ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2020)
82, *Part* 3, *pp.* 749–772

Optimal, two-stage, adaptive enrichment designs for randomized trials, using sparse linear programming

Michael Rosenblum,

Johns Hopkins Bloomberg School of Public Health, Baltimore, USA

Ethan X. Fang

Pennsylvania State University, University Park, USA

and Han Liu

Northwestern University, Evanston, USA

[Received May 2017. Final revision February 2020]

Summary. Adaptive enrichment designs involve preplanned rules for modifying enrolment 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 enrolment, 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 solve this optimization problem directly because of its non-convexity. 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 by using modern optimization techniques. We provide an R package that implements our method and is compatible with linear program solvers in several software languages. Our approach produces new, approximately optimal trial designs.

Keywords: Adaptive enrichment designs; Decision rules; Multiple testing; Optimization problems; Sparse linear programs

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. Standard designs may have low power if the treatment benefits only one subpopulation. Adaptive enrichment designs may be useful in this context.

As an example, consider the phase 3 randomized trial of a treatment for angiosarcoma called the 'TRC105 and pazopanib *versus* pazopanib alone in patients with advanced angiosarcoma trial' (which is known as the 'TAPPAS' trial) (Jones *et al.*, 2017; Mehta *et al.*, 2019). Data from an earlier trial provided suggestive evidence that the treatment may have a greater likelihood of benefiting the subpopulation who enter the trial with cutaneous lesions compared with those with non-cutaneous lesions. The TAPPAS trial design goals included having 80% power to detect

Address for correspondence: Michael Rosenblum, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205, USA.
E-mail: mrosen@jhu.edu

 1369–7412/20/82749

a benefit (hazard ratio of 0.55 for progression-free survival) in the combined population, and also 80% power to detect such a benefit only in the subpopulation with cutaneous lesions (Jones *et al.*, 2017). A two-stage, adaptive enrichment design was implemented. In stage 1, enrolment was from the combined population. There were four possible choices at the end of stage 1:

(a) continue enrolling the combined population,
(b) continue enrolling the combined population but expand the sample size,
(c) enrol only those in the subpopulation with cutaneous lesions or
(d) end the trial.

A challenge was how to select the rule that uses the stage one data to decide between options (a)–(d) to achieve the above power goals and to control type I error while minimizing the number of participants and trial duration. Two other examples of adaptive enrichment designs are the stroke treatment trials that were described by Albers *et al.* (2017) and Jovin *et al.* (2017).

We address trial design problems with similar types of power requirements as the TAPPAS trial, in that a minimum power is required for detecting treatment benefits not only in the combined population, but also in a prespecified subpopulation (and we additionally consider the complementary subpopulation). Like the TAPPAS trial, we focus on two-stage, adaptive enrichment designs with four options for stage 2 enrolment. Further similarities and differences between our approach and that of the TAPPAS trial are discussed later on.

Adaptive enrichment designs have been proposed, e.g. by Follmann (1997), Russek-Cohen and Simon (1997), Jennison and Turnbull (2007), Wang *et al.* (2007, 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) and 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) considered 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 5.2.

A two-stage, adaptive enrichment design consists of a decision rule for potentially modifying enrolment at the end of stage 1, and a multiple-testing procedure at the end of stage 2. The decision rule π_1 is a function from the stage 1 data to a finite set of possible enrolment choices for stage 2. The multiple-testing procedure π_2 is a function from the stage 1 and 2 data to the set of null hypotheses that are rejected. Fig. 1(a) shows the general structure of such designs. We put no restrictions on the functions π_1 and π_2 except that they are discretized, as described below. The resulting class of possible designs is quite large. Our goal is to construct new adaptive enrichment designs that minimize the expected sample size under constraints on power and type I error, over this class of possible designs. This is a non-convex 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 enrolment 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 more difficult to construct a representation as a sparse linear program, and the resulting linear program is more difficult 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.

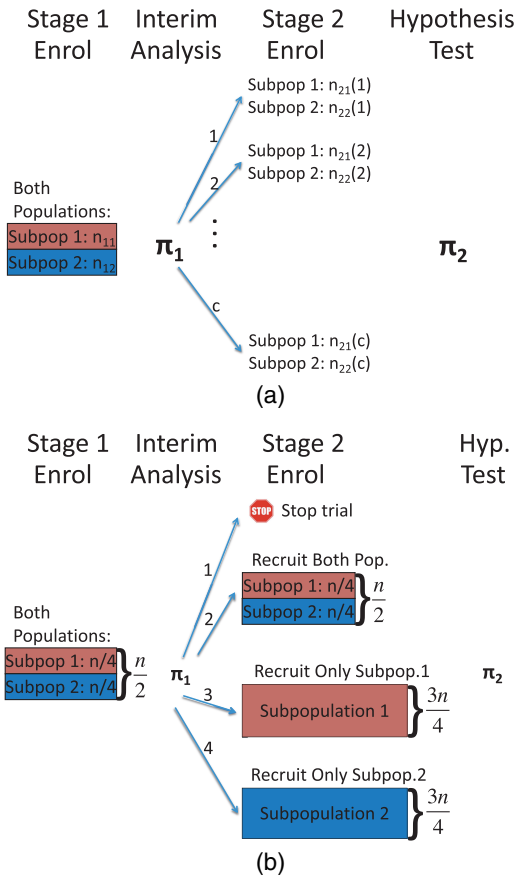

Fig. 1. (a) Generic, two-stage adaptive enrichment design (n_{ks} denotes the sample size in stage k for subpopulation s) and (b) example design with $c = 4$ possible stage 2 choices, denoted $\mathbf{n}^{(1b)}$

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

As in all of the above related work and as generally required by regulators for confirmatory adaptive trials (European Medicines Agency, 2007; Food and Drug Administration, 2016, 2019), we require subpopulations to be defined before the trial starts. Such a definition could be based on prior trial data and disease-specific knowledge. Freidlin and Simon (2005) and Lai *et al.* (2014) gave designs 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.

In our examples, the optimized designs substantially improve power compared with standard designs and some existing adaptive designs. A limitation is that our approach becomes computationally difficult or infeasible for more than two stages or subpopulations. Also, our approach requires that each participant's outcome is measured relatively soon after her or his enrolment. We focus on designs where the only allowed adaptations are to restrict enrolment for stage 2 to be from a single population or to stop the trial early as in Fig. 1(b). We also briefly consider designs with $c = 10$ prespecified options for stage 2 sample sizes in each subpopulation as in Fig. 1(a). We do not consider other adaptations such as allowing more flexible, data-dependent choices of the stage 2 sample size, changing randomization probabilities or modifying treatments for individuals on the basis of their outcomes over time.

In Section 2, we describe the general type of sequential decision problem that our proposed method can solve. Sections 3–6 are devoted to optimizing adaptive enrichment designs. In Section 3, we define our adaptive enrichment design optimization problem. In Section 4, we discretize the problem and transform it into a sparse linear program. The sparse linear program is solved in two examples in Section 5. Trade-offs between adaptive design methods and limitations of our approach are discussed in Section 6.

2. General sequential decision problem and our method for reducing it to a sparse linear program

2.1. Sequential decision problem

Consider the sequential decision problem that is defined by discrete states and actions s_k and a_k at each time $k = 1, \ldots, K$. Our set-up is similar to that of Bertsekas (2017), chapter 1.2. For computational reasons that are described below, our approach is likely to be feasible for only relatively small K such as 2 or 3. Let $\mathbf{t} = (s_1, a_1, \ldots, s_K, a_K)$ denote a typical trajectory, where each s_k is in the finite state space $\mathcal{S}_k$ and a_k is in the finite action space $\mathcal{A}_k$. Let $\{p_{\boldsymbol{\Delta}}(\mathbf{t}) : \boldsymbol{\Delta} \in \mathbb{R}^d\}$ denote the statistical model with unknown parameter $\boldsymbol{\Delta}$, where $p_{\boldsymbol{\Delta}}$ represents a probability mass function on the set of possible trajectories.

At each time $k = 1, \ldots, K$ in turn, nature draws the state at time k from the conditional distribution $p_{\boldsymbol{\Delta}}(s_k | s_1, a_1, \ldots, s_{k-1}, a_{k-1})$ given the history of states and actions $s_1, a_1, \ldots, s_{k-1}, a_{k-1}$ before time k, and then the statistician selects the action at time k as a function (denoted by π_k) of the history $s_1, a_1, \ldots, s_k$. The statistician's policy $\pi = (\pi_1, \ldots, \pi_K)$ represents her or his rule for selecting, for each time k and possible history $s_1, a_1, \ldots, s_k$, the action $a_k = \pi_k(s_1, a_1, \ldots, s_k)$. The set of all policies Π is defined as all $\pi = (\pi_1, \ldots, \pi_K)$ such that each π_k is a function from $\mathcal{T}_k = \mathcal{S}_1 \times \mathcal{A}_1 \times \ldots \times \mathcal{S}_k$ to actions $\mathcal{A}_k$. Denote the set of all trajectories by $\mathcal{T} = \mathcal{T}_K \times \mathcal{A}_K$. For any $\mathbf{t} \in \mathcal{T}$, let $p_{\boldsymbol{\Delta}}(\mathbf{t}) = \Pi_{k=1}^K p_{\boldsymbol{\Delta}}(s_k | s_1, a_1, \ldots, s_{k-1}, a_{k-1})$.

We next define the optimization problem. Consider a known loss function $L(\boldsymbol{\Delta}, \mathbf{t})$, representing the loss when the data-generating distribution is $p_{\boldsymbol{\Delta}}$ and the trajectory is $\mathbf{t}$. For a given parameter $\boldsymbol{\Delta}$ and policy π, let S_k denote the (random) state that is generated by $p_{\boldsymbol{\Delta}}$ at time k and let $A_k = \pi_k(S_1, A_1, \ldots, S_k)$, for each $k \leqslant K$; let $\mathbf{T} = (S_1, A_1, \ldots, S_K, A_K)$. Define the risk under loss function L, distribution $p_{\boldsymbol{\Delta}}$ and policy π as

$$R(L, \boldsymbol{\Delta}, \pi) = E_{p_{\boldsymbol{\Delta}}} L(\boldsymbol{\Delta}, \mathbf{T}) = \sum_{\mathbf{t} \in \mathcal{T}} \left[L(\boldsymbol{\Delta}, \mathbf{t}) p_{\boldsymbol{\Delta}}(\mathbf{t}) \prod_{k=1}^{K} \mathbf{1}\{\pi_k(s_1, a_1, \ldots, s_k) = a_k\} \right], \tag{1}$$

where $\mathbf{1}\{X\}$ is the indicator variable taking value 1 if X is true and 0 otherwise. The goal is to minimize the objective function, defined as the Bayes risk

$$\int R(L_0, \boldsymbol{\Delta}, \pi) \mathrm{d}\Lambda_0(\boldsymbol{\Delta}) = \sum_{\mathbf{t} \in \mathcal{T}} \left[\int L_0(\boldsymbol{\Delta}, \mathbf{t}) p_{\boldsymbol{\Delta}}(\mathbf{t}) \mathrm{d}\Lambda_0(\boldsymbol{\Delta}) \prod_{k=1}^{K} \mathbf{1}\{\pi_k(s_1, a_1, \ldots, s_k) = a_k\} \right], \tag{2}$$

for known loss function L_0 and distribution Λ_0 on the parameter space $\boldsymbol{\Delta} \in \mathbb{R}^d$. For each $j = 1, \ldots, J$, define the constraints $R(L_j, \boldsymbol{\Delta}^{(j)}, \pi) \leqslant \beta_j$ for known loss functions $L_1, \ldots, L_J$, parameter values $\boldsymbol{\Delta}^{(1)}, \ldots, \boldsymbol{\Delta}^{(J)}$ and scalars $\beta_1, \ldots, \beta_J$. Each constraint j could be generalized to incorporate an integral over the parameter space that is similar in form to the Bayes risk. The *sequential decision problem*, in its general form, is as follows.

Find a policy $\pi \in \Pi$ that minimizes the Bayes risk $\int R(L_0, \boldsymbol{\Delta}, \pi) \mathrm{d}\Lambda_0(\boldsymbol{\Delta})$ under the constraints $R(L_j, \boldsymbol{\Delta}^{(j)}, \pi) \leqslant \beta_j$ for each $1 \leqslant j \leqslant J$.

The inputs to the sequential decision problem are the state spaces, action spaces, statistical model, loss functions, parameter values $\boldsymbol{\Delta}^{(1)}, \ldots, \boldsymbol{\Delta}^{(J)}$, scalars $\beta_1, \ldots, \beta_J$ and the distribution Λ_0 defined above. It is a design problem in that the goal is to compute the optimal policy before any data are collected (analogous to how the decision rule for modifying enrolment and the multiple-testing procedure of the adaptive trial design need to be selected before the trial is started).

2.2. Transformation of sequential decision problem into sparse linear program

It follows from equation (2) that the objective function and constraints of the sequential decision problem can be represented in terms of the indicator variables $\mathbf{1}\{\pi_k(s_1, a_1, \ldots, s_k) = a_k\}$ for each $k \leqslant K, (s_1, a_1, \ldots, s_k) \in \mathcal{T}_k, a_k \in \mathcal{A}_k$. Any policy $\pi \in \Pi$ corresponds to a unique set of values of these variables. Unfortunately, the resulting problem is non-convex because of the product in equation (2), and so the problem is extremely difficult to solve.

To overcome this obstacle, we instead represent each policy $\pi \in \Pi$ by the set of indicator variables $\{v_\pi(\mathbf{t}) : \mathbf{t} \in \mathcal{T}\}$, where we define $v_\pi(\mathbf{t}) = \Pi_{k=1}^{K} \mathbf{1}\{\pi_k(s_1, a_1, \ldots, s_k) = a_k\}$, i.e. $v_\pi(\mathbf{t}) = 1$ if each action a_k in the trajectory $\mathbf{t} = (s_1, a_1, \ldots, s_K, a_K)$ is precisely what the policy π says to do given the observed history. For any policy $\pi \in \Pi$, the Bayes risk (2) can be written as the following linear function of the variables $v_\pi(\mathbf{t})$:

$$\int R(L_0, \boldsymbol{\Delta}, \pi) \mathrm{d}\Lambda_0(\boldsymbol{\Delta}) = \sum_{\mathbf{t} \in \mathcal{T}} v_\pi(\mathbf{t}) \int L_0(\boldsymbol{\Delta}, \mathbf{t}) p_{\boldsymbol{\Delta}}(\mathbf{t}) \mathrm{d}\Lambda_0(\boldsymbol{\Delta}). \tag{3}$$

The constraints $R(L_j, \boldsymbol{\Delta}^{(j)}, \pi) \leqslant \beta_j$ can be expressed similarly. Using an arbitrary ordering of the trajectories, we represent the set of variables $\{v_\pi(\mathbf{t}) : \mathbf{t} \in \mathcal{T}\}$ as the vector $\mathbf{v}_\pi$.

The above representation of the sequential decision problem has the computational advantage of being linear in the variables $v_\pi(\mathbf{t})$. However, this representation poses two important challenges:

(a) most $\{0, 1\}$-valued vectors of length $|\mathcal{T}|$ do not represent any policy $\pi \in \Pi$, since representing a policy imposes logical constraints on the entries of such a vector;
(b) there are many variables (one per trajectory $\mathbf{t} \in \mathcal{T}$).

We solve challenge (a) by using sparse linear constraints (i.e. linear constraints where most entries are 0) defined in expressions (6) and (7) below to represent the condition that an arbitrary $\{0, 1\}$-valued vector $\mathbf{v}$ of length $|\mathcal{T}|$ represents a policy $\pi \in \Pi$, i.e. the condition that $\mathbf{v} = \mathbf{v}_\pi$ for some $\pi \in \Pi$. Including these constraints in the optimization problem is equivalent to defining the search space as the set of all policies Π. Challenge (b) can be addressed by using computationally efficient algorithms for solving sparse linear programs that can handle, for example, 10^7 variables, as discussed in Section 2.3.

Let $\tilde{s}_2 \ldots, \tilde{s}_K$ denote states in $\mathcal{S}_2, \ldots, \mathcal{S}_K$ respectively, which can be chosen arbitrarily; for example $\tilde{s}_k$ could be defined as the first state in a list of the states $\mathcal{S}_k$. The states $\tilde{s}_2, \ldots, \tilde{s}_K$ are fixed in advance and used to define the binary integer program below. They can be thought of as reference states at each time. We prove in section D of the on-line supplementary material that the sequential decision problem from Section 2.1 is equivalent to the following binary integer program in the variables $\mathbf{v}$ with components denoted by $v(\mathbf{t})$ for each $\mathbf{t} \in \mathcal{T}$.

2.2.1. Binary integer program representation of sequential decision problem

$$\min_{\mathbf{v}} \sum_{\mathbf{t} \in \mathcal{T}} v(\mathbf{t}) \int L_0(\boldsymbol{\Delta}, \mathbf{t}) p_{\boldsymbol{\Delta}}(\mathbf{t}) \mathrm{d}\Lambda_0(\boldsymbol{\Delta}), \tag{4}$$

under the following constraints:

$$\sum_{\mathbf{t}\in\mathcal{T}} v(\mathbf{t}) L_j(\boldsymbol{\Delta}^{(j)}, \mathbf{t}) p_{\boldsymbol{\Delta}^{(j)}}(\mathbf{t}) \leqslant \beta_j, \qquad \text{for any } j \in \{1, \ldots, J\}; \tag{5}$$

$$\sum_{a_1,\ldots,a_K} v(s_1, a_1, \tilde{s}_2, a_2, \ldots, \tilde{s}_K, a_K) = 1, \qquad \text{for any } s_1 \in \mathcal{S}_1; \tag{6}$$

$$\sum_{a_k,\ldots,a_K} v(s_1, a_1, \ldots, s_k, a_k, \tilde{s}_{k+1}, \ldots, \tilde{s}_K, a_K) - v(s_1, a_1, \ldots, \tilde{s}_k, a_k, \tilde{s}_{k+1}, \ldots, \tilde{s}_K, a_K) = 0,$$
$$\text{for any } k : 2 \leqslant k \leqslant K, \text{ and any } (s_1, a_1, \ldots, s_k) \in \mathcal{T}_k; \tag{7}$$

$$v(\mathbf{t}) \in \{0, 1\}, \qquad \text{for all } \mathbf{t} \in \mathcal{T}. \tag{8}$$

Here we use the convention that the sum over each variable is with respect to its corresponding domain; for example Σ_{a_k} represents $\Sigma_{a_k \in \mathcal{A}_k}$. The only difference between the two terms in the sum in expression (7) is that the single variable s_k on the left-hand side is replaced by $\tilde{s}_k$ on the right-hand side. The objective function (3) and constraints $R(L_j, \boldsymbol{\Delta}^{(j)}, \pi) \leqslant \beta_j$ for $j \leqslant J$ of the sequential decision problem are encoded as the linear functions (4) and (5) respectively of $\mathbf{v}$.

There is a one-to-one correspondence between the set of vectors $\mathbf{v}$ that satisfy the constraints (6)–(8) and the set of policies Π. Given any $\mathbf{v}$ that satisfies constraints (6)–(8), we define the corresponding policy (denoted $\pi^{\mathbf{v}}$) as follows: for any $k \geqslant 1$ and $(s_1, a_1, \ldots, s_k, a_k) \in \mathcal{T}_k \times \mathcal{A}_k$,

$$\pi_k^{\mathbf{v}}(s_1, a_1, \ldots, s_k) = a_k \qquad \text{if and only if} \sum_{a_{k+1},\ldots,a_K} v(s_1, a_1, \ldots, s_k, a_k, \tilde{s}_{k+1}, a_{k+1}, \ldots, \tilde{s}_K, a_K) = 1.$$

If the sum in the above expression equals 0 for all $a_k \in \mathcal{A}_k$, then we let $\pi_k^{\mathbf{v}}(s_1, a_1, \ldots, s_k)$ be defined as an arbitrarily chosen element $\tilde{a}_k \in \mathcal{A}_k$; we show in the proof of theorem 1 below that this choice has no effect since in this case the sequence $s_1, a_1, \ldots, s_k$ can never occur.

Constraints (6)–(8) imply the following property: for any sequence of states $s_1, \ldots, s_K$, there is exactly one sequence of actions $a_1, \ldots, a_K$ for which $v(s_1, a_1, \ldots, s_K, a_K) = 1$. More than this property is required, however, to ensure that there exists a policy $\pi \in \Pi$ for which $v(\mathbf{t}) = \Pi_{k=1}^{K} \mathbf{1}\{\pi_k(s_1, a_1, \ldots, s_k) = a_k\}$ for all $\mathbf{t} \in \mathcal{T}$; we give an example showing this in section D.3 of the on-line supplementary material. Intuitively, the problem is that the aforementioned property does not enforce that the choice of action at time k depends only on previous (and not future) states and actions. Constraints (6)–(8) encode this, as proved in section D of the supplementary material. There, we prove that any choice of $\tilde{s}_2, \ldots, \tilde{s}_K$ leads to an equivalent definition of the binary integer program, and we also prove the following theorem.

Theorem 1. For any vector $\mathbf{v}$ that satisfies constraints (6)–(8), the corresponding $\pi^{\mathbf{v}}$ is a policy in Π; conversely, every policy $\pi \in \Pi$ is represented by some vector $\mathbf{v}$ that satisfies these constraints. For any optimal solution $\mathbf{v}$ to the binary integer program, the corresponding policy $\pi^{\mathbf{v}}$ is a well-defined, feasible, optimal solution to the sequential decision problem.

We relax the binary constraints (8) by replacing them by $v(\mathbf{t}) \geqslant 0$ for all $\mathbf{t} \in \mathcal{T}$, to make the problem computationally feasible. We prove in section D of the on-line supplementary material that the resulting linear program, which we call the sparse linear program, is equivalent to the sequential decision problem where the set of (deterministic) policies Π is replaced by the larger set of stochastic policies Π^*, defined as all $\pi^* = (\pi_1^*, \ldots, \pi_K^*)$ such that each π_k^* is a function that maps each $(s_1, a_1, \ldots, s_k) \in \mathcal{T}_k$ to a multinomial distribution on the set of actions $\mathcal{A}_k$. In other words, the action at time k is a random choice in $\mathcal{A}_k$ with probabilities encoded by $\pi_k^*(s_1, a_1, \ldots, s_k)$. Though such policies can be implemented by using a random-number generator, in many applications it is desirable to have a deterministic policy $\pi \in \Pi$, which can sometimes be obtained from a randomized policy by rounding; for example, in our adaptive

design application this was done and had a negligible effect since most components in the optimal solutions are $\{0, 1\}$ valued.

2.3. General form of sparse linear program and computational limitations

We describe the general form of the sparse linear program resulting from replacing the integrality constraints (8) in the binary integer program by the non-negativity constraints $v(\mathbf{t}) \geqslant 0$ for all $\mathbf{t} \in \mathcal{T}$. Let $\mathbb{R}_+$ denote the non-negative real numbers and let $w = |\mathcal{T}|$ denote the number of variables. The general form of the sparse linear program is

$$\min_{\mathbf{v} \in \mathbb{R}_+^w} \mathbf{c}^{\mathbf{T}}\mathbf{v} \qquad \text{subject to } \mathbf{A}^{(1)}\mathbf{v} \leqslant \mathbf{a}^{(1)}, \quad \mathbf{A}^{(2)}\mathbf{v} = \mathbf{a}^{(2)}, \tag{9}$$

for matrices $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ and vectors $\mathbf{c}$, $\mathbf{a}^{(1)}$ and $\mathbf{a}^{(2)}$. The matrix $\mathbf{A}^{(1)}$ has dimensions $J \times w$ and encodes constraints (5). Equality constraints (6) and (7) can be represented by $\mathbf{A}^{(2)}\mathbf{v} = \mathbf{a}^{(2)}$ where $\mathbf{A}^{(2)}$ has the following structure (where we describe only the non-zero elements):

$$\mathbf{A}^{(2)} = \left(\begin{array}{c} |\mathcal{S}_1| \text{ rows, each with } \prod_{k=1}^{K} |\mathcal{A}_k| \text{ 1s} \\ \hline |\mathcal{T}_2| \text{ rows, each with } \prod_{k=2}^{K} |\mathcal{A}_k| \text{ 1s, and same number of} -1\text{s} \\ \hline \vdots \\ \hline |\mathcal{T}_K| \text{ rows, each with } |\mathcal{A}_K| \text{1s, and same number of} -1\text{s} \end{array} \right).$$

The matrix $\mathbf{A}^{(1)}$ is typically dense (most entries are non-zero). The matrix $\mathbf{A}^{(2)}$ is sparse (most entries are 0) if the number of action sequences $\Pi_{k=1}^{K}|\mathcal{A}_k|$ is much smaller than the number of trajectories $|\mathcal{T}|$. In this case, 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 (4) and is dense. The vector $\mathbf{a}^{(1)} = (\beta_1, \ldots, \beta_J)^{\mathrm{T}}$, and the vector $\mathbf{a}^{(2)}$ consists of $|\mathcal{S}_1|$ 1s followed by all 0s.

There are important computational limitations to our approach for solving the sequential decision problem by transforming it into the above sparse linear program. We expect that, for problems with $|\mathcal{T}| \leqslant 10^7$ and $J \leqslant 500$, it will be computationally feasible to solve the sparse linear program. We solved problems that exceeded these thresholds for our adaptive enrichment design application (at $K = 2$), which is the focus of the remainder of the paper. Since the number of trajectories is $|\mathcal{T}_K \times \mathcal{A}_K|$, which grows (roughly) exponentially with K, we expect that our approach will only be computationally feasible for $K = 2$ or $K = 3$.

In section D.1 of the on-line supplementary material, we generalize the problem set-up above to allow the state space at each time $k \geqslant 2$ to depend on the history $s_1, a_1, \ldots, a_{k-1}$.

3. Adaptive enrichment design problem definition (non-discretized version)

3.1. Data, assumptions and 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 $s \in \{1, 2\}$; $p_1 + p_2 = 1$. Each enrolled participant is assigned to treatment ($r = 1$) or control ($r = 0$) with probability $\frac{1}{2}$.

For each subpopulation $s \in \{1, 2\}$ and stage $k \in \{1, 2\}$, we assume that exactly half the participants are randomized to each study arm $r \in \{0, 1\}$. 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 or his random study arm assignment by $R_{ksi} \in \{0, 1\}$ and outcome by $Y_{ksi} \in \mathbb{R}$. 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.

For clarity, we focus on normally distributed outcomes with a known common variance. Under regularity conditions, our results can be extended to different outcome distributions and unknown variances, as long as we use 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 $R_{ksi} = r$, the outcome $Y_{ksi} \sim N(\mu_{sr}, \sigma_s^2)$ and is independent of the data from all previously enrolled participants. We assume that each participant's outcome is observed soon after enrolment, 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 assume that the parameter $\boldsymbol{\Delta}$ is unknown, but that σ_1^2, σ_2^2 and p_1 are known. We discuss possible ways to deal with uncertainty in p_1 in Section 6.

Define the null hypotheses of no average treatment benefit in subpopulation 1, subpopulation 2 and the combined population respectively as $H_{01} : \Delta_1 \leqslant 0$, $H_{02} : \Delta_2 \leqslant 0$ and $H_{0C} : \Delta_C \leqslant 0$. 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 $\Delta_s \leqslant 0$; it contains H_{0C} if $p_1\Delta_1 + p_2\Delta_2 \leqslant 0$.

3.2. Two-stage, adaptive enrichment designs

In stage 1, n_{1s} participants are enrolled from each subpopulation s. In our examples, we set the stage 1 sample sizes n_{11} and n_{12} proportional to the subpopulation sizes p_1 and p_2; however, our general method does not require this. At the interim analysis following stage 1, a decision rule π_1 determines the number of participants to enrol from each subpopulation in stage 2 based on the stage 1 data. There are $c < \infty$ possible choices for stage 2 enrolment, denoted by the action set $\mathcal{A}_1 = \{1, \ldots, c\}$. Each action $a_1 \in \mathcal{A}_1$ represents a possible pair of stage 2 sample sizes denoted by $n_{21}(a_1)$ and $n_{22}(a_1)$ for subpopulations 1 and 2 respectively. At the end of stage 2, a multiple-testing procedure π_2 determines which subset (if any) of the null hypotheses to reject, based on the data from stages 1 and 2.

Define an adaptive design template, denoted by $\mathbf{n}$, to be the possible end of stage 1 decisions and corresponding sample sizes $\mathbf{n} = (\mathcal{A}_1, n_{11}, n_{12}, \{n_{21}(a_1), n_{22}(a_1)\}_{a_1 \in \mathcal{A}_1})$. A generic adaptive design template is depicted in Fig. 1(a). A specific example for $p_1 = \frac{1}{2}$ is given in Fig. 1(b), where, for a given $n > 0$, the stage 1 sample sizes satisfy $n_{11} = n_{12} = n/4$, and there are four choices for stage 2 enrolment: $a_1 = 1$, stop the trial, i.e. $n_{21}(1) = n_{22}(1) = 0$; $a_1 = 2$, enrol exactly as in stage 1, i.e. $n_{21}(2) = n_{22}(2) = n/4$; $a_1 = 3$, enrol from only subpopulation 1, i.e. $n_{21}(3) = 3n/4$ and $n_{22}(3) = 0$; $a_1 = 4$, enrol from only subpopulation 2, i.e. $n_{21}(4) = 0$ and $n_{22}(4) = 3n/4$. This adaptive design template, denoted $\mathbf{n}^{(1b)}$, is used in Section 5. It allows enrichment of subpopulation 1 ($a_1 = 3$) or subpopulation 2 ($a_1 = 4$), in which case the total enrolled from the enriched subpopulation is n ($n/4$ from stage 1 plus $3n/4$ from stage 2). This sample size choice was motivated by the problems in Section 5.

3.3. Sufficient statistics

Let N_{ks} denote the number enrolled during stage $k \in \{1, 2\}$ from subpopulation $s \in \{1, 2\}$. The stage 1 sample sizes are set in advance, whereas those in stage 2 are functions of the stage 1 data. For each subpopulation $s \in \{1, 2\}$ and stage $k \in \{1, 2\}$, define the z-statistic

$$Z_s^{(k)} = \left\{ \frac{\sum_{i=1}^{N_{ks}} Y_{ksi} R_{ksi}}{\sum_{i=1}^{N_{ks}} R_{ksi}} - \frac{\sum_{i=1}^{N_{ks}} Y_{ksi}(1 - R_{ksi})}{\sum_{i=1}^{N_{ks}} (1 - R_{ksi})} \right\} \left(\frac{4\sigma_s^2}{N_{ks}} \right)^{-1/2}, \tag{10}$$

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

$$Z_s^{(\mathrm{F})} = \left\{ \frac{\sum_{k=1}^{2} \sum_{i=1}^{N_{ks}} Y_{ksi} R_{ksi}}{\sum_{k=1}^{2} \sum_{i=1}^{N_{ks}} R_{ksi}} - \frac{\sum_{k=1}^{2} \sum_{i=1}^{N_{ks}} Y_{ksi}(1 - R_{ksi})}{\sum_{k=1}^{2} \sum_{i=1}^{N_{ks}} (1 - R_{ksi})} \right\} \left(\frac{4\sigma_s^2}{N_{1s} + N_{2s}} \right)^{-1/2}. \tag{11}$$

Let $\mathbf{Z}^{(k)} = (Z_1^{(k)}, Z_2^{(k)})$ for each stage $k \in \{1, 2\}$, and $\mathbf{Z}^{(\mathrm{F})} = (Z_1^{(\mathrm{F})}, Z_2^{(\mathrm{F})})$. The joint distribution of these random vectors is given in section A of the on-line supplementary materials. The first-stage z-statistics $\mathbf{Z}^{(1)}$ are bivariate normal, as are the second-stage statistics $\mathbf{Z}^{(2)}$ conditionally on the decision π_1 for stage 2 enrolment; the final z-statistic $Z_s^{(\mathrm{F})}$ for subpopulation s is a weighted combination of the corresponding first- and second-stage statistics, with each subpopulation s participant contributing equal information.

The distribution of the data X depends on three unknown parameters: our parameter of interest $\boldsymbol{\Delta} = (\Delta_1, \Delta_2)$ and the variation-independent nuisance parameters $\mu_{s1} + \mu_{s0}$ for each $s \in \{1, 2\}$. We prove the following theorem in section C of the on-line supplementary materials.

Theorem 2. Consider any fixed values of the nuisance parameters. Then $\mathbf{Z}^{(1)}$ is a minimal sufficient statistic at the end of stage 1. Also, for any end of stage 1 policy $\pi_1(\mathbf{Z}^{(1)})$, we have that $(\pi_1(\mathbf{Z}^{(1)}), \mathbf{Z}^{(\mathrm{F})})$ is a minimal sufficient statistic at the end of stage 2. Furthermore, the joint distribution of these statistics does not depend on the nuisance parameters.

We henceforth focus on decision rules π_1 that depend on the data only through $\mathbf{Z}^{(1)}$, and multiple-testing procedures π_2 that depend on the data only through $(\pi_1(\mathbf{Z}^{(1)}), \mathbf{Z}^{(\mathrm{F})})$. We refer to $\mathbf{Z}^{(1)}$ and $(\pi_1(\mathbf{Z}^{(1)}), \mathbf{Z}^{(\mathrm{F})})$ as the stage 1 and 2 sufficient statistics respectively. Because of the presence of nuisance parameters, these should be called 'specific sufficient for $\boldsymbol{\Delta}$' (Basu, 1978), but for conciseness we call them 'sufficient'. Let Π_1 denote the set of all possible stage 1 policies, i.e. all functions from the sample space $\mathbb{R}^2$ of the stage 1 sufficient statistics $\mathbf{Z}^{(1)}$ to $\mathcal{A}_1$.

At the end of stage 2, the multiple-testing procedure π_2 determines which (if any) null hypotheses are rejected. Let $\mathcal{A}_2$ denote the power set of $\{H_{01}, H_{02}, H_{0\mathrm{C}}\}$, except that we exclude the subset $\{H_{01}, H_{02}\}$ since if H_{01} and H_{02} are false then so is $H_{0\mathrm{C}}$. Let Π_2 denote the set of all possible stage 2 policies, i.e. all functions π_2 from the sample space $\mathcal{A}_1 \times \mathbb{R}^2$ of stage 2 sufficient statistics $(\pi_1(\mathbf{Z}^{(1)}), \mathbf{Z}^{(\mathrm{F})})$ to $\mathcal{A}_2$. The set of all policies is $\Pi = \Pi_1 \times \Pi_2$.

For a given adaptive design template $\mathbf{n}$, an adaptive enrichment design is defined as a pair $\pi = (\pi_1, \pi_2) \in \Pi$. Given $(p_1, \mathbf{n}, \pi, \boldsymbol{\Delta}, \sigma_1^2, \sigma_2^2)$, let $P_{\boldsymbol{\Delta}}$ denote the induced joint distribution of the statistics $(\mathbf{Z}^{(1)}, \pi_1(\mathbf{Z}^{(1)}), \mathbf{Z}^{(F)})$ and let $E_{\boldsymbol{\Delta}}$ denote expectation with respect to this distribution (where we suppress dependence on $p_1, \mathbf{n}, \pi, \sigma_1^2$ and σ_2^2 for clarity).

3.4. Adaptive enrichment design optimization problem

Our optimization problem can be represented by using the decision theory framework from Section 2 with $K = 2$ time points. The states s_1 and s_2 represent sufficient statistics at the end

of stages 1 and 2 respectively. Here we allow the state spaces $\mathcal{S}_1$ and $\mathcal{S}_2$ to be infinite; these are discretized in Section 4. The actions $a_1 \in \mathcal{A}_1$ and $a_2 \in \mathcal{A}_2$ represent how many to enrol in stage 2 from each subpopulation and the set of null hypotheses rejected after stage 2 respectively.

We consider loss functions L that are bounded, integrable functions of the treatment effect $\boldsymbol{\Delta}$, the enrolment decision a_1 and the set of hypotheses rejected a_2. For a given loss function $L(\boldsymbol{\Delta}, a_1, a_2)$, the risk at treatment effect vector $\boldsymbol{\Delta} \in \mathbb{R}^2$ is defined as $R(L, \boldsymbol{\Delta}, \pi) = E_{\boldsymbol{\Delta}} L[\boldsymbol{\Delta}, \pi_1(\mathbf{Z}^{(1)}), \pi_2\{\pi_1(\mathbf{Z}^{(1)}), \mathbf{Z}^{(\mathrm{F})}\}]$. By selecting an appropriate loss function L, the risk can be made to represent, for example, expected sample size, type I error, power or 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) $L^{\mathrm{SS}}(\boldsymbol{\Delta}, a_1, a_2) = n_{11} + n_{12} + n_{21}(a_1) + n_{22}(a_1)$; the corresponding risk at $\boldsymbol{\Delta} \in \mathbb{R}^2$ is the expected sample size under treatment effect vector $\boldsymbol{\Delta}$. A familywise type I error, i.e. rejecting one or more true null hypotheses, is encoded by the loss function $L^{\mathrm{FWE}}(\boldsymbol{\Delta}, a_1, a_2) = \mathbf{1}\{\mathcal{H}_{\mathrm{TRUE}}(\boldsymbol{\Delta}) \cap a_2 \neq \emptyset\}$. Similarly, a type II error for a null hypothesis $H \in \{H_{01}, H_{02}, H_{0\mathrm{C}}\}$ is encoded by the loss function $L^{(H)}(\boldsymbol{\Delta}, a_1, a_2) = \mathbf{1}\{H \notin a_2, H \notin \mathcal{H}_{\mathrm{TRUE}}(\boldsymbol{\Delta})\}$.

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 $Q = \{(0,0), (\Delta^{\min}, 0), (0, \Delta^{\min}), (\Delta^{\min}, \Delta^{\min})\}$, where $\Delta^{\min}$ represents the minimum, clinically meaningful treatment effect, which is user specified. Let Λ^{pm} denote this distribution with weight $\frac{1}{4}$ 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 centred at each point in Q and each having covariance matrix $c_\Lambda(\Delta^{\min})^2 \mathbf{I}_2$ for $\mathbf{I}_2$ the 2×2 identity matrix and constant $c_\Lambda > 0$.

3.4.1. *Adaptive design optimization problem*

Find the adaptive enrichment design $\pi \in \Pi$ minimizing the Bayes risk,

$$\int R(L_0, \boldsymbol{\Delta}, \pi) \mathrm{d}\Lambda_0(\boldsymbol{\Delta}), \tag{12}$$

under the familywise type I error constraints

$$P_{\boldsymbol{\Delta}}\{\pi_2 \text{ rejects any null hypotheses in } \mathcal{H}_{\mathrm{TRUE}}(\boldsymbol{\Delta})\} \leqslant \alpha, \text{ for any } \boldsymbol{\Delta} \in \mathbb{R}^2, \tag{13}$$

and power constraints

$$P_{\boldsymbol{\Delta}^{(m)}}\{\pi_2 \text{ rejects at least the null hypothesis } H^{(m)}\} \geqslant 1 - \beta_m, \tag{14}$$

for each $m = 1, \ldots, M$, where $1 - \beta_m$ is the required power, $\boldsymbol{\Delta}^{(m)} \in \mathbb{R}^2$ and $H^{(m)} \in \{H_{01}, H_{02}, H_{0\mathrm{C}}\}$ is a false null hypothesis under $\boldsymbol{\Delta}^{(m)}$, i.e. $H^{(m)} \notin \mathcal{H}_{\mathrm{TRUE}}(\boldsymbol{\Delta}^{(m)})$.

Constraints (13) represent strong control of the familywise type I error rate, i.e., for any pair of treatment effects Δ_1 and Δ_2, the probability of rejecting one or more true null hypotheses is at most α. An adaptive enrichment design $\pi \in \Pi$ is feasible if it satisfies all the constraints (13) and (14).

We informally refer to the distribution Λ_0 as a prior, with the understanding that our optimization problem uses the frequentist decision theory framework and the only role of Λ_0 is in defining the objective function (12). 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 (12) is replaced by the maximum over $\boldsymbol{\Delta}$ in a finite subset of $\mathbb{R}^2$.

3.5. Example optimization problems

We consider design problems with similar types of power and type I error requirements as in the TAPPAS trial that was described in Section 1. The TAPPAS trial had power goals for a single subpopulation and the combined population, whereas we additionally set a power goal for the complementary subpopulation. The trade-offs between these approaches were discussed by Freidlin *et al.* (2013). We also solved problems analogous to examples 1 and 2 below, except involving only the null hypotheses H_{01} and H_{0C} and only allowing enrichment of subpopulation 1, which is more similar to the TAPPAS trial set-up; see Section 6.

We give examples of goals below to illustrate our approach. We solve the following two example problems in Section 5, for values of $p_1, \mathbf{n}, \sigma_1^2, \sigma_2^2, \alpha, \beta, \Delta^{\min}$ and c_Λ defined there.

3.5.1. Example 1

Consider the problem of minimizing the expected sample size averaged over the four point masses in Q, under the type I error constraints (13) and the following power constraints for given type II error $\beta > 0$.

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

Condition 2. At $\boldsymbol{\Delta}^{(2)} = (0, \Delta^{\min})$, the power to reject H_{02} is at least $1 - \beta$.

Condition 3. At $\boldsymbol{\Delta}^{(3)} = (\Delta^{\min}, \Delta^{\min})$, the power to reject H_{0C} is at least $1 - \beta$.

This problem can be represented by setting $(L_0, \Lambda_0) = (L^{SS}, \Lambda^{pm})$.

3.5.2. Example 2

We modify the above example by replacing the prior Λ^{pm} by Λ^{mix}.

4. Discretization of adaptive design optimization problem and transformation into sparse linear program

4.1. Overview

The adaptive design optimization problem is extremely difficult to solve directly. This is because the optimization is over the very large class of decision rules Π_1 and multiple-testing procedures Π_2, and involves infinitely many familywise type I error constraints (13).

We propose a novel approach to solving a discretized version of the above problem, involving three 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 non-convex and so is still extremely difficult to solve. Step 2, which is handled in Section 4.4, involves reparameterizing this problem so that it can be represented as a sparse linear program: a class of problems that is much easier to solve than non-convex problems. This reparameterization uses our general method from Section 2.2. The third step is to apply large-scale optimization methods to solve the sparse linear program, which is described for our examples in Section 5.2.

4.2. Definition of discretized problem and class of designs Π^{DISC}

The first of the above steps is to discretize the adaptive design optimization problem. This involves partitioning the sample space $\mathbb{R}^2$ of the first-stage z-statistics $\mathbf{Z}^{(1)}$ into a finite set of rectangles, and similarly partitioning the sample space of the z-statistics $\mathbf{Z}^{(F)}$ at the end of stage 2. One approach to constructing a partition, as in Rosenblum *et al.* (2014), is to start with a

box $B = [-b, b] \times [-b, b]$ for a given integer $b > 0$. We partition the box into rectangles each having side lengths $\tau = (\tau_1, \tau_2)$ such that b/τ_s is an integer for each $s \in \{1, 2\}$. For each pair of integers j and j', define the rectangle $R_{j,j'} = [j\tau_1, (j+1)\tau_1) \times [j'\tau_2, (j'+1)\tau_2)$. Define the set of such rectangles in the bounded region B as $\mathcal{R}_B = \{R_{j,j'} : j, j' \in \mathbb{Z}, R_{j,j'} \subset B\}$. Lastly, define the partition $\mathcal{R} = \mathcal{R}_B \cup \{\mathbb{R}^2 \setminus B\}$ of $\mathbb{R}^2$. Even though $\mathbb{R}^2 \setminus B$ is not a rectangle, for conciseness we still call $\mathcal{R}$ a partition of rectangles.

Let the finite state spaces $\mathcal{S}_1$ and $\mathcal{S}_2$ (for stages 1 and 2 respectively) denote partitions of $\mathbb{R}^2$ into rectangles. We restrict to the subclass Π_1^{DISC} of decision rules $\pi_1 \in \Pi_1$ that depend on the data only through the rectangle that contains the first-stage z-statistics, i.e. decision rules $\pi_1 \in \Pi_1$ such that, for any $s_1 \in \mathcal{S}_1$ and $\mathbf{z}^{(1)}, \mathbf{z}^{(1)'} \in s_1$, we have $\pi_1(\mathbf{z}^{(1)}) = \pi_1(\mathbf{z}^{(1)'})$.

Similarly, we restrict to the subclass Π_2^{DISC} of multiple-testing procedures $\pi_2 \in \Pi_2$ that depend on the data only through the end of stage 1 decision a_1 and the rectangle in $\mathcal{S}_2$ that contains the cumulative statistics $\mathbf{Z}^{(\text{F})}$ at the end of stage 2, i.e. we restrict to $\pi_2 \in \Pi_2$ such that, for any $a_1 \in \mathcal{A}_1$, $s_2 \in \mathcal{S}_2$, $\mathbf{z}^{(\text{F})} \in s_2$ and $\mathbf{z}^{(\text{F})'} \in s_2$, we have $\pi_2(a_1, \mathbf{z}^{(\text{F})}) = \pi_2(a_1, \mathbf{z}^{(\text{F})'})$. Define the class of discretized adaptive enrichment designs as $\Pi^{\text{DISC}} = \Pi_1^{\text{DISC}} \times \Pi_2^{\text{DISC}}$.

It remains to discretize the set $\boldsymbol{\Delta} \in \mathbb{R}^2$ in the type I error constraints (13) by selecting a finite subset $G \subseteq \mathbb{R}^2$. Define the boundaries of the null spaces for H_{01}, H_{02} and $H_{0\text{C}}$ to be $\{(0, \Delta_2) : \Delta_2 \in \mathbb{R}\}$, $\{(\Delta_1, 0) : \Delta_1 \in \mathbb{R}\}$ and $\{(\Delta_1, \Delta_2) \in \mathbb{R}^2 : p_1\Delta_1 + p_2\Delta_2 = 0\}$ respectively. Let G denote a grid of points on the union of these boundaries; an example is given in Section 5.2. This choice of G is based on the conjecture that the active constraints in expression (13) will be on the null space boundaries. We demonstrate that, by a careful selection of G, the solutions to the discretized problem in our examples satisfy constraints (13) at all $\boldsymbol{\Delta} \in \mathbb{R}^2$, if we solve the discretized problem by using a slightly smaller α than the required value in expression (13).

The *discretized problem* is defined as the adaptive design optimization problem from Section 3.4 restricted to policies Π^{DISC} and using only the type I error constraints (13) for $\boldsymbol{\Delta} \in G$. From here on, we fix $\mathcal{S}_1$, $\mathcal{S}_2$ and G, and focus on solving the discretized problem.

4.3. (Non-convex) representation of discretized problem

The discretized problem can be represented in terms of the variables $\mathbf{1}\{\pi_1(s_1) = a_1\}$ and $\mathbf{1}\{\pi_2(s_1, a_1, s_2) = a_2\}$ for each trajectory $\mathbf{t} = (s_1, a_1, s_2, a_2) \in \mathcal{T}$ where $\mathcal{T} = \mathcal{S}_1 \times \mathcal{A}_1 \times \mathcal{S}_2 \times \mathcal{A}_2$. Any policy $\pi \in \Pi^{\text{DISC}}$ corresponds to a unique set of values of these variables. It follows from equation (2) that the Bayes risk (12) and constraints (14) can be represented in terms of these variables. It follows from equation (1) that each type I error constraint (13), which equals the risk $R(L^{\text{FWE}}, \boldsymbol{\Delta}, \pi)$, can be represented in terms of these variables.

Unfortunately, the resulting problem is non-convex because of the products of variables in equations (1) and (2). It is computationally intractable to solve, since only *ad hoc* methods exist for solving non-convex optimization problems and even if a local minimum is found there is no general way to determine whether it is the global minimum. Our solution is to transform this problem into a sparse linear program by applying our method from Section 2.2 at $K = 2$.

4.4. Transformation of discretized problem into linear program

Define the new variables $v(\mathbf{t}) = \mathbf{1}\{\pi_1(s_1) = a_1, \pi_2(s_1, a_1, s_2) = a_2\}$ for each trajectory $\mathbf{t} \in \mathcal{T}$. We denote the set of $\{0, 1\}$-valued variables $\{v(\mathbf{t}) : \mathbf{t} \in \mathcal{T}\}$ by $\mathbf{v}$, which we consider as a vector of length $|\mathcal{T}|$. By replacing the product term $\Pi_{k=1}^{2} \mathbf{1}\{\pi_k(s_1, a_1, \ldots, s_k) = a_k\}$ in equations (1) and (2) by the new variable $v(\mathbf{t})$, we have that the risk and the Bayes risk respectively are linear functions of $\mathbf{v}$. A challenge is that most vectors in $\{0, 1\}^{|\mathcal{T}|}$ do not represent any policy $\pi \in \Pi^{\text{DISC}}$. To handle

this, we add the sparse linear constraints (6)–(8). By theorem 1, these constraints enforce that $\mathbf{v}$ represents a unique policy $\pi \in \Pi^{\text{DISC}}$ and every policy $\pi \in \Pi^{\text{DISC}}$ is represented by a vector $\mathbf{v} \in \{0,1\}^{|\mathcal{T}|}$ satisfying these constraints.

It follows from theorem 1 that the discretized problem is equivalent to the binary integer program (4)–(8) from Section 2.2 at $K=2$, where the objective function (12) is represented by equation (4) and the type I error (13) and power (14) constraints respectively are represented by the following constraints of type (5):

$$\sum_{\mathbf{t}\in\mathcal{T}} v(\mathbf{t})\, p_{\boldsymbol{\Delta}}(\mathbf{t}) L^{\text{FWE}}(\boldsymbol{\Delta},\mathbf{t}) \leqslant \alpha, \qquad \text{for each } \boldsymbol{\Delta} \in G; \tag{15}$$

$$\sum_{\mathbf{t}\in\mathcal{T}} v(\mathbf{t})\, p_{\boldsymbol{\Delta}^{(m)}}(\mathbf{t}) L^{H^{(m)}}(\boldsymbol{\Delta}^{(m)},\mathbf{t}) \leqslant \beta_m, \qquad \text{for each } m \in \{1,\ldots,M\}. \tag{16}$$

The binary integer program is sparse since the vast majority of elements of the corresponding constraint matrix are 0, which follows from Section 2.3.

Define the *sparse linear program* as the binary integer program (4)–(8) tailored to represent the discretized problem as described in the previous paragraph, except with the integrality constraints (8) replaced by $v(\mathbf{t}) \geqslant 0$ for all $\mathbf{t} \in \mathcal{T}$ to make the problem computationally feasible. This is equivalent to constraining each variable $v(\mathbf{t})$ to be in the unit interval rather than being $\{0,1\}$ valued. We prove in section D of the on-line supplementary material that the sparse linear program is equivalent to the discretized problem where the set of (deterministic) policies Π^{DISC} is replaced by the larger set of stochastic policies Π^* defined in Section 2.2. The importance of representing the discretized problem over Π^* as the sparse linear program is that we have derived a computationally feasible approximation of the original adaptive design optimization problem (12)–(14). This relies on the fact that even very large, sparse linear programs can be computationally feasible to solve.

For any trajectory $\mathbf{t} = (s_1, a_1, s_2, a_2) \in \mathcal{T}$, the probability $p_{\boldsymbol{\Delta}}(\mathbf{t})$ that appears in the binary integer program satisfies $p_{\boldsymbol{\Delta}}(\mathbf{t}) = \Pr_{\boldsymbol{\Delta},a_1}\{\mathbf{Z}^{(1)} \in s_1, \mathbf{Z}^{(\text{F})} \in s_2\}$, where $\Pr_{\boldsymbol{\Delta},a_1}$ represents the multivariate normal distribution on the z-statistics $(\mathbf{Z}^{(1)}, \mathbf{Z}^{(\text{F})})$ under $\boldsymbol{\Delta}$ in the design that always enrols in stage 2 according to action a_1. This distribution is given in section A of the on-line supplementary material and can be evaluated by using the multivariate normal distribution function using the R package `mvtnorm` (Genz and Bretz, 2009). We also show at the end of section E of the supplementary material how the integral in the objective function (4) can similarly be computed for examples 1 and 2.

5. Solutions to examples 1 and 2

5.1. Problem definition

We solve the optimization problems in examples 1 and 2 from Section 3.5 over the class of discretized adaptive enrichment designs Π^{DISC}. The problem inputs depend on $p_1, \mathbf{n}, \sigma_1^2, \sigma_2^2, \alpha, \beta, \Delta^{\min}$ and σ_Λ^2, which we specify next. Let $p_1 = \frac{1}{2}$ and $\alpha = 0.05$ and assume a common variance $\sigma^2 = \sigma_1^2 = \sigma_2^2$.

We use the adaptive design template $\mathbf{n}^{(1\text{b})}$ defined in Section 3.2 and depicted in Fig. 1(b); the corresponding sample sizes are functions of n, i.e. the total sample size under $a_1 = 2$ (where both subpopulations are enrolled during stage 2). This adaptive design template allows enrichment of subpopulation 1 ($a_1 = 3$) or subpopulation 2 ($a_1 = 4$), in which case the total enrolled from the enriched subpopulation is n. We next describe the intuition for this choice of sample sizes. The power conditions 1–3 in Section 3.5.1 require the same power $1-\beta$ to reject $H_{0\text{C}}$ when $\Delta_1 = \Delta_2 = \Delta_{\min}$ as to reject H_{0s} when $\Delta_s = \Delta_{\min}$ and $\Delta_{s'} = 0$, 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 $a_1 = 2$ equals the information for Δ_s under $a_1 = 2 + s$, for each $s \in \{1, 2\}$, i.e. it is possible to generate the same information for the parameter of interest in each of conditions 1–3 by a corresponding choice for stage 2 enrolment.

For each of examples 1 and 2, the optimal solution to the adaptive design optimization problem depends on the inputs $(\sigma^2, \Delta^{\min}, n)$ only through the non-centrality parameter $\zeta = (n/8)^{1/2}\Delta^{\min}/\sigma$, as proved in section E of the on-line supplementary material. We set $\zeta = 2^{1/2}\Phi^{-1}(1 - 0.05) \approx 2.33$, for Φ the standard normal cumulative distribution function; for any $\sigma^2 > 0$ and $\Delta_{\min} > 0$, this is equivalent to setting $n = 16\sigma^2\{\Phi^{-1}(1 - 0.05)\}^2(\Delta^{\min})^{-2}$.

We use n defined above 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 condition 3 in Section 3.5.1 at $1 - \beta = 0.95$. In contrast, our optimization problem has the more stringent set of power conditions 1–3, 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 conditions 1–3 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 for examples 1 and 2 respectively.

We next set the constant c_Λ, which is used to define the covariance matrix $c_\Lambda(\Delta^{\min})^2\mathbf{I}^2$ in Λ^{mix}. As a benchmark, compare the distribution of $\mathbf{Z}^{(1)}$ under

(a) a point mass at $\boldsymbol{\Delta} = (\Delta^{\min}, \Delta^{\min})$ *versus*
(b) a bivariate normal distribution on $\boldsymbol{\Delta}$ centred at $(\Delta^{\min}, \Delta^{\min})$ with covariance matrix $c_\Lambda(\Delta^{\min})^2\mathbf{I}^2$.

In scenario (a), $\mathbf{Z}^{(1)}$ is bivariate normal with covariance matrix $\mathbf{I}^2$. In scenario (b), $\mathbf{Z}^{(1)}$ is bivariate normal with the same mean as in (a) and with covariance matrix $(1 + c_\Lambda\zeta^2/2)\mathbf{I}^2$. We set $c_\Lambda = \zeta^{-2}$ so that the latter covariance is 50% more than the former.

5.2. Implementation, discretization and iterative selection of G

To solve each sparse linear program, we used the IBM CPLEX solver (`https://www.ibm.com/support/knowledgecenter/SSSA5P_12.7.0/ilog.odms.studio.help/pdf/usrcplex.pdf`), version 12.4. To take advantage of the 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 10^{-10}.

We describe the discretization and two-step approach that we used to solve the discretized problem corresponding to example 2; the problem in example 1 had a similar structure and was solved analogously. In step 1, the sparse linear program was constructed by using the following discretization: $\mathcal{S}_1$ 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 action a_1, the multiple-testing procedure partition $\mathcal{S}_2$ consisted of unit squares covering $[-6, 7] \times [-6, 7]$ except that for $a_1 \neq 1$ (i.e. $a_1 \neq$ STOP) we replaced all squares in the lower left quadrant $[-6, 0] \times [-6, 0]$ by a single large square. Our use of different state spaces $\mathcal{S}_2$ depending on the action a_1 involves a minor extension of our general method from Section 2, which we present in section D of the on-line supplementary material.

We define G to be the 541 type I error constraints corresponding to the pairs of non-centrality parameters $\{(n/8)^{1/2}\sigma^{-1}\}(\Delta_1, \Delta_2)$ in the set $\{(x, y) : [x = 0, y \in \{-9, -8.9, \ldots, 9\}]$, or $[x \in \{-9, -8.9, \ldots, 9\}, y = 0]$ or $[x \in \{-9, -8.9, \ldots, 9\}, y = -x]\}$, which are grids along the

boundaries of the null spaces for H_{01}, H_{02} and H_{0C} respectively. This resulted in $w \approx 10^6$ variables in $\mathbf{v}$ and approximately 10^5 equality constraints in $\mathbf{A}^{(2)}$. We call the solution to the above sparse linear program the 'step 1' solution.

In step 2, we used features of the step 1 solution to refine the choice of G and the discretization in $\mathcal{S}_1$ and $\mathcal{S}_2$; we then solved the resulting discretized problem and iterated this refinement process. The refinement of G involved using the dual of the step 1 solution to identify the active type I error constraints approximately; we then augmented G by points Δ concentrated in small neighbourhoods of these active constraints. Further augmentation of G was done as described in section B.1 of the on-line supplementary material. A finer discretization was obtained by iteratively breaking some rectangles in $\mathcal{S}_1$ 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 enrolment. 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 $\mathcal{S}_2$. We incorporated additional constraints as described in section G 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 2 is denoted by $\mathbf{v}^{\text{opt}}$. This solution 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 is deterministic (non-randomized) except on a small fraction of rectangles; for each such rectangle, we rounded the corresponding fractional values. Each enrolment decision was set to the action a_1 in $\mathcal{A}_1$ with the largest corresponding probability. Each null hypothesis is rejected if the corresponding probability is above a threshold that depends on the end of stage 1 decision; the threshold is 0.5 when $a_1 \in \{1, 2\}$ and 0.9 otherwise. We picked these thresholds by examining the fractional parts of the solution, which occurred on the boundaries of decision and rejection regions, and then used trial and error to select thresholds that maintain strong control of the familywise type I error rate while changing expected sample size and power at conditions 1–3 in Section 3.5.1 by a negligible amount (each by less than 1%). The resulting policy $(\pi_1^{\text{opt}}, \pi_2^{\text{opt}})$ is depicted in Fig. 2. Strong control of the familywise type I error rate, i.e. that constraints (13) hold for all $\boldsymbol{\Delta} \in \mathbb{R}^2$, was verified as described in section B of the supplementary materials.

Hampson and Jennison (2015) solved 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 constrains type I error only at the global null hypothesis $\boldsymbol{\Delta} = (0, 0)$ to control the familywise type I error constraints at all other values of $\boldsymbol{\Delta}$ also; the approach of Wason and Jaki (2012) has a similar requirement. This requirement does not hold for our example problems. For example, when we solve our optimization problem in example 2 over Π^{DISC} but replacing the set of familywise type I error constraints by the single constraint at the global null hypothesis, the resulting optimal design has familywise type I error 0.953 at non-centrality parameters $(0, 3.57)$ and $(3.57, 0)$; these correspond to one subpopulation benefiting from treatment and no effect for the other subpopulation. This shows the need for including more type I error constraints than the global null hypothesis.

5.3. Optimal solution for example 2

We present the optimal adaptive enrichment design $(\pi_1^{\text{opt}}, \pi_2^{\text{opt}}) \in \Pi^{\text{DISC}}$ for the problem in example 2, which was computed by using sparse linear programming as described above; the solution to example 1 was qualitatively similar. We separately solved each sparse linear program

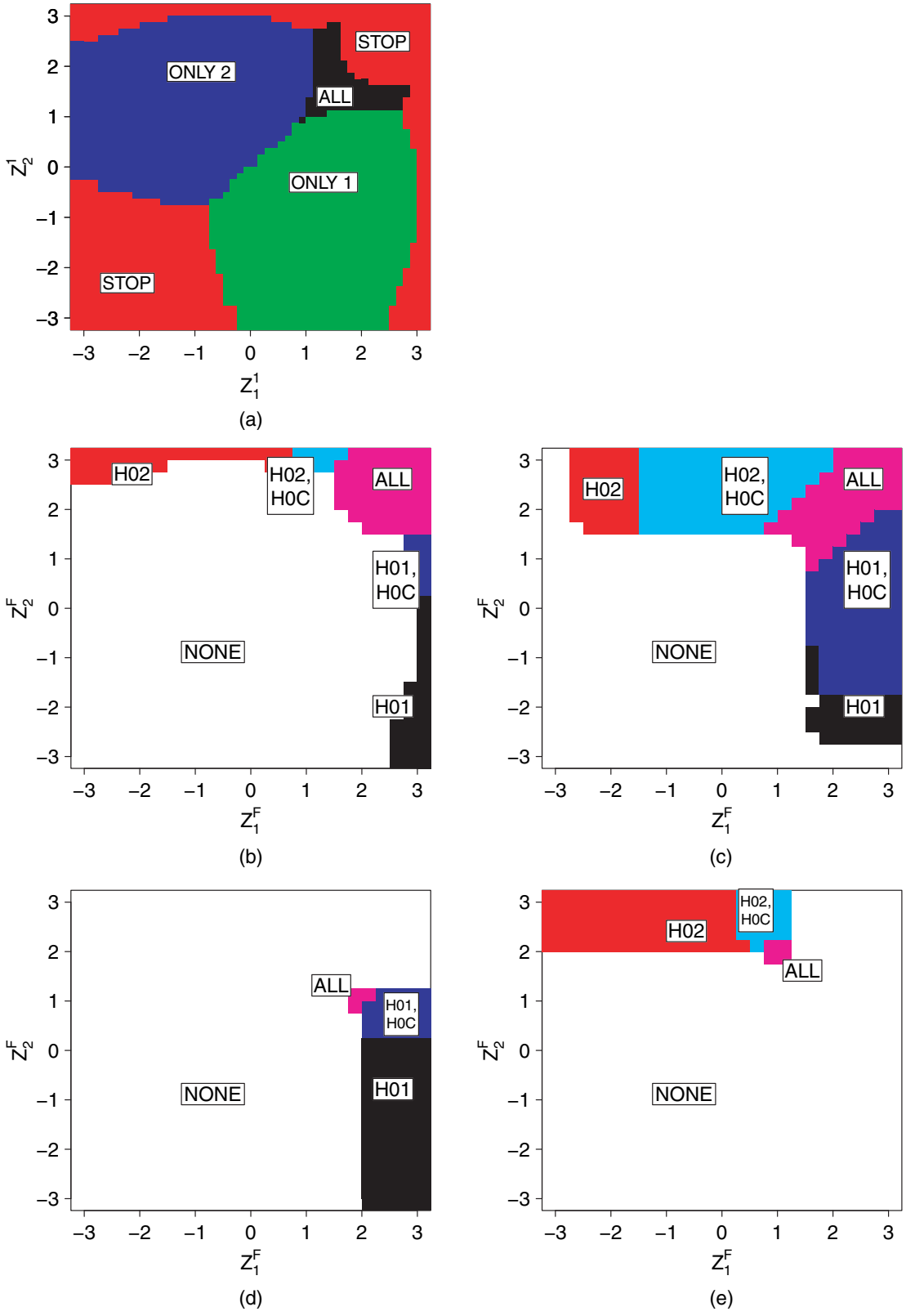

Fig. 2. Optimal design $(\pi_1^{opt}, \pi_2^{opt})$ for the discretized problem in example 2 at $1-\beta=0.82$ (stage 2 enrolment choices 'STOP', 'ALL', 'ONLY 1' and 'ONLY 2' correspond to $\pi_1^{opt}=1,2,3,4$): (a) decision rule π_1^{opt} for stage 2 enrolment (z-statistics correspond to $\mathbf{Z}^{(1)}$); (b)–(e) rejection regions of π_2^{opt} after each decision π_1^{opt} (z-statistics correspond to $\mathbf{Z}^{(F)}$); (b) $\pi_1^{opt} \equiv$ STOP; (c) $\pi_1^{opt} \equiv$ ALL; (d) $\pi_1^{opt} \equiv$ ONLY 1; (e) $\pi_1^{opt} \equiv$ ONLY 2

at every power constraint threshold $\beta \in \{0.01, \ldots, 0.99\}$, where $1-\beta$ represents the required power in each power constraint 1–3 in Section 3.5.1. Larger values of $1-\beta$ correspond to stricter constraints. Our results show that the problems are feasible, i.e. the type I error and power constraints 1–3 can be simultaneously satisfied, if and only if $1-\beta < 0.83$.

For the case of $1-\beta = 0.82$, Fig. 2 depicts the optimal solution $(\pi_1^{\text{opt}}, \pi_2^{\text{opt}}) \in \Pi^{\text{DISC}}$ to example 2. We first focus on Fig. 2(a), which represents the decision rule π_1^{opt}. The different regions correspond to the four possible stage 2 enrolment choices from the adaptive design template $\mathbf{n}^{(1b)}$. The top right and bottom left regions (in red) of Fig. 2(a) correspond to stopping the trial after stage 1 (i.e. $\pi_1^{\text{opt}} = 1$, marked 'STOP'). 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), whereas the bottom left region represents stopping early for futility (since no null hypothesis is rejected if $\mathbf{Z}^{(1)}$ is in this region). The black region marked 'ALL' represents the choice $\pi_1^{\text{opt}} = 2$ to continue enrolment from both subpopulations in stage 2. Intuitively, this occurs when the stage 1 z-statistics for each subpopulation both indicate a non-negligible, positive signal that is not sufficiently strong to allow outright rejection of any null hypothesis; this motivates the investment of stage 2 enrolment from both subpopulations, to determine which (if any) null hypotheses to reject. The green and blue regions marked 'ONLY 1' (representing $\pi_1^{\text{opt}} = 3$) and 'ONLY 2' (representing $\pi_1^{\text{opt}} = 4$) respectively represent choosing stage 2 enrolment to be only from the corresponding subpopulation.

Figs 2(b)–2(e) represent the multiple-testing procedure π_2^{opt} that is used after each of the four enrolment choices. For each possible value of the enrolment decision π_1^{opt}, the corresponding plot shows the mapping from the final z-statistics $\mathbf{Z}^{(\text{F})}$ to the set of null hypotheses rejected. Figs 2(b), 2(c), 2(d) and 2(e) correspond to $\pi_1^{\text{opt}} = 1, 2, 3, 4$ respectively. Each plot has a white region where nothing is rejected (marked 'NONE') and coloured regions where specified null hypotheses are rejected.

The plot of π_2^{opt} for $\pi_1^{\text{opt}} \equiv \text{STOP}$ in Fig. 2(b) has coloured regions whose union is approximately identical to the red 'STOP' region in the upper right of Fig. 2(a). This means that, when the first-stage z-statistics are in the red 'STOP' region in the upper right of Fig. 2(a), at least one null hypothesis will be rejected by π_2^{opt} (since, when $\pi_1^{\text{opt}} \equiv \text{STOP}$, the first-stage z-statistics $\mathbf{Z}^{(1)}$ are identical to the final z-statistics $\mathbf{Z}^{(\text{F})}$). 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 π_2^{opt} compared with π_1^{opt}: a choice we made to reduce the computational requirements for solving the optimization problem.

Next, consider the plot of π_2^{opt} for $\pi_1^{\text{opt}} \equiv \text{ALL}$ in Fig. 2(c). This is qualitatively similar to the plot of π_2^{opt} for $\pi_1^{\text{opt}} \equiv \text{STOP}$, 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 enrolment decision $\pi_1^{\text{opt}} \equiv \text{ALL}$ incorporate twice as much data as under $\pi_1^{\text{opt}} \equiv \text{STOP}$ 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 because more information is 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 π_2^{opt} for $\pi_1^{\text{opt}} \equiv \text{ALL}$ where we might have expected red and black (i.e. extensions of these regions) respectively. We conjecture that this is due to the very small joint probability of $\pi_1^{\text{opt}}(\mathbf{Z}^{(1)}) \equiv \text{ALL}$ and π_2^{opt} being in these white areas; these 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 that is used in solving the sparse linear program. Asymmetries in the plots, which occur at some rectangles on or near the boundaries of different regions, may

have a similar explanation; some arise from minor differences that are accentuated because of rounding.

Consider the plot of π_2^{opt} for $\pi_1^{\text{opt}} \equiv$ ONLY 1 in Fig. 2(d). No null hypothesis is rejected when $Z_2^{(\text{F})} > 1.25$. In fact, it is not possible to have both $\pi_1^{\text{opt}}(\mathbf{Z}^{(1)}) \equiv$ ONLY 1 and $Z_2^{(\text{F})} > 1.25$. This is a consequence of the green 'ONLY 1' region in Fig. 2(a) being below the horizontal line $Z_2^{(1)} = 1.25$, as explained next. Since the enrolment decision $\pi_1^{\text{opt}} \equiv$ ONLY 1 occurs precisely when $\mathbf{Z}^{(1)}$ is in the green 'ONLY 1' region in Fig. 2(a), and since $Z_2^{(1)} = Z_2^{(\text{F})}$ whenever $\pi_1^{\text{opt}} \equiv$ ONLY 1 (because no new subpopulation 2 data are collected in stage 2), it is not possible to have $\pi_1^{\text{opt}}(\mathbf{Z}^{(1)}) \equiv$ ONLY 1 and $Z_2^{(\text{F})} > 1.25$. The plot of π_2^{opt} for $\pi_1^{\text{opt}} \equiv$ ONLY 2 in Fig. 2(e) is (approximately) a symmetric version of the plot for $\pi_1^{\text{opt}} \equiv$ ONLY 1.

The decision rule π_1^{opt} in Fig. 2(a) continues to enrol subpopulation 1 when $\mathbf{Z}^{(1)} = (2.9, 0)$ but stops the trial when $\mathbf{Z}^{(1)} = (2.9, -3)$. We conjecture that, because of the power constraint condition 1 in Section 3.5.1 and the prior distribution, there is a greater incentive to enrol subpopulation 1 in stage 2 in the former case. A related phenomenon is that in Fig. 2(b) ($\pi_1^{\text{opt}} \equiv$ STOP), H_{01} is rejected when $\mathbf{Z}^{(\text{F})} = (2.9, -3)$ but nothing is rejected if the second component is increased to 0. Rejecting nothing in the latter case is a result of the shape of Fig. 2(a), which makes it impossible to have $\pi_1^{\text{opt}} \equiv$ STOP and $\mathbf{Z}^{(\text{F})} = (2.9, 0)$ (since this would require $\mathbf{Z}^{(1)} = (2.9, 0)$ but then π_1^{opt} would continue enrolling from subpopulation 1).

5.4. Comparison with some alternative designs

For each problem in Section 5.1, we compare the performance of optimal designs from different classes. Some of these classes use much simpler decision rules and/or multiple-testing procedures than Π^{DISC}. All designs below (except in the final two paragraphs of this subsection) use the same template from Section 5.1 i.e. $\mathbf{n}^{(1b)}$ with n defined in Section 5.1.

Let $\pi_1^{\text{STD}} \in \Pi_1^{\text{DISC}}$ denote the decision rule corresponding to the standard (non-adaptive) design that always enrols from both subpopulations in stage 2, i.e. $\pi_1^{\text{STD}} = 2$ for all values of the stage 1 statistics. This is equivalent to a design with no interim analysis that enrols n participants, with $p_s n$ from each subpopulation (where each $p_s = \frac{1}{2}$ in our case). Define the class of standard (non-adaptive) designs to be $\Pi^{\text{STD}} = \{(\pi_1^{\text{STD}}, \pi_2) : \pi_2 \in \Pi_2^{\text{DISC}}\}$.

We next define a class Π^{COMB} of adaptive enrichment designs based on the p-value combination approach of Bauer (1989), Bauer and Köhne (1994) and 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, for example, Bretz *et al.* (2006), Schmidli *et al.* (2006), Jennison and Turnbull (2007), Brannath *et al.* (2009), Jenkins *et al.* (2011) and Boessen *et al.* (2013). This general approach was used in the TAPPAS trial as well. Since it is an open problem how to optimize over the class of all possible designs that can be constructed by using this approach, we instead define a low dimensional, simple class Π^{COMB} of such designs. The full description of Π^{COMB} is given in section F of the on-line supplementary material, but we summarize the key features. The multiple-testing procedure, denoted π_2^{pv}, uses the Dunnett intersection test (Dunnett, 1955; Jennison and Turnbull, 2007), with p-values combined across stages by 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, for example, 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 $\pi_1^{(t_{\text{c}}, t_{\text{i}})}(\mathbf{Z}^{(1)})$ as follows: if the multiple-testing procedure π_2^{pv} rejects any null hypothesis at the end of stage 1, stop the entire trial; otherwise, if the combined population statistic $(Z_1^{(1)} + Z_2^{(1)})/\sqrt{2} > t_{\text{c}}$, enrol both subpopulations in stage 2; otherwise, enrol in stage 2 from each subpopulation s for which $Z_s^{(1)} > t_{\text{i}}$. Let

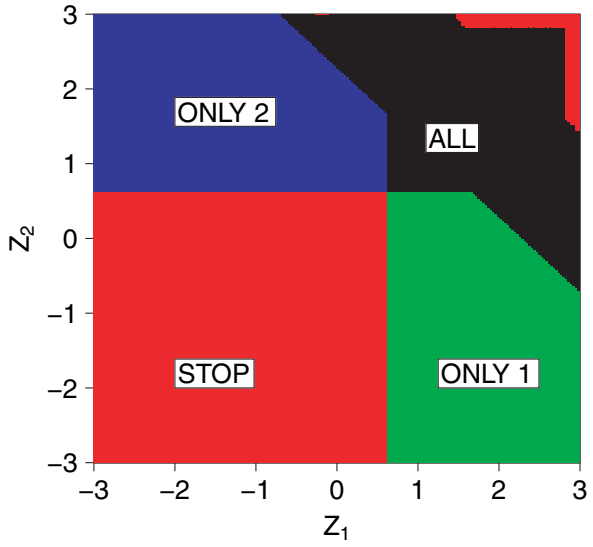

Fig. 3. Enrolment decision rule $\pi_1^{(t_c,t_i)}$ for $(t_c, t_i) = (1.6, 0.6)$, which corresponds to the optimal design over Π^{COMB} for the problem in example 2 under the power constraints 1–3 at $1-\beta = 0.74$ (the z-statistics in the plot correspond to first-stage statistics $\mathbf{Z}^{(1)}$; stage 2 enrolment choices 'STOP', 'ALL', 'ONLY 1' and 'ONLY 2' correspond to decisions 1, 2, 3 and 4 respectively, from the adaptive design template $\mathbf{n}^{(1b)}$): ■, stopping the trial at the end of stage 1

$\Pi^{COMB} = \{(\pi_1^{(t_c,t_i)}, \pi_2^{pv}) : (t_c, t_i) \in (-3, -2.9, \ldots, 3) \times (-3, -2.9, \ldots, 3)\}$. An example of the decision rule $\pi_1^{(t_c,t_i)}$ is depicted in Fig. 3. Each design in Π^{COMB} strongly controls the familywise type I error rate. This holds even if the end of stage 1 decision rule is not followed, which is a general property of using the p-value combination approach for multiple testing; this property does not generally hold for designs in Π^{DISC}.

A referee suggested a hybrid class of designs that incorporates features from both Π^{DISC} and Π^{COMB}. Specifically, it uses the end of stage 1 decision rules Π_1^{DISC}, but instead of allowing for optimization of the multiple-testing procedure it uses the p-value combination test π_2^{pv} at the end of stage 2. Define the hybrid class $\Pi^{HYBRID} = \{(\pi_1, \pi_2^{pv}) : \pi_1 \in \Pi_1^{DISC}\}$. Each design in this class maintains the advantageous feature of the p-value combination test that the familywise type I error rate is strongly controlled at level 0.05 even if the end of stage 1 decision rule is not followed.

We next compare the expected sample size of the optimal design in each of the classes Π^{DISC}, Π^{STD}, Π^{COMB} and Π^{HYBRID} as we vary the power constraint $1-\beta$. Let ESS denote the value of the objective function (12), which represents the expected sample size with respect to the corresponding prior. For each of examples 1 and 2 and each value of $1-\beta$ in the top row of Table 1, we solved the adaptive design optimization problem from Section 5.1 over each class of designs Π^{DISC}, Π^{STD}, Π^{COMB} and Π^{HYBRID}. For all except Π^{COMB}, we used the sparse linear programming method from Section 4.4. For Π^{COMB}, we did an exhaustive search over the set of thresholds (t_c, t_i) given above.

We first compare the optimal designs over Π^{DISC} *versus* Π^{STD}. The problems in examples 1 and 2 are infeasible for the class Π^{STD} whenever the power constraint $1-\beta > 0.65$, i.e. it is not possible to satisfy simultaneously the type I error constraints and power constraint conditions 1–3 in Section 3.5.1; in contrast, the problem is feasible for the class Π^{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 Π^{DISC} (which uses adaptive design template $\mathbf{n}^{(1b)}$); these

Table 1. Performance comparison of optimal adaptive designs from different classes, for examples 1 and 2†

	Results for the following power constraints $1-\beta$:						
	58%	*62%*	*66%*	*70%*	*74%*	*78%*	*82%*
Example 1							
Minimum ESS over Π^{DISC}	$0.65n$	$0.69n$	$0.74n$	$0.79n$	$0.85n$	$0.92n$	$1.04n$
Minimum ESS over Π^{HYBRID}	$0.74n$	$0.78n$	$0.82n$	$0.86n$	$0.92n$	$1.07n$	×
Minimum ESS over Π^{COMB}	$0.86n$	$0.89n$	$0.92n$	$0.97n$	$1.01n$	×	×
Example 2							
Minimum ESS over Π^{DISC}	$0.64n$	$0.67n$	$0.72n$	$0.76n$	$0.81n$	$0.88n$	$1.00n$
Minimum ESS over Π^{HYBRID}	$0.73n$	$0.76n$	$0.79n$	$0.83n$	$0.88n$	$1.04n$	×
Minimum ESS over Π^{COMB}	$0.89n$	$0.92n$	$0.95n$	$0.98n$	$1.01n$	×	×
Example 2, for design classes with 10 end of stage 1 enrolment options							
Minimum ESS over $\Pi^{\mathrm{DISC,10}}$	$0.59n$	$0.63n$	$0.67n$	$0.72n$	$0.78n$	$0.86n$	$0.98n$
Minimum ESS over $\Pi^{\mathrm{DISC,10,FS}}$	$0.55n$	$0.60n$	$0.66n$	$0.72n$	$0.78n$	$0.86n$	$0.95n$
Total stage 1 sample size for optimal design in $\Pi^{\mathrm{DISC,10,FS}}$	$0.25n$	$0.31n$	$0.38n$	$0.44n$	$0.50n$	$0.50n$	$0.56n$

†The top two sections compare the three classes of adaptive designs Π^{DISC}, Π^{HYBRID} and Π^{COMB}; the bottom section compares the augmented design classes $\Pi^{\mathrm{DISC,10}}$ and $\Pi^{\mathrm{DISC,10,FS}}$ only for example 2. The symbol '×' indicates that no design in the class satisfies the type I error constraints and power constraints 1–3 in Section 3.5.1 at the required power threshold $1-\beta$.

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 our problems.

The top two sections of Table 1 give the optimal ESS for the adaptive design optimization problems in examples 1 and 2, comparing the classes of adaptive designs Π^{DISC}, Π^{HYBRID} and Π^{COMB}. At all values of $1-\beta$ that we considered, the minimum value of ESS was substantially lower for the optimal design over Π^{DISC} compared with the optimal design over Π^{COMB}. For example, in example 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 Π^{COMB} at $1-\beta \geqslant 0.78$, i.e. it is not possible to satisfy simultaneously the type I error constraints and power constraints 1–3; in contrast, the problem is feasible for the class Π^{DISC} up to power threshold $1-\beta = 0.82$. One reason that the designs Π^{COMB} achieve lower power is that their maximum type I error over $\boldsymbol{\Delta} \in \mathbb{R}^2$ is less than 0.05; for example, it is 0.039 for the optimal such design for example 2 at $1-\beta = 0.74$.

We next examine optimal designs from Π^{HYBRID}. Each design in this class maintains the advantageous feature of the *p*-value combination test that the familywise type I error rate is strongly controlled at level 0.05, even if the end of stage 1 decision rule is not followed. This added flexibility comes at the cost of larger expected sample size compared with the optimal design in Π^{DISC}. Consider examples 1 and 2 and each power threshold between 58% and 74%. The expected sample size of the optimal design in Π^{HYBRID} is roughly halfway between that of Π^{DISC} and Π^{COMB} (being a little closer to the former than to the latter). This shows that just over half of the improvement in expected sample size comparing Π^{DISC} with Π^{COMB} can be attributed to the improved end of stage 1 decision rule in the former. The optimal design in Π^{HYBRID} can achieve the 78% power threshold (like Π^{DISC} but not Π^{COMB}) but not the 82% power threshold (only achieved by Π^{DISC}).

We next evaluate the effect of adding more stage 2 enrolment options to the template $\mathbf{n}^{(1b)}$ in Π^{DISC}. Define the class $\Pi^{\text{DISC},10}$ to be Π^{DISC} except with the following 10 options for stage 2 enrolment: (0,0), (1,1), (2,2), (3,3), (0,1), (0,2), (0,3), (1,0), (2,0) and (3,0), where each pair (x, y) represents stage 2 sample sizes $(n_{21}, n_{22}) = (xn/4, yn/4)$; for example, (3,0) represents $3n/4$ enrolled from subpopulation 1 and none from subpopulation 2 in stage 2. The pairs (0,0), (1,1), (3,0) and (0,3) represent the original four enrolment choices shown in Figs 1(b)–1(e). The other choices allow enrolment of both subpopulations or a single subpopulation at different sample sizes. The expected samples sizes of the optimal designs over $\Pi^{\text{DISC},10}$ for example 2 are shown in the bottom third of Table 1. Compared with Π^{DISC}, the expected sample sizes are somewhat lower, with the relative difference decreasing from 8% to 2% as the power constraint is increased from 58% to 82%.

We next explored the effect of the first-stage sample sizes. We modified the class $\Pi^{\text{DISC},10}$ by setting (n_{11}, n_{12}) to be the original $(n/4, n/4)$ multiplied by each of the following nine scaling factors: 0.5, 0.625, 0.75, 0.875, 1, 1.125, 1.25, 1.375 and 1.5. Let $\Pi^{\text{DISC},10,\text{FS}}$ denote the union of the resulting nine classes of designs (each corresponding to $\Pi^{\text{DISC},10}$ but with first-stage sample size scaled by one of the above factors). To optimize over this larger class $\Pi^{\text{DISC},10,\text{FS}}$, we constructed nine sparse linear programs, solved them separately and selected the solution with the smallest expected sample size. The results, shown in Table 1, indicate only small improvements compared with using the original first-stage sample sizes. The first-stage total sample size $n_{11} + n_{12}$ for the optimal design in $\Pi^{\text{DISC},10,\text{FS}}$ is shown in the last row of Table 1; this should be compared with the original, total first-stage sample size $0.5n$. Unsurprisingly, smaller first-stage sample sizes are optimal for less stringent power constraints.

6. Discussion

Our general approach from Section 2 outputs a stochastic policy. In our applications, most components of the optimal policy were deterministic ($\{0, 1\}$ valued) and the remaining fractional components were rounded, leading to a negligible effect on expected sample size and type I–II error. This is not guaranteed to occur in general.

We solved problems that are analogous to examples 1 and 2, except involving only the null hypotheses H_{01} and H_{0C} and allowing enrichment of subpopulation 1 only. The reductions in expected sample size comparing the optimal designs from Π^{DISC} to Π^{COMB} were roughly similar in magnitude to those for the original problems involving the three null hypotheses H_{01}, H_{02} and H_{0C}. Full details are given in section H of the on-line supplementary material.

The optimal adaptive enrichment designs from Π^{DISC} that were 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 $\mathbf{n}$. The class of hybrid designs Π^{HYBRID} may be a useful efficiency or flexibility compromise, because they achieve some of the expected sample size reductions of Π^{DISC} while retaining strong control of the familywise type I error rate even if the end of stage 1 decision rule is not followed. Alternatively, when the added value of the optimal design in Π^{DISC} is substantial, its decision rule could be approximated by replacing the discretized regions in Fig. 2 by simpler curves.

We assumed that subpopulation proportions are known in advance. To deal with uncertainty in these proportions, one could build in constraints on power and type I error that need to hold over a range of these values; this would define a new optimization problem that could be approximated by our approach, e.g. by using a grid of values of the subpopulation proportions

over a prespecified range. This is an area for future research. Another future research direction is to optimize over different prior distributions by using priors elicited from subject matter experts, to investigate sensitivity to this choice.

A limitation of our approach, as stated in Section 1, is that we assume that each participant's outcome is measured relatively soon after her or his enrolment. In contrast, the TAPPAS trial involved survival outcomes, which lead to correlations between statistics computed by using stage 1 data and statistics computed by using stage 2 data; this is because some participants will contribute information to both stages. Our approach can be modified to handle such correlations by incorporating them into the joint distribution of statistics in section A of the on-line supplementary material; there, we describe how the only change in implementing our method would be to include these correlations as inputs in each evaluation of the multivariate normal distribution function.

The main feature that made our adaptive design problem challenging was the multiple constraints (on power and type I error), which cannot generally be handled by backward induction methods. This situation arises more generally when it is desired to optimize average case performance under constraints on worst-case performance (or when the goal is simply to minimize the maximum expected loss). Our method can incorporate such constraints, and the objective function can be modified to represent minimax problems. This may be useful in problems where there is substantial *a priori* uncertainty about parameter values and the goal is to ensure good performance over a range of such values.

Our general method in Section 2 has potential applications to other two-stage experimental design problems where statistics at each stage have one or two discrete components or real-valued components that can be discretized by using the approach in Section 4.2. We briefly discuss potential applications to two-stage, non-linear regression problems from Abdelbasit and Plackett (1983), section 4, and Lane *et al.* (2014), and to the group screening problem of Lewis and Dean (2001), in section J of the on-line supplementary material.

We created an R package implementing our method, which is described in section I of the supplementary material.

Acknowledgements

This work was funded by the Patient-Centered Outcomes Research Institute (grant ME-1306-03198) and the US Food and Drug Administration (grant 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 these organizations.

Han Liu was supported by National Science Foundation grants BIGDATA 1840866, CAREER 1841569, TRIPODS 1740735 and grants DARPA-PA-18-02-09-QED-RML-FP-003, along with an Alfred P. Sloan Fellowship and a PECASE award.

Ethan X. Fang was supported by National Science Foundation grants DMS-1820702 and DMS-1953196.

References

Abdelbasit, K. M. and Plackett, R. L. (1983) Experimental design for binary data. *J. Am. Statist. Ass.*, **78**, 90–98.

Albers, G. W., Lansberg, M. G., Kemp, S., Tsai, J. P., Lavori, P., Christensen, S., Mlynash, M., Kim, S., Hamilton, S., Yeatts, S. D., Palesch, Y., Bammer, R., Broderick, J. and Marks, M. P. (2017) A multicenter randomized controlled trial of endovascular therapy following imaging evaluation for ischemic stroke (defuse 3). *Int. J. Stroke*, **12**, 896–905.

Basu, D. (1978) On partial sufficiency: a review. *J. Statist. Planng Inf.*, **2**, 1–13.
Bauer, P. (1989) Multistage testing with adaptive designs (with discussion). *Biometr. Inform. Med. Biol.*, **20**, 130–148.
Bauer, P. and Köhne, K. (1994) Evaluations of experiments with adaptive interim analyses. *Biometrics*, **50**, 1029–1041.
Bertsekas, D. P. (2017) *Dynamic Programming and Optimal Control*, 4th edn, vol. 1. New York: Springer.
Boessen, R., van der Baan, F., Groenwold, R., Egberts, A., Klungel, O., Grobbee, D., Knol, M. and Roes, K. (2013) Optimizing trial design in pharmacogenetics research: comparing a fixed parallel group, group sequential, and adaptive selection design on sample size requirements. *Pharm. Statist.*, **12**, 366–374.
Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M. and Racine-Poon, A. (2009) Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statist. Med.*, **28**, 1445–1463.
Bretz, F., Schmidli, H., König, F., Racine, A. and Maurer, W. (2006) Confirmatory seamless phase ii/iii clinical trials with hypotheses selection at interim: general concepts. *Biometr. J.*, **48**, 623–634.
Dunnett, C. W. (1955) A multiple comparison procedure for comparing several treatments with a control. *J. Am. Statist. Ass.*, **50**, 1096–1121.
European Medicines Agency (2007) Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. European Medicines Agency, London. (Available from `https://www.ema.europa.eu/documents/scientificguideline/reflection-paper-methodological-issues-confirmatory-clinical-trials-planned-adaptive-design-en.pdf`.)
Follmann, D. (1997) Adaptively changing subgroup proportions in clinical trials. *Statist. Sin.*, **7**, 1085–1102.
Food and Drug Administration (2016) Guidance for industry: Adaptive designs for medical device clinical studies. Food and Drug Administration, Whiteoak. (Available from `https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm446729.pdf`.)
Food and Drug Administration (2019) Guidance for industry: Adaptive design clinical trials for drugs and biologics. Food and Drug Administration, Whiteoak. (Available from `https://www.fda.gov/media/78495/download`.)
Food and Drug Administration and European Medicines Agency (1998) E9 statistical principles for clinical trials. *Report CPMP/ICH/363/96*. European Medicines Agency, London.
Freidlin, B. and Simon, R. (2005) Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clin. Cancer Res.*, **11**, 7872–7878.
Freidlin, B., Sun, Z., Gray, R. and Korn, E. L. (2013) Phase iii clinical trials that integrate treatment and biomarker evaluation. *J. Clin. Oncol.*, **31**, 3158–3161.
Friede, T., Parsons, N. and Stallard, N. (2012) A conditional error function approach for subgroup selection in adaptive clinical trials. *Statist. Med.*, **31**, 4309–4320.
Genz, A. and Bretz, F. (2009) *Computation of Multivariate Normal and t Probabilities*. Heidelberg: Springer.
Götte, H., Donica, M. and Mordenti, G. (2015) Improving probabilities of correct interim decision in population enrichment designs. *J. Biopharm. Statist.*, **25**, 1020–1038.
Graf, A. C., Posch, M. and Koenig, F. (2015) Adaptive designs for subpopulation analysis optimizing utility functions. *Biometr. J.*, **57**, 76–89.
Hampson, L. V. and Jennison, C. (2015) Optimizing the data combination rule for seamless phase II/III clinical trials. *Statist. Med.*, **34**, 39–58.
Hochberg, Y. and Tamhane, A. C. (1987) *Multiple Comparison Procedures*. New York: Wiley Interscience.
Jenkins, M., Stone, A. and Jennison, C. (2011) An adaptive seamless phase ii/iii design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharm. Statist.*, **10**, 347–356.
Jennison, C. and Turnbull, B. W. (2007) Adaptive seamless designs: selection and prospective testing of hypotheses. *J. Biopharm. Statist.*, **17**, 1135–1161.
Jones, R. L., Attia, S., Mehta, C. R., Liu, L., Sankhala, K. K., Robinson, S. I., Ravi, V., Penel, N., Stacchiotti, S., Tap, W. D., Alvarez, D., Yocum, R., Theuer, C. P. and Maki, R. G. (2017) TAPPAS: an adaptive enrichment phase 3 trial of TRC105 and pazopanib versus pazopanib alone in patients with advanced angiosarcoma (AAS). *J. Clin. Oncol.*, **35**, suppl., article TPS11081.
Jovin, T. G., Saver, J. L., Ribo, M., Pereira, V., Furlan, A., Bonafe, A., Baxter, B., Gupta, R., Lopes, D., Jansen, O., Smith, W., Gress, D., Hetts, S., Lewis, R. J., Shields, R., Berry, S. M., Graves, T. L., Malisch, T., Rai, A., Sheth, K. N., Liebeskind, D. S. and Nogueira, R. G. (2017) Diffusion-weighted imaging or computerized tomography perfusion assessment with clinical mismatch in the triage of wake up and late presenting strokes undergoing neurointervention with Trevo (DAWN) trial methods. *Int. J. Stroke*, **12**, 641–652.
Krisam, J. and Kieser, M. (2015) Optimal decision rules for biomarker-based subgroup selection for a targeted therapy in oncology. *Int. J. Molec. Sci.*, **16**, 10354–10375.
Lai, T. L., Lavori, P. W. and Liao, O. Y.-W. (2014) Adaptive choice of patient subgroup for comparing two treatments. *Contemp. Clin. Trials*, **39**, 191–200.
Lane, A., Yao, P. and Flournoy, N. (2014) Information in a two-stage adaptive optimal design. *J. Statist. Planng Inf.*, **144**, 173–187.
Lehmacher, W. and Wassmer, G. (1999) Adaptive sample size calculations in group sequential trials. *Biometrics*, **55**, 1286–1290.

Lewis, S. M. and Dean, A. M. (2001) Detection of interactions in experiments on large numbers of factors (with discussion). *J. R. Statist. Soc.* B, **63**, 633–672.
Marcus, R., Peritz, E. and Gabriel, K. R. (1976) On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**, 655–660.
Mehta, C. R., Liu, L. and Theuer, C. (2019) An adaptive population enrichment phase III trial of TRC105 and pazopanib versus pazopanib alone in patients with advanced angiosarcoma (TAPPAS trial). *Ann. Oncol.*, **30**, 103–108.
O'Brien, P. and Fleming, T. (1979) A multiple testing procedure for clinical trials. *Biometrics*, **35**, 549–556.
Rosenblum, M. and van der Laan, M. J. (2011) Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika*, **98**, 845–860.
Rosenblum, M., Liu, H. and Yen, E.-H. (2014) Optimal tests of treatment effects for the overall population and two subpopulations in randomized trials, using sparse linear programming. *J. Am. Statist. Ass.*, **109**, 1216–1228.
Russek-Cohen, E. and Simon, R. M. (1997) Evaluating treatments when a gender by treatment interaction may exist. *Statist. Med.*, **16**, 455–464.
Schmidli, H., Bretz, F., Racine, A. and Maurer, W. (2006) Confirmatory seamless phase ii/iii clinical trials with hypotheses selection at interim: applications and practical considerations. *Biometr. J.*, **48**, 635–643.
Stallard, N., Hamborg, T., Parsons, N. and Friede, T. (2014) Adaptive designs for confirmatory clinical trials with subgroup selection. *J. Biopharm. Statist.*, **24**, 168–187.
Wang, S. J., Hung, H. and O'Neill, R. T. (2009) Adaptive patient enrichment designs in therapeutic trials. *Biometr. J.*, **51**, 358–374.
Wang, S. J., O'Neill, R. T. and Hung, H. (2007) Approaches to evaluation of treatment effect in randomized clinical trials with genomic subsets. *Pharm. Statist.*, **6**, 227–244.
Wason, J. M. S. and Jaki, T. (2012) Optimal design of multi-arm multi-stage trials. *Statist. Med.*, **31**, 4269–4279.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material for Optimal, two stage, adaptive enrichment designs for randomized trials, using sparse linear programming'.