JOHNS HOPKINS
BLOOMBERG
SCHOOL of PUBLIC HEALTH

Johns Hopkins University, Dept. of Biostatistics Working Papers

12-7-2016

STOCHASTIC OPTIMIZATION OF ADAPTIVE ENRICHMENT DESIGNS FOR TWO SUBPOPULATIONS

Aaron Fisher
Harvard T.H. Chan School of Public Health, aafisher@hsph.harvard.edu

Michael Rosenblum
Johns Hopkins University Bloomberg School of Public Health

Stochastic Optimization of Adaptive Enrichment Designs for Two Subpopulations

Aaron Fisher

Harvard T.H. Chan School of Public Health, 677 Huntington Avenue Boston, MA 02115, USA

E-mail: aafisher@hsph.harvard.edu

Michael Rosenblum

Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

for the Alzheimer's Disease Neuroimaging Initiative

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at:
http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Summary.
An adaptive enrichment design is a randomized trial that allows enrollment criteria to be modified at interim analyses, based on a preset decision rule. When there is prior uncertainty regarding treatment effect heterogeneity, these trial designs can provide improved power for detecting treatment effects in subpopulations. We present a simulated annealing approach to search over the space of decision rules and other parameters for an adaptive enrichment design. The goal is to minimize the expected number enrolled or expected duration, while preserving the appropriate power and Type I error rate. We also explore the benefits of parallel computation in the context of this goal. We find that optimized designs can be substantially more efficient than simpler designs using Pocock or O'Brien-Fleming boundaries.

Keywords: Clinical Trials, Optimization, Simulated Annealing, Treatment Effect Heterogeneity

1. Introduction

Prior uncertainty regarding treatment effect heterogeneity can pose a challenge to trial designers. If the treatment only benefits a subset of the population, standard clinical trials enrolling from the entire population may have low power. However, if the entire population benefits, a standard trial enrolling only one subpopulation will not provide any information about the complementary population.

These issues can be mitigated with the use of an adaptive enrichment trial design. Such a design involves a decision rule for early stopping of participant accrual in different population subsets based on interim analyses (Wang et al., 2009). For example, early stopping can occur if there is strong evidence early in the trial of the treatment's benefit or harm for a subpopulation. The design also includes a multiple testing procedure, since there is one null hypothesis to test for each population of interest.

We aim to optimize the enrollment modification rule and multiple testing procedure for an adaptive enrichment design. The goal is to minimize either the expected number enrolled or expected trial duration, under constraints on power and the Type I error rate. We focus on designs that are guaranteed to strongly control the familywise Type I error rate, i.e., where the probability is at most α that one or more true null hypotheses is rejected, regardless of the (unknown) data generating distribution.

General approaches exist for constructing optimal designs for simpler problems, such as those involving a single null hypothesis (Eales and Jennison, 1992; Hampson and Jennison, 2013). Hampson and Jennison (2015) extend this approach to handle multiple hypotheses in two-stage designs, but the resulting designs are not guaranteed to strongly control the familywise Type I error rate (which must be checked by simulation). Thall et al. (1988) perform a 2-dimensional grid search to minimize the expected number enrolled of a 2-stage trial comparing the effects of several treatments. Krisam and Kieser (2015), Graf et al. (2015), and Rosenblum et al. (2016), consider different adaptive designs involving two subpopulations, which they optimize over at most a few parameters. In contrast to this related work, our aim is to search over more flexible, higher dimensional families of designs. For trials involving two subpopulations, optimal 2-stage designs can be found via sparse linear programming (Rosenblum et al., 2014), but this approach becomes computationally infeasible for more than two stages.

The optimization problems we consider are challenging in that no existing approach is guaranteed to find the global optimum. The main difficulty is that there are many design parameters to optimize over, as well as multiple constraints. The parameters in our adaptive enrichment designs include the following (plus additional parameters in some settings): the number of stages; per-stage sample sizes; and, an efficacy and futility boundary for each population at each stage. For example, in the case of 2 subpopulations and 5 stages, there are over 30 design parameters. To the best of our knowledge, we are the first to address the problem of optimizing adaptive enrichment designs with more than just a few parameters and more than 2 stages. Additionally, we explore a two step optimization strategy, where first an unstructured search is conducted, and then its results are used to define a more structured search.

While our approach based on simulated annealing (described below) does not ensure that a global optimum is found, we show that it can substantially reduce the expected number enrolled compared to simpler designs. In one of our examples, the optimized adaptive design reduces the expected number enrolled by approximately 32% compared to simpler adaptive and non-adaptive designs; the cost is a 22% increase in maximum number enrolled. In another example, there is a longer follow-up time to measurement of the primary outcome, compared to the enrollment rate; the benefits of adaptation on expected number enrolled are meager since by the time sufficient information has accrued to make a useful decision, most of the total enrollment has already been completed.

For trial design problems where no existing approach is guaranteed to find an optimal solution, one may turn to general-purpose, approximate methods such as simulated annealing (SA). Wason et al. (2012) apply SA to optimize the worst-case expected number enrolled using a group sequential design, with penalties added to the objective function for violations of either Type I or Type II error constraints. Wason and Jaki (2012) extend these results by applying SA to optimize a multi-arm, multi-stage trial where several treatments groups are compared against a shared, single control group.

Our optimization problem differs from that of Wason and Jaki (2012) in that our futility boundaries are non-binding (which is typically preferred by regulators such as the U.S. Food and Drug Administration, as noted by Liu and Anderson (2008)), and our designs allow continuation after one null hypothesis is rejected (so other hypotheses may be rejected at later stages). Also, in our optimization procedure, we include a final adjustment step after the SA algorithm to ensure that power constraints are met. Without this, the optimal solution is generally not guaranteed to have the desired power. Another difference is that we apply a parallelized version of SA. These and other differences between our implementation of SA and that of Wason and Jaki (2012) are discussed in Section 4.

Over the course of developing the general optimization approach in this paper, we also applied it in (Rosenblum et al., 2016) to a different problem than discussed here. The focus there was on illustrating a novel multiple testing procedure, rather than on the optimization, which was summarized in a paragraph and not otherwise mentioned. Unlike here, Rosenblum et al. (2016) did not compare the optimized design to other designs in order to investigate the value added from the optimization.

In Section 2, we introduce motivating data examples based on a new surgical intervention for stroke, and on a hypothetical intervention for preventing progression to Alzheimer's disease. In Section 3, we introduce a class of adaptive enrichment designs, referred to hereafter as "adaptive designs." We discuss how efficacy boundaries can be constructed by incorporating either the covariance of the test statistics (Rosenblum et al., 2016), or by using alpha-reallocation (Maurer and Bretz, 2013). We also introduce different levels of trial design complexity, which balance design flexibility versus simplicity. In Section 4, we outline our approach for optimization. In Section 5, we explore the performance of each type of trial, and compare to simpler trial designs using approximate O'Brien Fleming boundaries (O'Brien and Fleming, 1979) or Pocock boundaries (Pocock, 1977). We end with a discussion of future work.

2. Applications

2.1. Application 1: Surgical Treatment of Stroke (MISTIE)

We first describe an example of planning a Phase III trial of a surgical treatment for stroke, which was also considered by Rosenblum et al. (2016). The treatment is called Minimally-Invasive Surgery Plus rt-PA for Intracerebral Hemorrhage (MISTIE), and is described in detail by Morgan et al. (2008). Each participant had his/her functional disability measured 180 days from enrollment, based on the modified Rankin Scale (mRS). The primary outcome was the indicator of having mRS ≤ 3.

In planning the Phase III trial, the investigators were interested in two subpopulations defined by size of intraventricular hemorrhage (IVH) at baseline. "Small IVH" participants are defined to have IVH volume less than 10ml and not requiring a catheter for intracranial pressure monitoring. The remaining participants are called "large IVH". The Phase II trial only recruited small IVH participants. A preliminary analysis of the data resulted in an estimated treatment effect of approximately 12.1%. Knowledge of the underlying biology of these types of brain hemorrhage suggested a possible benefit for those with large IVH as well. However, there was greater uncertainty about the treatment effect in the large IVH subpopulation. Investigators inquired about the possibility of running a phase III trial that included both small IVH and large IVH participants (called subpopulation 1 and 2, respectively), but with the option to stop a subpopulation's accrual (using a preplanned rule) if interim data indicated that a benefit was unlikely. The proportion of participants in subpopulation 1 was projected to be 0.33; the enrollment rate was projected to be 420 participants per year from the combined population.

2.2. Application 2: Alzheimer's Disease Neuroimaging Initiative (ADNI)

We also consider an example using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early Alzheimer's disease. We focus on subpopulations defined by a participant's apolipoprotein E (APOE) $\epsilon 4$ allele carrier status, which is associated with increased risk of late onset Alzheimer's disease (Sadigh-Eteghad et al., 2012). Non-carriers of the APOE $\epsilon 4$ allele are called subpopulation 1, and carriers of at least one allele are called subpopulation 2. Clinical investigators, who were planning a Phase III trial of a new treatment to prevent progression from mild cognitive impairment to Alzheimer's disease, suspected that there may be treatment effect heterogeneity across carrier status. The proportion in subpopulation 1, based on the ADNI data, is 46.9%.

The primary outcome is the 2-year change in Clinical Dementia Rating Sum of Boxes score (CDR-SB), an aggregate measure of symptom severity. Enrollment would include those with baseline CDR-SB ≥ 0.5, which is indicative of mild cognitive impairment. The enrollment rate was projected to be 500 participants per year from the combined population.

3. Adaptive Trial Designs

3.1. Notation, Hypotheses, and Statistics

We consider two subpopulations that partition the overall population. Let $j = 1, 2, C$ be an index respectively denoting subpopulation 1, subpopulation 2, or the combined population. The treatment effect for a population is defined as the difference between the mean outcome for treatment and control.

Outcomes may be continuous, binary, or on any scale that allows the treatment effect to be estimated with a difference in means z-statistic. The data collected for each participant is a vector (S, A, Y) representing his/her subpopulation, study arm assignment, and outcome, respectively. We assume each participant's data vector is an independent, identically distributed draw from an unknown joint distribution on (S, A, Y). We assume that study arm assignment is randomized with probability $\frac{1}{2}$ independent of the subpopulation, and that each participant's outcome is measured after a delay of d years from enrollment.

Let π_j denote the proportion of the combined population in subpopulation j (with $\pi_C = 1$ by convention). Let δ_j denote the average treatment effect in population j. It follows that $\delta_C = \pi_1\delta_1 + \pi_2\delta_2$. Let H_1, H_2, and H_C respectively be the null hypotheses of no average treatment benefit in populations 1, 2 and C, i.e., $H_j : \delta_j \leq 0$. Each corresponding alternative hypothesis has the form $\delta_j > 0$. Let σ_{tj}^2 and σ_{cj}^2 denote the outcome variances in population j under treatment and control, respectively, with all variances assumed known. The combined population outcome variance is a function of π_1 and the outcome variances for each subpopulation. Define the global null hypothesis $H_0 : \delta_1 = \delta_2 = \delta_C = 0$; though this is not of primary interest, we refer to it in defining the different multiple testing procedures below.

Analysis timing is based on the cumulative number of outcomes observed from each subpopulation, referred to as the sample size. Because participant outcomes are measured with delay, there can be enrolled participants at interim analyses whose outcomes are not yet measured. These participants are referred to as "pipeline participants." Pipeline participants do not contribute to test statistic calculations, although they always contribute to the number enrolled in the trial.

Our designs have $K > 0$ stages, each concluding with an analysis. These analyses may lead to stopping accrual in one or both subpopulations, according to a set of predefined rules that are functions of the data collected so far. To stop accrual means to stop enrollment and follow-up of pipeline participants.

We assume enrollment is proportional to subpopulation size π_j, and enrollment is uniform over time. Analysis k occurs when the k^{th} stage is completed, defined to be the first time when $\pi_j n_k$ additional participant outcomes have been measured from each subpopulation j for which accrual has not been previously stopped. The terms n_k are design parameters that are not modified during the trial. Let $N := \sum_{k=1}^{K} n_k$ denote the maximum total sample size. The decision at the end of stage k takes as input the following cumulative test-statistic:

$$Z_j^{(k)} := \hat{\delta}_j \left(\frac{\sigma_{cj}^2 + \sigma_{tj}^2}{\frac{1}{2}\sum_{k'=1}^{k} n_{k'}\pi_j} \right)^{-\frac{1}{2}},$$

for each population $j \in \{1, 2, C\}$ that is still being enrolled, where $\pi_C = 1$, and $\hat{\delta}_j$ denotes the difference in sample means estimator for δ_j based on the accrued data. The combined population statistic $Z_C^{(k)}$ is undefined if one or more subpopulations had accrual stopped early at a previous stage. Throughout, we make the asymptotic approximation that the above statistics have the canonical multivariate normal distribution from Jennison and Turnbull (1999, Chapter 3) with covariances:

$$Cov\left(Z_j^{(k)}, Z_j^{(l)}\right) = \left(\frac{\sum_{k'=1}^{k} n_{k'}}{\sum_{l'=1}^{l} n_{l'}} \right)^{1/2}, \qquad Cov\left(Z_1^{(k)}, Z_2^{(l)}\right) = 0,$$

$$Cov\left(Z_j^{(k)}, Z_C^{(l)}\right) = \left\{ \pi_j \left(\frac{\sigma_{cj}^2 + \sigma_{tj}^2}{\sigma_{cC}^2 + \sigma_{tC}^2} \right) \left(\frac{\sum_{k'=1}^{k} n_{k'}}{\sum_{l'=1}^{l} n_{l'}} \right) \right\}^{1/2},$$

for any $j = 1, 2, C$, and stages k and l such that $1 \leq k \leq l \leq K$. These covariances depend only on the variables $\sigma_{cj}^2, \sigma_{tj}^2$, n_k, and π_j. The mean vector for the z-statistics is a function of these variables and the subpopulation treatment effects δ_1, δ_2.

At the analysis just after stage k, each statistic $Z_j^{(k)}$ is compared against an efficacy boundary $e_j^{(k)}$ and futility boundary $f_j^{(k)}$. If $Z_j^{(k)} > e_j^{(k)}$, H_j is rejected. Whenever H_1 and H_2 are both rejected, we automatically reject H_C as well. Rejecting any null hypothesis implies that the hypothesis remains rejected in all future stages. If $Z_j^{(k)} \leq f_j^{(k)}$, then accrual in population j is stopped for futility. We use non-binding futility boundaries, i.e., strong control of the familywise Type I error rate is asymptotically guaranteed for all designs in this paper, even if futility boundaries are ignored and both subpopulations continue accrual through the end of the last stage K.

For each subpopulation $j \in \{1, 2\}$, its accrual continues until one (or more) of the following occurs: H_j is rejected ($Z_j^{(k)} > e_j^{(k)}$), subpopulation j is stopped for futility ($Z_j^{(k)} \leq f_j^{(k)}$), the combined population is stopped for futility ($Z_C^{(k)} \leq f_C^{(k)}$). Under this setup, rejecting the combined population null hypothesis H_C does not imply stopping all accrual; further tests of H_1 and H_2 may still be conducted. Since $Z_C^{(k)}$ is only defined if the combined population accrual is not stopped early before stage k, we do not conduct the test $Z_C^{(k)} > e_C^{(k)}$ at stages after accrual for at least one subpopulation has stopped; it is still possible to reject H_C at future stages, if both subpopulation null hypotheses are rejected.

We compare two methods for calculating efficacy boundaries. The first uses an error spending approach based on the covariances of the statistics $Z_j^{(k)}$ (Rosenblum et al., 2016). The second method involves a graphical approach to reallocating Type I error to the remaining null hypotheses after some null hypotheses have been rejected (Bretz et al., 2009; Maurer and Bretz, 2013). We refer to these two types of multiple testing procedures as $\mathscr{H}^{COV}$ and $\mathscr{H}^{MB}$, respectively. For both, Type I error is calculated under the assumption that futility boundaries are never adhered to (to give the worst-case Type I error under non-binding futility boundaries). Power, expected number enrolled, and expected duration are calculated under the assumption that futility boundaries are adhered to.

3.2. Multiple Testing Procedure 1: Covariance Approach

Rosenblum et al. (2016) propose a method for efficacy boundary calculation that incorporates the covariance matrix among the statistics (which is assumed known), both across stages and across populations. We refer to this approach as $\mathscr{H}^{COV}$. A key feature of this approach is that in order to guarantee strong control of the familywise Type I error rate, it is sufficient to control the familywise Type I error rate under the global null hypothesis H_0 of no treatment effect in any subpopulation.

The first step of this approach is to prespecify an ordering for the null hypotheses. For example, the ordering H_1, H_2, H_C implies that we always first test H_1, then H_2, and finally H_C, at each stage where all three hypotheses are still being tested. If only a subset of hypotheses are still being tested at a given stage, the same ordering applies to the remaining hypotheses. The second step is to prespecify nonnegative values $\alpha_j^{(k)}$ for each population $j \in \{1, 2, C\}$ and stage $k : 1 \leq k \leq K$, that sum to the desired familywise Type I error rate α (e.g., $\alpha = 0.025$). The efficacy boundaries $e_j^{(k)}$ are iteratively computed at the end of each stage k by solving the following for each $j \in \{1, 2, C\}$ according to the above ordering:

$$P_{H_0}\{Z_j^{(k)} > e_j^{(k)};\ \text{and}\ Z_{j'}^{(k')} \leq e_{j'}^{(k')}\ \text{for all}\ k', j'\ \text{such that}\ (k', j') \prec (k, j)\} = \alpha_j^{(k)}, \tag{1}$$

using the known covariance for the test statistics, where $(k', j') \prec (k, j)$ means that either $k' < k$ or ($k' = k$ and j' precedes j in the ordering).

The required inputs to the multiple testing procedure $\mathscr{H}^{COV}$ are nonnegative values $\alpha_j^{(k)}$ that sum to α. We also consider simpler designs, called "structured", that restrict these values to have a particular form. Specifically, we consider setting the $\alpha_j^{(k)}$ allocated to each stage k and hypothesis H_j to be $w_j \times (a_k - a_{k-1})$, where $a_k = \left(\sum_{k'=1}^{k} n_{k'}/N\right)^{\rho_{ej}}$, $a_0 = 0$, $\rho_{ej} > 0$, and $w_j = \sum_{k=1}^{K} \alpha_j^{(k)}/\alpha$. This is the power family of error spending functions, described by Jennison and Turnbull (1999). In this way, the search space is reduced from the $3 \times K$ values for the $\alpha_j^{(k)}$, to the six design parameters w_j and ρ_{ej} for $j = 1, 2, C$. For a group sequential design testing only one null hypothesis, setting ρ_{ej} to be 1 or 3 results in efficacy boundaries that are similar to those of Pocock (1977) or O'Brien and Fleming (1979) respectively (Jennison and Turnbull, 1999). We refer to the optimization problem where the search space consists of designs that have no restrictions on $\alpha_j^{(k)}$ (except that they are nonnegative and sum to α) as "unstructured."

3.3. Multiple Testing Procedure 2: Alpha-Reallocation Approach

Maurer and Bretz (2013) propose a procedure that reallocates alpha from a rejected null hypothesis to remaining null hypotheses; the result is a lowering of the efficacy boundaries for the remaining null hypotheses, thereby increasing power. In contrast to the approach in the previous subsection, the Maurer and Bretz (2013) procedure does not directly use the covariance among statistics for different hypotheses (though it uses the covariance among statistics from the same hypothesis across stages). We summarize this reallocation procedure, referred to as $\mathscr{H}^{MB}$; a more detailed description is in the supplemental materials.

The first step in $\mathscr{H}^{MB}$ is for investigators to prespecify a weighted Bonferroni procedure to adjust for multiple testing across the hypotheses H_1, H_2 and H_C, with corresponding weights w_1, w_2 and w_C. Specifically, for each $j \in \{1, 2, C\}$, the null hypothesis H_j is tested using efficacy boundaries $e_j^{(k)}, k = 1, \ldots, K$, corresponding to a standard group sequential design for a single null hypothesis with total Type I error $w_j\alpha$; the fraction of $w_j\alpha$ allocated to each stage is prespecified, and efficacy boundaries are created using an error spending function analogous to (1) except only considering a single null hypothesis. These boundaries are used until at least one null hypothesis is rejected, at which point reallocation occurs, as described next.

After a null hypothesis H_j is rejected, its weight w_j is reallocated across the remaining null hypotheses. This reallocation follows a graphical procedure of Bretz et al. (2009), which must be prespecified. Each node in the graph corresponds to a null hypothesis, and there is a directed edge between each pair of nodes. The edge from H_i to H_j has a transition weight $g_{ij} \in [0, 1]$ that indicates the proportion of the weight at node i reallocated to H_j after rejection of H_i. Each rejection also requires updating the edges g_{ij}, as transitions to the rejected hypotheses become defunct. We use g_{ij} to refer to the initial transition weight, before any hypothesis has been rejected. Maurer and Bretz (2013) prove that their procedure strongly controls the familywise Type I error rate.

The required inputs to the multiple testing procedure $\mathscr{H}^{MB}$ are nonnegative values $\alpha_j^{(k)}$ that sum to α (just as for $\mathscr{H}^{COV}$), as well as the initial transition weight g_{ij} for each edge. Each initial node weight is set as $w_j = \sum_{k=1}^{K} \alpha_j^{(k)}/\alpha$. These inputs correspond to the unstructured design using $\mathscr{H}^{MB}$; a structured design, with fewer inputs, is defined analogously as in the previous subsection.

4. Optimization Problem

4.1. Search Space

Let $\mathscr{D}$ denote a trial design, which consists of the list of design parameters necessary to fully specify the decision rule and testing procedure for a trial, i.e., the following: the maximum number of stages K (which we restrict to be at most 10), the maximum sample size per stage $\{n_k\}_{k \leq K}$ (each n_k being a positive integer), alpha allocations $\{\alpha_j^{(k)} \geq 0 : j \in \{1, 2, C\}, k \leq K\}$ (which sum to α), futility boundaries $\{f_j^{(k)} \in \mathbb{R} : j \in \{1, 2, C\}, k \leq K\}$, a hypothesis testing framework ($\mathscr{H}^{MB}$ or $\mathscr{H}^{COV}$), and the alpha reallocation rule if using $\mathscr{H}^{MB}$. The class of trial designs satisfying the above restrictions is the search space for the unstructured optimization problem. The structured optimization problem has a reduced search space as described at the end of Section 3.2.

4.2. Constraints

We next define the power and Type I error constraints. Let $\underline{\boldsymbol{\delta}} = (\delta^{(1)}, \delta^{(2)})$ denote a vector of possible values for the treatment effect in each subpopulation, and let $\delta^{\min} > 0$ denote the minimum value of the treatment effect that is clinically meaningful. We consider the following values for $\underline{\boldsymbol{\delta}}$:

$$\underline{\boldsymbol{\delta}}^{(0)} = (0,0); \qquad \underline{\boldsymbol{\delta}}^{(1)} = (\delta^{\min}, 0); \qquad \underline{\boldsymbol{\delta}}^{(2)} = (0, \delta^{\min}); \qquad \text{and} \qquad \underline{\boldsymbol{\delta}}^{(C)} = (\delta^{\min}, \delta^{\min}).$$

Let $1 - \beta_j(\underline{\boldsymbol{\delta}}', \mathscr{D})$ be the power of the design $\mathscr{D}$ to reject at least H_j by the end of the trial when (δ_1, δ_2) is equal to the vector $\underline{\boldsymbol{\delta}}'$. Let $P_{\delta_1,\delta_2,\mathscr{D}}$ and $E_{\delta_1,\delta_2,\mathscr{D}}$, respectively, denote probability and expectation with respect to the z-statistics defined in Section 3.1 for design $\mathscr{D}$ under treatment effects (δ_1, δ_2); we assume this distribution of z-statistics is multivariate normal with covariance matrix given in Section 3.1.

We impose the following constraints on power and familywise Type I error:

(a) $1 - \beta_1(\underline{\boldsymbol{\delta}}^{(1)}, \mathscr{D}) \geq 0.8$; (b) $1 - \beta_2(\underline{\boldsymbol{\delta}}^{(2)}, \mathscr{D}) \geq 0.8$; (c) $1 - \beta_C(\underline{\boldsymbol{\delta}}^{(C)}, \mathscr{D}) \geq 0.8$;
(d) Strong control of the familywise Type I error rate, i.e.,

$$\sup_{\delta_1,\delta_2 \in \mathbb{R}} P_{\delta_1,\delta_2,\mathscr{D}}(\text{reject one or more true null hypotheses}) \leq \alpha = 0.025.$$

Constraints (a)-(c), respectively, represent having at least 80% power to reject H_1 when the treatment only benefits subpopulation 1 (at level $\delta^{\min}$), at least 80% power to reject H_2 when the treatment only benefits subpopulation 2, and at least 80% power to reject the combined population null hypothesis H_C when the treatment benefits both subpopulations.

4.3. *Objective Function*

The objective function for the optimization problem is the expected number enrolled. (We also consider the expected trial duration.) The expected number enrolled is computed with respect to a prespecified distribution Λ on the treatment effects (δ_1, δ_2). We refer to this as the prior distribution on the treatment effects. However, all of our designs have guaranteed asymptotic, familywise Type I error control without regard to this prior, i.e., it holds for any possible pair (δ_1, δ_2).

We aim to minimize the following objective function (representing expected number enrolled) over the space of trial designs $\mathscr{D}$ that satisfy constraints (a)-(d):

$$E_\Lambda\{\tilde{n}(\mathscr{D})\} = \int_{\delta_1,\delta_2} E_{\delta_1,\delta_2,\mathscr{D}}(\text{total participants enrolled}) d\Lambda(\delta_1, \delta_2), \tag{2}$$

where $\tilde{n}(\mathscr{D})$ denotes the inner expectation on the right side of (2), which represents the expected number enrolled under $P_{\delta_1,\delta_2,\mathscr{D}}$. In this paper, we set Λ to be a discrete distribution with equal mass at $\underline{\boldsymbol{\delta}}^{(0)}, \underline{\boldsymbol{\delta}}^{(1)}, \underline{\boldsymbol{\delta}}^{(2)}$, and $\underline{\boldsymbol{\delta}}^{(C)}$. Under such a prior, $E_\Lambda\{\tilde{n}(\mathscr{D})\}$ is the expected number enrolled across these four scenarios. While minimizing expected number enrolled is our primary goal, we also consider the problem of minimizing expected duration – the expected time from the start of enrollment until all accrual has stopped. This expectation is taken with respect to the same prior for the treatment effects.

The power constraints (a)-(c), Type I error constraints (d), and objective function (2) depend on the design parameters (encoded in $\mathscr{D}$) and the population parameters $\sigma^2_{cj}, \sigma^2_{tj}$, π_j, $\delta^{\min}$. This is because these parameters are sufficient to determine the distribution of the z-statistics in Section 3.1, and all designs here use the data only through these z-statistics. For a given application (such as the MISTIE and ADNI applications described below), it suffices to specify the population parameters $\sigma^2_{cj}, \sigma^2_{tj}$, π_j, $\delta^{\min}$, and the class of designs to be searched over, in order to fully define the corresponding optimization problem. The problem also depends on the prior Γ, defined above.

Due to the difficulty in directly solving the optimization problem (2) under the power and Type I error constraints, we instead define an unconstrained optimization problem where the constraints are incorporated as penalty terms as in (Wason and Jaki, 2012; Wason et al., 2012). The unconstrained objective function we aim to minimize over $\mathscr{D}$ is

$$J(\mathscr{D}) := E_\Lambda\{\tilde{n}(\mathscr{D})\} + \lambda \textstyle\sum_{j \in \{1,2,C\}} \left[0.8 - \{1 - \beta_j(\underline{\boldsymbol{\delta}}^{(j)}, \mathscr{D})\}\right]^3_+, \tag{3}$$

where λ is a positive tuning parameter (set here to 10^6), and $(x)_+ = \max\{x, 0\}$. The first term can also be replaced with the expected trial duration. If any of the power constraints in Section 4.2 are violated, the objective function will incur a severe penalty. The exponent in the penalty term is meant to allow second order differentiability of $J(\mathscr{D})$ with respect to the power of the trial. This exponent is not necessary, but is potentially useful for some of the approaches discussed in Section 5.

Evaluating $J(\mathscr{D})$ requires the calculation of several multidimensional integrals. Due to the computational obstacle of these calculations, we instead estimate $J(\mathscr{D})$ via simulation. We used 10,000 simulation iterations, such that the Monte Carlo standard error for estimating a power close to 0.80 is approximately $\{0.8(1 - 0.8)/10000\}^{1/2} \approx 0.004$.

Since we parametrize the trial in terms of alpha allocations $\alpha_j^{(k)}$ that sum to $\alpha = 0.025$, all of our proposed designs are asymptotically guaranteed to control the familywise Type I error as proved in (Rosenblum et al., 2016; Maurer and Bretz, 2013), and it is not necessary to penalize for violations of the required familywise Type I error rate in the manner of (Wason and Jaki, 2012; Wason et al., 2012).

4.4. Optimization Using Simulated Annealing

We search for $\text{argmin}_{\mathscr{D}} J(\mathscr{D})$ using simulated annealing (SA). The general form of SA is as follows. Given a trial design $\mathscr{D}$ as a reference point, SA randomly perturbs $\mathscr{D}$ in order to generate a new proposal design $\mathscr{D}'$. If $J(\mathscr{D}') < J(\mathscr{D})$ then the proposal is "accepted," and $\mathscr{D}'$ becomes the new reference point. If $J(\mathscr{D}') > J(\mathscr{D})$, then $\mathscr{D}'$ is accepted according to a certain probability, and discarded otherwise. The nonzero probability of exploring undesirable regions of the parameter space allows SA to avoid becoming stuck at a local minimum. As the algorithm progresses, new proposal designs $\mathscr{D}'$ are taken from a closer neighborhood around the reference design, and the probability of accepting inferior designs decreases. Both of these changes are modulated by a parameter known as the "temperature," which decreases with each iteration. We use the variant of SA implemented in the `optim` function in `R`, which is based on the algorithm of (Bélisle, 1992). We implemented SA in parallel across 100 nodes, each starting with a different random seed. Our implementation is "embarrassingly parallel" in that each node runs the SA algorithm independent of the others (i.e., without communication between nodes); when the SA search terminates for all nodes, we select the best design found, denoted $\mathscr{D}_{SA}$.

The search space for $\mathscr{D}$ is defined in Section 4.1. Separate searches are performed for the two hypothesis testing frameworks $\mathscr{H}^{MB}$ and $\mathscr{H}^{COV}$. One difficulty is that the dimension of this search space changes with the value of K, since greater values of K require additional sample sizes, efficacy boundaries, and futility boundaries. We give details on our method to address this issue in the supplemental materials.

The SA algorithm allows design parameters to take any real values, which may violate the constraints on our search space of feasible designs. In particular, since the alpha allocated to each test at each stage must be bounded between 0 and α, we instead use SA to search for the logit transform of the alpha allocated, i.e., $\log\{\alpha_j^{(k)}/(1 - \alpha_j^{(k)})\}$. We then transform proposed values for $\text{logit}(\alpha_j^{(k)})$ back to the (0,1) interval, and rescale them to sum to α. In the same way, we search over the logit of the graph transition weights $g_{ij} \in [0, 1]$ when using the framework $\mathscr{H}^{MB}$. Non-negative and integer constraints (e.g. for n_k and K) are achieved by truncating and rounding respectively. We restricted the designs to have at most $K = 10$ stages. Additionally, rather than searching for each individual n_k, we search across the space for N and separately search over the proportion of N allocated to each stage. Under this parametrization, the maximum sample size can naturally be changed without affecting the efficacy boundaries, as the efficacy boundaries depend only on the relative sample sizes n_k/N for each k.

Penalized approaches such as (Wason and Jaki, 2012; Wason et al., 2012), or approaches based on (3), will not necessarily guarantee that the resulting optimized design meets the power constraints in Section 4.2, as there may be cases where a small penalty is outweighed by a larger reduction in expected number enrolled. For the designs proposed by (Wason and Jaki, 2012; Wason et al., 2012), these concerns also apply to Type I error control.

To address the above issue, we built in an extra step to correct for cases where, after the SA algorithm completes, the resulting design $\mathscr{D}_{SA}$ fails to satisfy one or more of the power constraints. This step involves starting with $\mathscr{D}_{SA}$, and increasing only the total sample size parameter N. A binary search over N is conducted to find the smallest value such that the constraints in Section 4.2 are met. During this search, all other elements of $\mathscr{D}_{SA}$ are held constant. When implementing the SA procedure in parallel, we apply this extra step after SA completes in each node. These supplemental searches also reduce the danger of choosing the tuning parameter λ in (3) to be too small. Our specific use of binary search is motivated by our empirical experience of $E_\Lambda\{\tilde{n}(\mathscr{D})\}$ being monotonically increasing in N for a variety of tested scenarios, and by the fact that the power constraints can always be satisfied by a sufficient increase to N as long as each $\alpha_j^{(k)} > 0$.

In order to derive designs that are simpler to interpret and perform approximately optimally, we propose a two-step procedure for discovering efficacy and futility boundaries. First, we optimize as above, and refer to the resulting design as "unstructured". Based on the resulting design, we next construct a lower dimensional parametrization that has a simpler form, and solve the same optimization problem in this restricted space. If the value of the objective function is very close to that attained in the unrestricted case, we report the simpler "structured" solution along with the "unstructured" one, as the former may be easier to communicate. We discuss our specific choice of structured boundaries in Section 5.

4.5. Comparison Designs

We compare optimized adaptive designs against three types of simpler designs, denoted "non-optimized single stage designs," "optimized single stage designs," and "non-optimized multistage designs." The term "optimized" means that design parameters are optimized using our SA approach (or by grid search for the simplest designs). Single stage (i.e., $K = 1$) designs can be optimized in terms of their multiple testing procedure design parameters, such as $\alpha_j^{(1)}$ and g_{ij}. We define non-optimized single stage designs as trials with equal alpha allocation (i.e., each $\alpha_j^{(1)} = \alpha/3$) and reallocation (i.e., each $g_{ij} = 1/2$). We define optimized single stage designs as trials where the alpha allocations and reallocations are computed either through a grid search (for $\mathscr{H}^{COV}$) or through SA (for $\mathscr{H}^{MB}$).

We define non-optimized multistage designs as 5-stage trials with each n_k equal to a common value, equal alpha allocation and reallocation (i.e., each $g_{ij} = 1/2$) across hypotheses, futility boundaries set equal to zero, and the initial alpha allocations across stages set according to the structured alpha spending function in Section 3.2, with ρ_{ej} set equal to either 1 or 3 for all j. These settings for ρ_{ej} result in boundaries similar to those of Pocock (1977) or O'Brien and Fleming (1979), respectively (Jennison and Turnbull, 1999). We refer to these sets of boundaries as Pocock and O'Brien-Fleming boundaries, respectively. For all comparison designs, the maximum sample size was selected to be the smallest value that satisfied the power constraints in Section 4.2.

5. Results

5.1. Overview

We compare the performance of optimized adaptive designs versus the simpler designs as described in Section 4.5. We find that optimized designs can offer substantial reductions in expected number enrolled for trials where the delay time between enrollment and outcome measurement is relatively small compared to the enrollment rate. The reason is that in such trials, enrollment has not yet been exhausted before enough information has accrued to make a useful decision about changing enrollment criteria. The MISTIE example illustrates such a case, while the ADNI example demonstrates the opposite situation.

Single stage (non-adaptive) designs generally have a lower maximum number enrolled than group sequential designs and adaptive enrichment designs, but at the cost of a higher expected number enrolled due to their lack of ability to stop early. We illustrate this tradeoff in the MISTIE and ADNI examples.

5.2. MISTIE Example

Before presenting results for the MISTIE example, we list the inputs to the corresponding trial design problem that are used to set the power constraints and objective function in Sections 4.2–4.3. The enrollment rate and subpopulation proportions are given in Section 2.1. Recall the primary outcome, measured at 180 days ($d = 1/2$ year), is the indicator of having mRS score 3 or less (called a successful outcome). The probability of a successful outcome after 180 days was projected to be 0.290 under control, for each subpopulation. Investigators aimed to satisfy the power constraints listed in Section 3 for $\delta^{\min} = 0.122$. The variance of the outcome under control is assumed to be $\sigma^2_{c1} = \sigma^2_{c2} = 0.290(1 - 0.290)$, and the variance of the outcome under treatment is assumed to be $\sigma^2_{t2} = \sigma^2_{t1} = 0.412(1 - 0.412)$.

The first row of Figure 1 shows the z-statistic boundaries and per-stage sample sizes for the optimized adaptive designs using $\mathscr{H}^{COV}$ and $\mathscr{H}^{MB}$, respectively. These boundaries are the result of the unstructured search described in the last paragraph of Section 4.4. For $\mathscr{H}^{MB}$, the boundaries shown are those before any alpha-reallocation has taken place (referred to as "initial"). Initial efficacy boundaries for $\mathscr{H}^{COV}$ and $\mathscr{H}^{MB}$ are similar, each roughly resembling Pocock boundaries. Futility boundaries are similar across hypothesis testing frameworks as well, with futility boundaries for H_1 or H_2 being highest at the midpoint of the trial, and futility boundaries for H_C remaining low throughout the trial. Within a given design, symmetry between the futility boundaries for H_1 and H_2 is not necessarily to be expected, as $\pi_1 = 0.33, \pi_2 = 0.67$.

We implemented the 2-step procedure described in the last paragraph of Section 4.4. Step 1 was the optimization over unstructured designs leading to the sample sizes and boundaries in the top row of Figure 1. Based on these results, we proposed the following structured form for the futility boundaries in step 2: set $f_j^{(k)} = c_j + l_j \times \left(\sum_{k'=1}^{k} n_{k'}/N\right)^{\rho_{fj}}$, where c_j, l_j, and ρ_{fj} are unrestricted design parameters for $j = 1, 2, C$. This form encompasses boundaries similar to those of Pocock (1977) and O'Brien and Fleming (1979), with additional shift parameters c_j to capture the behavior discovered in the first row of Figure 1. We also restrict to the structured efficacy boundaries described in Section 3.2. We reapplied our SA procedure optimizing over this restricted set of parameters that includes c_j, l_j, ρ_{fj}, w_j, ρ_{ej} (and K, n_k, $\mathscr{H}^{MB}$ or $\mathscr{H}^{MB}$, g_{ij}). Starting values for both steps of this search are given in the supplemental materials. The resulting optimized designs are shown in the bottom row of Figure 1. These structured designs have expected number enrolled within approximately 1% of the unstructured optimized designs.

Figure 2 shows the distribution of the number enrolled in the MISTIE example for the optimized structured multistage designs (i.e., those in the bottom row of Figure 1), and for the simpler comparison designs described in Section 4.5. The distributions for multistage designs are shown as violin plots, and the fixed number enrolled for the optimized and non-optimized single stage designs are shown as horizontal lines. All distributions are calculated based on the prior distribution for the treatment effects in Section 4.2. Our optimized adaptive designs using $\mathscr{H}^{COV}$ and $\mathscr{H}^{MB}$, respectively, have an expected number enrolled of 1006 and 981, with maximum number enrolled 2002 and 1762. Non-optimized single stage designs enroll 1875, regardless of whether $\mathscr{H}^{COV}$ or $\mathscr{H}^{MB}$ is used. Optimized single stage designs using $\mathscr{H}^{COV}$ and $\mathscr{H}^{MB}$, respectively, have number enrolled 1447 and 1443. Non-optimized multistage designs with Pocock boundaries have expected number enrolled 1562 and 1531 for $\mathscr{H}^{COV}$ and $\mathscr{H}^{MB}$, respectively; when using O'Brien-Fleming boundaries, the analogous results are 1682 and 1652. In summary, the optimized adaptive designs have a substantially lower expected number enrolled than all the simpler designs, but have higher maximum number enrolled than the single stage designs. Specifically, the optimized adaptive design using $\mathscr{H}^{MB}$ has 32% (462 participants) lower expected number enrolled but 22% (319 participants) greater maximum number enrolled versus the the best competitor (the optimized single stage design) among those we considered.

We next examine the progress of the SA algorithm versus the number of search iterations. Figure 3 shows the objective function of the best design discovered so far across iterations of the parallel, unstructured SA search algorithm, for the MISTIE example. For each optimization problem, the search procedure was parallelized across 100 computing nodes. Each node was set to run for 5000 iterations or 72 hours, whichever occurred first. The curves in Figure 3 represent the trajectory of the cumulative minimum objective function value found by each parallel node. Quartiles with

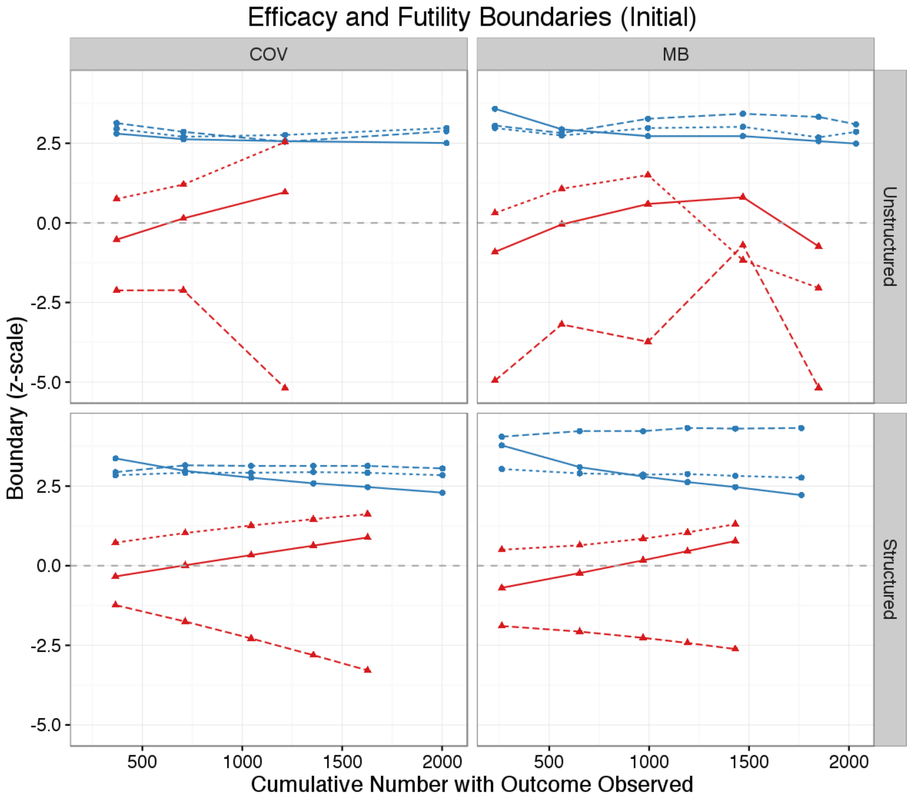

Fig. 1. MISTIE Example: Efficacy and Futility Boundaries for Optimized Adaptive Designs. The efficacy boundaries $e_j^{(k)}$ and futility boundaries $f_j^{(k)}$ for four different optimized adaptive designs (one in each panel) are shown. Boundaries for H_1, H_2, and H_C are respectively shown as solid, dotted, and dashed lines. Efficacy boundaries (blue) are marked by circles, and futility boundaries (red) are marked by triangles. Boundaries are on the z-scale, and the horizontal axis represents the sample size at the time of the interim analysis (if there is no early stopping). Each of the 2 columns of panels corresponds to a different hypothesis testing framework, with $\mathscr{H}^{COV}$ on the left and $\mathscr{H}^{MB}$ on the right. The top and bottom rows of panels show boundaries for the designs resulting from optimizing using the unstructured and structured methods, respectively, described in the last paragraph of Section 4.4. For $\mathscr{H}^{MB}$, the boundaries shown represent those before alpha reallocation; the alpha reallocation rules are given in the supplemental materials, along with the initial alpha allocations for all four designs.

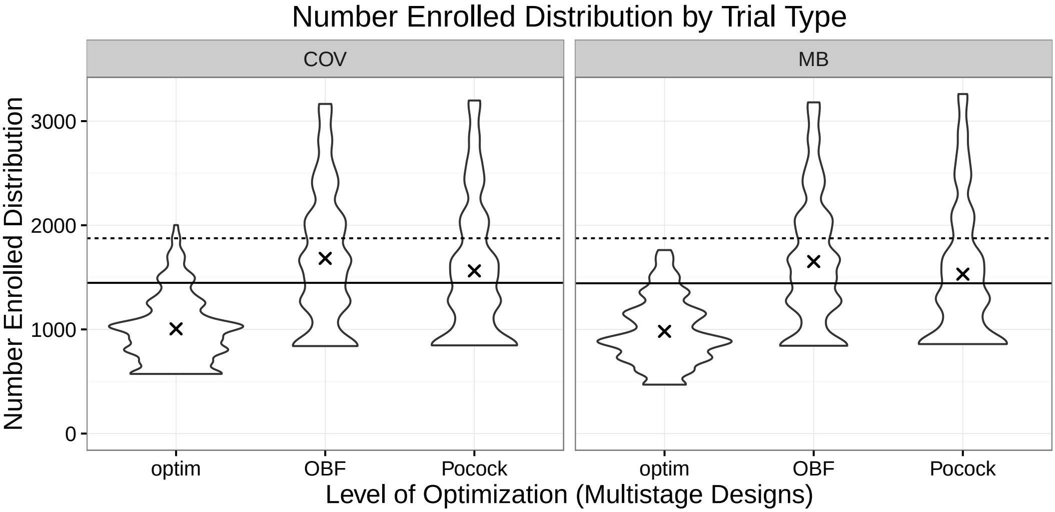

Fig. 2. MISTIE Example: Distributions of Number Enrolled for Different Designs. Violin plots depict the approximate distributions for three types of multistage designs: our optimized adaptive designs with structured boundaries (optim), non-optimized multistage design with O'Brien-Fleming boundaries (OBF), and non-optimized multistage design with Pocock boundaries (Pocock). These violin shapes represent smoothed histograms of the distribution of simulated number enrolled. The distributions are taken with respect to the prior for the treatment effects given in Section 4.3, with the mean number enrolled for each design marked "×". Horizontal lines show the deterministic number enrolled from two types of single stage designs: non-optimized (dotted lines) and optimized (solid lines). Each panel corresponds to a different hypothesis testing framework, with $\mathscr{H}^{COV}$ on the left and $\mathscr{H}^{MB}$ on the right.

respect to the distribution of final objective function values are shown as horizontal lines. The figure is approximate in that no binary search corrections have yet been made to guarantee that power constraints are met. (See Section 4.4.) The most notable increases in performance occur in the early stages of SA, after which the distribution of performance across nodes remains relatively constant. This implies that a reduced number of search iterations might achieve similar performance if the temperature parameter of the search was set to decrease more slowly.

The quartile lines in Figure 3 can be used to roughly approximate performance in cases where fewer computing resources would be available. For instance, if only 5 parallel nodes had been available, the probability of achieving a result below the first quartile would be approximately $(1 - 0.75^5) = 76\%$. This implies that there is a roughly 24% chance of getting a substantially larger expected sample size if the search used only 5 nodes compared to 100 nodes. This suggests that at least some parallelization is useful in the search. It is an area of future research to optimize the degree of parallelization in the search procedure.

5.3. ADNI Example

We next consider the ADNI data example. Inputs to the trial design problem, described next, were based on values observed in the ADNI data. The enrollment rate and subpopulation proportions are given in Section 2.2. Recall the primary outcome is the change in CDR-SB score measured at $d = 2$ years compared to baseline. The sample variance in the 2-year change in CDR-SB was approximately 3.35 for subpopulation 1, and 3.61 for subpopulation 2. We set $\sigma^2_{c1} = \sigma^2_{t1} = 3.35$ and $\sigma^2_{c2} = \sigma^2_{t2} = 3.61$. The sample average change in CDR-SB in the data for the combined population was 1.41. The minimum, clinically important treatment effect was set at a 30% relative reduction in this CDR-SB change, i.e., $\delta^{\min} = 1.41 \times 0.3 = 0.42$.

We optimized designs analogously as for the MISTIE example. In contrast to that example, no

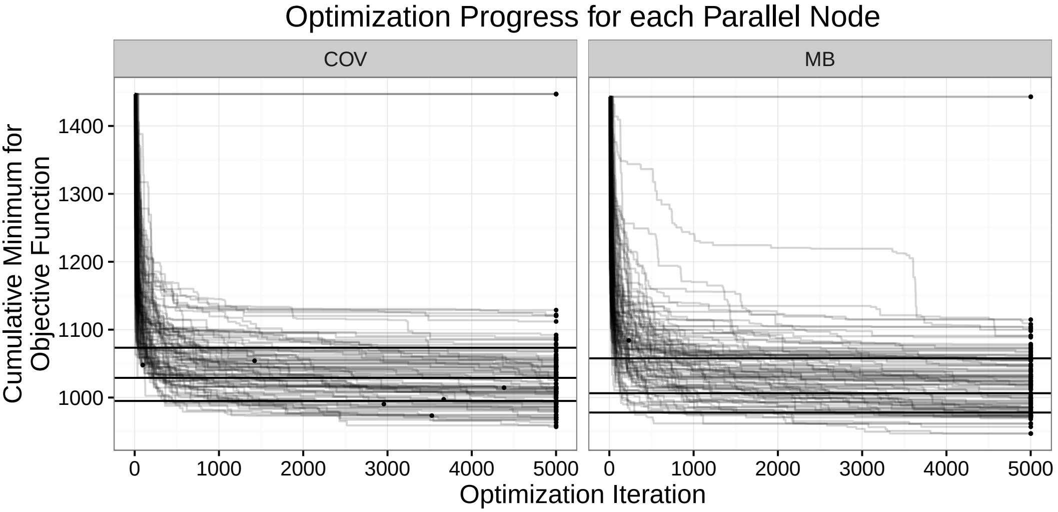

Fig. 3. MISTIE Example: Objective Function Progress versus Number of Search Iterations. Each curve shows the trajectory of the cumulative minimum value of the objective function discovered by a parallel computing node. Black dots show the terminal of each node's trajectory. For the approximately 2% of nodes that did not complete 5000 iterations of SA within 72 hours, these dots mark the last iteration completed. Horizontal lines show the 0.25, 0.5, and 0.75 quantiles, respectively, for the distribution of final objective function values across the 100 parallel nodes. Each panel corresponds to a different hypothesis testing framework, with $\mathscr{H}^{COV}$ on the left and $\mathscr{H}^{MB}$ on the right.

adaptive trial in our search in the ADNI example had lower expected number enrolled relative to an optimized single stage design. As mentioned above, this can be largely attributed to the high enrollment required before any outcomes are measured. In such a case, it still may be possible to reduce the trial's expected duration using an adaptive enrichment design; such reductions can result from stopping the trial for efficacy or futility before all participants have their outcomes measured.

We next optimized multistage designs for the ADNI example using trial duration rather than number enrolled in the objective function. Figure 4 shows performance comparisons for the ADNI example analogous to Figure 2, but with the vertical axis showing trial duration rather than number enrolled. Relative to an optimized single stage design, the optimized multistage designs reduced expected duration by 5.4% for $\mathscr{H}^{COV}$, and 7.4% for $\mathscr{H}^{MB}$. However, this comes at the cost of a higher maximum duration. Relative to an optimized 1-stage design, the optimized multi-stage designs increase maximum duration from 5.07 to 5.84 years for $\mathscr{H}^{COV}$, and from 5.02 to 5.75 years for $\mathscr{H}^{MB}$.

5.4. *Alternative Optimization Algorithms*

We also compared the performance of SA against other optimization algorithms available in the `optim` function in `R`. For each combination of testing procedure ($\mathscr{H}^{MB}$ or $\mathscr{H}^{COV}$), application (ADNI or MISTIE) and boundary form (structured or unstructured), each optimization method was allowed to run on 250 parallel nodes for either 4 hours or 2500 iterations, whichever occurred first. Rather than searching for the optimal number of stages (K), we fixed K within a node at either 2, 3, 4, 5, or 6. These values for K were evenly distributed such that each unique configuration was allotted $250/5 = 50$ parallel nodes with different starting seeds. The minimum objective function value across all 250 parallel nodes was recorded for comparison.

SA outperformed gradient methods such as BFGS, L-BFGS-B, and Conjugate Gradient by 3-6% in the ADNI example for expected duration and 7-27% in the MISTIE example for expected number enrolled. Nelder-Mead and SA performed much more similarly, with Nelder-Mead outperforming SA by approximately 1% in the ADNI example, and SA outperforming Nelder-Mead by approximately

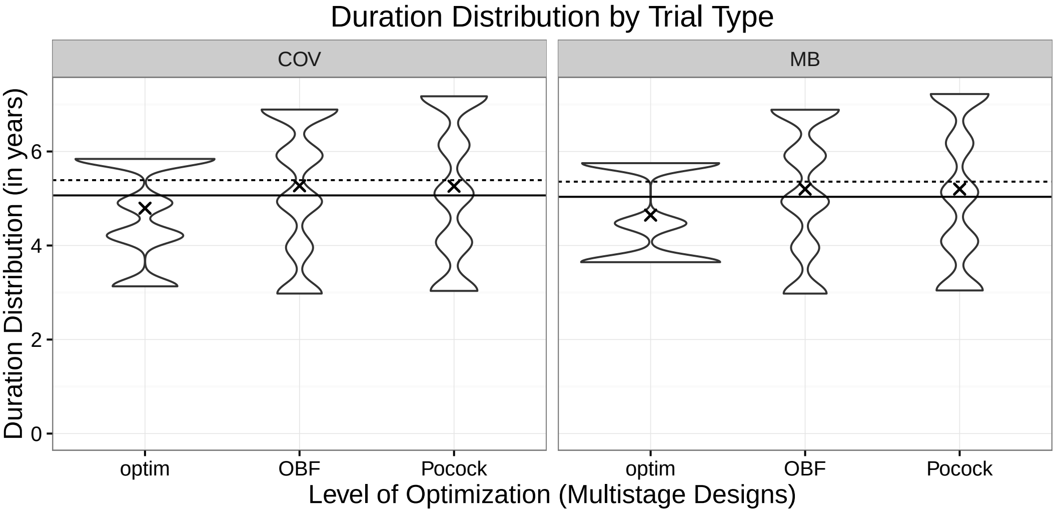

Fig. 4. ADNI Example: Distribution of Trial Durations. Violin plots show the trial duration distributions for three types of multistage designs: our optimized adaptive designs with structured boundaries (optim), non-optimized multistage design with O'Brien-Fleming boundaries (OBF), and non-optimized multistage design with Pocock boundaries (Pocock). The duration distributions are taken with respect to the prior for the treatment effects described in Section 4.2, with the mean duration for each design marked "×". Horizontal lines show the deterministic durations from two types of single stage designs: non-optimized (dotted lines) and optimized (solid lines). Each panel corresponds to a different hypothesis testing framework, with $\mathscr{H}^{COV}$ on the left and $\mathscr{H}^{MB}$ on the right.

1% in the MISTIE example.

We also compared against a version of SA where the objective function for the current design $\mathscr{D}$ is re-evaluated (using a new Monte Carlo sample) at each comparison to a new candidate design $\mathscr{D}'$, as discussed in the conclusion of Branke et al. (2008). Such an approach doubles the number of simulations, but decreases the probability that the algorithm becomes stuck at an inferior design whose objective function value is initially underestimated due to Monte Carlo error. This altered SA algorithm improved over gradient based methods, but was outperformed by both Nelder-Mead and by standard SA.

6. Discussion

We showed empirical evidence that SA can produce multistage adaptive enrichment designs with substantially lower expected number enrolled compared to optimized single stage designs, and to non-optimized multistage designs that use approximate Pocock or O'Brien-Fleming boundaries. Relative to single stage designs, optimized multistage designs come at the cost of increases in maximum number enrolled. There is an analogous tradeoff between expected and maximum number enrolled comparing standard group sequential designs versus single stage designs (Eales and Jennison, 1992). In the MISTIE example, the most striking difference between optimized and non-optimized trials is the futility boundaries, while the other parameters are roughly similar. We conjecture that this change in futility boundary is an important driver of improved trial performance.

In both the MISTIE and ADNI examples, we compared covariance-based ($\mathscr{H}^{COV}$) and alpha-reallocation-based ($\mathscr{H}^{MB}$) multiple testing procedures, based on optimizing multistage designs using each procedure. The resulting designs were similar in both their design parameters and their performance.

The optimized designs in this paper are the best designs found by the SA algorithm we used.

These could be local optima. It is an open problem to determine how close these are to global optima for the problems we addressed. This is because no current method exists to find the global optimum in these problems due to the relatively large number of design parameters being simultaneously searched over.

One exciting area of future work is to modify the search algorithm to actively account for the Monte Carlo simulation error in each objective function evaluation. Some optimization methods leverage noise present in the objective function, or add noise to the objective function (Kushner, 1987; Maryak and Chin, 2001), in order to increase the probability of reaching a global minimum. In the specific context of SA, (Fink, 1998; Branke et al., 2008) argue that noise in the objective function is analogous to having a higher temperature parameter.

Acknowledgments

This work was supported by the Patient-Centered Outcomes Research Institute (ME-1306-03198) and the U.S. Food and Drug Administration (HHSF223201400113C). We thank Daniel Hanley and Michela Gallagher for providing summary statistics from the MISTIE and ADNI data, respectively.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

This work is solely the responsibility of the authors and does not represent the views of the above people and agencies.

References

Bélisle, C. J. (1992). Convergence theorems for a class of simulated annealing algorithms on R^d. *Journal of Applied Probability 29*(4), 885–895. 4.4

Branke, J., S. Meisel, and C. Schmidt (2008). Simulated annealing in the presence of noise. *Journal of Heuristics 14*(6), 627–654. 5.4, 6

Bretz, F., W. Maurer, W. Brannath, and M. Posch (2009). A graphical approach to sequentially rejective multiple test procedures. *Statist. Med. 28*(4), 586. 3.1, 3.3

Eales, J. D. and C. Jennison (1992). An improved method for deriving optimal one-sided group sequential tests. *Biometrika 79*(1), 13–24. 1, 6

Fink, T. M. (1998). *Inverse protein folding, hierarchical optimisation and tie knots.* Ph. D. thesis, University of Cambridge. 6

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

Hampson, L. V. and C. Jennison (2013). Group sequential tests for delayed responses (with discussion). *J. R. Statist. Soc. B 75*(1), 3–54. 1

Hampson, L. V. and C. Jennison (2015). Optimizing the data combination rule for seamless phase ii/iii clinical trials. *Statist. Med. 34*(1), 39–58. 1

Jennison, C. and B. W. Turnbull (1999). *Group Sequential Methods with Applications to Clinical Trials.* Chapman and Hall/CRC Press. 3.1, 3.2, 4.5

Krisam, J. and M. Kieser (2015). Optimal decision rules for biomarker-based subgroup selection for a targeted therapy in oncology. *International Journal of Molecular Sciences 16*(5), 10354–10375. 1

Kushner, H. (1987). Asymptotic global behavior for stochastic approximation and diffusions with slowly decreasing noise effects: global minimization via monte carlo. *SIAM Journal on Applied Mathematics 47*(1), 169–185. 6

Liu, Q. and K. M. Anderson (2008). On adaptive extensions of group sequential trials for clinical investigations. *J. Amer. Statist. Assoc 103*(484), 1621–1630. 1

Maryak, J. L. and D. C. Chin (2001). Global random optimization by simultaneous perturbation stochastic approximation. In *American Control Conference, 2001. Proceedings of the 2001*, Volume 2, pp. 756–762. IEEE. 6

Maurer, W. and F. Bretz (2013). Multiple testing in group sequential trials using graphical approaches. *Statistics in Biopharmaceutical Research 5*(4), 311–320. 1, 3.1, 3.3, 4.3

Morgan, T., M. Zuccarello, R. Narayan, P. Keyl, K. Lane, and D. Hanley (2008). Preliminary findings of the minimally-invasive surgery plus rtPA for intracerebral hemorrhage evacuation (MISTIE) clinical trial. *Acta Neurochir Suppl. 105*, 147–51. 2.1

O'Brien, P. C. and T. R. Fleming (1979). A multiple testing procedure for clinical trials. *Biometrics 35*(3), 549–556. 1, 3.2, 4.5, 5.2

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika 64*, 191–199. 1, 3.2, 4.5, 5.2

Rosenblum, M., X. Fang, and H. Liu (2014). Optimal, two stage, adaptive enrichment designs for randomized trials using sparse linear programming. *Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 273. http://biostats.bepress.com/jhubiostat/paper273 109*, 1216–1228. 1

Rosenblum, M., B. Luber, R. E. Thompson, and D. Hanley (2016). Group sequential designs with prospectively planned rules for subpopulation enrichment. *Statistics in Medicine 35*(21), 3776–3791. sim.6957. 1, 2.1, 3.1, 3.2, 4.3

Rosenblum, M., T. Qian, Y. Du, H. Qiu, and A. Fisher (2016). Multiple testing procedures for adaptive enrichment designs: combining group sequential and reallocation approaches. *Biostatistics 17*(4), 650–662. 1

Sadigh-Eteghad, S., M. Talebi, and M. Farhoudi (2012). Association of apolipoprotein E epsilon 4 allele with sporadic late onset Alzheimer's disease. *A meta-analysis. Neurosciences (Riyadh) 17*(4), 321–326. 2.2

Thall, P. F., R. Simon, and S. S. Ellenberg (1988). Two-stage selection and testing designs for comparative clinical trials. *Biometrika 75*(2), 303–310. 1

Wang, S. J., H. Hung, and R. T. O'Neill (2009). Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal 51*, 358–374. 1

Wason, J. and T. Jaki (2012). Optimal design of multi-arm multi-stage trials. *Statist. Med. 31*(30), 4269–4279. 1, 4.3, 4.3, 4.4

Wason, J., A. P. Mander, and S. G. Thompson (2012). Optimal multistage designs for randomised clinical trials with continuous outcomes. *Statist. Med. 31*(4), 301–312. 1, 4.3, 4.3, 4.4