9-20-2017

COMPARISON OF ADAPTIVE RANDOMIZED TRIAL DESIGNS FOR TIME-TO-EVENT OUTCOMES THAT EXPAND VERSUS RESTRICT ENROLLMENT CRITERIA, TO TEST NON-INFERIORITY

Josh Betz
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Jon Arni Steingrimsson
Department of Biostatistics, Brown School of Public Health

Tianchen Qian
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

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

Comparison of Adaptive Randomized Trial Designs for Time-to-Event Outcomes that Expand Versus Restrict Enrollment Criteria, to Test Non-Inferiority

Josh Betz, Jon Arni Steingrimsson, Tianchen Qian, Michael Rosenblum

September 20, 2017

1 Abstract

Adaptive enrichment designs involve preplanned rules for modifying patient enrollment criteria based on data accrued in an ongoing trial. These designs may be useful when it is suspected that a subpopulation, e.g., defined by a biomarker or risk score measured at baseline, may benefit more from treatment than the complementary subpopulation. We compare two types of such designs, for the case of two subpopulations that partition the overall population. The first type starts by enrolling the subpopulation where it is suspected the new treatment is most likely to work, and then may expand inclusion criteria if there is early evidence of a treatment benefit. The second type starts by enrolling from the overall population and then may selectively restrict enrollment if sufficient evidence accrues that the treatment is not benefiting a subpopulation. We construct two-stage designs of each type that guarantee strong control of the familywise Type I error rate, asymptotically. We then compare performance of the designs from each type under different scenarios; the scenarios mimic key features of a completed non-inferiority trial of HIV treatments. Performance criteria include power, sample size, Type I error, estimator bias, and confidence inteval coverage probability.

Keywords: qualitative interaction, treatment effect heterogeneity, trial optimization

2 Introduction

Our trial design optimization problem is motivated by the Prospective Evaluation of Antiretrovirals in Resource Limited Settings (PEARLS) study of the AIDS Clinical Trials Group (ACTG Trial A5175) Campbell et al. (2012). This was a randomized, non-inferiority trial that enrolled 1,571 HIV positive participants, each randomized to one of three HIV treatments: ATV+DDI+FTC, EFV+3TC-ZDV, or EFV+FTC-TDF, referred to as treatment arms A, B, and C, respectively. The primary outcome was time to the composite endpoint of virologic failure, HIV disease progression (AIDS), or death. The treatment in arm A had

potential benefits over the other treatments for women, in that it did not require stopping medication during pregnancy.

At an interim analysis, a comparison between overall event rates in arms A and B showed evidence of inferiority of arm A, which was then stopped. However, there was evidence of a difference in treatment effects for men and women, in the direction of arm A being worse for men. This left open the question of whether the treatment in arm A, which as described above has a potential advantage for women, may have been just as effective as arm B for women only. This motivates exploring whether an adaptive enrichment design with subpopulation-specific rules for early stopping could have better answered this question. We construct new adaptive enrichment designs for non-inferiority trials tailored to answering this question and evaluate these designs in simulations. Our focus is on arms A and B only.

Our approach may be relevant for problems where it is suspected that there is a qualitative interaction; in the non-inferiority trial context, this means inferiority of treatment A versus B for one subpopulation, but non-inferiority for the complementary subpopulation. We assume throughout that there are two subpopulations of interest that partition the overall population, and that these are prespecified in the study protocol.

Liu et al. (2010) present a two-stage adaptive design that enrolls one subpopulation in the first stage; a decision is made after stage 1 to terminate the trial or to enroll both subpopulations in stage 2. We build on this approach, where our novel contributions include the following: optimizing such designs, comparing them to adaptive designs that start by enrolling the combined population and that may restrict enrollment, and considering non-inferiority trials.

Russek-Cohen and Simon (1997) consider two-stage designs for two subpopulations (men and women). These designs start by enrolling both subpopulations in stage 1 and then may stop the trial completely at the interim analysis or continue follow-up for both subpopulations in stage 2. In contrast, our adaptive designs allow early stopping of accrual for the overall population or for a single subpopulation.

Related work on optimizing adaptive enrichment designs includes, e.g., Graf et al. (2015), Krisam and Kieser (2015), Götte et al. (2015), Rosenblum et al. (2016). These involved optimizing over 2 or 3 design parameters; in contrast, we optimize over many more design parameters. Our approach to handle this challenge is to optimize using simulated annealing. Related work using simulated annealing for trial design optimization in different contexts include Wason and Jaki (2012) and Fisher and Rosenblum (2016). The former consider multi-arm trials involving a single population; the latter consider adaptive enrichment designs for continuous or binary treatments in the context of superiority trials. In contrast, we consider time-to-event outcomes and non-inferiority trials. We also compare designs that start enrolling both subpopulations and potentially restrict enrollment versus designs that start enrolling the favored subpopulation and potentially expand enrollment. Rosenblum et al. (2017) optimize over two stage adaptive enrichment designs that do not have the features in the previous two sentences.

We next discuss limitations of our approach. Duration may be long, especially if hazard rates are low relative to enrollment rates, since then there is a substantial delay between

enrollment and information accrual. The same problem would occur if either subpopulation is a relatively small proportion of the overall population. These problems would affect almost any design that has a power requirement for such a subpopulation, since it would take substantial time for sufficient information to accrue for that subpopulation. Another limitation is that our trial design optimization problem is high dimensional (i.e., the number of design parameters to be optimized is relatively large), and there does not yet exist a computationally feasible algorithm for computing the global optimum design. Instead, we apply the general purpose optimization method of simulated annealing to search over design classes, with the goal of improving performance compared to standard designs.

3 Prospective Evaluation of Antiretrovirals in Resource Limited Settings (PEARLS) Trial

The primary outcome, called the failure time, was time to the first of virologic failure, HIV disease progression (AIDS), or death. The main efficacy result from Campbell et al. (2012) regarding treatments A and B is the following: "Comparing ATV+DDI+FTC to EFV+3TC+ZDV, during a median follow-up of 81 wk there were 108 failures (21%) among 526 participants assigned to ATV+DDI+FTC and 76 (15%) among 519 participants assigned to EFV+3TC-ZDV (HR 1.51, CI 1.122.04; p=0.007)." However, when the same comparison of treatments A and B was stratified by sex, according to Campbell et al. (2012), "Men randomized to ATV+DDI+FTC had higher risk of treatment failure compared to men randomized to EFV+3TC-ZDV (HR 2.14, CI 1.423.42) but a difference in regimen efficacy was not detected in women."

In addition to the above efficacy results, treatment A had a lower risk of the safety endpoint compared to treatment B. According to Campbell et al. (2012), "Women randomized to ATV+DDI+FTC had lower risk of a safety endpoint compared to women randomized to EFV+3TC-ZDV (HR 0.56, CI 0.42–0.74)." This supports the conjecture that treatment A, if it is non-inferior to treatment B for women for the primary efficacy outcome, could be preferable due to its better safety profile for women. This motivated our comparison of different trial designs to learn about treatment effects in subpopulations, in the context of non-inferiority trials.

4 Trial Design Optimization Problem

4.1 Assumptions

We let subpopulation 1 denote women and subpopulation 2 denote men. The trial is assumed to be group sequential, that is, with predetermined interim analysis times. If the trial is not stopped early, the study enrolls participants until a prespecified time c. Each participant is followed until the first of the following occurs: failure or the end of study. We assume that enrollment is uniform until it is stopped. We assume that the only form of censoring is

administrative censoring, i.e., due to reaching the end of the study. For each subpopulation $s \in \{1, 2\}$ by treatment $a \in \{A, B\}$ pair, we assume a constant hazard rate λ_{sa} over time. The asymptotic, joint distribution of the statistics used by our designs (based on a proportional hazards model defined in Section 5.2) is multivariate normal with the canonical mean and covariance structure from Jennison and Turnbull (1999, Chapter 3.1); throughout, we approximate the joint distribution of statistics using this limit distribution.

In our data generating distributions, we mimic key features from the PEARLS trial. We use the observed proportions $p_1 = 0.47, p_2 = 0.53$ of the subpopulations of women and men, respectively. To mimic the hazard rate under treatment B observed in the PEARLS trial, we set $\lambda_{sB} = 0.08$ for each $s \in \{1, 2\}$. To mimic the observed enrollment rate for the combined population (1571 participants over 2.17 years), we set the combined population enrollment rate per year to be $1571/2.17 \approx 724$; we also consider the case where the enrollment rate is half that from the PEARLS trial, i.e., 362 participants per year. We assume enrollment in each subpopulation is proportional to the subpopulation size, i.e., enrollment for $s = 1$ is $0.47 * 724$ per year and for $s = 2$ is $0.53 * 724$ per year.

4.2 Null Hypotheses Tested

We test for non-inferiority of a single treatment (arm A) versus control (arm B) in each of two disjoint subpopulations. The outcome of interest is time-to-event. The event of interest is a failure time, and so lower event rates are desirable.

We define the two subpopulations to be women and men, denoted by $s = 1, 2$, respectively. The non-inferiority margin is defined, as in the PEARLS trial, as a hazard ratio HR (treatment A hazard rate divided by treatment B hazard rate) at most 1.35. Define the null hypothesis of inferiority of treatment A to treatment B for each subpopulation as:

- H_{01}: hazard ratio for subpopulation 1 (women) at least the non-inferiority margin 1.35.
- H_{02}: hazard ratio for subpopulation 2 (men) at least the non-inferiority margin 1.35.

The alternative hypothesis for each subpopulation is non-inferiority at margin 1.35, i.e., hazard ratio < 1.35. Our goal is to construct a trial design, consisting of an accrual modification rule and multiple testing procedure for $\{H_{01}, H_{02}\}$, that minimizes expected sample size under power and Type I error constraints.

4.3 Scenarios

The performance criterion (minimizing expected sample size) and power constraints are with respect to the four scenarios defined as follows:

1. Treatment A is equivalent to treatment B (hazard ratio 1) for each subpopulation (women and men).
2. Treatment A is equivalent to treatment B (hazard ratio 1) for subpopulation 1 (women), but A is inferior to B (hazard ratio 1.35) for subpopulation 2 (men).

3. Treatment A is equivalent to treatment B (hazard ratio 1) for subpopulation 1 (women), but A is highly inferior to B (hazard ratio 2.14) for subpopulation 2 (men).

4. Treatment A is inferior to treatment B (hazard ratio 1.35) for each subpopulation (women and men).

We chose the hazard ratio 1.35 since this is the preplanned, non-inferiority margin in the PEARLS trial. The hazard ratio 2.14 is the estimated hazard ratio comparing treatments A to B in men, from the PEARLS trial. We think both are relevant to consider.

4.4 Type I Error and Power Constraints

Familywise Type I error constraints: We require strong control of the familywise Type I error rate, that is, the probability of rejecting one or more true null hypotheses must be at most $\alpha = 0.05$, asymptotically. This must hold regardless of the hazard rates in each arm for each subpopulation. In scenario 1, both null hypotheses are false, so it is not possible to make a Type I error; in scenarios 2 and 3 only H_{02} is true, so the Type I error constraint is that the probability of rejecting H_{02} is at most 0.05; in scenario 4, both null hypotheses are true, so the Type I error constraint is that the probability of rejecting one or more null hypotheses is at most 0.05. *Strong control* of the familywise Type I error rate means that not only should the familywise Type I error be at most 0.05 in the above 4 scenarios, but also that this must hold for any possible values of the hazard rates.

Power Constraints: The following are the power constraints required of each design:

1. In scenario 1, power at least 0.8 to reject H_{01} and power at least 0.8 to reject H_{02}.

2. In scenarios 2 and 3, power at least 0.8 to reject H_{01}.

3. Scenario 4: no constraints (since both null hypotheses are true).

4.5 Objective Function

Our performance goal is represented by an objective function, which is the expected sample size. Expectation is defined with respect to the distribution that assigns equal weight to each scenario 1-4. This distribution can be informally thought of as a prior. However, it is only used in defining the objective functions below, and is not used otherwise; in particular, the familywise Type I error constraints are not with respect to this distribution, since they are required to hold regardless of the hazard rates for each subpopulation by arm combination.

Expectation is over two sources of randomness: the scenario and the statistics conditioned on the scenario. The former is drawn from the equal weight distribution on scenarios 1-4; the latter, asymptotically, has multivariate normal distribution with mean vector and covariance matrix depending on the scenario and the information accrued at each analysis.

4.6 Statement of Trial Design Optimization Problem

For a given class of trial designs (defined in the next section), the trial design optimization problem is to minimize expected sample size under the familywise Type I error and power constraints from Section 4.4. Three classes of designs are defined below; we optimize over each class of designs and then compare the optimal design from each class.

5 Classes of Trial Designs Optimized Over

5.1 Overview

We consider the following three classes of designs:

1. $\mathcal{D}_{ONE-STAGE}$: Standard (non-adaptive) designs consisting of a single stage
2. $\mathcal{D}_{ADAPTIVE,START-BOTH}$: Adaptive enrichment design that starts enrolling both subpopulations
3. $\mathcal{D}_{ADAPTIVE,START-SUBPOP.1}$: Adaptive enrichment design that starts enrolling only subpopulation 1 (women); enrollment of the complementary subpopulation (men) starts (if it gets started at all) immediately after the first interim analysis.

Each enrolled participant is randomized to treatment A or B with probability 1/2. Each participant is followed until they experience an event or until information accrual is stopped for their subpopulation. Stopping enrollment and stopping information accrual are different; the former means that no new participants (from a given subpopulation) are enrolled in the trial, while the latter means enrollment is stopped (if not stopped previously) and follow-up is stopped.

The primary reason for stopping information accrual of a subpopulation before the trial is completed is ethical: if treatment A is demonstrated to be inferior to arm B in subpopulation 2, then the treatment should be switched to arm B for all subpopulation 2 participants in the inferior arm. In the PEARLS trial (which did not have rules for stopping subpopulations), the entire arm A was stopped early for inferiority and the participants in that arm were switched to treatment B. We use simulations that mimic features of the PEARLS trial to explore the tradeoffs involved in using adaptive versus standard designs to achieve the goals in Section 4.

Each design is defined by the following parameters, each of which must be preplanned in the study protocol:

1. Number of stages K. We set $K = 2$ throughout.
2. Analysis times $t_1 < \cdots < t_K$.
3. Maximum number N_s to enroll from each subpopulation $s \in \{1, 2\}$.

4. Rule for terminating each subpopulation's accrual (which can only occur at the pre-planned interim analyses).

5. For the class of designs that starts by enrolling only subpopulation 1, the rule that determines if subpopulation two enrollment will start just after the first interim analysis.

6. Multiple testing procedure for the null hypotheses $\{H_{01}, H_{02}\}$.

5.2 Common Features of the Three Design Classes

The decision rules we consider can make the following modification at any interim analysis: for each subpopulation $s \in \{1, 2\}$ whose accrual has continued until the current analysis time, accrual can be allowed to continue into the next stage or can be stopped. We do not allow restarting enrollment or accrual for a subpopulation once it has already been stopped. For the design that delays the start of subpopulation $s = 2$ enrollment until stage 2, this is the only opportunity to start this subpopulation's enrollment; the decision rule can decide at the first interim analysis to never start this subpopulation's enrollment.

For each subpopulation $s \in \{1, 2\}$, each design prespecifies the maximum number N_s that can be enrolled. If subpopulation s enrollment is stopped before N_s have enrolled, then all accrual (follow-up) is stopped for subpopulation s as well.

The decision rule for modifying accrual and the multiple testing procedure use a Wald test statistic from a proportional hazards model computed separately for each subpopulation based on cumulative data available at a given analysis. Let $Z_{s,k}$ denote the Wald statistic corresponding to subpopulation $s \in \{1, 2\}$ based on all of that subpopulation's data accrued up through the end of stage k. This Wald statistic, on the z-score scale, is defined as the standardized difference between the natural log of the non-inferiority margin 1.35 and the estimated coefficient on the indicator of assignment to arm A in a Cox proportional hazards model with intercept and a main term for this indicator. Ceteris paribus, large, positive values of the Wald statistic are more likely for smaller hazard ratios (the hazard rate for treatment A divided by that for treatment B), which favor the alternative hypothesis (of non-inferiority) compared to the null hypothesis (of inferiority). The reason for this choice (which has positive and negative reversed compared to typical usage for non-inferiority designs) is so that efficacy and futility boundaries on the z-scale can be more easily interpreted (analogous to the case of continuous or binary outcomes). Formal definitions of the statistics $Z_{s,k}$ are given in Section B of the Appendix.

Each decision rule for modifying accrual at interim analysis k has the following form: if the cumulative statistic $Z_{s,k}$ for subpopulation s is above an efficacy boundary $e_{s,k}$ (in which case H_{0s} is rejected) or below a futility boundary $f_{s,k}$ (in which case we fail to reject H_{0s}), accrual for that subpopulation is stopped; otherwise, accrual continues. The main challenge is to determine the interim analysis times and efficacy/futility boundaries $\{e_{s,k}, f_{s,k}\}$ for which the corresponding design satisfies the requirements on power and Type I error at the minimum cost in terms of expected sample size.

All of our designs use error-spending functions (G. Lan and DeMets, 1983) to determine the efficacy boundaries at each stage. These error-spending functions are allowed to be quite general; they are not restricted to standard families of error spending functions such as the power family described by Jennison and Turnbull (1999).

All futility boundaries are non-binding. That is, even if these boundaries are not adhered to, the familywise Type I error rate is controlled (in the strong sense) at level 0.05. Non-binding futility boundaries have the advantage of often being preferred by regulators (Liu and Anderson, 2008); a disadvantage is that power may be increased by instead using binding futility boundaries.

We restrict the duration of all of our trial designs to be at most 8 years. The reason for imposing this restriction is that once enrollment has been completed, a longer trial duration will only lead to more information accrual (i.e., more events occurring) with no cost in terms of added sample size; therefore, if duration were not restricted, the optimal design (in terms of expected sample size) in any of our classes would set duration equal to infinity. We set the maximum duration to be 8 years since that is sufficient for each class of designs to meet all of the power and Type I error constraints. It is an area for future research to solve our optimization problem at different values of maximum duration, to determine tradeoffs between expected sample size and duration within each design class.

Let the design parameter $t_{MAX-ENROLL}$ denote the maximum time at which enrollment is allowed to proceed for each subpopulation. We optimize over this parameter (as well as other parameters) within each of the three design classes. In each of the first two design classes, we set the subpopulation s maximum sample size N_s equal to the product of $t_{MAX-ENROLL}$, the subpopulation proportion p_s, and the combined population enrollment rate (defined in Section 4.1). The definition is similar for the third design class, except that for N_2 we replace $t_{MAX-ENROLL}$ in the previous sentence by $t_{MAX-ENROLL}$ minus the time t_1 of the first interim analysis (which is the time when subpopulation 2 enrollment can be started).

5.3 Class of Standard (Non-Adaptive) Designs $\mathcal{D}_{ONE-STAGE}$

This class of designs consists of a single stage. The multiple testing procedure is based on the Bonferroni multiplicity correction with α partitioned between the two null hypotheses. We allow α to be split unequally between the two null hypotheses, which may be useful due to the asymmetry of the problem (e.g., in subpopulation proportions). The design parameters to be optimized are $t_{MAX-ENROLL}$ (which determines the sample size in each subpopulation) and the partitioning of α between the two null hypotheses.

5.4 Adaptive Enrichment Designs Enrolling Both Subpopulations in Stage 1: $\mathcal{D}_{ADAPTIVE,START-BOTH}$

This class of designs is defined by the following actions. At the start of the trial, enroll participants from both subpopulations. At the analysis after each stage k, for each subpopulation s the statistic $Z_{s,k}$ is used to test for efficacy and futility. If $Z_{s,k} > e_{s,k}$ for the efficacy

boundary $e_{s,k}$, then the null hypothesis H_{0s} is rejected and accrual from subpopulation s is stopped. If $Z_{s,k} < f_{s,k}$ for the futility boundary $f_{s,k}$, then accrual from subpopulation s is stopped and the procedure fails to reject H_{0s}. If a null hypothesis is rejected at an interim analysis, it stays rejected throughout the trial.

The efficacy boundaries are calculated using an error spending approach. Define the alpha allocation across subpopulations and stages as $\{\alpha_{s,k} \geq 0 : s = 1, 2; k = 1, \ldots, K\}$ satisfying $\sum_{s=1}^{2} \sum_{k=1}^{K} \alpha_{s,k} = \alpha$, where $\alpha = 0.05$ is the familywise Type I error rate. The alpha allocation will be optimized (along with other design parameters described below).

Let H_0 denote the global null hypothesis that the hazard ratio for subpopulation s equals the non-inferiority margin 1.35, for each $s \in \{1, 2\}$. Under H_0, the statistics $\{Z_{s,k} : s = 1, 2; k = 1, \ldots, K\}$ are asymptotically, multivariate normal with mean 0 and covariance matrix derived in the Appendix. We are not interested in testing H_0; the only reason for defining it is for use below in computing the efficacy boundaries $e_{s,k}$.

For each subpopulation $s \in \{1, 2\}$, the efficacy boundaries $\{e_{s,k} : k = 1, \ldots, K\}$ are calculated sequentially. First $e_{s,1}$ is calculated by finding the smallest $e_{s,1}$ satisfying

$$P_{H_0}(Z_{s,1} > e_{s,1}) \leq \alpha_{s,1}.$$

The efficacy boundaries $\{e_{s,k} : 1 < k \leq K\}$ are then calculated sequentially. The efficacy boundary $e_{s,k}$ is defined as the smallest $e_{s,k}$ satisfying

$$P_{H_0}(Z_{s,k} > e_{s,k} \text{ and for all } k' < k, Z_{s,k'} \leq e_{s,k'}) = \alpha_{s,k}. \tag{1}$$

We use the convention that $e_{s,k} = \infty$ if $\alpha_{s,k} = 0$.

To improve power, we use alpha reallocation to lower the efficacy boundaries for one subpopulation if the null hypothesis for the other subpopulation gets rejected, using ideas similar to those in Liu and Anderson (2008); Maurer and Bretz (2013). Specifically, if the null hypothesis H_{0s} is rejected at any stage $k \leq K$, then to calculate the stage K efficacy boundary $e_{s',K}$ for the other subpopulation $s' \neq s$, replace the final stage value $\alpha_{s',K}$ on the right side of (1) by $\tilde{\alpha}_{s',k} = \alpha_{s',K} + \sum_{k'=1}^{K} \alpha_{s,k'}$. Reallocation of alpha from one null hypothesis to the other is only done after rejection of the former; no reallocation occurs when a subpopulation is stopped for futility. The above multiple testing procedure strongly controls the familywise Type I error rate asymptotically; this follows since the statistics $\{Z_{s,k} : s = 1, 2; k = 1, \ldots, K\}$ have asymptotic joint distribution equal to the canonical joint distribution (Jennison and Turnbull, 1999, p.49).

The set of design parameters for the class $\mathcal{D}_{ADAPTIVE,START-BOTH}$ that are optimized include the following for $K = 2$:

1. the analysis times $t_1 < \cdots < t_K$;
2. the maximum enrollment time $t_{MAX-ENROLL}$;
3. the alpha allocation $\{\alpha_{s,k} \geq 0 : s = 1, 2; k = 1, \ldots, K\}$ that sums to $\alpha = 0.05$;
4. the futility boundaries $\{f_{s,k} \in \mathbb{R} : s = 1, 2; k = 1, \ldots, K - 1\}$.

We restrict the search space by requiring $t_1 \geq 0.5$ and $t_2 = 8$. The reason was that we wanted to avoid a decision with too little information accrued, and also there is no loss (in terms of the objective function of expected sample size) to continuing follow-up of all enrolled participants to the maximum allowed duration of 8 years.

5.5 Adaptive Enrichment Designs Enrolling Only Subpopulation 1 in Stage 1: $\mathcal{D}_{ADAPTIVE,START-SUBPOP.1}$

The decision rule for accrual modification and multiple testing procedure are similar in structure to those in Section 5.4, except that only subpopulation 1 is enrolled in stage 1; enrollment for subpopulation 2 can only be started at interim analysis 1. The alpha allocation, efficacy boundary construction, and rule for alpha reallocation are defined exactly as in Section 5.4 except that we require $\alpha_{2,1} = 0$ since there is not yet any data from subpopulation 2 at analysis 1; the impact is that $e_{2,1} = \infty$, i.e., there is no early stopping of subpopulation 2 for efficacy at the first interim analysis. For the same reason, we set the futility boundary $f_{2,1} = -\infty$.

At the analysis after each stage $k = 1, \ldots, K$ in which subpopulation 1 accrual has continued, the statistic $Z_{1,k}$ is used to test for efficacy and futility exactly as described above for $\mathcal{D}_{ADAPTIVE,START-BOTH}$. Specifically, if $Z_{1,k} > e_{1,k}$ then the null hypothesis H_{01} is rejected and accrual from subpopulation 1 is stopped, while if $Z_{1,k} < f_{1,k}$ then accrual from subpopulation 1 is stopped and the procedure fails to reject H_{01}; otherwise accrual continues.

At the first interim analysis, enrollment of subpopulation 2 is started if $Z_{1,1} > \theta$, where θ is a predefined threshold; otherwise, if $Z_{1,1} \leq \theta$ then subpopulation 2 enrollment is never started. If subpopulation 2 enrollment is started at analysis 1, then at each analysis $k > 1$ where subpopulation 2 accrual has continued, the statistic $Z_{2,k}$ is used to test for efficacy and futility. The null hypothesis H_{02} is rejected if $Z_{2,k} > e_{2,k}$, while if $Z_{2,k} < f_{2,k}$ then accrual for subpopulation 2 is stopped for futility; otherwise accrual continues.

The design parameters for the class $\mathcal{D}_{ADAPTIVE,START-SUBPOP.1}$ to be optimized are the same as for $\mathcal{D}_{ADAPTIVE,START-BOTH}$ except that we have the additional parameter $\theta \in \mathbb{R}$ and we set $\alpha_{2,1} = 0$ and $f_{2,1} = -\infty$.

6 Results

6.1 Description of Optimal Designs from Each Class and Their Performance: Enrollment Rate 724 Per Year

For each class of adaptive enrichment designs, we solved the optimization problem from Section 4.6 using enrollment rate 724 participants per year. We consider the case of 362 participants per year in the next subsection. The optimization algorithms we used, including simulated annealing, are described in Section A of the Appendix. We present the results of these optimizations below. We refer to the design returned by the optimization algorithm that has the smallest expected sample size among those that satisfy the power and Type I

error constraints as the optimal design in a given class. These may be local optima rather than the global optimum solution; finding the global optimum is an open problem in our context, due to the objective function and constraints being nonconvex functions of the design parameters.

For the class $\mathcal{D}_{ONE-STAGE}$, the optimal design set $t_{MAX-ENROLL} = 1.97, \alpha_{1,1} = 0.88, \alpha_{2,1} = 0.12$. The expected sample size is 1426. The optimal design from each of the classes $\mathcal{D}_{ADAPTIVE,START-BOTH}$ and $\mathcal{D}_{ADAPTIVE,START-SUBPOP.1}$ had expected sample size no lower than 1426. We were not able to detect any value added from our adaptive designs for this problem.

6.2 Description of Optimal Designs from Each Class and Their Performance: Enrollment Rate 362 Per Year

We also solved the above optimization problem where we set the enrollment rate to be 724/2=362 participants per year. The reason was to show the impact of a slower enrollment rate, which gives more opportunity for adaptation later in the trial to impact the expected sample size. The results are given below.

For the class $\mathcal{D}_{ONE-STAGE}$, the optimal design set $t_{MAX-ENROLL} = 4.70, \alpha_{1,1} = 0.88, \alpha_{2,1} = 0.12$. The familywise type I error under each scenario 2, 3, 4 is 0.041, 0.000, 0.050, respectively. The power to reject H_{01} under each scenario 1, 2, 3 is 0.81, 0.80, 0.80, respectively. The power to reject H_{02} under scenario 1 is 0.80. The sample size is 1702.

For the class $\mathcal{D}_{ADAPTIVE,START-BOTH}$, the optimal design at $K = 2$ had the following design parameters:

1. the analysis times $(t_1, t_2) = (3.4, 8)$.
2. the maximum enrollment time $t_{MAX-ENROLL} = 4.97$;
3. the alpha allocation $\alpha_{1,1} = 0.15$, $\alpha_{2,1} = 0.01$,$\alpha_{1,2} = 0.74$, $\alpha_{2,2} = 0.10$.
4. the futility boundaries $f_{1,1} = -2.1$,$f_{2,1} = -0.74$.

The expected sample size is 1660 and the maximum sample size is 1799. The familywise type I error under each scenario 2, 3, 4 is 0.041, 0.000, 0.051, respectively. The power to reject H_{01} under each scenario 1, 2, 3 is 0.81, 0.80, 0.80, respectively. The power to reject H_{02} under scenario 1 is 0.80. For estimation of the subpopulation 1 hazard ratio, the bias was at most 0.02 and the coverage probability of each nominal 95% confidence interval (constructed ignoring the adaptive nature of the trial design and using all data available at the end of the trial) was at least 0.93, for each scenario. The analogous quantities for subpopulation 2 are 0.08 for bias and 92% for confidence interval coverage. We conjecture that the performance degradation for subpopulation 2 is due to the corresponding futility boundary being closer to 0, which results in early stopping of this subpopulation with probability approximately 4%, 23%, 96%, 23% in scenarios 1, 2, 3, 4, respectively. This leads to the reduction in expected sample size compared to the single stage design described above, but also induces bias.

For the class $\mathcal{D}_{ADAPTIVE,START-SUBPOP.1}$, the optimal design at $K = 2$ had the following design parameters:

1. the analysis times $(t_1, t_2) = (0.5, 8)$.
2. the maximum enrollment time $t_{MAX-ENROLL} = 5.39$;
3. the alpha allocation $\alpha_{1,1} = 0.02$, $\alpha_{1,2} = 0.75$, $\alpha_{2,2} = 0.23$.
4. the futility boundary $f_{1,1} = -2.8$
5. The threshold $\theta = -2.8$ to start enrollment of subpopulation 2 just after interim analysis 1.

The expected sample size is 1850 and the maximum sample size is 1951. The familywise type I error under each scenario 2, 3, 4 is 0.042, 0.000, 0.049, respectively. The power to reject H_{01} under scenarios 1, 2, 3 is 0.83, 0.81, 0.81, respectively. The power to reject H_{02} under scenario 1 is 0.80.

In summary, for the enrollment rate 362 per year, the optimal design from $\mathcal{D}_{ADAPTIVE,START-BOTH}$ has 42 fewer expected participants compared to the optimal 1-stage design, but at the cost of having 97 more participants in the worst-case (maximum sample size). There was no advantage provided by the class $\mathcal{D}_{ADAPTIVE,START-SUBPOP.1}$.

7 Discussion

The added value, if any, from our adaptive designs can be highly dependent on the enrollment rate versus the event rate. In general, we expect that slower enrollment rates and higher event rates will lead to more value added from the adaptive designs compared to single stage (non-adaptive) designs. This is because the adaptive designs require sufficient information (which is roughly proportional to the total number of events pooling both arms) to have accrued in order to make an informed decision about modifying enrollment criteria before enrollment ends.

There was no benefit to the adaptive designs in terms of expected sample size for the original optimization problem, i.e., the problem where the enrollment rate was 724 participants per year. However, for the problem with half that enrollment rate, there was a benefit to adaptation in terms of reduced expected sample size; the cost is an increase in the maximum sample size. Such a tradeoff between expected sample size and maximum sample size also occurs when comparing group sequential designs (for a single population) to single stage designs. There is also a tradeoff in that adaptation leads to increased bias and a reduction in confidence interval coverage compared to the standard design, for one of the subpopulations. It may be possible to reduce bias and increase confidence interval coverage by adapting methods from, e.g., Posch et al. (2005); Rosenblum (2013); Kunzmann et al. (2017).

For both optimization problems, the class of adaptive designs $\mathcal{D}_{ADAPTIVE,START-SUBPOP.1}$ provided no added value. We conjecture that if other optimization criteria are considered, such as the total participant-time, then this class may provide advantages. This is an open area for future research.

Limitations of our adaptive enrichment designs include that they only allow enrichment to a single subpopulation and they do not incorporate adaptations such as sample size re-assessment for a given stage or adaptation of randomization probabilities. We assumed that the subpopulation definitions were determined before the trial started; this requires both prior data and domain-specific knowledge as to who is more likely to benefit from treatment. In cases where these are not available, the proposed adaptive enrichment designs would not be appropriate.

We assumed proportional hazards and that the hazard rate for each subpopulation by arm combination does not change over time. An area of future research is to examine the robustness of our results to deviations from these assumptions. Similarly, we only considered administrative censoring; an area for future research is to consider different censoring distributions, e.g., differential censoring by study arm and subpopulation, and censoring that changes over time. Another area for future research is to consider estimators that leverage information in prognostic baseline variables to improve precision, e.g., Lu and Tsiatis (2011).

Acknowledgments

This work was supported by the Patient-Centered Outcomes Research Institute (ME-1306-03198), the U.S. Food and Drug Administration (HHSF223201400113C), and the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number UM1 AI068634, UM1 AI068636 and UM1 AI106701. This publication's contents are solely the responsibility of the authors and do not represent the official views of the above agencies. We thank Victor DeGruttola for his valuable input on this project.

References

Campbell, T. B., L. M. Smeaton, N. Kumarasamy, T. Flanigan, K. L. Klingman, C. Firnhaber, B. Grinsztejn, M. C. Hosseinipour, J. Kumwenda, U. Lalloo, M. M. Cynthia Riviere, Jorge Sanchez, K. Supparatpinyo, S. Tripathy, A. I. Martinez, A. Nair, A. Walawander, L. Moran, Y. Chen, W. Snowden, J. F. Rooney, J. Uy, R. T. Schooley, V. D. Gruttola, and J. G. H. for the PEARLS study team of the ACTG (2012). Efficacy and safety of three antiretroviral regimens for initial treatment of HIV-1: a randomized clinical trial in diverse multinational settings. *PLoS medicine 9*(8), e1001290.

Fisher, A. and M. Rosenblum (2016). Stochastic optimization of adaptive enrichment designs for two subpopulations. *Johns Hopkins University, Dept. of Biostatistics Working Papers. http://biostats.bepress.com/jhubiostat/paper279*.

G. Lan, K. K. and D. L. DeMets (1983). Discrete sequential boundaries for clinical trials. *Biometrika 70*(3), 659–663.

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

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

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

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.

Kunzmann, K., L. Benner, and M. Kieser (2017). Point estimation in adaptive enrichment designs. *Statistics in Medicine*. http://dx.doi.org/10.1002/sim.7412.

Liu, A., C. Liu, Q. Li, K. F. Yu, and V. W. Yuan (2010). A threshold sample-enrichment approach in a clinical trial with heterogeneous subpopulations. *Clinical Trials 7*(5), 537–545. PMID: 20685769.

Liu, Q. and K. M. Anderson (2008). On adaptive extensions of group sequential trials for clinical investigations. *Journal of the American Statistical Association 103*(484), 1621–1630.

Lu, X. and A. A. Tsiatis (2011). Semiparametric estimation of treatment effect with time-lagged response in the presence of informative censoring. *Lifetime and Data Analysis 17*(4), 566–593.

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

Posch, M., F. Koenig, M. Branson, W. Brannath, C. Dunger-Baldauf, and P. Bauer (2005). Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine 24*(24), 3697–3714.

Rosenblum, M. (2013). Confidence intervals for the selected population in randomized trials that adapt the population enrolled. *Biometrical Journal 55*(3), 322–340.

Rosenblum, M., X. Fang, and H. Liu (2017). Optimal, two stage, adaptive enrichment designs for randomized trials using sparse linear programming. *Johns Hopkins University, Dept. of Biostatistics Working Papers. http://biostats.bepress.com/jhubiostat/paper273*.

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.

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

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

A Optimization Method

We describe the algorithm used for optimizing each design class. For each of the classes $\mathcal{D}_{ADAPTIVE,START-BOTH}$ and $\mathcal{D}_{ADAPTIVE,START-SUBPOP.1}$, there are multiple parameters to optimize over. We used the generic optimization algorithm of simulated annealing implemented in the R package optim. This is not guaranteed to find the global optimum solution; computing the global optimum for our problems is currently an open problem.

For a given value of K, the initial values of the design parameters (used to start the search conducted by simulated annealing) were set as follows:

1. the analysis time $t_k = 8k/K$ for each $k \leq K$.
2. the maximum enrollment time $t_{MAX-ENROLL} = 8$.
3. the alpha allocation $\{\alpha_{s,k} \geq 0 : s = 1, 2; k = 1, \ldots, K\}$ set all $\alpha_{s,k}$ to a common value such that they sum to 0.05 for $\mathcal{D}_{ADAPTIVE,START-BOTH}$. For $\mathcal{D}_{ADAPTIVE,START-SUBPOP.1}$, where we set $\alpha_{2,1} = 0$, all other $\alpha_{s,k}$ are set to a common value such that they sum to 0.05.
4. the futility boundaries $\{f_{s,k} \in \mathbb{R} : s = 1, 2; k = 1, \ldots, K-1\}$ are set to -3, except that for $\mathcal{D}_{ADAPTIVE,START-SUBPOP.1}$ we set $f_{2,1} = -\infty$.
5. for $\mathcal{D}_{ADAPTIVE,START-SUBPOP.1}$, we set $\theta = -3$.

For the class of standard (non-adaptive) designs $\mathcal{D}_{ONE-STAGE}$ from Section 5.3, there are only 2 parameters to optimize over: $t_{MAX-ENROLL}$ and the partition of α between the two null hypotheses. We applied simulated annealing to optimize over these 2 parameters. Another approach for these simple designs would be to use a binary search.

B Definitions of null hypotheses and statistics

For each subpopulation $s = 1, 2$ and treatment arm $a = A, B$, the corresponding hazard rate is denoted λ_{sa}. The inferiorty null hypothesis $H_{0s} : \lambda_{sA}/\lambda_{sB} \geq NI$ where NI denotes

the non-inferiority margin. This null hypothesis can equivalently be written as $\log(NI) - \log(\lambda_{sA}/\lambda_{sB}) \leq 0$.

Let $\hat{\beta}_{s,k}$ denote the estimated coefficient on the indicator of assignment to arm A in a proportional hazards model with intercept and a main term for this indicator, based on all data at analysis k for subpopulation s. Under the proportional hazards assumption and appropriate regularity conditions, $\hat{\beta}_{s,k}$ is a consistent, asymptotically normal estimator for $\log(\lambda_{sA}/\lambda_{sB})$. The Wald statistic for subpopulation s at interim analysis k is defined as

$$Z_{s,k} = \frac{\log(NI) - \hat{\beta}_{s,k}}{\sqrt{Var(\hat{\beta}_{s,k})}}. \tag{2}$$

For each subpopulation s, the vector of statistics $Z_{s,k}$ has, asymptotically, the canonical joint distribution described in (Jennison and Turnbull, 1999, Ch. 3.1). That is, asymptotically, the vector $(Z_{s,1}, \ldots, Z_{s,k})$ converges in distribution to a multivariate normal with mean $(\log(NI) - \log(\lambda_{sA}/\lambda_{sB}))\sqrt{\mathcal{I}_{s,k}}$ and covariance $Cov(Z_{s,k}, Z_{s,k'}) = \sqrt{\mathcal{I}_{s,k}\mathcal{I}_{s,k'}^{-1}}$ with $k \leq k'$, where $\mathcal{I}_{s,k}$ is the information for subpopulation s at analysis k. As the two subpopulations are disjoint, we have $Cov(Z_{1,k}, Z_{2,k'}) = 0$ for all k, k'. This fully specifies the asymptotic distribution of the test statistics. When the number of participants at risk is approximately equal in each treatment arm at each time, the information $\mathcal{I}_{s,k}$ can be approximated by 1/4 the number of events in subpopulation s at analysis k; we use this approximation throughout. Below, we give formulas calculating the expected number of events.

C Calculating the Expected Number of Events

As stated in Section 4.1, we assume that the failure time distribution in each subpopulation $s = 1, 2$ and treatment arm $a = A, B$ is exponential with mean $\lambda_{sa}^{-1} > 0$. The corresponding hazard rate is given by λ_{sa}. Hence, the log hazard ratio for subpopulation s is $\text{HR}_s = \log(\lambda_{sA}/\lambda_{sB})$ for each $s \in \{1, 2\}$.

For simplicity of presentation, we consider one subpopulation at a time and suppress the subpopulation indicator. Let τ denote the maximum follow-up time for a participant; in the main paper we assumed that τ equals the study duration (so that each enrolled participant who does not experience the failure event is followed until the end of the study); however, we present our results in greater generality below where it is possible to set τ less than the study duration, e.g., each participant is followed for 1 year after her/his enrollment. The expected number of events is the expected number of participants enrolled before time t multiplied by the probability of experiencing an event before time t. The expected number of participants enrolled before time t is calculated as the accrual rate multiplied by $\min(t, c)$.

We next derive formulas for the probability of an event occurring before time t for a randomly selected participant who is enrolled before time t, drawn from a hypothetical population with constant hazard rate $\lambda > 0$. These formulas can then be used to determine the expected number of events for each arm by subpopulation combination at each analysis

time t_k, which determines the corresponding information $\mathcal{I}_{s,k}$ used in the previous section (for computing the asymptotic means and covariances of the statistics $Z_{s,k}$).

We consider six different cases.

- **Assume that** $c > t > \tau$**.** Let R denote the enrollment time, then by the uniform enrollment assumption $R \sim \text{unif}[0, t]$. As each individual is followed for a maximum of τ time units the follow-up time for the individual at time t is given by $\min(\tau, t - R)$. For a participant that has been followed up for τ time units at time t the probability of experiencing an event is $1-e^{-\lambda\tau}$. For a participant that has been in the study less than τ time-units the length in the study, denoted by K, follows an uniform distribution on $[0, \tau]$. We have that

$$\begin{aligned} P(T < K) &= \int_0^\tau \frac{1}{\tau}(1 - e^{-\lambda x})dx \\ &= 1 - \frac{(1 - e^{-\lambda\tau})}{\lambda\tau} \end{aligned}$$

  Therefore we have for an interim analysis at time t that the probability of an event is given by

$$\begin{aligned} P(\text{event}) &= P(t - R > \tau)(1 - e^{-\lambda\tau}) + P(t - R \leq \tau)\left(1 - \frac{(1 - e^{-\lambda\tau})}{\lambda\tau}\right) \\ &= \frac{t - \tau}{t}(1 - e^{-\lambda\tau}) + \frac{\tau}{t}\left(1 - \frac{(1 - e^{-\lambda\tau})}{\lambda\tau}\right) \end{aligned}$$

- **Assume that** $c > \tau > t$. Here the analysis is performed before the maximum follow-up. Hence by previous calculations,

$$P(\text{event}) = 1 - \frac{(1 - e^{-\lambda t})}{\lambda t}.$$

- **Assume that** $t \geq c > \tau$. Define k as the number of time-units after end of enrollment that the analysis takes place, that is, $k = t - c$. By the assumptions made $0 < k \leq \tau$. For participants enrolled in the interval $[0, c+k-\tau]$ the follow up time is τ time-units. The probability of being followed up for τ time-units is by the uniform enrollment assumption $\frac{c+k-\tau}{c}$. For participants enrolled for τ time-units $P(\text{event}) = 1 - e^{-\lambda\tau}$.

  For participants enrolled in the time-interval $[c + k - \tau, c]$ the follow-up time follows an uniform distribution on $[k, \tau]$. Therefore for participants in that group

$$P(\text{event}) = 1 - \frac{e^{-\lambda k} - e^{-\lambda\tau}}{(\tau - k)\lambda}.$$

  Combining the above gives

$$P(\text{event}) = \frac{c + k - \tau}{c}(1 - e^{-\lambda\tau}) + \frac{\tau - k}{c}\left(1 - \frac{e^{-\lambda k} - e^{-\lambda\tau}}{(\tau - k)\lambda}\right)$$

- **Assume that** $\tau \geq c > t$. All participants have been in the study for less than τ units. Hence, as before we have

$$P(\text{event}) = 1 - \frac{(1 - e^{-\lambda t})}{\lambda t}.$$

- **Assume that** $\tau > t \geq c$. Follow up time is uniform on $[t - c, t]$. Similar calculations to before give that

$$P(\text{event}) = 1 - \frac{e^{-\lambda(t-c)} - e^{-\lambda t}}{c\lambda}.$$

- **Assume that** $\tau + c \geq t \geq \tau > c$. For participants enrolled in $[0, t - \tau]$ the time in study is τ time-units so $P(\text{event}) = 1 - e^{-\lambda\tau}$. The probability of having follow-up time of τ years is $(t - \tau)/c$. For other participants the follow-up time is uniform on the interval $[t - c, \tau]$, and the probability of having follow-up time in that time-interval is $(c - t + \tau)/c$. Similar calculations to before give that for participants enrolled in the interval $]t - \tau, c]$ we have

$$P(\text{event}) = 1 - \frac{e^{-\lambda(t-c)} - e^{-\lambda\tau}}{(\tau - t + c)\lambda}.$$

  Combining this gives

$$P(\text{event}) = \frac{t - \tau}{c}(1 - e^{-\lambda\tau}) + \frac{c - t + \tau}{c}\left(1 - \frac{e^{-\lambda(t-c)} - e^{-\lambda\tau}}{(\tau - t + c)\lambda}\right).$$