5-4-2012

CONFIDENCE INTERVALS FOR THE SELECTED POPULATION IN RANDOMIZED TRIALS THAT ADAPT THE POPULATION ENROLLED

Michael Rosenblum
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, mrosenbl@jhsph.edu

Confidence intervals for the selected population in randomized trials that adapt the population enrolled

Michael Rosenblum[*]
Department of Biostatistics, Room E3616
Johns Hopkins Bloomberg School of Public Health
Baltimore, MD, 21205

April 29, 2012

Abstract

It is a challenge to design randomized trials when it is suspected that a treatment may benefit only certain subsets of the target population. In such situations, trial designs have been proposed that modify the population enrolled based on an interim analysis, in a preplanned manner. For example, if there is early evidence that the treatment only benefits a certain subset of the population, enrollment may then be restricted to this subset. At the end of such a trial, it is desirable to draw inferences about the selected population. We focus on constructing confidence intervals for the average treatment effect in the selected population. Confidence interval methods that fail to account for the adaptive nature of the design may fail to have the desired coverage probability. We provide a new procedure for constructing confidence intervals having at least 95% coverage probability, uniformly over a large class of possible data generating distributions. We prove an optimality property for our confidence interval procedure in terms of minimizing the average confidence interval widths.

1 Introduction

We consider the problem of constructing confidence intervals in randomized trial designs that involve a preplanned rule for changing enrollment criteria based on an interim analysis. Such trial designs have been proposed in situations where a baseline risk factor is conjectured to be predictive of treatment benefit (Follmann, 1997; Russek-Cohen and Simon, 1997; Wang et al., 2007; Jennison and Turnbull, 2007; Wang et al., 2009). As an example of such a predictor, in trials of trastuzumab for treating metastatic breast cancer, the level of overexpression of human epidermal growth factor receptor-2 (HER2) has been shown to be predictive of treatment benefit (Slamon et al., 2001).

[*] Author: e-mail: mrosenbl@jhsph.edu

As another example, Kirsch et al. (2008) present suggestive evidence that for a certain class of antidepressants, a patient's initial severity of depression may be predictive of treatment benefit.

For concreteness, we focus on two-stage, randomized trial designs of a single treatment versus control, where the overall population is partitioned into two, prespecified subpopulations. In the designs we consider, a decision is made at an interim analysis to possibly restrict the enrollment criteria for the second stage, based on data in the first stage; this is called an enrichment design. The decision rule in such designs must be specified before the study starts. We allow the population enrolled in stage two to be subpopulation 1, subpopulation 2, or the combined population. We do not allow changes to the randomization probabilities, total sample size, or number of treatments. However, it may be possible to apply our general method for constructing confidence intervals to some designs with these additional types of adaptation, as we discuss in Section 6.

We focus on the problem of constructing confidence intervals for the average treatment effect in the population selected for enrollment in stage two. Standard confidence interval procedures that ignore the adaptive nature of the design may fail to have the desired coverage probability. The main contributions of this paper are: demonstrating a general method for constructing confidence intervals that have at least 95% coverage probability, uniformly over a large class of possible data generating distributions, for this problem; and, proving these confidence intervals have an optimality property in terms of their average width. To the best of our knowledge, our confidence interval procedure is the first to have this optimality property, for the enrichment designs we consider.

The confidence intervals we present are centered at the difference $\hat{\Delta}_S$ in sample means between treatment and control arms for the selected population, using all data from both stages from that population. We compute the minimum factor c by which the standard confidence interval centered at $\hat{\Delta}_S$ must be expanded in order to have, asymptotically, at least 95% coverage probability, uniformly over a large class of data generating distributions. Computing this constant is not trivial, since it is not a priori clear, for a given decision rule, what the least favorable data generating distribution is, i.e., which distribution requires the largest constant c in order for the corresponding confidence interval procedure to have coverage probability at least 95%. We show how to compute the least favorable distribution and the corresponding minimum factor c.

For the enrichment designs we consider, the ratio of average width of our confidence intervals, compared to naive confidence intervals that ignore the adaptive nature of the design, is never more than 1.1. Thus, for the adaptive designs we consider, at most a 10% inflation of the standard confidence interval width is required in order to ensure at least 95% coverage probability. However, in many cases that we expect to occur in practice, the required inflation is at most 5%, as we describe in Section 5.

Our results are asymptotic, as the sample sizes in both stages of the design go to infinity. However, our confidence interval coverage is at least 95% at all sample sizes in the special case that the outcome is normally distributed.

In Section 2, we present related work. We describe the setup of our problem and the type of adaptive designs we consider, in Section 3. We then give a confidence interval procedure that has uniform coverage probability at least 95% for these designs, and that has an optimality property, in Section 4. The distribution of confidence interval widths from our procedure is examined in Section 5. We discuss limitations and directions for future research in Section 6.

2 Related Work

We focus on two-stage designs that allow changes to a trial's enrollment criteria after an interim analysis; our goal is to construct a confidence interval for the average treatment effect in the population selected for enrollment in stage two. We now describe related work.

Designs that make changes to enrollment criteria based on preplanned rules include those of, e.g., (Follmann, 1997; Russek-Cohen and Simon, 1997; Wang et al., 2007; Jennison and Turnbull, 2007; Wang et al., 2009; Rosenblum and van der Laan, 2011). However, the aforementioned papers focus on hypothesis testing rather than on confidence intervals. It is not clear how to invert the multiple testing procedures from these methods to construct a valid confidence interval for the selected population, due to the multiple hypotheses involved as well as the adaptive nature of the design.

Repeated confidence intervals for a single population, under various rules for modifying a trial's total sample size, have been constructed, e.g., by (Jennison and Turnbull, 1984; Lehmacher and Wassmer, 1999; Brannath et al., 2006). Posch et al. (2005) give simultaneous confidence intervals for adaptive designs with multiple treatments. That is, they give a simultaneous set of confidence intervals for all the treatments. Analogous ideas could be applied to our setting of adaptive designs with multiple subpopulations. However, the focus of our paper is constructing confidence intervals with minimum width for the selected population, rather than simultaneous confidence intervals for all subpopulations considered.

Sampson and Sill (2005) and Wu et al. (2010) consider drop the loser designs, that is, designs in which the treatment with the largest estimated treatment effect at the end of stage one is selected for continued study in stage two. They provide confidence intervals for the selected treatment that have conservative coverage probability for such designs. These methods could be extended to trial designs that, instead of continuing the best performing treatment, continue the population with the largest estimated treatment effect. However, in the context of designs that adapt the population enrolled, it may not be ideal to always continue the population with largest estimated treatment effect. For example, if the overall population had a large estimated benefit, but a small subpopulation had a slightly larger estimated benefit, it may be unwise to give up on the overall population and only continue enrollment from the small subpopulation. Therefore, drop the loser designs may not be ideal for changing enrollment criteria. The designs we consider below involve decision rules that are tailored to population selection, and that allow continued enrollment of the combined population even when it doesn't have the largest estimated treatment benefit.

3 Problem setup

3.1 Assumptions on data generating distribution

For each subject i, we collect the following vector of data: (S_i, W_i, A_i, Y_i), where S_i is the subpopulation (1 or 2), W_i is the stage of the trial in which the subject is enrolled (1 or 2), A_i is the study arm assignment (1 indicating the treatment arm and 0 indicating the control arm), and Y_i is the outcome. We allow the outcome variable Y to be discrete or continuous valued.

The definition of the subpopulations must be a prespecified function of variables measured prior to randomization. We assume the two subpopulations are disjoint, and together make up the combined population. For example, subpopulation 1 could be defined as those having a certain biomarker positive at baseline, and subpopulation 2 would then be the biomarker negative population. For each $s \in \{1, 2\}$, let π_s denote the proportion of the overall population in subpopulation s. We assume that in stage one, the proportion of subjects enrolled in the study from each subpopulation $s \in \{1, 2\}$ is the same as the corresponding population proportion π_s. We denote the total sample size in stage j by n_j, for $j \in \{1, 2\}$; these are fixed at the beginning of the study. We assume π_1, π_2, n_1, n_2 are known, non-random quantities. We also assume that neither stage completely dominates the total sample size, in that we assume the fraction of the sample in stage one, $t_1 = n_1/(n_1 + n_2)$, is between 0.01 and 0.99.

We assume that each subject is enrolled with probability $1/2$ to treatment or control, independent of the subpopulation S_i and stage W_i. For simplicity in what follows, we assume that for each subpopulation and each stage, exactly half the subjects are assigned to the treatment arm ($A_i = 1$), and half to the control arm ($A_i = 0$). This can be approximately guaranteed by using stratified block randomization.

We denote the unknown outcome distribution for each subpopulation $s \in \{1, 2\}$ and study arm $a \in \{0, 1\}$ by Q_{sa}, We assume conditioned on the set of patient subpopulations $\{S_i\}$, study arm assignments $\{A_i\}$, and trial stages at enrollment $\{W_i\}$, that the outcome Y_i for each subject i is an independent draw from the unknown outcome distribution $Q_{S_i A_i}$.

We denote subpopulation 1, subpopulation 2, and the combined population by the subscripts $1, 2$, and $*$, respectively. Denote the mean outcome for subpopulation $s \in \{1, 2\}$ under assignment to arm $a \in \{0, 1\}$ by $\mu(Q_{sa})$, and denote the corresponding variance by $\sigma^2(Q_{sa})$. We assume the variances $\sigma^2(Q_{sa})$ are known.

We make no assumptions on the forms of the outcome distributions Q_{sa} except that their support is contained in an interval $[-M, M]$, for some $M > 0$, and that the variance of each Q_{sa} is at least a (small) constant $\tau > 0$. In particular, the means, variances, and other features of these distributions may differ across treatment arms and subpopulations. We fix $M > 0, \tau > 0$, and define $\mathcal{Q}$ to be the class of data generating distributions $Q = (Q_{10}, Q_{11}, Q_{20}, Q_{21})$ for which each Q_{sa} has support contained in the interval $[-M, M]$, and the variance of each Q_{sa} is at least $\tau > 0$. We assume each subject's outcome Y_i is measured relatively quickly after enrollment, so that all outcomes in stage one can be used to determine the enrollment criteria in stage two.

3.2 Definition of average treatment effects

For each subpopulation $s \in \{1, 2\}$, define the average treatment effect for subpopulation s, on the risk difference scale, by $\Delta_s(Q) = \mu(Q_{s1}) - \mu(Q_{s0})$. Similarly, for the combined population, define the average treatment effect on the risk difference scale by $\Delta_*(Q) = \pi_1 \Delta_1(Q) + (1 - \pi_1)\Delta_2(Q)$. For clarity of notation, we sometimes suppress dependence on Q, e.g., writing Δ_* instead of $\Delta_*(Q)$.

Let S denote the population selected to be enrolled in stage two. $S = 1$ indicates population 1 is enrolled in stage two, $S = 2$ indicates subpopulation 2 is enrolled in stage two, and $S = *$ indicates both subpopulations are enrolled in stage two in the same proportions as in stage 1. The total number of subjects enrolled in stage two is set at n_2, regardless of which population is selected to be enrolled in stage two. Below, we describe the set of decision rules for determining S as a function of stage one data, which will use the statistics defined next.

3.3 Statistics used in decision rule and confidence interval procedure

For each subpopulation $s \in \{1,2\}$ and stage $w \in \{1,2\}$, we denote the difference between the sample means under treatment and under control by

$$\hat{\Delta}_s^{(w)} = \left(\frac{\sum_{\{i:S_i=s,W_i=w,A_i=1\}} Y_i}{|\{i : S_i = s, W_i = w, A_i = 1\}|} \right) - \left(\frac{\sum_{\{i:S_i=s,W_i=w,A_i=0\}} Y_i}{|\{i : S_i = s, W_i = w, A_i = 0\}|} \right),$$

where $|B|$ denotes the number of elements in the set B. Similarly, for the combined population, in stage $w \in \{1,2\}$, we denote the difference in the sample means under treatment and under control by

$$\hat{\Delta}_*^{(w)} = \left(\frac{\sum_{\{i:W_i=w,A_i=1\}} Y_i}{|\{i : W_i = w, A_i = 1\}|} \right) - \left(\frac{\sum_{\{i:W_i=w,A_i=0\}} Y_i}{|\{i : W_i = w, A_i = 0\}|} \right) = \pi_1 \hat{\Delta}_1^{(w)} + (1-\pi_1)\hat{\Delta}_2^{(w)}.$$

For the population S selected for enrollment in stage two, let $\hat{\Delta}_S$ denote the difference in sample means between treatment and control arms, pooling all subjects in both stages of the trial in population S. On the event $S = *$, that is, if the combined population is enrolled in stage two, then $\hat{\Delta}_S$ equals

$$\left(\frac{\sum_{\{i:A_i=1\}} Y_i}{|\{i : A_i = 1\}|} \right) - \left(\frac{\sum_{\{i:A_i=0\}} Y_i}{|\{i : A_i = 0\}|} \right) = \frac{n_1 \hat{\Delta}_*^{(1)} + n_2 \hat{\Delta}_*^{(2)}}{n_1 + n_2},$$

while if enrollment in stage two consists of n_2 subjects from only one of the subpopulations, i.e., if $S = s \in \{1,2\}$, then $\hat{\Delta}_S$ equals

$$\left(\frac{\sum_{\{i:S_i=s,A_i=1\}} Y_i}{|\{i : S_i = s, A_i = 1\}|} \right) - \left(\frac{\sum_{\{i:S_i=s,A_i=0\}} Y_i}{|\{i : S_i = s, A_i = 0\}|} \right) = \frac{\pi_s n_1 \hat{\Delta}_s^{(1)} + n_2 \hat{\Delta}_s^{(2)}}{\pi_s n_1 + n_2}.$$

We define the stage one z-statistics for subpopulation 1, subpopulation 2, and the combined population, respectively, as:

$$Z_1^{(1)} = \frac{\sqrt{\pi_1 n_1}}{\sigma_1(Q)} \hat{\Delta}_1^{(1)}, \qquad Z_2^{(1)} = \frac{\sqrt{\pi_2 n_1}}{\sigma_2(Q)} \hat{\Delta}_2^{(1)}, \qquad Z_*^{(1)} = \frac{\sqrt{n_1}}{\sigma_*(Q)} \hat{\Delta}_*^{(1)},$$

where for each $s \in \{1,2\}$, $\sigma_s^2(Q) = 2\{\sigma^2(Q_{s0}) + \sigma^2(Q_{s1})\}$, and $\sigma_*^2(Q) = \pi_1 \sigma_1^2(Q) + (1 - \pi_1)\sigma_2^2(Q)$.

Let ρ_s denote the covariance between $Z_s^{(1)}$ and $Z_*^{(1)}$ for each $s \in \{1,2\}$. We then have

$$\rho_s = \sqrt{\pi_s}\sigma_s(Q)/\sigma_*(Q), \ \rho_1^2 + \rho_2^2 = 1, \text{ and } Z_*^{(1)} = \rho_1 Z_1^{(1)} + \rho_2 Z_2^{(1)}.$$

3.4 Set of decision rules

For concreteness, throughout this paper we focus on the following type of decision rule for stage two enrollment, for a prespecified constant d:

> If $Z_*^{(1)} > d$, in stage two, continue enrolling from the combined population, in the same proportions as in stage one. Else, enroll exclusively from the subpopulation corresponding to the larger of $Z_1^{(1)}, Z_2^{(1)}$ (with ties broken arbitrarily, e.g., selecting subpopulation 1).

The general method we give below can be applied to construct confidence intervals for other types of rules than above, which we discuss in Section 6.

4 Confidence interval procedure with uniform coverage probability

4.1 Goal of confidence interval procedure

The goal in this paper is to provide a procedure CI for constructing confidence intervals with at least 95% coverage for the treatment effect in the population selected to be enrolled in stage two of the trial. The treatment effect for the population selected to be enrolled in stage two is Δ_S. This is a random quantity, since it depends on the selected population S. The confidence interval procedure can use the following as input: all the data from both stages of the trial, π_1, π_2, n_1, n_2, and the outcome variances $\sigma^2(Q_{sa})$ for each subpopulation $s \in \{1, 2\}$ and study arm $a \in \{0, 1\}$, since the variances $\sigma^2(Q_{sa})$ are assumed known.

We seek a function CI from the set of data in both stages of the trial to an interval in $\mathbb{R}$ that, asymptotically as the sample sizes in both stages go to infinity, has at least 95% coverage probability, uniformly over all possible data generating distributions $\mathcal{Q}$. Formally, we require our procedure CI to satisfy, for fixed decision rule threshold d and subpopulation proportions π_1, π_2:

$$\liminf_{n_1, n_2 \to \infty} \inf_{Q \in \mathcal{Q}} P_{Q,n_1,n_2}(\Delta_S \in CI) \geq 0.95, \tag{1}$$

where P_{Q,n_1,n_2} is the probability distribution defined by data generating distribution Q and stage one and two sample sizes n_1 and n_2, respectively.

Though the above criterion is asymptotic, our confidence interval procedure has at least 95% coverage probability at any sample size, in the special case in which the outcome distributions Q_{sa} are each normal distributions. We note that standard, normal-based confidence intervals for fixed (non-adaptive) designs have at least 95% coverage only asymptotically, or in special cases such as when outcomes are normally distributed.

Our goal is to construct a confidence interval procedure CI satisfying (1). The probability in (1) is a type of average coverage probability, in that we can write it as a convex combination of coverage probabilities conditioned on each possible enrollment decision:

$$P_{Q,n_1,n_2}(\Delta_S \in CI) = \sum_{s \in \{1,2,*\}} P_{Q,n_1,n_2}(S = s) P_{Q,n_1,n_2}(\Delta_s \in CI \mid S = s).$$

Alternatively, one may wish to have the stronger guarantee of asymptotically conservative *conditional* coverage probability:

$$\liminf_{n_1,n_2\to\infty} \inf_{Q\in\mathcal{Q}} \inf_{s\in\{1,2,*\}} P_{Q,n_1,n_2}(\Delta_s \in CI \mid S = s) \geq 0.95. \tag{2}$$

We focus on constructing confidence intervals that satisfy (1). However, our general method can be used to determine how much one would need to expand these confidence intervals to meet the stricter condition (2).

4.2 Confidence interval procedure satisfying (1)

Let N_S denote the total sample size, at the end of the trial, for the selected population S. This is a random variable that equals $n_1 + n_2$ if the combined population is selected, else it equals $\pi_s n_1 + n_2$ if subpopulation $s \in \{1, 2\}$ is selected.

We define the naive 95% confidence interval that ignores the adaptive nature of the design as

$$\left[\hat{\Delta}_S - z_{0.975}\sigma_S/\sqrt{N_S}, \hat{\Delta}_S + z_{0.975}\sigma_S/\sqrt{N_S}\right], \tag{3}$$

for $z_{0.975}$ the 0.975 quantile of the standard normal distribution. In general this will fail to have 95% coverage probability.

Our modified confidence interval, CI, is defined as

$$CI = \left[\hat{\Delta}_S - c(\pi_1, t_1, \rho_1)z_{0.975}\sigma_S/\sqrt{N_S}, \hat{\Delta}_S + c(\pi_1, t_1, \rho_1)z_{0.975}\sigma_S/\sqrt{N_S}\right], \tag{4}$$

for $c(\pi_1, t_1, \rho_1)$ the solution to an optimization problem we define in (6) below. We provide a short program in R to compute an approximation to $c(\pi_1, t_1, \rho_1)$ in the Supplementary Materials, and we explore the features of the function c in Section 5. The value $c(\pi_1, t_1, \rho_1)$ can be thought of as the smallest multiplicative factor by which the naive confidence interval (3) that ignores the adaptive nature of the design needs to be expanded, in order to maintain at least 95% coverage probability, uniformly over $\mathcal{Q}$, as sample size goes to infinity.

We prove in the Appendix that the confidence interval CI has the desired property (1), that is, it has asymptotic coverage probability at least 95%, uniformly over the class of data generating distributions $\mathcal{Q}$. Furthermore, we prove that our expansion factor c is smallest possible, at every value of (π_1, t_1, ρ_1). More precisely, we prove that for any continuous function c' that maps (π_1, t_1, ρ_1) into $\mathbb{R}$, if the confidence interval CI in (4), except replacing c by c', has property (1), then $c'(\pi_1, t_1, \rho_1) \geq c(\pi_1, t_1, \rho_1)$ at every (π_1, t_1, ρ_1).

4.3 Construction of expansion factor c

We first give the intuition behind our construction, given below, of the minimum expansion factor $c(\pi_1, t_1, \rho_1)$ that guarantees asymptotic, uniform coverage probability at least 95%, as expressed in (1). Fix π_1, t_1, ρ_1, and consider a candidate value for the expansion factor $c(\pi_1, t_1, \rho_1)$. Consider large sample sizes n_1, n_2, and any data generating distribution $Q \in \mathcal{Q}$ for which $\rho_1(Q) = \rho_1$. The coverage probability of CI can be decomposed as:

$$P_Q\left(\Delta_S \in CI\right) = \sum_{s \in \{1,2,*\}} P_Q(\Delta_S \in CI, S = s). \tag{5}$$

We show in the Appendix that each term $P_Q(\Delta_S \in CI, S = s)$ in this sum is approximately the probability that a multivariate normal distribution G_s is in a certain rectangular region R_s; furthermore, the mean and covariance of G_s and the boundaries of the rectangular region depend on π_1, t_1, ρ_1, c, d, and depend on Q only through the non-centrality parameters $\delta_{s'} = EZ_{s'}^{(1)} = \sqrt{\pi_{s'} n_1}\Delta_{s'}/\sigma_{s'}$, for $s' \in \{1, 2\}$. We choose the expansion factor $c(\pi_1, t_1, \rho_1)$ to be the smallest value such that no matter what the values of the non-centrality parameters δ_1, δ_2, the sum of the multivariate normal probabilities $\sum_{s\in\{1,2,*\}} P(G_s \in R_s)$ is at least 0.95.

We now formally define the function c. Fix the threshold d used in the decision rule from Section 3.4. For each $s \in \{1, 2, *\}$, let $G_s(\pi_1, t_1, \rho_1)$ denote the multivariate normal random vector in $\mathbb{R}^3$ with mean vector $(0, 0, 0)$ and covariance matrix the function of (π_1, t_1, ρ_1) given in (9)-(11) in the Appendix. For any $\tilde{c} \in \mathbb{R}$, define the following disjoint, rectangular regions of $\mathbb{R}^3$:

$$\begin{aligned}
R_1(\rho_1, \tilde{c}, d, \delta_1, \delta_2) &= [-\tilde{c}z_{0.975}, \tilde{c}z_{0.975}] \times (-\infty, d - (\rho_1\delta_1 + \rho_2\delta_2)] \times [(\delta_2 - \delta_1)/\sqrt{2}, \infty). \\
R_2(\rho_1, \tilde{c}, d, \delta_1, \delta_2) &= [-\tilde{c}z_{0.975}, \tilde{c}z_{0.975}] \times (-\infty, d - (\rho_1\delta_1 + \rho_2\delta_2)] \times (-\infty, (\delta_2 - \delta_1)/\sqrt{2}). \\
R_*(\rho_1, \tilde{c}, d, \delta_1, \delta_2) &= [-\tilde{c}z_{0.975}, \tilde{c}z_{0.975}] \times (d - (\rho_1\delta_1 + \rho_2\delta_2), \infty) \times \mathbb{R}.
\end{aligned}$$

Define

$$c(\pi_1, t_1, \rho_1) = \inf\left\{\tilde{c} > 0 : \inf_{\tilde{\delta}_1, \tilde{\delta}_2 \in \mathbb{R}} \sum_{s\in\{1,2,*\}} P\left[G_s(\pi_1, t_1, \rho_1) \in R_s(\rho_1, \tilde{c}, d, \tilde{\delta}_1, \tilde{\delta}_2)\right] \geq 0.95\right\}. \tag{6}$$

In Appendix A.3, we explain how to solve the above non-linear optimization problem, and therefore compute $c(\pi_1, t_1, \rho_1)$, to any desired accuracy, in the following sense: for any π_1, t_1, ρ_1, and any tolerance $\epsilon > 0$, we show how to compute a value c_ϵ satisfying the following two conditions:

i. $c_\epsilon \geq c(\pi_1, t_1, \rho_1)$, which implies the confidence interval procedure CI using c_ϵ satisfies the asymptotic, uniform coverage probability condition (1);

ii. c_ϵ is no larger than ϵ plus the expansion threshold required to obtain uniform coverage probability at least $0.95 + \epsilon$; that is, c_ϵ is no greater than ϵ plus (6) with 0.95 replaced by $0.95 + \epsilon$.

We give a brief overview of the main ideas used in Appendix A.3 to compute $c(\pi_1, t_1, \rho_1)$ to any desired accuracy, in the above sense. For any given vector (π_1, t_1, ρ_1), any candidate value $\tilde{c}$, and any $\tilde{\delta}_1, \tilde{\delta}_2$, one can compute the sum in the right hand side of (6) using statistical software such as the `mvtnorm` package in `R`. To compute $\inf_{\tilde{\delta}_1, \tilde{\delta}_2 \in \mathbb{R}}$ of this sum to any desired accuracy, one can compute this sum at each pair $(\tilde{\delta}_1, \tilde{\delta}_2)$ in a sufficiently fine grid of values, as described in

Appendix A.3. One can then use binary search over candidate values of $\tilde{c}$ to determine the smallest value for which the minimum coverage probability is at least 95%, yielding an approximation to $c(\pi_1, t_1, \rho_1)$.

It turns out that the right hand side of (6) does not depend on d, the threshold used in the decision rule from Section 3.4. This justifies our writing c as a function only of the variables π_1, t_1, ρ_1. This is advantageous in that once $c(\pi_1, t_1, \rho_1)$ is computed for a given d, it is unnecessary to recompute it for any other value of d. We also learn that there is no advantage in adjusting d in the hope of obtaining shorter confidence intervals with uniform coverage probability (assuming we are using the general type of confidence interval in (4)). The reason for no dependence of the right hand side of (6) on d is that for fixed $\rho_1, \tilde{c}, d$, the set of triples of regions generated by varying $\tilde{\delta}_1, \tilde{\delta}_2$:

$$\left\{\Big(R_1(\rho_1, \tilde{c}, d, \tilde{\delta}_1, \tilde{\delta}_2), R_2(\rho_1, \tilde{c}, d, \tilde{\delta}_1, \tilde{\delta}_2), R_3(\rho_1, \tilde{c}, d, \tilde{\delta}_1, \tilde{\delta}_2)\Big)\right\}_{\tilde{\delta}_1, \tilde{\delta}_2 \in \mathbb{R}}$$

does not depend on d. This can be seen since for any d, d', the triple of regions corresponding to $(\rho_1, \tilde{c}, d, \tilde{\delta}_1, \tilde{\delta}_2)$ is identical to that corresponding to $\Big(\rho_1, \tilde{c}, d', \tilde{\delta}_1 + (d' - d)/(\rho_1 + \rho_2), \tilde{\delta}_2 + (d' - d)/(\rho_1 + \rho_2)\Big)$.

5 Confidence interval widths

The function c defined in (6) represents the multiplicative factor by which the naive confidence interval (3) must be expanded in order to achieve at least 95% coverage, uniformly over the class of possible data generating distributions $\mathcal{Q}$. When $c = 1$, this indicates no change is needed to the naive confidence interval that ignores the adaptive nature of the design. In general, the width of the naive confidence interval must be increased by $100(c-1)\%$ to maintain at least 95% coverage, uniformly over $\mathcal{Q}$. We note that the value of the expansion factor $c(\pi_1, t_1, \rho_1)$ is always at least 1, which we prove in Appendix A.1.

The arguments to the function c are π_1, t_1, ρ_1, which represent the proportion of the population in subpopulation 1, the proportion of the sample in stage one, and the covariance of the first stage statistics $Z_1^{(1)}, Z_*^{(1)}$ (which is a function of π_1 and the variances $\sigma_1^2(Q), \sigma_2^2(Q)$ of the outcome for each subpopulation), respectively. At different values of these arguments, the required expansion factor $c(\pi_1, t_1, \rho_1)$ can differ.

In Figure 1, we plot the value of the expansion factor $c(\pi_1, t_1, \rho_1)$, showing how it varies as we change each of π_1, t_1, and ρ_1. We initialized the values of (π_1, t_1, ρ_1) to $(1/2, 1/2, 1/\sqrt{2})$, respectively, which corresponds to equal subpopulation sizes, equal sample sizes in each stage of the trial, and equal variances in the outcome distribution for each subpopulation. We then changed one variable at a time, and looked at the effect on $c(\pi_1, t_1, \rho_1)$. In each plot in Figure 1, we computed c at 20 points with equally spaced values on the horizontal axis ranging from 0.01 to 0.99.

In Figure 1a, we vary π_1, the proportion of the population in subpopulation 1. The minimum value of c is 1.03 (here and below, we round all values to two decimal places), which occurs at $\pi_1 = 1/2$; c increases as π_1 moves farther from $1/2$. A similar pattern occurs in Figure 1c, where

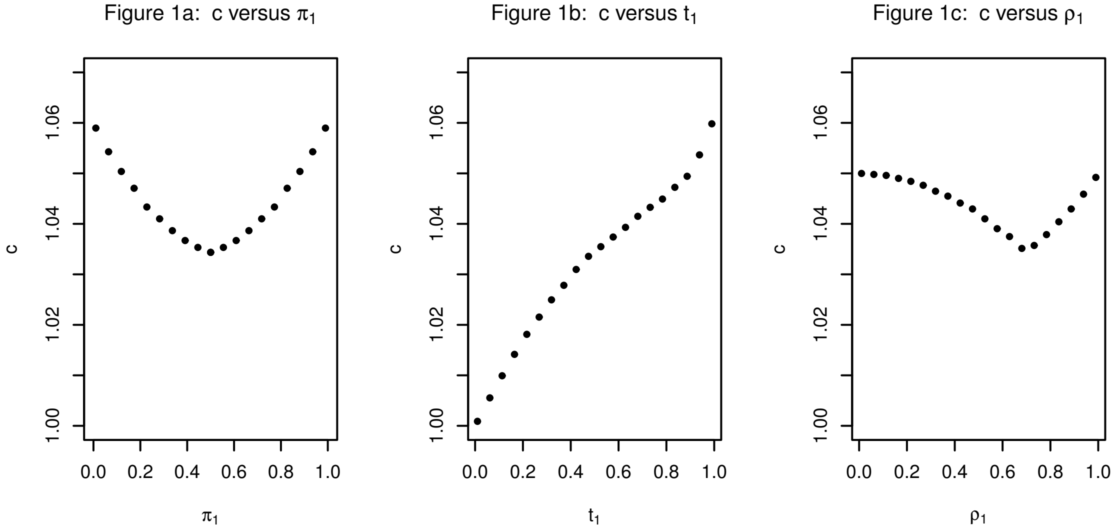

Figure 1: Plots of the expansion factor c versus the proportion of the population in subpopulation 1 (π_1), the proportion of the sample in stage one (t_1), and the covariance (ρ_1) of the first stage statistics $Z_1^{(1)}, Z_*^{(1)}$, respectively. In each plot, the vector (π_1, t_1, ρ_1) is initialized to $(1/2, 1/2, 1/\sqrt{2})$, and then one component of the vector (π_1, t_1, ρ_1) is varied while the other two are held fixed.

we vary ρ_1, the covariance of the first stage statistics $Z_1^{(1)}, Z_*^{(1)}$. The minimum value of c is 1.04, which occurs at $\rho_1 = 1/\sqrt{2}$, and c increases as ρ_1 moves farther from $1/\sqrt{2}$.

In Figure 1b, we varied t_1, the proportion of the sample in stage one. As this proportion increases from 0.01 to 0.99, the expansion factor c increases from 1 to approximately 1.06. Intuitively, the limit as $t_1 \downarrow 0$ corresponds to the limit design in which the stage two population is selected instantaneously after the trial starts, i.e., before any data is collected; therefore, the naive confidence interval (3) has the desired coverage probability since the population is selected independent of the data, and so no expansion is required. Also, the limit as $t_1 \uparrow 1$ corresponds to the limit design in which the population of interest is selected after the trial is over (but based on the trial data); this is a fixed design with data-dependent parameter selection, and requires inflation of the standard confidence interval width to obtain coverage probability at least 95%.

In order to explore the range of values the function c can take as (π_1, t_1, ρ_1) varies, we did a grid search, computing c at each point in the grid. We used the grid $D' = D \times D \times D$, for $D = \{0.01, 0.1, 0.2., \ldots, 0.8, 0.9, .99\}$. The largest value of c was 1.10. Also, the smallest coverage probability of the naive confidence interval (3) that ignores the adaptive nature of the design was 93.5%. We emphasize that before our analysis, which uses the general method for determining the least favorable distribution described in the Appendix, it was unknown how much under-coverage the naive confidence interval can have for our set of enrichment designs.

If we restrict our grid search to those points in D' satisfying $\pi_1, \rho_1 \in [0.2, 0.8]$ and $t_1 \leq 0.5$, then the maximum value of c is 1.05. We think such values of π_1, ρ_1 will be common in practice, as long as neither subpopulation is very small compared to the other, and as long as the variances of the outcome distributions in both subpopulations are of the same order of magnitude. However, this might be violated if the outcome is binary and one population has much smaller risk than the other. Having $t_1 \leq 0.5$ can be achieved if the enrollment decision is planned to be made before half the total number of subjects are enrolled.

6 Discussion

The general method we used can be applied to construct confidence intervals for other types of decision rules than those in Section 3.4. In general, a decision rule based on the first stage statistics $Z_1^{(1)}, Z_2^{(1)}, Z_*^{(1)}$ can be constructed by partitioning $\mathbb{R}^3$ into three regions D_1, D_2, D_*, where the rule is that if $(Z_1^{(1)}, Z_2^{(1)}, Z_*^{(1)}) \in D_s$ then one enrolls from population s in stage two. If the regions D_s have simple boundaries and are straightforward to integrate a multivariate normal distriubution over using standard software, then the method in this paper can be used. However, an important limitation of the method is that it may not be feasible to implement for complex rules.

It may be possible to extend the above results to designs that allow prespecified changes to the randomization probabilities and/or total sample size, but this is an area of future research.

In this paper we considered only prespecified decision rules for which population to enroll in stage two. Another interesting set of designs are flexible designs, where there is no prespecified decision rule. Posch et al. (2005) and Brannath et al. (2006) present confidence interval procedures for such designs that have uniform coverage probability under any possible adaptation within the class of adaptations they allow. Confidence intervals with uniform coverage probability will in general need larger widths if a flexible design is used, compared to if a prespecified decision rule is used. It is in interesting open problem to determine the price, in terms of wider confidence interval widths, that must be paid if one does not specify the decision rule for the design in advance.

Our confidence interval procedure is optimal, as described in Section 4.2, among those that are symmetric and centered at the difference $\hat{\Delta}_S$ in sample means between treatment and control arms for the selected population, using all data from both stages from that population. However, it may be possible to reduce the confidence interval widths by considering asymmetric confidence intervals, or by centering at a different value. This is an open problem for future research.

This research was supported by the National Institutes of Health, U.S.A., and by the U.S. Food and Drug Administration's Partnership in Applied Comparative Effectiveness Science (PACES).

Appendix

In Appendix A.1, we prove the confidence interval procedure (4) from Section 4.2 satisfies (1), i.e., it has asymptotic, uniform coverage probability at least 95%. Then, in Appendix A.2, we prove that this function c is smallest possible, at every value of (π_1, t_1, ρ_1), in that for any continuous function c' that maps (π_1, t_1, ρ_1) into $\mathbb{R}$, if the confidence interval CI in (4), except replacing c

by c', has property (1), then $c'(\pi_1, t_1, \rho_1) \geq c(\pi_1, t_1, \rho_1)$ at every (π_1, t_1, ρ_1). Lastly, in Appendix A.3, we show how to compute $c(\pi_1, t_1, \rho_1)$, for given values of (π_1, t_1, ρ_1), to any desired accuracy, using statistical software.

A.1. Proof that confidence interval procedure (4) satisfies (1)

We prove the confidence interval procedure (4) from Section 4.2 satisfies (1), i.e., it has asymptotic, uniform coverage probability at least 95%.

We first show that the value of the expansion factor $c(\pi_1, t_1, \rho_1)$ is always at least 1. To show this, consider any $\tilde{c} < 1$. The sum in (6) converges to $\Phi(\tilde{c}z_{0.975}) - \Phi(-\tilde{c}z_{0.975})$ as $\tilde{\delta}_1, \tilde{\delta}_2$ both go to infinity, which follows by the form of the regions R_1, R_2, R_* in Section 4.3. Therefore, the infimum of the sum in (6) over $\tilde{\delta}_1, \tilde{\delta}_2 \in \mathbb{R}$ is at most $\Phi(\tilde{c}z_{0.975}) - \Phi(-\tilde{c}z_{0.975})$, which is less than 0.95 for $\tilde{c} < 1$. This implies the inequality to the right of the colon in (6) does not hold for $\tilde{c} < 1$, and therefore $c(\pi_1, t_1, \rho_1) \geq 1$.

We now prove the confidence interval procedure (4) satisfies (1). Fix the decision rule threshold d from Section 3.4 and the subpopulation proportions π_1, π_2. For any sample sizes n_1, n_2, and any data generating distribution $Q \in \mathcal{Q}$, the probability that the confidence interval CI contains $\Delta_S(Q)$ is

$$P_Q(\Delta_S(Q) \in CI) = \sum_{s \in \{1,2,*\}} P_Q(\Delta_S(Q) \in CI, S = s). \tag{7}$$

We consider the terms in the sum on the right hand side separately for each of $s \in \{1, 2, *\}$. First, consider the term corresponding to $s = 1$. We suppress the dependence on Q for notational clarity.

$$\begin{aligned}
& P\left(\Delta_S \in CI, S = 1\right) \\
= \;& P\left(\Delta_1 \in CI, S = 1\right) \\
= \;& P\left(\Delta_1 \in CI, Z_*^{(1)} \leq d, Z_1^{(1)} \geq Z_2^{(1)}\right) \qquad (8) \\
= \;& P\left(\Delta_1 \in [\hat{\Delta}_1 - cz_{0.975}\sigma_1/\sqrt{N_S}, \hat{\Delta}_1 + cz_{0.975}\sigma_1/\sqrt{N_S}], Z_*^{(1)} \leq d, Z_1^{(1)} \geq Z_2^{(1)}\right) \\
= \;& P\left(\hat{\Delta}_1 \in [\Delta_1 - cz_{0.975}\sigma_1/\sqrt{N_S}, \Delta_1 + cz_{0.975}\sigma_1/\sqrt{N_S}], Z_*^{(1)} \leq d, Z_1^{(1)} \geq Z_2^{(1)}\right) \\
= \;& P\left(\frac{\sqrt{N_S}}{\sigma_1}\left\{\hat{\Delta}_1 - \Delta_1\right\} \in [-cz_{0.975}, cz_{0.975}], Z_*^{(1)} \leq d, Z_1^{(1)} \geq Z_2^{(1)}\right) \\
= \;& P\left(\frac{\sqrt{\pi_1 n_1 + n_2}}{\sigma_1}\left\{\hat{\Delta}_1 - \Delta_1\right\} \in [-cz_{0.975}, cz_{0.975}], Z_*^{(1)} \leq d, Z_1^{(1)} \geq Z_2^{(1)}\right) \\
= \;& P\Bigg(\frac{\sqrt{\pi_1 n_1 + n_2}}{\sigma_1}\left\{\hat{\Delta}_1 - \Delta_1\right\} \in [-cz_{0.975}, cz_{0.975}], \\
& \qquad Z_*^{(1)} - \frac{\sqrt{n_1}\Delta_*}{\sigma_*} \leq d - \frac{\sqrt{n_1}\Delta_*}{\sigma_*}, \\
& \qquad \frac{1}{\sqrt{2}}\left\{\left(Z_1^{(1)} - \frac{\sqrt{\pi_1 n_1}\Delta_1}{\sigma_1}\right) - \left(Z_2^{(1)} - \frac{\sqrt{\pi_2 n_1}\Delta_2}{\sigma_2}\right)\right\} \geq \frac{\sqrt{\frac{\pi_2 n_1}{2}}\Delta_2}{\sigma_2} - \frac{\sqrt{\frac{\pi_1 n_1}{2}}\Delta_1}{\sigma_1}\Bigg),
\end{aligned}$$

where (8) follows from our decision rule from Section 3.4.

Denote the random variables on the right hand side of the last equality above by X_1, X_2, X_3, respectively, that is,

$$\begin{aligned} X_1^{(1)} &= \frac{\sqrt{\pi_1 n_1 + n_2}}{\sigma_1}\left\{\hat{\Delta}_1 - \Delta_1\right\}, \\ X_2^{(1)} &= Z_*^{(1)} - \frac{\sqrt{n_1}\Delta_*}{\sigma_*}, \\ X_3^{(1)} &= \frac{1}{\sqrt{2}}\left\{\left(Z_1^{(1)} - \frac{\sqrt{\pi_1 n_1}\Delta_1}{\sigma_1}\right) - \left(Z_2^{(1)} - \frac{\sqrt{\pi_2 n_1}\Delta_2}{\sigma_2}\right)\right\}. \end{aligned}$$

The vector $(X_1^{(1)}, X_2^{(1)}, X_3^{(1)})$ has mean $(0,0,0)$ and covariance matrix

$$\Sigma_1 := \begin{bmatrix} 1 & \rho_1 f_1 & f_1/\sqrt{2} \\ \rho_1 f_1 & 1 & -\rho' \\ f_1/\sqrt{2} & -\rho' & 1 \end{bmatrix}, \tag{9}$$

for $f_s = \sqrt{\frac{\pi_s n_1}{\pi_s n_1 + n_2}} = \sqrt{\frac{\pi_s t_1}{\pi_s t_1 + 1 - t_1}}$ and $\rho' = (\rho_2 - \rho_1)/\sqrt{2}$. Summarizing the above argument, we have shown

$$P\left(\Delta_S \in CI, S = 1\right) = P_{Q,n_1,n_2}\left\{(X_1^{(1)}, X_2^{(1)}, X_3^{(1)}) \in R_1(\rho_1, c, d, \delta_1, \delta_2)\right\},$$

for R_1 defined in Section 4.3. We will show below that the vector $(X_1^{(1)}, X_2^{(1)}, X_3^{(1)})$ converges to a multivariate normal distribution, uniformly over $\mathcal{Q}$ as $(n_1, n_2) \to \infty$, in a sense we define formally below.

Consider the term in the sum in (7) corresponding to $S = 2$. Analogous arguments as above show that

$$P\left(\Delta_S \in CI, S = 2\right) = P_{Q,n_1,n_2}\left\{(X_1^{(2)}, X_2^{(2)}, X_3^{(2)}) \in R_2(\rho_1, c, d, \delta_1, \delta_2)\right\},$$

for

$$\begin{aligned} X_1^{(2)} &= \frac{\sqrt{\pi_2 n_1 + n_2}}{\sigma_2}\left\{\hat{\Delta}_2 - \Delta_2\right\}, \\ X_2^{(2)} &= Z_*^{(1)} - \frac{\sqrt{n_1}\Delta_*}{\sigma_*}, \\ X_3^{(2)} &= \frac{1}{\sqrt{2}}\left\{\left(Z_1^{(1)} - \frac{\sqrt{\pi_1 n_1}\Delta_1}{\sigma_1}\right) - \left(Z_2^{(1)} - \frac{\sqrt{\pi_2 n_1}\Delta_2}{\sigma_2}\right)\right\}. \end{aligned}$$

The vector $(X_1^{(2)}, X_2^{(2)}, X_3^{(2)})$ has mean $(0,0,0)$ and covariance matrix

$$\Sigma_2 := \begin{bmatrix} 1 & \rho_2 f_2 & -f_2/\sqrt{2} \\ \rho_2 f_2 & 1 & -\rho' \\ -f_2/\sqrt{2} & -\rho' & 1 \end{bmatrix}. \tag{10}$$

Lastly, consider the term in the sum in (7) corresponding to $S = *$. Analogous arguments as above show that

$$P\left(\Delta_S \in CI, S = *\right) = P_{Q,n_1,n_2}\left\{(X_1^{(*)}, X_2^{(*)}, X_3^{(*)}) \in R_*(\rho_1, c, d, \delta_1, \delta_2)\right\},$$

where

$$(X_1^{(*)}, X_2^{(*)}) \;\; = \;\; \left(\frac{\sqrt{n_1+n_2}}{\sigma_*}\left\{\hat{\Delta}_* - \Delta_*\right\}, Z_*^{(1)} - \frac{\sqrt{n_1}\Delta_*}{\sigma_*}\right),$$

and we let $X_3^{(*)}$ denote a random variable that is independent of $(X_1^{(*)}, X_2^{(*)})$, and that has a standard normal distribution. We add in this exogenous variable so that all of our random vectors below have three components, which allows a simpler exposition. The vector $(X_1^{(*)}, X_2^{(*)}, X_3^{(*)})$ has mean $(0, 0, 0)$ and covariance matrix

$$\Sigma_* := \left[\begin{array}{ccc} 1 & \sqrt{t_1} & 0 \\ \sqrt{t_1} & 1 & 0 \\ 0 & 0 & 1 \end{array}\right]. \tag{11}$$

By the uniform central limit theorem of Götze (1991), under the assumptions in Section 3.1, we have for each $s \in \{1, 2, *\}$

$$\lim_{n_1,n_2\to\infty} \sup_{Q\in\mathcal{Q},C\in\mathcal{C}} \left| P_{Q,n_1,n_2}\left\{\Sigma_s^{-1/2}(X_1^{(s)}, X_2^{(s)}, X_3^{(s)})^t \in C\right\} - \int_C d\Phi^3 \right| = 0, \tag{12}$$

where $\mathcal{C}$ denotes the set of all Borel measurable, convex subsets of $\mathbb{R}^3$, Φ^3 is the distribution function of the multivariate normal distribution in $\mathbb{R}^3$ with zero mean and covariance matrix the identity matrix, and the superscript t indicates the transpose. Since we showed above that for each $s \in \{1, 2, *\}$, $P\left(\Delta_S \in CI, S = s\right) = P_{Q,n_1,n_2}\left\{(X_1^{(s)}, X_2^{(s)}, X_3^{(s)}) \in R_s(\rho_1, c, d, \delta_1, \delta_2)\right\}$, for $R_s(\rho_1, c, d, \delta_1, \delta_2)$ the rectangular region in $\mathbb{R}^3$ defined in (4.3), the above display (12) implies

$$\lim_{n_1,n_2\to\infty} \sup_{Q\in\mathcal{Q}} \left| P\left(\Delta_S \in CI, S = s\right) - \int_{\Sigma_s^{-1/2} R_s(\rho_1,c,d,\delta_1,\delta_2)} d\Phi^3 \right| = 0. \tag{13}$$

We point out that $\Sigma_s, \rho_1, c, \delta_1, \delta_2$ are each functions of π_1, n_1, n_2, and Q. The integral in (13) equals $P\left[G_s(\pi_1, t_1, \rho_1) \in R_s(\rho_1, c, d, \delta_1, \delta_2)\right]$, for G_s a multivariate normal random vector with mean vector $(0, 0, 0)$ and covariance matrix Σ_s. Therefore,

$$\lim_{n_1,n_2\to\infty} \sup_{Q\in\mathcal{Q}} \left| P\left(\Delta_S \in CI, S = s\right) - P\left[G_s(\pi_1, t_1, \rho_1) \in R_s(\rho_1, c, d, \delta_1, \delta_2)\right] \right| = 0. \tag{14}$$

Combining the above with (7) implies

$$\lim_{n_1,n_2\to\infty} \sup_{Q\in\mathcal{Q}} \left| P\left(\Delta_S \in CI\right) - \sum_{s\in\{1,2,*\}} P\left[G_s(\pi_1, t_1, \rho_1) \in R_s(\rho_1, c, d, \delta_1, \delta_2)\right] \right| = 0. \tag{15}$$

This implies

$$\liminf_{n_1,n_2\to\infty}\inf_{Q\in\mathcal{Q}} P\left(\Delta_S\in CI\right)=\liminf_{n_1,n_2\to\infty}\inf_{Q\in\mathcal{Q}}\sum_{s\in\{1,2,*\}} P\left[G_s(\pi_1,t_1,\rho_1)\in R_s(\rho_1,c,d,\delta_1,\delta_2)\right]. \quad (16)$$

Recall that by the definition of c in (6), we have

$$\sum_{s\in\{1,2,*\}} P\left[G_s(\pi_1,t_1,\rho_1)\in R_s(\rho_1,c,d,\delta_1,\delta_2)\right]\geq 0.95,$$

which together with (16) proves that the confidence interval procedure (4) satisfies (1), completing the proof. □

A.2. Proof that c as defined in (6) is smallest possible

We prove that for any continuous function c' that maps (π_1,t_1,ρ_1) into $\mathbb{R}$, if the confidence interval CI in (4), except replacing c by c', has property (1), then $c'(\pi_1,t_1,\rho_1)\geq c(\pi_1,t_1,\rho_1)$ at every (π_1,t_1,ρ_1).

Consider any continuous function c' that maps (π_1,t_1,ρ_1) into $\mathbb{R}$, for which the confidence interval CI in (4), except replacing c by c' (which we refer to in what follows as confidence interval procedure CI'), has property (1). For sake of contradiction, assume there is a vector of values $(\tilde{\pi}_1,\tilde{t}_1,\tilde{\rho}_1)$, for which $c'(\tilde{\pi}_1,\tilde{t}_1,\tilde{\rho}_1)<c(\tilde{\pi}_1,\tilde{t}_1,\tilde{\rho}_1)$. Then by the definition of the function c in (6), there exist values $\tilde{\delta}_1,\tilde{\delta}_2\in\mathbb{R}$ for which

$$\sum_{s\in\{1,2,*\}} P\left[G_s(\tilde{\pi}_1,\tilde{t}_1,\tilde{\rho}_1)\in R_s(\tilde{\rho}_1,c'(\tilde{\pi}_1,\tilde{t}_1,\tilde{\rho}_1),d,\tilde{\delta}_1,\tilde{\delta}_2)\right]<\theta<0.95, \quad (17)$$

for some θ.

We first consider the case in which all components of $(\tilde{\pi}_1,\tilde{t}_1,\tilde{\rho}_1)$ are rational numbers. We will construct a sequence of sample sizes $(n_1^{(k)},n_2^{(k)})$ tending to infinity and data generating distributions $Q^{(k)}\in\mathcal{Q}$ for which the confidence interval procedure CI' has coverage probability less than θ for each k. Since $(\tilde{\pi}_1,\tilde{t}_1,\tilde{\rho}_1)$ are assumed to be rational numbers, there exist positive integers j_1,j_2 such that $\tilde{t}_1=j_1/(j_1+j_2)$, and such that both $j_1\tilde{\pi}_1$ and $j_1(1-\tilde{\pi}_1)$ are positive integers. For each positive integer k, define:

i. the kth pair of first and second stage sample sizes to be $n_1^{(k)}=kj_1, n_2^{(k)}=kj_2$;

ii. the outcome distribution for subpopulation 1, $Q_{1a}^{(k)}$, for each $a\in\{0,1\}$, to be a normal distribution with mean $2a\tilde{\delta}_1\{(1-\tilde{\pi}_1)\tilde{\rho}_1^2/(\tilde{\pi}_1 n_1^{(k)})\}^{1/2}$ and variance $(1-\tilde{\pi}_1)\tilde{\rho}_1^2$;

iii. the outcome distribution for subpopulation 2, $Q_{2a}^{(k)}$, for each $a\in\{0,1\}$, to be a normal distribution with mean $2a\tilde{\delta}_2\{\tilde{\pi}_1(1-\tilde{\rho}_1^2)/((1-\tilde{\pi}_1)n_1^{(k)})\}^{1/2}$ and variance $\tilde{\pi}_1(1-\tilde{\rho}_1^2)$.

Define the kth data generating distribution $Q^{(k)} = \left(Q_{10}^{(k)}, Q_{11}^{(k)}, Q_{20}^{(k)}, Q_{21}^{(k)}\right)$. We constructed the sample sizes and data generating distributions above so that for each k, the covariance of the corresponding stage one z-statistics $Z_1^{(1)}, Z_*^{(1)}$ equals $\tilde{\rho}_1$, which follows since this covariance equals

$$\begin{aligned}
& \sqrt{\tilde{\pi}_1}\sigma_1(Q^{(k)})/\sigma_2(Q^{(k)}) \\
= & \left[\frac{\tilde{\pi}_1\left(\sigma^2(Q_{10}^{(k)}) + \sigma^2(Q_{11}^{(k)})\right)}{\tilde{\pi}_1\left(\sigma^2(Q_{10}^{(k)}) + \sigma^2(Q_{11}^{(k)})\right) + \tilde{\pi}_1\left(\sigma^2(Q_{20}^{(k)}) + \sigma^2(Q_{21}^{(k)})\right)}\right]^{1/2} \\
= & \left[\frac{2\tilde{\pi}_1(1-\tilde{\pi}_1)\tilde{\rho}_1^2}{2\tilde{\pi}_1(1-\tilde{\pi}_1)\tilde{\rho}_1^2 + 2(1-\tilde{\pi}_1)\tilde{\pi}_1(1-\tilde{\rho}_1^2)}\right]^{1/2} \\
= & \ \tilde{\rho}_1,
\end{aligned}$$

and so that the corresponding non-centrality parameters equal $\tilde{\delta}_1, \tilde{\delta}_2$, which follows since the non-centrality parameter corresponding to subpopulation 1 equals

$$\begin{aligned}
\sqrt{\tilde{\pi}_1 n_1^{(k)}}\Delta_1(Q^{(k)})/\sigma_1(Q^{(k)}) &= \sqrt{\tilde{\pi}_1 n_1^{(k)}}\left[2\tilde{\delta}_1\{(1-\tilde{\pi}_1)\tilde{\rho}_1^2/(\tilde{\pi}_1 n_1^{(k)})\}^{1/2}\right] / \left[4(1-\tilde{\pi}_1)\tilde{\rho}_1^2\right]^{1/2} \\
&= \tilde{\delta}_1,
\end{aligned}$$

and a similar argument shows the non-centrality parameter corresponding to subpopulation 2 equals $\tilde{\delta}_2$.

By construction, the outcome distributions are normal, and by the assumptions in Section 3.1, this implies that for each $s \in \{1, 2, *\}$, the vector of statistics $(X_1^{(s)}, X_2^{(s)}, X_3^{(s)})$ defined in Appendix A.1 has a multivariate normal distribution. We then have, by the arguments in Appendix A.1, that for each $s \in \{1, 2, *\}$, under confidence interval procedure CI',

$$\begin{aligned}
& P_{Q^{(k)}, n_1^{(k)}, n_2^{(k)}}(\Delta_S \in CI', S = s) \\
= & \ P_{Q^{(k)}, n_1^{(k)}, n_2^{(k)}}\left\{(X_1^{(s)}, X_2^{(s)}, X_3^{(s)}) \in R_s(\tilde{\rho}_1, c'(\tilde{\pi}_1, \tilde{t}_1, \tilde{\rho}_1), d, \tilde{\delta}_1, \tilde{\delta}_2)\right\} \\
= & \ P\left[G_s(\tilde{\pi}_1, \tilde{t}_1, \tilde{\rho}_1) \in R_s(\tilde{\rho}_1, c'(\tilde{\pi}_1, \tilde{t}_1, \tilde{\rho}_1), d, \tilde{\delta}_1, \tilde{\delta}_2)\right].
\end{aligned}$$

Therefore, by (7) and (17), we have

$$\begin{aligned}
P_{Q^{(k)}, n_1^{(k)}, n_2^{(k)}}(\Delta_S \in CI') &= \sum_{s \in \{1,2,*\}} P\left[G_s(\tilde{\pi}_1, \tilde{t}_1, \tilde{\rho}_1) \in R_s(\tilde{\rho}_1, c'(\tilde{\pi}_1, \tilde{t}_1, \tilde{\rho}_1), d, \tilde{\delta}_1, \tilde{\delta}_2)\right]. \\
&< \theta \\
&< 0.95.
\end{aligned}$$

We have shown the coverage probability at each $Q^{(k)}, n_1^{(k)}, n_2^{(k)}$ is less than θ. Since $\theta < 0.95$, this shows the confidence interval procedure CI' does not have uniform coverage probability property (1). This completes the proof for the case in which all components of $(\tilde{\pi}_1, \tilde{t}_1, \tilde{\rho}_1)$ are rational numbers. In general, the vector $(\tilde{\pi}_1, \tilde{t}_1, \tilde{\rho}_1)$ can be approximated by a sequence of vectors with rational components, and applying the above arguments at each vector in this sequence, and using the assumed continuity of c', the general result follows. □

A.3. Computing $c(\pi_1, t_1, \rho_1)$ to any desired accuracy

We describe how to compute the expansion factor c at a given vector (π_1, t_1, ρ_1), to any desired accuracy, in the following sense: for any π_1, t_1, ρ_1, and any tolerance $\epsilon > 0$, we show how to compute a value c_ϵ satisfying conditions (i) and (ii) in Section 4.3.

We first outline the algorithm for computing c_ϵ, and then give the details of the procedure below. The outer loop in the procedure is a binary search over candidate values $\tilde{c}$ for $c(\pi_1, t_1, \rho_1)$. We initialize the lower bound for candidate values $\tilde{c}$ of $c(\pi_1, t_1, \rho_1)$ to be 1, since as argued in Appendix A.1, $c(\pi_1, t_1, \rho_1) \geq 1$; we describe how we initialize the corresponding upper bound below. For a given candidate $\tilde{c}$, we compute an approximation (as described below) of

$$\inf_{\tilde{\delta}_1, \tilde{\delta}_2 \in \mathbb{R}} \sum_{s \in \{1,2,*\}} P\left[G_s(\pi_1, t_1, \rho_1) \in R_s(\rho_1, \tilde{c}, d, \tilde{\delta}_1, \tilde{\delta}_2)\right] \tag{18}$$

and see how it compares to $0.95 + \epsilon$. If the approximation to the above display is less than $0.95 + \epsilon$, we set $\tilde{c}$ to be the new lower bound in our search; else, we set $\tilde{c}$ to be the new upper bound in our search. We then take the midpoint of the current lower and upper bounds in our search as the next candidate value of $\tilde{c}$, and iterate the above procedure until the upper and lower bounds of our search are less than ϵ apart; we let c_ϵ be the upper bound after the search terminates.

We now describe how we initialize the upper bound for candidate values $\tilde{c}$ of $c(\pi_1, t_1, \rho_1)$ in the above search. We compute the approximation (as described below) to (18) at $\tilde{c} = 1.1, (1.1)^2, (1.1)^3, \ldots$, proceeding until the first time the approximation to (18) is greater than $0.95 + \epsilon$, at which point we initialize the upper bound to the corresponding value of $\tilde{c}$.

We now describe how we approximate (18) to accuracy $\pm\epsilon$, for given $\pi_1, t_1, \rho_1, \tilde{c}$. We do a grid search, where we compute the summation in (18), for each pair $(\tilde{\delta}_1, \tilde{\delta}_2) \in H(\epsilon)$, where $H(\epsilon)$ is a sufficiently fine grid of points in $\mathbb{R}^2$ that we define below; we then output the minimum value found, i.e.,

$$\min_{(\tilde{\delta}_1, \tilde{\delta}_2) \in H(\epsilon)} \sum_{s \in \{1,2,*\}} P\left[G_s(\pi_1, t_1, \rho_1) \in R_s(\rho_1, \tilde{c}, d, \tilde{\delta}_1, \tilde{\delta}_2)\right]. \tag{19}$$

Each computation of the summation in the above display can be computed by statistical software for computing the distribution function of a multivariate normal distribution; in our computations, we used the `mvtnorm` package in `R`. R code for this computation is given in the Supplementary Materials.

It remains to define the grid $H(\epsilon) \subset \mathbb{R}^2$, so that (19) is within ϵ of (18). We make the change of variables $x_1 = d - (\rho_1 \tilde{\delta}_1 + \rho_2 \tilde{\delta}_2)$, $x_2 = (\tilde{\delta}_2 - \tilde{\delta}_1)/\sqrt{2}$ to make the following explanation clearer. Also, for each $s \in \{1, 2, *\}$, define

$$h_s(x_1, x_2) = P\left[G_s(\pi_1, t_1, \rho_1) \in R_s(\rho_1, \tilde{c}, d, \tilde{\delta}_1(x_1, x_2), \tilde{\delta}_2(x_1, x_2)\right], \tag{20}$$

where $\tilde{\delta}_j(x_1, x_2)$ is the value of $\tilde{\delta}_j$ according to the above change of variables. With a slight abuse of notation, we additionally define h_s for x_1, x_2 taking values $-\infty$ or ∞, by which we mean the corresponding limit of (20), e.g., for $x_1 \in \mathbb{R}$,

$$h_s(x_1, -\infty) = \lim_{x_2' \to -\infty} h_s(x_1, x_2'); \text{ and } h_s(\infty, \infty) = \lim_{x_1' \to \infty, x_2' \to \infty} h_s(x_1', x_2').$$

Define the vector $g = (g_1, g_2)$ to be the gradient with respect to x_1, x_2 of $\sum_{s\in\{1,2,*\}} h_s(x_1, x_2)$. For for each $j \in \{1, 2\}$ define $\tilde{g}_j = \sup_{x_1,x_2\in\mathbb{R}} |g_j|$; these values can be upper bounded using the method in Section F of the Supplementary Materials of (Rosenblum and van der Laan, 2011). Then for any two points (x_1, x_2) and (x_1', x_2') in $\mathbb{R}^2$, by the mean value theorem,

$$\left| \sum_{s\in\{1,2,*\}} h_s(x_1, x_2) - \sum_{s\in\{1,2,*\}} h_s(x_1', x_2') \right| \leq |x_1 - x_1'|\tilde{g}_1 + |x_2 - x_2'|\tilde{g}_2. \tag{21}$$

Define the grid of values of (x_1, x_2) to be $H'(\epsilon) = H_1'(\epsilon) \times H_1'(\epsilon)$, for

$$H_1'(\epsilon) = \{-\infty\} \cup \{\Phi^{-1}(\epsilon/6), \Phi^{-1}(\epsilon/6) + \gamma, \ldots, \Phi^{-1}(\epsilon/6) + k\gamma, \Phi^{-1}(1 - \epsilon/6)\} \cup \{\infty\},$$

where we set $\gamma = \epsilon/(2\max\{\tilde{g}_1, \tilde{g}_2\})$ and $k = \lfloor\{\Phi^{-1}(1 - \epsilon/6) - \Phi^{-1}(\epsilon/6)\}/\gamma\rfloor$. Define the grid $H(\epsilon)$ to be that corresponding to $H'(\epsilon)$, inverting the change of variables above that maps pairs $(\tilde{\delta}_1, \tilde{\delta}_1)$ to (x_1, x_2).

We now show that the above grid leads to an approximation of (18) to accuracy $\pm\epsilon$, i.e., that

$$\min_{(x_1,x_2)\in H'(\epsilon)} \sum_{s\in\{1,2,*\}} h_s(x_1, x_2) - \inf_{(x_1,x_2)\in\mathbb{R}^2} \sum_{s\in\{1,2,*\}} h_s(x_1, x_2) \leq \epsilon. \tag{22}$$

It suffices to show for every pair $(x_1, x_2) \in \mathbb{R}^2$, that there is a point (y_1, y_2) in the grid $H'(\epsilon)$ such that

$$\sum_{s\in\{1,2,*\}} h_s(x_1, x_2) \geq \sum_{s\in\{1,2,*\}} h_s(y_1, y_2) - \epsilon. \tag{23}$$

For each $(x_1, x_2) \in \mathbb{R}^2$, we construct such a point (y_1, y_2), as follows: for each $j \in \{1, 2\}$, define

$$y_j = \begin{cases} -\infty, & \text{if } x_j < \Phi^{-1}(\epsilon/6) \\ \arg\min_{z\in H_1'(\epsilon)} |z - x_j| & \text{if } \Phi^{-1}(\epsilon/6) \leq x_j \leq \Phi^{-1}(1 - \epsilon/6) \\ \infty, & \text{if } x_j > \Phi^{-1}(1 - \epsilon/6). \end{cases}$$

That is, if x_j is contained in the interval $[\Phi^{-1}(\epsilon/6), \Phi^{-1}(1 - \epsilon/6)]$, we set y_j to be the closest point in the grid $H_1'(\epsilon)$; otherwise, we set y_j to be $\pm\infty$, where the sign equals that of x_j. We will show the point $(y_1, y_2) \in H'(\epsilon)$ satisfies (23).

We have for any $(x_1, x_2) \in \mathbb{R}^2$ for which $x_1 \in [\Phi^{-1}(\epsilon/6), \Phi^{-1}(1 - \epsilon/6)]$, that y_1 is the nearest point to x_1 in the one-dimensional grid $H_1'(\epsilon)$, and we have

$$\left| \sum_{s\in\{1,2,*\}} h_s(x_1, x_2) - \sum_{s\in\{1,2,*\}} h_s(y_1, x_2) \right| \leq \epsilon/2, \tag{24}$$

which follows by (21) and our choice of the grid width γ. An analogous result holds for any $(x_1, x_2) \in \mathbb{R}^2$ for which $x_2 \in [\Phi^{-1}(\epsilon/6), \Phi^{-1}(1 - \epsilon/6)]$.

For any $x_1 \in \mathbb{R} \cup \{-\infty, \infty\}$, for any $x_2 < \Phi^{-1}(\epsilon/6)$, it follows by the form of the regions R_1, R_2, R_* defined in Section 4.3 that for any $s \in \{1, 2, *\}$,

$$|h_s(x_1, x_2) - h_s(x_1, -\infty)| \leq \epsilon/6. \tag{25}$$

Similarly, for any $x_1 \in \mathbb{R} \cup \{-\infty, \infty\}$, for any $x_2 > \Phi^{-1}(1-\epsilon/6)$, for any $s \in \{1, 2, *\}$, we have

$$|h_s(x_1, x_2) - h_s(x_1, \infty)| \leq \epsilon/6. \tag{26}$$

Analogous results hold if we interchange x_1 and x_2.

By (24), (25), and (26), we have

$$\left| \sum_{s \in \{1,2,*\}} h_s(x_1, x_2) - \sum_{s \in \{1,2,*\}} h_s(y_1, x_2) \right| \leq \epsilon,$$

which proves (23). This completes the verification of (22), that the above grid is sufficiently fine so as to guarantee (19) is within ϵ of (18).

We now prove that the binary search algorithm in the second paragraph of this appendix leads to a value c_ϵ satisfying the approximation conditions (i) and (ii) in Section 4.3. The algorithm maintains, at every iteration, that the value of (19) is at least $0.95 + \epsilon$ for $\tilde{c}$ equal to the upper bound in the search. Then by (22), we have (18) is at least 0.95 at $\tilde{c} = c_\epsilon$, which implies condition (i).

The lower bound for candidate values $\tilde{c}$ of $c(\pi_1, t_1, \rho_1)$ in our binary search algorithm is initialized to 1, and as argued in beginning of Appendix A.1, the value of (18) at $\tilde{c} = 1$ is at most 0.95. By the structure of the above algorithm, whenever the lower bound for candidate values $\tilde{c}$ of $c(\pi_1, t_1, \rho_1)$ is updated, the corresponding value of (19) must be less than $0.95 + \epsilon$. It follows from (19) being greater or equal to (18) at any $\tilde{c}$, that the value of (18) is less than $0.95 + \epsilon$ for $\tilde{c}$ equal to the lower bound in the search, at any iteration of the search. Therefore, the lower bound in the search is always at most the value of (6) with 0.95 replaced by $0.95 + \epsilon$. By the termination condition of the binary search, i.e., that the upper and lower bounds be within ϵ of each other, we have c_ϵ is at most ϵ plus the value of (6) with 0.95 replaced by $0.95 + \epsilon$, which shows condition (ii) is satisfied. We have demonstrated that c_ϵ satisfies the approximation conditions (i) and (ii) in Section 4.3.

The above binary search algorithm assumed that computation of the terms in the summation in (19), at any fixed values of $\tilde{\delta}_1, \tilde{\delta}_2$, can be done without error. Since computing each such term involves integration, in practice there will be some error, which can be made small by integrating over a fine partition. The above algorithm can be modified to handle such error, if each term in the summation in (19) is computed to within $\pm\epsilon/4$, if the binary search algorithm uses $0.95 + \epsilon/2$ as a threshold instead of $0.95 + \epsilon$, and if the grid $H(\epsilon)$ in (19) is replaced by the finer grid $H(\epsilon/4)$. Then the approximation conditions (i) and (ii) in Section 4.3 hold for the modified algorithm.

References

Brannath, W., F. König, and P. Bauer (2006). Estimation in flexible two stage designs. *Statistics in Medicine 25*, 3366–3381.

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

Götze, F. (1991). On the rate of convergence in the multivariate clt. *The Annals of Probability 19*(2), 724–739.

Jennison, C. and B. W. Turnbull (1984). Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials 5*, 33–45.

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

Kirsch, I., B. J. Deacon, T. B. Huedo-Medina, A. Scoboria, T. J. Moore, and B. T. Johnson (2008, 02). Initial severity and antidepressant benefits: A meta-analysis of data submitted to the food and drug administration. *PLoS Med 5*(2), e45.

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

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. *Statist. Med. 24*, 3697–3714.

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

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

Sampson, A. R. and M. W. Sill (2005). Drop-the-losers design: Normal case. *Biometrical Journal 47*(3), 257–268.

Slamon, D. J., B. Leyland-Jones, S. Shak, H. Fuchs, V. Paton, A. Bajamonde, T. Fleming, W. Eiermann, J. Wolter, M. Pegram, J. Baselga, and L. Norton (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *New England Journal of Medicine 344*(11), 783–792.

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

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

Wu, S. S., W. Wang, and M. C. K. Yang (2010). Interval estimation for drop-the-losers designs. *Biometrika 97*(2), 405–418.