$ in Section 5.1 with $b = 5$, $\tau = \tau_1 = \tau_2 = 0.02$, and $|\mathscr{S}'| - 1 = 6$, we have $n_\upsilon = |\mathscr{R}|(|\mathscr{S}'| - 1) = 1{,}506{,}006$, $n_d = |G\curlywedge + 1 = 106$ (where $G'=G'_{new}$ defined in Section 6.2), and $n_s = |\mathscr{R}| + n_\upsilon = 1{,}757{,}007$. Then $\mathbf{A}$ is a 1,757,113 × 1,506,006 matrix with structure:

$$\mathbf{A}=\left[\begin{array}{c} \overbrace{\qquad\qquad}^{n_d=106\text{rows in which most elements are non-zero}} \\ \left.\begin{array}{l} 1111110000000000000000000000\ldots \\ 0000001111110000000000000000\ldots \\ 0000000000001111110000000000\ldots \\ \vdots \end{array}\right\} |\Re|=251{,}001\text{rows of the form:} \\ \hline -\mathbf{I}_{1,506,006},\text{that is, the negative}n_\upsilon\times n_\upsilon\text{identify matrix} \end{array}\right],$$

$\mathbf{b}$ is a vector with $n_d + n_s = 1{,}757{,}113$ components (comp.) as follows:

$$\mathbf{b}^T=\begin{pmatrix} |G'|=105\text{comp.} & 1\text{comp.} & |\Re|=251{,}001\text{comp.} & n_\upsilon=1{,}506{,}006\text{comp.} \\ \alpha,\alpha,\ldots,\alpha, & -(-1-\beta), & 1,1,\ldots,1 & 0,0,\ldots,0 \end{pmatrix},$$

and $\mathbf{c}$ is a vector with $n_\upsilon = 1{,}506{,}006$ components.

The problem scale of (18) is quite large. In particular, the constraint matrix $\mathbf{A}$ has $\approx 2.6 \times 10^{12}$ entries. However, we can solve (18) by exploiting the sparsity structure of $\mathbf{A}$. We use a projected subgradient descent method, which consists of a subgradient descent step and a projection step, where the solution at iteration $k + 1$ is $\mathbf{x}^{(k+1)} = P_s(\mathbf{x}^{(k)} - \delta_k\mathbf{g}^{(k)})$, where $P_s(.)$

means projection onto the feasible region determined by the sparse constraints, δ_k is a step size, and $\mathbf{g}^{(k)}$ is the subgradient of $\mathbf{x}_k$, defined as

$$\mathbf{g}^{(k)} = \begin{cases} \mathbf{c}, & \text{if for all} i=1,\ldots,n_d, \mathbf{a}_i^T\mathbf{x}^{(k)} \leq b_i, \\ -\mathbf{a}_{i'}, & \text{otherwise, where} i' \text{is a randomly selected index in} \{i: \mathbf{a}_i^T\mathbf{x}^{(k)} > b_i\}. \end{cases}$$

The projection operator $P_s(.)$ can be applied in $\mathcal{O}(n_\upsilon)$ oating point operations (ops) by computing the projection in $|\mathscr{R}|$ independent subsystems, each with $|\mathscr{S}'| - 1$ variables. Checking violations of n_d dense constraints together with the projection costs at most $\mathcal{O}(n_\upsilon(n_d+1))$ flops per iteration. The projected subgradient descent method above is guaranteed to converge to the optimum of (18) (Boyd et al., 2004). However, it may take a large number of iterations to achieve a high precision solution. In our implementation, we continue until an iteration k' is reached where the proportion improvement in the objective function value is smaller than 10^{-3}; we then use $\mathbf{x}^{(k')}$ as the initial point in a parametric simplex solver (Vanderbei, 2010). Though each iteration of a parametric simplex solver runs in superlinear time, for our problem it only requires a few iterations to move from $\mathbf{x}^{(k')}$ to a very precise optimal solution. Our solutions all had duality gap at most 10^{-8} showing they are within 10^{-8} of the true optimal solution to the discretized problem.

10 Discussion

An area of future research is to consider a variety of optimization criteria, and to find a multiple testing procedure (if one exists) that simultaneously has good performance under each criterion. For example, one may specify a finite set of pairs of loss functions and priors, with each pair determining an objective function of the form (2). Our general method can be adapted to minimize the maximum of these objective functions, under the constraints (3) and (4), as described in Section J of the Supplementary Materials.

Though the discretized problem involved optimizing over the class of randomized multiple testing procedures $\mathscr{M}_{\mathscr{R}}$, the optimal solutions in all our examples were in $\mathscr{M}_{\text{det}}$. This is interesting, since there is no a priori guarantee that there exists an optimal solution that is deterministic, since the problem involves the large class of constraints (11). If the optimal solution to a problem is not deterministic, it might be possible to learn from its structure to find a close approximation that is deterministic; this is an area for future research.

An important question posed by a reviewer is what to do if, for given L and Λ, the optimal solution to the constrained Bayes optimization problem does not have monotonicity properties (a)–(d). Then Theorem 1 would not apply, and the active constraints would not be guaranteed to be in G. How to handle this situation is an area for future research, but we briey describe two approaches that could be tried. The first approach is to augment G' to additionally include points (δ_1, δ_2) outside of G. For example, one could include a grid of points on the subset of B where at least one null hypothesis is true. Intuitively, if each active constraint in the original problem is closely approximated by a constraint in the discretized problem, one may expect the solution to the discretized problem to be approximately

feasible and optimal. A limitation is that the more constraints one adds, the more computationally difficult the discretized problem becomes.

A second way to handle the above situation is to restrict attention to the subclass of procedures that satisfy monotonicity properties (a)–(d). In Section M of the Supplementary Materials, we generalize the definitions of the these properties to randomized multiple testing procedures, and denote the subclass satisfying these properties by $\mathscr{W}_{mon} \subset \mathscr{W}$. We show our general method can be adapted to solve the constrained Bayes optimization problem restricted to procedures in $\mathscr{W}_{mon}$, by encoding each monotonicity property as a set of sparse constraints in the discretized problem. A limitation is that the optimal solution restricted to procedures $\mathscr{W}_{mon}$ may have worse performance compared to the optimal solution over $\mathscr{W}$.

An additional area of future research is to apply our methods to construct optimal testing procedures for trials comparing more than two treatments. Other potential applications include optimizing seamless Phase II/Phase III designs and adaptive enrichment designs.

Though we focused on two subpopulations, it may be possible to extend our approach to three or four subpopulations. This is an area for future research. However, with more than this many populations, our approach will likely be computationally infeasible. This is because the number of variables in the discretized linear program grows with the fineness of the discretization as well as the number of components in the sufficient statistic for the problem. One strategy for reducing the computational burden in larger problems is to start by solving the problem at a relatively coarse discretization; one can then use the structure of the resulting solution to inform where to set constraints when solving the problem at a finer discretization. An example of this strategy was used in Section 6.2.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Banerjee A, Tsiatis AA. Adaptive two-stage designs in phase II clinical trials. Statistics in Medicine. 2006; 25(19):3382–3395. [PubMed: 16479547]
2. Bergmann, B.; Hommel, G. Improvements of general multiple test procedures for redundant systems of hypotheses. In: Bauer, P.; Hommel, G.; Sonnemann, E., editors. Multiple Hypothesenprüfung–Multiple Hypotheses Testing. Springer: Berlin; 1988. p. 100-115.
3. Boyd S, Xiao L, Mutapcic A. Subgradient methods, Lecture notes of EE392o, Stanford University, Autumn Quarter. 2003–2004 http://www.stanford.edu/class/ee364b/notes/subgrad_method_notes.pdf.
4. Eales JD, Jennison C. An improved method for deriving optimal one-sided group sequential tests. Biometrika. 1992; 79(1):13–24.
5. Fätkenheuer G, Nelson M, Lazzarin A, Konourina I, Hoepelman AI, Lampiris H, Hirschel B, Tebas P, Raffi F, Trottier B, Bellos N, Saag M, Cooper DA, Westby M, Tawadrous M, Sullivan JF, Ridgway C, Dunne MW, Felstead S, Mayer H, van der Ryst E. Subgroup analyses of maraviroc in previously treated R5 HIV-1 infection. New England Journal of Medicine. 2008; 359(14):1442–1455. [PubMed: 18832245]

6. FDA and EMEA. E9 statistical principles for clinical trials. U.S. Food and Drug Administration: CDER/CBER. European Medicines Agency: CPMP/ICH/363/96. 1998. http://www.fda.gov/cder/guidance/index.htm.
7. Hampson LV, Jennison C. Group sequential tests for delayed responses. J. RStatist. Soc. B. 2013; 75(1):1–37.
8. Hochberg, Y.; Tamhane, AC. Multiple Comparison Procedures. New York: Wiley Inter-science; 1987.
9. Jennison C. Efficient group sequential tests with unpredictable group sizes. Biometrika. 1987; 74(1): 155–165.
10. Katlama C, Haubrich R, Lalezari J, Lazzarin A, Madruga JV, Molina J-M, Schechter M, Peeters M, Picchio G, Vingerhoets J, Woodfall B, De Smedt G. DUET-1, DUET-2 study groups. Efficacy and safety of etravirine in treatment-experienced, HIV-1 patients: pooled 48 week analysis of two randomized, controlled trials. AIDS. 2009; 23(17):2289–2300. [PubMed: 19710593]
11. Lehmann, EL.; Romano, JP. Testing Statistical Hypotheses. Springer: 2005.
12. Romano JP, Shaikh A, Wolf M. Consonance and the closure method in multiple testing. The International Journal of Biostatistics. 2011; 7(1)
13. Rosenbaum PR. Testing hypotheses in order. Biometrika. 2008; 95(1):248–252.
14. Sonnemann, E.; Finner, H. Vollständigkeitssätze für multiple testprobleme. In: Bauer, P.; Hommel, G.; Sonnemann, E., editors. Multiple Hypothesenprüfung. Springer: Berlin; 1988. p. 121-135.
15. Song Y, Chi GYH. A method for testing a prespecified subgroup in clinical trials. Statistics in Medicine. 2007; 26(19):3535–3549. [PubMed: 17266164]
16. Vanderbei, R. Linear Programming: Foundations and Extensions. Springer: 2010.

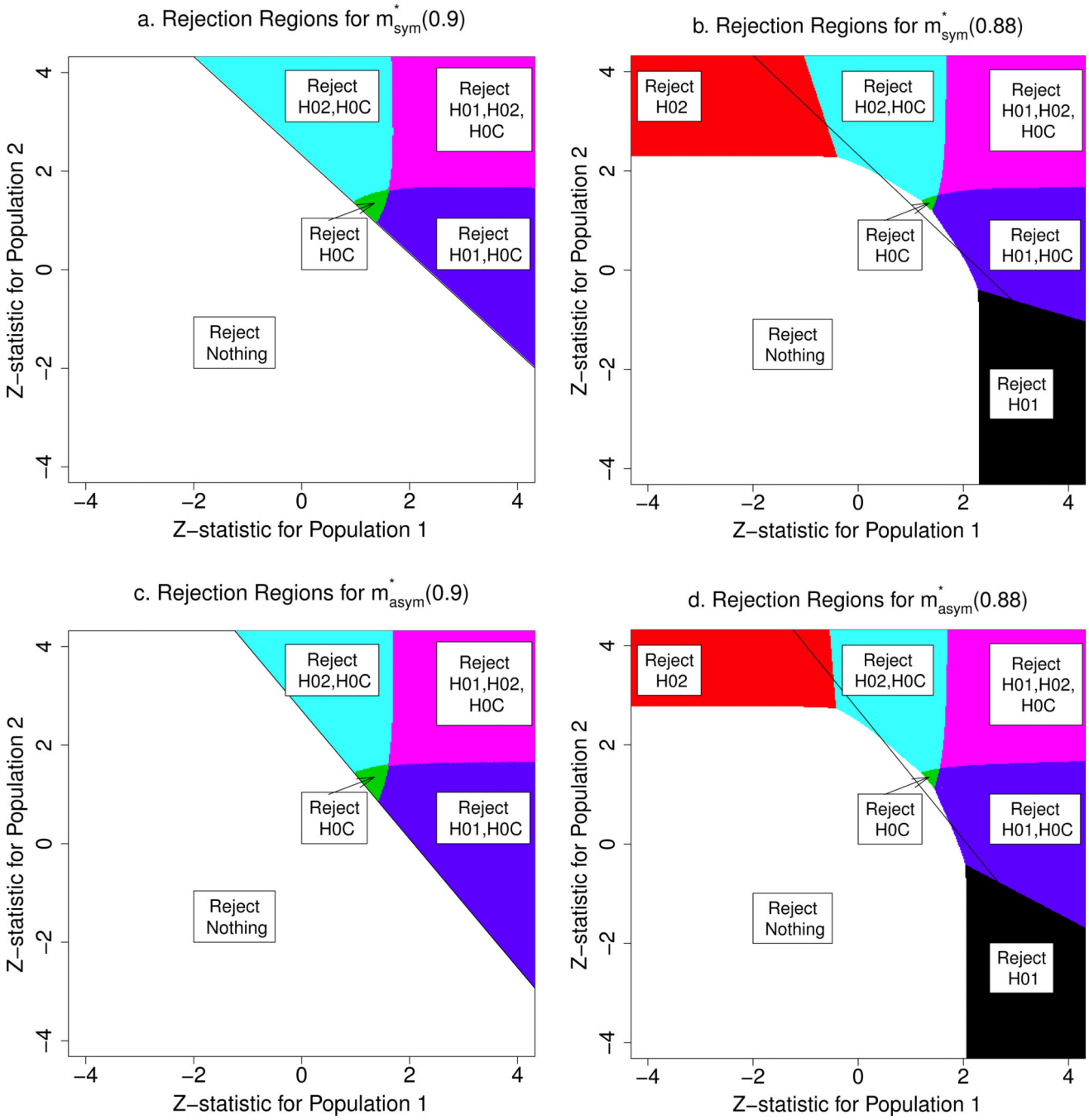

Figure 1.

Optimal multiple testing procedures, for the symmetric case (a) and (b), and for the asymmetric case (c) and (d). In each plot, the black line is the boundary of RUMP for the corresponding case.

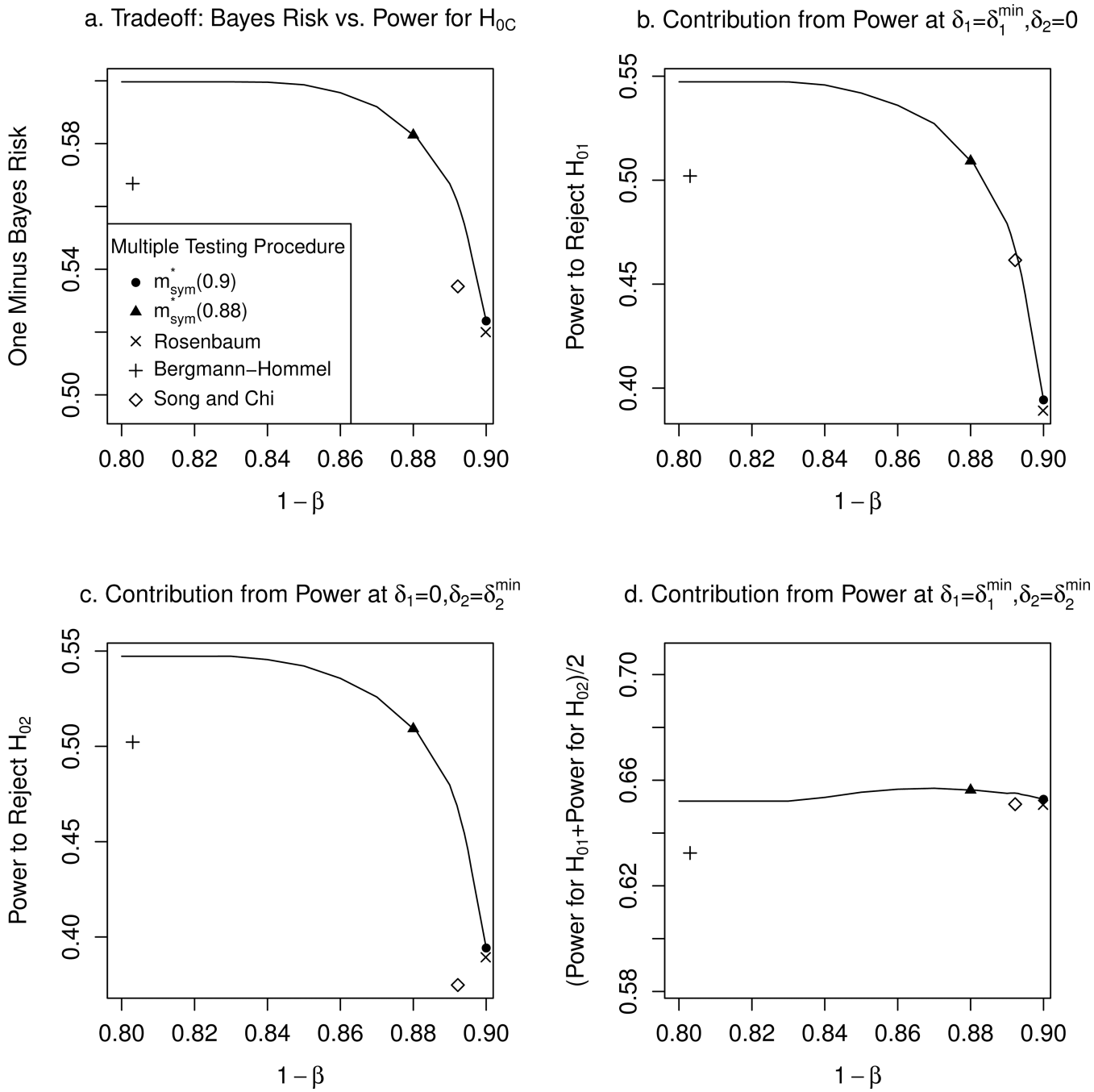

Figure 2.

Optimal tradeoff between Bayes risk and power constraint 1 − β on H_{0C}, for symmetric case, i.e., $p_1 = p_2 = 1/2$ and prior Λ_1. In (a), we give one minus the Bayes risk on the vertical axis, so that in all four plots above, larger values represent better performance.

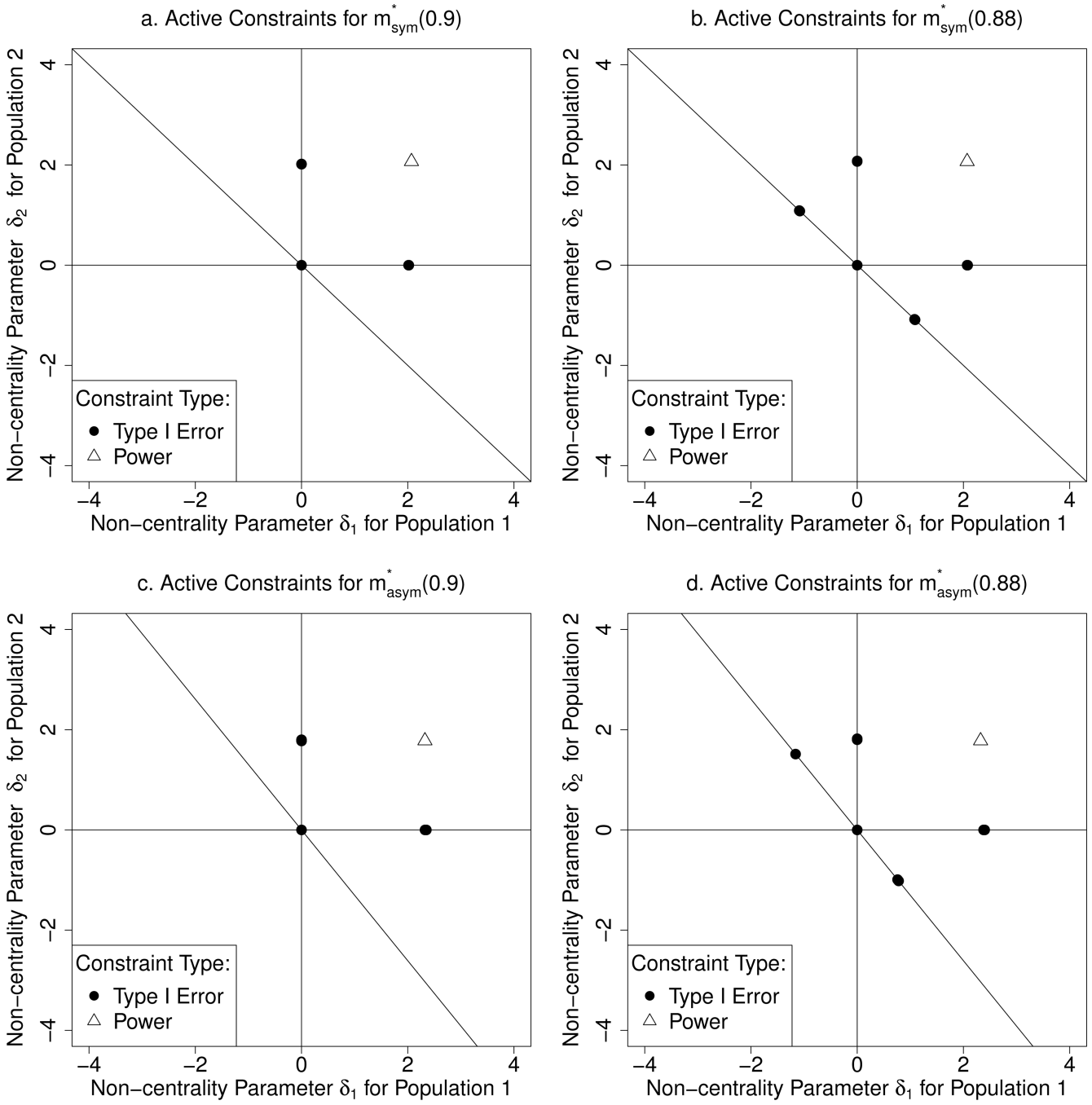

Figure 3.

Active constraints for optimal procedures $\mathbf{m}^*_{\mathrm{sym}}(0.9)$, $\mathbf{m}^*_{\mathrm{sym}}(0.88)$, $\mathbf{m}^*_{\mathrm{sym}}(0.9)$; and $\mathbf{m}^*_{\mathrm{asym}}(0.88)$. Lines indicate boundaries of null spaces for H_{01},H_{02},H_{0C}.

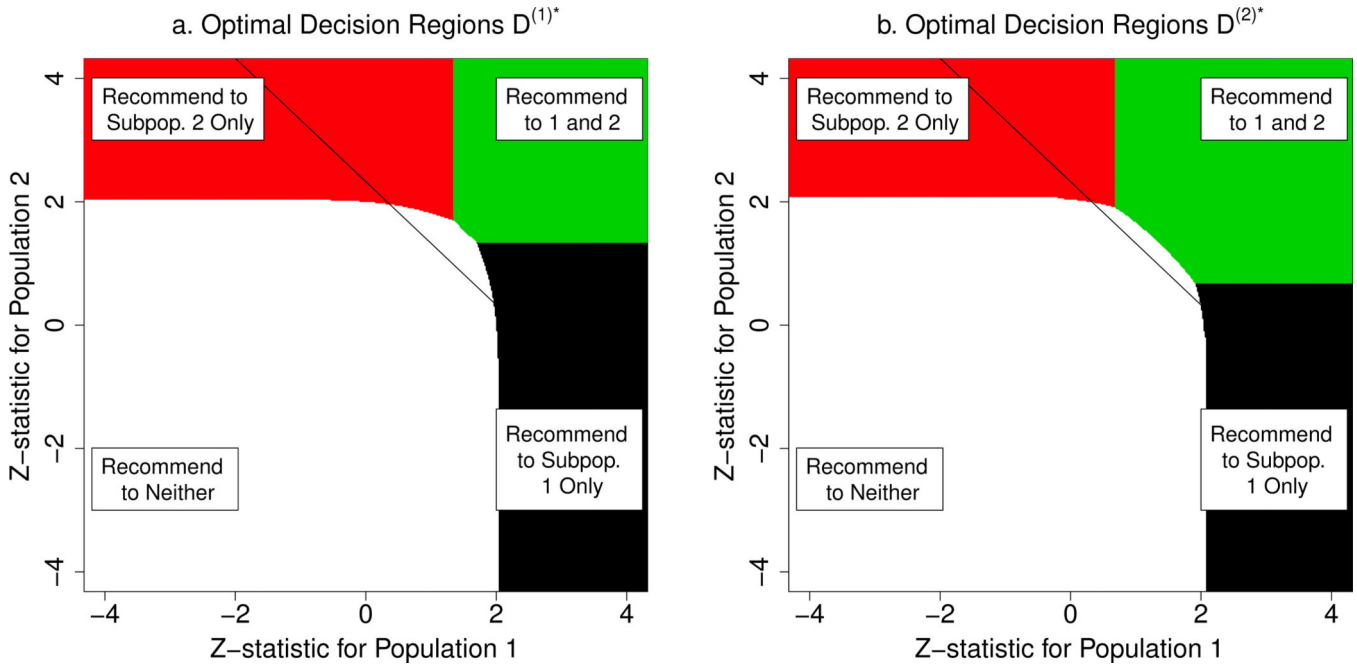

Figure 4.

Optimal decision regions for symmetric case ($p_1 = 1/2$, $\Lambda = \Lambda_1$), for oss functions (a) $L_D^{(1)}$ and (b) $L_D^{(1)}$. For comparability to Figure 1, we included the solid line representing the boundary of the rejection region for $M_{H_{0C}}^{\text{UMP}}$.

Table 1

Bayes risk and power for optimal multiple testing procedures in symmetric and asymmetric cases, at $1 - \beta = 0.9$ and $1 - \beta = 0.88$.

	Symmetric Case		Asymmetric Case	
	$\mathbf{m}^*_{\mathrm{sym}}(0.9)$	$\mathbf{m}^*_{\mathrm{sym}}(0.88)$	$\mathbf{m}^*_{\mathrm{asym}}(0.9)$	$\mathbf{m}^*_{\mathbf{asym}}\mathbf{(0.88)}$
One Minus Bayes Risk	0.52	0.58	0.67	0.71
Power for H_{01} at $(\delta_1^{\min}, 0)$	0.39	0.51	0.55	0.67
Power for H_{02} at $(0, \delta_2^{\min})$	0.39	0.51	0.25	0.30
[Power H_{01} at $(\delta_1^{\min}, \delta_2^{\min})$ + Power H_{02} at $(\delta_1^{\min}, \delta_2^{\min})$]/2	0.65	0.66	0.64	0.64
Power for H_{0C} at $(\delta_1^{\min}, \delta_2^{\min})$	0.90	0.88	0.90	0.88

Table 2

Probabilities of Different Recommendations by Optimal Decision Procedures $D^{(1)*}$ and $D^{(2)*}$, at three alternatives (Alt). The optimal recommendation (Rec.) at each alternative is in bold type.

Alt:	$(\delta_1, \delta_2) = (\delta_1^{\min}, 0)$				$(\delta_1, \delta_2) = (\delta_1^{\min}, \delta_2^{\min})$				**(δ_1, δ_2) = (0, 0)**			
Rec.:	**∅**	**{1}**	**{2}**	**{1, 2}**	**∅**	**{1}**	**{2}**	**{1, 2}**	**∅**	**{1}**	**{2}**	**{1, 2}**
$D^{(1)*}$	0.45	**0.48**	0.01	0.06	0.16	0.14	0.14	**0.57**	**0.95**	0.02	0.02	0.01
$D^{(2)*}$	0.46	**0.37**	0	0.17	0.13	0.4	0.4	**0.78**	**0.95**	0.01	0.01	0.02