Received: 13 September 2017 | Revised: 16 March 2018 | Accepted: 30 April 2018

DOI: 10.1002/sim.7831

RESEARCH ARTICLE

WILEY Statistics in Medicine

# Point estimation following two-stage adaptive threshold enrichment clinical trials

Peter K. Kimani[1] | Susan Todd[2] | Lindsay A. Renfro[3] | Nigel Stallard[1]

[1]Warwick Medical School, University of Warwick, Coventry CV4 7AL, UK

[2]Department of Mathematics and Statistics, University of Reading, Reading RG6 6AX, UK

[3]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN 55905, USA

**Correspondence**
Peter K. Kimani, Warwick Medical School, University of Warwick, Coventry CV4 7AL, UK.
Email: peter.k-u.kimani@warwick.ac.uk

**Funding information**
UK Medical Research Council, Grant/Award Number: MR/N028309/1

Recently, several study designs incorporating treatment effect assessment in biomarker-based subpopulations have been proposed. Most statistical methodologies for such designs focus on the control of type I error rate and power. In this paper, we have developed point estimators for clinical trials that use the two-stage adaptive enrichment threshold design. The design consists of two stages, where in stage 1, patients are recruited in the full population. Stage 1 outcome data are then used to perform interim analysis to decide whether the trial continues to stage 2 with the full population or a subpopulation. The subpopulation is defined based on one of the candidate threshold values of a numerical predictive biomarker. To estimate treatment effect in the selected subpopulation, we have derived unbiased estimators, shrinkage estimators, and estimators that estimate bias and subtract it from the naive estimate. We have recommended one of the unbiased estimators. However, since none of the estimators dominated in all simulation scenarios based on both bias and mean squared error, an alternative strategy would be to use a hybrid estimator where the estimator used depends on the subpopulation selected. This would require a simulation study of plausible scenarios before the trial.

**KEYWORDS**

biomarker, multistage, personalized medicine, subgroup or subpopulation selection, targeted therapy

## 1 | INTRODUCTION

An area of recent interest in development of new therapies is stratified medicine, which involves using a biomarker to stratify patients into subgroups to distinguish those with the best likelihood of responding to particular treatments. If a biomarker has two levels, it is common to refer to one level as biomarker negative and the other as biomarker positive. We consider predictive biomarkers that allow the possibility of differences in treatment effects in different subpopulations, that is, a treatment by biomarker interaction effect.[1]

Advances in genetics have played a key role in stratified medicine, where biomarkers are based on genes. This has led to targeted therapies, where investigators determine a target subset of patients (subpopulation) and develop a drug (a targeted therapy) expected to be more efficacious than the control for these patients, and is possibly not beneficial to others. The target subpopulation may consist of patients with a certain gene (specifically a gene containing a certain allele) or platform of genes (specifically certain alleles corresponding to multiple genes). However, genes are not the only characteristics that are used to define a subset of patients. Examples of other biomarkers in the cancer setting include the size of tumor, protein level in the blood, and graded scores. When the clinical utility of the biomarker is not very strong or clear from previous studies, the biomarker stratified design may be used to test the effect of an experimental treatment. In this design, a trial enrolls patients from the full population but with provision for analyses of outcomes from the subpopulation.

One methodological challenge in stratified medicine is how to design and analyze efficient clinical trials that incorporate identification of the subpopulation that will benefit from the experimental treatment. An efficient design in late phase clinical trials is the two-stage adaptive enrichment design.[2] In stage 1, patients are recruited from the full population and data are used to perform an interim analysis to decide whether, in stage 2, enrollment will be from the full population or the subpopulation. The final confirmatory analysis uses data from both stages. Although the design is efficient because stage 1 data are used for subpopulation selection and confirmatory analysis, the latter is complex because of inclusion of subpopulation selection data.

We consider the case of a continuous (or a graded score) biomarker where the cut-off value to distinguish between biomarker positive and negative patients is not definite from previous trials. Consequently, several candidate cut-off values are possible, with trial data used to determine the cut-off value. Simon and Simon[2] refer to such a design that includes threshold determination as an adaptive threshold enrichment design. We give examples of clinical trials where this design can be used in Section 2.1.

Subpopulation selection based on the treatment effect can be advantageous because using an appropriate rule, the subgroup is selected in the case where there is apparent benefit in the subgroup and not in its complement (qualitative interaction) such as was observed by Mok et al.[3] The full population is selected if there is apparent benefit in the full population including when the drug benefits the subgroup and its complement with different magnitudes (quantitative interaction) such as was observed in Tran et al.[4] A subpopulation selection based on a hypothesis test for interaction only would not be able to distinguish between the two types of interactions.

Previous research that considers analysis of adaptive threshold enrichment trials focuses on control of type I error rate and power with less emphases on point estimation.[2,5] Recently, Li et al[6] have derived expressions for the biases of estimators that ignore the adaptation but do not propose point estimators that account for subpopulation selection. Kimani et al[7] and Kunzmann et al[8] have developed estimators for a setting analogous to a single fixed cut-off value. However, these estimators do not allow for using stage 1 data to determine the cut-off value in an adaptive threshold enrichment trial.

A setting similar to an adaptive threshold enrichment design is that of treatment selection, where a control is compared to multiple experimental treatments, with stage 1 data used to select the experimental treatment to test further in stage 2.[9-16] Although several point estimators for this setting exist, they cannot be applied directly in adaptive threshold enrichment clinical trials because the correlation structure of the stage 1 sample means used for selection is different.

In this paper, we develop estimators that account for subpopulation selection following adaptive threshold enrichment trials using the principles that have been used to obtain point estimators that account for treatment selection. Two unbiased estimators build on the works by Kimani et al[7] and Robertson et al.[17] Two estimators build on the works by Whitehead[18] and Stallard and Todd[10] and involve deriving the bias function to calculate bias and subtracting bias from the naive estimator. The last is a shrinkage estimator and builds on the works by Hwang[19] and Carreras and Brannath.[14]

# 2 | DESCRIPTION OF THE SETTING AND NAIVE ESTIMATION

## 2.1 | Motivation and notation

A condition where continuous biomarkers are tested and so the adaptive threshold design may be used is depression. Examples of continuous predictive biomarkers in depression are protein levels in the blood and an electrophysiological measure.[20] While introducing notation, we describe features of clinical trials that are key in our methodology based on the setting of depression.

Patients' outcomes will be assumed to be normally distributed with a known standard deviation $\sigma$. In the context of depression, Uher et al[20] perform simulations to give a guidance of the treatment effect size to be sought when predictive biomarkers are evaluated. One outcome measure they consider that is widely used in trials is the Hamilton Rating Scale for Depression (HRSD) score and is usually assumed to be normally distributed. For a trial of a prespecified duration of treatment, the aim may be to estimate the mean difference (experimental arm minus control arm) in HRSD scores between two interventions at the final follow-up visit. Based on two trials,[21,22] the standard deviation of HRSD scores may be taken to be 7, that is, $\sigma = 7$.

We will consider trials that allow stopping for futility at an interim analysis if the observed treatment difference is less than some value $b$ that we refer to as the futility boundary. The UK NICE guidelines recommend that an intervention for depression should demonstrate a difference of at least 3 HRSD points[20] to be considered superior to its comparator. Therefore, at an interim analysis, the treatment may be deemed not to warrant further testing if the observed mean difference $< 2$ (slightly less than the recommended value of 3), that is, $b = 2$.

We assume that a single continuous biomarker is used to identify the patients who benefit from a new intervention. We assume that in regard to biomarker values, there is monotonicity in treatment effect so that a higher biomarker value leads to a bigger treatment effect or a higher biomarker value leads to a smaller treatment effect. For ease of notation, we use the latter to develop methodology. Note that, if a higher biomarker value leads to a bigger treatment effect, the biomarker values can be transformed by multiplying by $-1$.

Using some biomarker threshold values, the full population ($F$) is partitioned into distinct partitions. For example, if $F$ is subdivided into four partitions, the candidate threshold values $c_1$, $c_2$, $c_3$, and $c_4$ are such that patients in partitions 1, 2, 3, and 4 have biomarker values less than $c_1$, between $c_1$ and $c_2$, between $c_2$ and $c_3$, and between $c_3$ and $c_4$, respectively. The true mean differences in partitions 1 to 4 are denoted by $\delta_1$, $\delta_2$, $\delta_3$, and $\delta_4$, respectively. We denote the number of partitions by $K$ so that, in this case, $K = 4$. We refer to the parts of $F$ below threshold values $c_1$, $c_2$, $c_3$, and $c_4$ as subpopulations $S_1$, $S_2$, $S_3$, and $S_4$. Note for $K = 4$, $S_K = S_4 = F$, and $S_1$, $S_2$, $S_3$, and $S_4$ consist of partition 1, partitions 1 and 2, partitions 1 to 3, and partitions 1 to 4, respectively. The true mean differences in $S_1$, $S_2$, $S_3$, and $S_4$ are denoted by $\theta_1$, $\theta_2$, $\theta_3$, and $\theta_4$, respectively. If, as expected, a higher biomarker value leads to a smaller treatment effect, then $\delta_1 \geq \delta_2 \geq \delta_3 \geq \delta_4$ and $\theta_1 \geq \theta_2 \geq \theta_3 \geq \theta_4$.

We assume that the threshold values $c_1, \ldots, c_K$ are prespecified. There are different ways for the choice of the thresholds values. For $K = 4$, quartiles may be used so that the prevalences for $S_1$ to $S_4$ are $p_1 = 0.25$, $p_2 = 0.50$, $p_3 = 0.75$, and $p_4 = 1$, respectively. Consequently, the partitions have equal prevalence (0.25) since if we set $p_0 = 0$, $p_i - p_{i-1} = 0.25$ ($i = 1, \ldots, 4$). In some instances, the threshold values are chosen based on aspects such as biological activity so that the prevalences for partitions are not equal. Figure 1 summarizes the partitioning of $F$ for any $K \geq 3$.

## 2.2 | Hypothetical two-stage adaptive threshold enrichment clinical trial

Predictive assessment of continuous biomarkers can been done in single-stage clinical trials.[23,24] The alternative is to use the two-stage adaptive threshold enrichment design, which is more efficient as more resources can be focused on the subpopulation that is most likely to benefit from the new treatment.[24] The design has been used in recent trials with time-to-event (progression-free survival) outcome data.[5,25,26] As we propose in this paper, the design can be similarly used in trials with normally distributed outcome data. We note in Section 6 that the methods developed in this paper can be adapted for time-to-event outcome data.

We describe the form of the adaptive threshold enrichment design that we consider based on a hypothetical trial for depression, where for example protein level is used to partition $F$ into quartiles. In stage 1, the trial recruits $n_{11} = 90$, $n_{12} = 90$, $n_{13} = 90$, and $n_{1K} = n_{14} = 90$ patients in partitions 1 to 4. The number of patients in $S_1$ to $S_4$ are $m_{11} = 90$, $m_{12} = 180$, $m_{13} = 270$, and $m_{1K} = n_{14} = 360$, respectively, since $m_{1i} = \sum_{i'=1}^{i} n_{1i'}$ ($i = 1, \ldots, 4$). For simplicity, we assume that, in each partition, the 90 patients are equally split between the control and the experimental treatment. The outcome of interest is HRSD score and is assumed to be normally distributed with $\sigma = 7$. Let $\tau_{11}^2 = 4\sigma^2/n_{11}$, $\tau_{12}^2 = 4\sigma^2/n_{12}$,

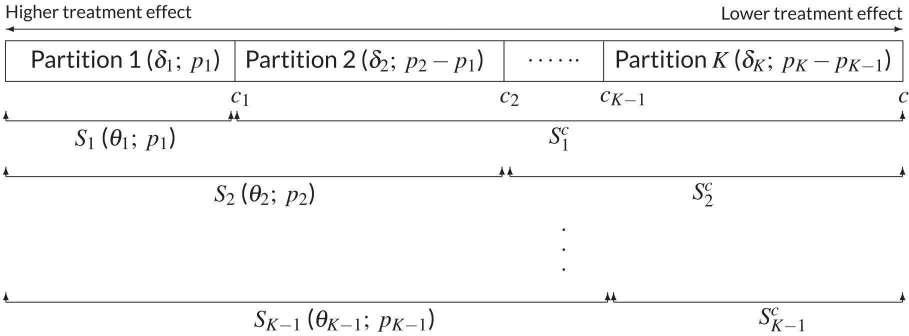
Higher treatment effect
Lower treatment effect
Partition 1 ($\delta_1$; $p_1$)
Partition 2 ($\delta_2$; $p_2 - p_1$)
······
Partition $K$ ($\delta_K$; $p_K - p_{K-1}$)
$c_1$
$c_2$
$c_{K-1}$
$c_K$
$S_1$ ($\theta_1$; $p_1$)
$S_1^c$
$S_2$ ($\theta_2$; $p_2$)
$S_2^c$
$S_{K-1}$ ($\theta_{K-1}$; $p_{K-1}$)
$S_{K-1}^c$

**FIGURE 1** Partitioning of the full population. Partitions to the left are expected to have bigger treatment effects. The pairs in the brackets are true mean differences and prevalences for partitions and candidate subpopulations

$\tau_{13}^2 = 4\sigma^2/n_{13}$, and $\tau_{14}^2 = 4\sigma^2/n_{14}$, the stage 1 sample mean differences in partitions 1 to 4 are $\bar{X}_{11} \sim N(\delta_1, \tau_{11}^2)$, $\bar{X}_{12} \sim N(\delta_2, \tau_{12}^2)$, $\bar{X}_{13} \sim N(\delta_2, \tau_{13}^2)$, and $\bar{X}_{14} \sim N(\delta_4, \tau_{14}^2)$, respectively. Let $\sigma_{11}^2 = 4\sigma^2/m_{11}$, $\sigma_{12}^2 = 4\sigma^2/m_{12}$, $\sigma_{13}^2 = 4\sigma^2/m_{13}$, and $\sigma_{14}^2 = 4\sigma^2/m_{14}$, the stage 1 sample means in $S_1$ to $S_4$ are $\bar{Y}_{11} \sim N(\theta_1, \sigma_{11}^2)$, $\bar{Y}_{12} \sim N(\theta_2, \sigma_{12}^2)$, $\bar{Y}_{13} \sim N(\theta_3, \sigma_{13}^2)$, and $\bar{Y}_{14} \sim N(\theta_4, \sigma_{14}^2)$, respectively. If the number of patients in a partition is not equally split between the control and the experimental treatment, the expressions for $\tau_{11}^2$ to $\tau_{14}^2$ and $\sigma_{11}^2$ to $\sigma_{14}^2$ are different. Note that, in this hypothetical trial, $\tau_{11}^2 = \cdots = \tau_{14}^2 = \sigma_{11}^2 = 2.178$, $\sigma_{12}^2 = 1.089$, $\sigma_{13}^2 = 0.726$, and $\sigma_{14}^2 = 0.544$. The random vectors $\bar{\mathbf{X}}_1 = (\bar{X}_{11}, \bar{X}_{12}, \bar{X}_{13}, \bar{X}_{14})'$ and $\bar{\mathbf{Y}}_1 = (\bar{Y}_{11}, \bar{Y}_{12}, \bar{Y}_{13}, \bar{Y}_{14})'$ have a linear relationship and are multivariate normal with mean vectors $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3, \delta_4)'$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)'$, respectively (see supplementary material). Hence, selection rules based on observed values for $\bar{\mathbf{X}}_1$ can be restated using the observed values for $\bar{\mathbf{Y}}_1$ and vice versa.

Since a higher biomarker value is expected to lead to lower treatment effect, the largest subpopulation for which the observed stage 1 sample mean difference (in HRSD scores) is $\geq b$ is selected to continue to stage 2. If the observed stage 1 sample mean differences in $S_1$, $S_2$, $S_3$, and $S_4 = F$ are all less than $b$, the trial stops for futility. Note that the selected subpopulation is a random variable determined by observed stage 1 data. We use lower case $s (s \in \{1, \dots, 4\})$ as the index for the "observed" selected subpopulation, with $S_s (s \in \{1, \dots, 4\})$ denoting the selected subpopulation. At the end of stage 2, the primary objective is to obtain an estimate for $\theta_s$, using an estimator that has good properties such as being mean unbiased and having small mean squared error (MSE).

Suppose that the stage 1 observed sample mean differences in partitions 1 to 4 are $\bar{x}_{11} = 3$, $\bar{x}_{12} = 2$, $\bar{x}_{13} = 0.8$, and $\bar{x}_{14} = 0$ so that $S_1$ to $S_4$ stage 1 observed sample mean differences are $\bar{y}_{11} = 3$, $\bar{y}_{12} = 2.5$, $\bar{y}_{13} = 1.93$, and $\bar{y}_{14} = 1.45$. Subpopulation 2 would be selected, that is, $S_s = S_2$, since it is the largest subpopulation with observed mean difference of at least 2 points, so that $\theta_s = \theta_2$.

In stage 2, the trial recruits $n_{21} = 120$ and $n_{22} = 120$ patients in partitions 1 and 2, respectively. The number of patients in $S_1$ and $S_2$ are $m_{21} = 120$ and $m_{22} = 240$, respectively, since $m_{2i} = \sum_{i'=1}^{i} n_{2i'}$ $(i = 1, \dots, s)$. The sample sizes $n_{21}$ and $n_{22}$ and, hence, $m_{21}$ and $m_{22}$, should be prespecified in advance for example by fixing the total stage 2 sample size and the ratio of allocation to the selected partitions. Let $\tau_{21}^2 = 4\sigma^2/n_{21}$ and $\tau_{22}^2 = 4\sigma^2/n_{22}$, the stage 2 sample mean differences in partitions 1 and 2 are $\bar{X}_{21} \sim N(\delta_1, \tau_{21}^2)$ and $\bar{X}_{22} \sim N(\delta_2, \tau_{22}^2)$, respectively. Let $\sigma_{21}^2 = 4\sigma^2/m_{21}$ and $\sigma_{22}^2 = 4\sigma^2/m_{22}$, the stage 2 sample means in $S_1$ and $S_2$ are $\bar{Y}_{21} \sim N(\theta_1, \sigma_{21}^2)$ and $\bar{Y}_{22} \sim N(\theta_2, \sigma_{22}^2)$, respectively. For this hypothetical trial, $\tau_{21}^2 = \tau_{22}^2 = \sigma_{21}^2 = 1.633$ and $\sigma_{22}^2 = 0.817$. Table 1 summarizes the notation we have introduced for any $K \geq 3$. When a subscript in a notation includes two indices, the first corresponds to stage and the second to partition or subpopulation.

Suppose that, in stage 2, the observed sample mean differences in partitions 1 and 2 are $\bar{x}_{21} = 3.0$ and $\bar{x}_{22} = 2.4$. Consequently, the stage 2 observed sample mean difference for $S_2$ is $\bar{y}_{22} = 2.7$. The naive estimate for $\theta_2$ is the two-stage sample mean difference for $S_2$ given by $\hat{\theta}_{2,N} = (m_{12}\bar{y}_{12} + m_{22}\bar{y}_{22})/(m_{21} + m_{22}) = 2.614$. We describe in Section 2.4 that the naive estimates are biased because they ignore subpopulation selection. The aim of this paper is to develop estimators that adjust for subpopulation selection. The estimators are based on the selection rule described for the hypothetical trial, which we state for any $K \geq 3$ partitions in the next section, and are conditional on the observed ordering of stage 1 data.

## 2.3 | Selection rule

We derive estimators that are unbiased or with small bias conditional on the following specific selection rule. Other selection rules are considered in the discussion. Let $b$ denote a futility boundary. The trial stops after stage 1 if $\bar{y}_{1i} < b$

**TABLE 1** Summary of notation

| | | Stage 1 Partitions | | | | | Stage 2 Partitions | | |
|---|---|---|---|---|---|---|---|---|---|
| **Measure** | **Subgroup** | **1** | **2** | … | ***K* − 1** | ***K*** | **1** | … | ***s* ∈ {1, … , *K*}** |
| Upper threshold | | $c_1$ | $c_2$ | … | $c_{K-1}$ | $c_K$ | $c_1$ | … | $c_s$ |
| Sample size | Partition | $n_{11}$ | $n_{12}$ | … | $n_{1,K-1}$ | $n_{1K}$ | $n_{21}$ | … | $n_{2S}$ |
| | Subpopulation | $m_{11}$ | $m_{12}$ | … | $m_{1,K-1}$ | $m_{1K}$ | $m_{21}$ | … | $m_{2s}$ |
| Sample variance | Partition | $\tau_{11}^2$ | $\tau_{12}^2$ | … | $\tau_{1,K-1}^2$ | $\tau_{1K}^2$ | $\tau_{21}^2$ | … | $\tau_{2s}^2$ |
| | Subpopulation | $\sigma_{11}^2$ | $\sigma_{12}^2$ | … | $\sigma_{1,K-1}^2$ | $\sigma_{1K}^2$ | $\sigma_{21}^2$ | … | $\sigma_{2s}^2$ |
| True mean | Partition | $\delta_1$ | $\delta_2$ | … | $\delta_{K-1}$ | $\delta_K$ | $\delta_1$ | … | $\delta_s$ |
| | Subpopulation | $\theta_1$ | $\theta_2$ | … | $\theta_{K-1}$ | $\theta_K$ | $\theta_1$ | … | $\theta_s$ |
| Sample mean | Partition | $\bar{X}_{11}$ | $\bar{X}_{12}$ | … | $\bar{X}_{1,K-1}$ | $\bar{X}_{1K}$ | $\bar{X}_{21}$ | … | $\bar{X}_{2s}$ |
| | Subpopulation | $\bar{Y}_{11}$ | $\bar{Y}_{12}$ | … | $\bar{Y}_{1,K-1}$ | $\bar{Y}_{1K}$ | $\bar{Y}_{21}$ | … | $\bar{Y}_{2s}$ |

for all $i$ $(i = 1, \dots, K)$. The trial continues to stage 2 with the full population $(S_K)$ if $\bar{y}_{1K} \geq b$ and with subpopulation $S_s \in \{1, \dots, K-1\}$ if $\bar{y}_{1s} \geq b$ and $\bar{y}_{1i} < b$ for all $i \in \{s+1, \dots, K\}$. Thus, as shown in the supplementary material, subpopulation $S_s (s \in \{1, \dots, K-1\})$ is selected if $b \leq \bar{y}_{1s} < u$, where

$$u = \min\left\{\frac{p_{s+1}b - (p_{s+1} - p_s)\bar{x}_{1,s+1}}{p_s}, \frac{p_{s+2}b - \sum_{i=s+1}^{s+2}(p_i - p_{i-1})\bar{x}_{1i}}{p_s}, \dots, \frac{p_K b - \sum_{i=s+1}^{K}(p_i - p_{i-1})\bar{x}_{1i}}{p_s}\right\}.$$

Equivalently, subpopulation $S_s (s \in \{1, \dots, K-1\})$ is selected if for all $i' \in \{1, \dots, s\}$, $v_{i'} \leq \bar{x}_{1i'} < w_{i'}$, where $v_{i'} = \frac{1}{p_{i'} - p_{i'-1}}\left(p_s \cdot b - \sum_{\substack{i=1\\ i \neq i'}}^{s}(p_i - p_{i-1})\bar{x}_{1i}\right)$ and

$$w_{i'} = \min\left\{\frac{p_{s+1}b - \sum_{\substack{i=1\\ i\neq i'}}^{s+1}(p_i - p_{i-1})\bar{x}_{1i}}{p_{i'} - p_{i'-1}}, \frac{p_{s+2}b - \sum_{\substack{i=1\\ i\neq i'}}^{s+2}(p_i - p_{i-1})\bar{x}_{1i}}{p_{i'} - p_{i'-1}}, \dots, \frac{p_K b - \sum_{\substack{i=1\\ i\neq i'}}^{K}(p_i - p_{i-1})\bar{x}_{1i}}{p_{i'} - p_{i'-1}}\right\}.$$

### 2.4 | Naive estimation

For the selected subpopulation $S_s (s \in \{1, \dots, K\})$, define $t_s = m_{1s}/(m_{1s} + m_{2s})$. The naive estimator for $\theta_s$ that ignores subpopulation selection is

$$\hat{\theta}_{s,N} = t_s \bar{Y}_{1s} + (1 - t_s)\bar{Y}_{2s}. \tag{1}$$

This is biased because the first term in (1) includes data used in the selection. Let $\mathbf{1}_{[S_s]}$ and $\text{Prob}(S_s)$ denote the indicator and probability of selecting $S_s$, respectively. The conditional bias is

$$\text{Bias}(\hat{\theta}_{s,N}) = t_s\left\{\frac{\sum_{i=1}^{s}(p_i - p_{i-1})E\left[\bar{X}_{1i}\mathbf{1}_{[S_s]}\right]}{p_s \cdot \text{Prob}(S_s)} - \theta_s\right\}. \tag{2}$$

Using the joint density for $\bar{\mathbf{X}}_1$ or $\bar{\mathbf{Y}}_1$ to compute $\text{Prob}(S_s)$ and $\sum_{i=1}^{s}(p_i - p_{i-1})E\left[\bar{X}_{1i}\mathbf{1}_{[S_s]}\right]$ is computationally time consuming because the limits of integration for each element in the vector depend on the values of the other elements. To overcome this, we use $\mathbf{Z} = (Z_1, \dots, Z_K)'$, where $Z_1 = \bar{X}_{11}$ and $Z_{i'} = \sum_{i=1}^{i'}(p_i - p_{i-1})\bar{X}_{1i}$ $(i' = 2, \dots, K)$. The density for $\mathbf{Z}$ and the expressions for $\text{Prob}(S_s)$ and $\sum_{i=1}^{s}(p_i - p_{i-1})E\left[\bar{X}_{1i}\mathbf{1}_{[S_s]}\right]$ are provided in the supplementary material.

## 3 | ESTIMATORS THAT ACCOUNT FOR SUBPOPULATION SELECTION

### 3.1 | Unbiased estimators

#### 3.1.1 | General principles of obtaining unbiased estimators

One technique to account for subpopulation selection is Rao-Blackwellization. By the Rao-Blackwell theorem, conditional on a sufficient and complete statistic based on stages 1 and 2 data, the expected value of a conditionally unbiased estimator from the stage 2 data is the uniformly minimum variance conditional unbiased estimator (UMVCUE). We consider two methods for obtaining unbiased estimators for $\theta_s$: deriving an UMVCUE for $\theta_s$ directly or, because the relationship between $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ is linear, deriving the UMVCUE for each $\delta_i$ $(i = 1, \dots, s)$ and using a linear function to obtain an unbiased (though not necessarily minimum variance) estimator for $\theta_s$. The latter builds on the work by Kimani et al.[7] The former would involve correlated stage 1 statistics in the vector $\bar{\mathbf{Y}}_1$ and builds on the work by Robertson et al.[17]

#### 3.1.2 | Uniformly minimum variance unbiased estimator following the work of Robertson et al (2016a)

The UMVCUE for $\theta_s$ is the expected value of $\bar{Y}_{2s}$ conditional on a sufficient and complete statistic. As before, let $\hat{\theta}_{s,N}$ denote the naive estimator for $\theta_s$ given by expression (1) and $U$ be as $u$ in Section 2.3 with $\bar{x}_{1,s+1}, \dots, \bar{x}_{1K}$ replaced with $\bar{X}_{1,s+1}, \dots, \bar{X}_{1K}$. Following the work of Robertson et al,[17] the UMVCUE for $\theta_s$ is

$$\hat{\theta}_{s,UMV} = \hat{\theta}_{s,N} - \frac{\sigma_{2s}^2}{\sqrt{\sigma_{1s}^2 + \sigma_{2s}^2}} \frac{\phi\left(f(b)\right) - \phi\left(f(U)\right)}{\Phi\left(f(b)\right) - \Phi\left(f(U)\right)}, \tag{3}$$

where $f(b) = \frac{\sqrt{\sigma_{1s}^2+\sigma_{2s}^2}}{\sigma 1_s^2}\left(\hat{\theta}_{s,N} - b\right)$, $f(U) = \frac{\sqrt{\sigma_{1s}^2+\sigma_{2s}^2}}{\sigma 1_s^2}\left(\hat{\theta}_{s,N} - U\right)$, and $\phi(.)$ and $\Phi(.)$ denote the density and distribution functions of a standard normal, respectively.

### 3.1.3 | Unbiased estimator following the work of Kimani et al (2015)

The UMVCUE for $\delta_{i'}$ $(i' = 1, \dots, s)$ is the expected value of $\bar{X}_{2i'}$ conditional on a sufficient and complete statistic. Let $\hat{\delta}_{i',N} = (n_{1i'}\bar{X}_{1i'} + n_{2i'}\bar{X}_{2i'})/(n_{1i'} + n_{2i'})$ $(i' = 1, \dots, s)$ denote the naive estimator for $\delta_{i'}$. Furthermore, let $V_{i'}$ and $W_{i'}$ be as $v_{i'}$ and $w_{i'}$ in Section 2.3 with $\bar{x}_{11}, \dots, \bar{x}_{1K}$ replaced with $\bar{X}_{11}, \dots, \bar{X}_{1K}$. Following the work of Kimani et al,[7] the UMVCUE for $\delta_{i'}$ $(i' = 1, \dots, s)$ is

$$\hat{\delta}_{i',UMVCUE} = \hat{\delta}_{i',N} - \frac{\tau_{2i'}^2}{\sqrt{\tau_{1i'}^2 + \tau_{2i'}^2}} \frac{\phi\left(f(V_{i'})\right) - \phi\left(f(W_{i'})\right)}{\Phi\left(f(V_{i'})\right) - \Phi\left(f(W_{i'})\right)},$$

where $f(V_{i'}) = \frac{\sqrt{\tau_{1i'}^2+\tau_{2i'}^2}}{\tau_{1i'}^2}\left(\hat{\delta}_{i',N} - V_{i'}\right)$ and $f(W_{i'}) = \frac{\sqrt{\tau_{1i'}^2+\tau_{2i'}^2}}{\tau_{1i'}^2}\left(\hat{\delta}_{i',N} - W_{i'}\right)$. Consequently, the unbiased estimator for $\theta_s$ is

$$\hat{\theta}_{s,U} = \sum_{i=1}^{s} \frac{(p_i - p_{i-1})\hat{\delta}_{i,UMVCUE}}{p_s}. \tag{4}$$

## 3.2 | Bias-adjusted estimators

### 3.2.1 | An overview of bias-adjusted estimation

Another technique to account for subpopulation selection would be to utilize the fact that we can calculate bias of the naive estimate using expression (2). The naive estimate is then adjusted by subtracting the bias. However, expression (2) is a function of $\boldsymbol{\delta}$ (or equivalently $\boldsymbol{\theta}$), the vector of the unknown treatment effects. To overcome this, we estimate bias, and hence, bias-adjusted estimators obtained in this way are not necessarily mean unbiased.

### 3.2.2 | Single-iteration bias-adjusted estimator

We consider two bias-adjusted estimators. For the first one, the bias is estimated based on the observed sample mean differences $\hat{\delta}_{i,N} = \left(n_{1i}\bar{x}_{1i} + n_{2i}\bar{x}_{2i}\mathbf{1}_{[i\le s]}\right) / \left(n_{1i} + n_{2i}\mathbf{1}_{[i\le s]}\right)$ $(i = 1, \dots, K)$. Let $\hat{\boldsymbol{\delta}} = (\hat{\delta}_{1,N}, \dots, \hat{\delta}_{K,N})'$ and $b_{\theta_s}(\hat{\boldsymbol{\delta}})$ denote the bias estimator for $\theta_s$ obtained by replacing $\boldsymbol{\delta}$ with $\hat{\boldsymbol{\delta}}$ in expression (2) to get an adjusted estimator for $\theta_s$ of

$$\hat{\theta}_{s,SI} = \hat{\theta}_{s,MLE} - b_{\theta_s}(\hat{\boldsymbol{\delta}}). \tag{5}$$

We will refer to this estimator as the single-iteration bias-adjusted estimator.

### 3.2.3 | Multiple-iteration bias-adjusted estimator

For the second bias-adjusted estimator, the bias is estimated iteratively.[10,13,18] Let $\hat{\theta}_i$ $(i = 1, \dots, K)$ denote the naive estimator for $\theta_i$ and $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_K)'$. The biases for the naive estimators depend on $\boldsymbol{\theta}$ and we denote bias for $\hat{\theta}_i$ $(i = 1, \dots, K)$ by $b_i(\boldsymbol{\theta})$ and the vector $(b_1(\boldsymbol{\theta}), \dots, b_K(\boldsymbol{\theta}))$ by $\boldsymbol{b}(\boldsymbol{\theta})$. The second adjusted estimator, which we refer to as multiple-iteration bias-adjusted estimator is obtained by solving $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} - \boldsymbol{b}(\tilde{\boldsymbol{\theta}})$ iteratively. Using similar notation, alternatively, one could solve $\tilde{\boldsymbol{\delta}} = \hat{\boldsymbol{\delta}} - \boldsymbol{b}(\tilde{\boldsymbol{\delta}})$ and then use the relationship between $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ to obtain a bias-adjusted estimate for $\theta_s$. For the simulations in Section 5, we solve $\tilde{\boldsymbol{\delta}} = \hat{\boldsymbol{\delta}} - \boldsymbol{b}(\tilde{\boldsymbol{\delta}})$ and with an accuracy of 0.001, convergence was achieved in almost all simulated trials. Suppose that the solution is obtained at iteration $\boldsymbol{r}$ and let $b_i(\tilde{\boldsymbol{\delta}}_{\boldsymbol{r}})$ denote the bias for $\hat{\delta}_i$ when $\boldsymbol{\delta}$ is taken to be $\tilde{\boldsymbol{\delta}}_{\boldsymbol{r}}$, then the multiple-iteration adjusted estimate for $\delta_i$ is $\hat{\hat{\delta}}_{i,MI} = \hat{\delta}_i - b_i(\tilde{\boldsymbol{\delta}}_{\boldsymbol{r}})$ and the multiple-iteration bias-adjusted estimator for $\theta_s$ is

$$\hat{\theta}_{s,MI} = \sum_{i=1}^{s} \frac{(p_i - p_{i-1})\hat{\hat{\delta}}_{i,MI}}{p_s}. \tag{6}$$

The details of calculating $b_i(\tilde{\boldsymbol{\delta}}_{\boldsymbol{r}})$ are given in the supplementary materials.

## 3.3 | Shrinkage estimators

### 3.3.1 | General principles for shrinkage estimation

A third technique for accounting for subpopulation selection is to use shrinkage methods. Hwang[19] considered the case of estimating a treatment mean after ordering independent sample means in a single-stage trial for $K \geq 4$. A subpopulation selection rule that corresponds to Hwang's case is that of selecting only one partition based on some ordering of $\bar{x}_{11}, \dots, \bar{x}_{1K}$. We initially consider Hwang's selection rule and denote the selected partition by $s_H (s_H \in \{1, \dots, K\})$. Hwang assigns a common normal prior distribution $N(\mu, \nu^2)$ to each $\delta_i (i = 1, \dots, K)$. The posterior mean for $\delta_{s_H}$, its Bayes estimator, is $C\bar{X}_{1s_H} + (1 - C)\mu$, where $C = 1 - 2\sigma^2/(2\sigma^2 + n\nu^2)$ and $n$ is stage 1 sample size in each intervention in each partition. Replacing the unknown $\mu$ and $C$ with their unbiased estimators $\bar{Y}_{1K} = \sum_{i=1}^{K} \bar{X}_{1j}/K$ and $\hat{C} = 1 - 2(K-3)\sigma^2/[n \sum_{j=1}^{K} (\bar{X}_{1j} - \bar{Y}_{1K})^2]$, respectively, gives the empirical Bayes estimator. Let $\hat{C}_+ = \max\{0, \hat{C}\}$, Hwang indicates that a better estimator, which we refer to as the shrinkage estimator, is $\hat{\hat{\delta}}_{s_H,B_1} = \hat{C}_+\bar{X}_{1s_H} + (1 - \hat{C}_+)\bar{Y}_{1K}$.

Carreras and Brannath[14] extended the work to two-stage trials. Define $t_{s_H} = n_{1s_H}/(n_{1s_H} + n_{2s_H})$ to be the proportion of stage 1 data. The two-stage shrinkage estimator for $\delta_{s_H}$ is $\hat{\hat{\delta}}_{s_H,B} = t_{s_H}\hat{\hat{\delta}}_{s_H,B_1} + (1 - t_{s_H})\bar{X}_{2S}$. For $K < 4$, Carreras and Brannath propose defining $\hat{C} = 1 - 2(K-1)\sigma^2/[n \sum_{i=1}^{K} (\bar{X}_{1i} - \bar{Y}_{1K})^2]$. Using the fact that the estimator of Hwang[19] applies for all parameters $\delta_i$ $(i = 1, \dots, K)$ and that its examination by Carreras and Brannath showed that it works for any rule used to pick the parameters on which to make inference, in Sections 3.3.2 and 3.3.3, we extend this work to give two shrinkage estimators for the subpopulation selection rule in Section 2.3.

### 3.3.2 | First shrinkage estimator

As in unbiased estimation, we consider both combining shrinkage estimators for treatment effects in partitions to obtain an estimator for $\theta_s$ and directly obtaining a shrinkage estimator for $\theta_s$. From Section 3.3.1, the shrinkage estimator for $\delta_i$ $(i = 1, \dots, s)$ is $\hat{\hat{\delta}}_{i,L} = t_s\left[\hat{C}_+\bar{X}_{1i} + (1 - \hat{C}_+)\bar{Y}_{1K}\right] + (1 - t_s)\bar{X}_{2i}$, where $\hat{C}_+ = \max\{0, \hat{C}\}$ and for $K \geq 4$, $\hat{C} = 1 - 2(K - 3)\sigma^2/[n \sum_{j=1}^{K} (\bar{X}_{1j} - \bar{Y}_{1K})^2]$, whereas for $K < 4$, $\hat{C} = 1 - 2(K-1)\sigma^2/[n \sum_{j=1}^{K} (\bar{X}_{1j} - \bar{Y}_{1K})^2]$. The first shrinkage estimator for $\theta_s$ is

$$\hat{\theta}_{s,L_1} = \sum_{i=1}^{s} \frac{(p_i - p_{i-1})\hat{\hat{\delta}}_{i,L}}{p_s}. \tag{7}$$

### 3.3.3 | Second shrinkage estimator

The second shrinkage estimator, which we denote by $\hat{\theta}_{s,L_2}$, involves using the entire parameter vector $\boldsymbol{\theta}$. A multivariate normal prior for $\boldsymbol{\theta}$ is specified and updated with the data $\bar{\mathbf{Y}}_1$. The resulting posterior is multivariate normal with nonzero covariance, and hence, the iterative procedure of Morris[27] and Brüncker et al[28] is utilized to obtain $\hat{\theta}_{s,L_2}$ (see supplementary material).

# 4 | WORKED EXAMPLE

We use data from the hypothetical trial for depression in Section 2.2 to demonstrate how to compute the naive ($\hat{\theta}_{2,N}$), the UMVCUE ($\hat{\theta}_{2,UMV}$), the unbiased ($\hat{\theta}_{2,U}$), the single-iteration bias-adjusted ($\hat{\theta}_{2,SI}$), the multiple-iteration bias-adjusted ($\hat{\theta}_{2,MI}$), the first shrinkage ($\hat{\theta}_{2,L_1}$), and the second shrinkage ($\hat{\theta}_{2,L_2}$) estimates. We also use the example to demonstrate differences among the various estimates in a single trial. The data and the various estimates are summarized in Table 2. The explicit computations for the various estimates and the R program used are provided in the supplementary material.

**TABLE 2** Worked example data and estimates

| | | Data and Summary Measures | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Stage 1 Partitions | | | | Stage 2 Partitions | | Estimating $\theta_2$ | |
| **Measure** | **Subgroup** | **1** | **2** | **3** | **4** | **1** | $s = 2$ | **Estimator** | **Estimate** |
| Sample size | Partition | $n_{11} = 90$ | $n_{12} = 90$ | $n_{13} = 90$ | $n_{14} = 90$ | $n_{21} = 120$ | $n_{22} = 120$ | $\hat{\theta}_{2,N}$ | 2.614 |
| | Subgroup | $m_{11} = 90$ | $m_{12} = 180$ | $m_{13} = 270$ | $m_{14} = 360$ | $m_{21} = 120$ | $m_{22} = 240$ | $\hat{\theta}_{2,UMV}$ | 2.839 |
| Sample variance | Partition | $\tau_{11}^2 = 2.178$ | $\tau_{12}^2 = 2.178$ | $\tau_{13}^2 = 2.178$ | $\tau_{14}^2 = 2.178$ | $\tau_{21}^2 = 1.633$ | $\tau_{22}^2 = 1.633$ | $\hat{\theta}_{2,U}$ | 2.965 |
| | Subgroup | $\sigma_{11}^2 = 2.178$ | $\sigma_{12}^2 = 1.089$ | $\sigma_{13}^2 = 0.726$ | $\sigma_{14}^2 = 0.544$ | $\sigma_{21}^2 = 1.633$ | $\sigma_{22}^2 = 0.817$ | $\hat{\theta}_{2,SI}$ | 2.633 |
| Sample mean | Partition | $\bar{x}_{11} = 3$ | $\bar{x}_{12} = 2$ | $\bar{x}_{13} = 0.8$ | $\bar{x}_{14} = 0$ | $\bar{x}_{21} = 3$ | $\bar{x}_{22} = 2.4$ | $\hat{\theta}_{2,MI}$ | 2.666 |
| | Subgroup | $\bar{y}_{11} = 3$ | $\bar{y}_{12} = 2.5$ | $\bar{y}_{13} = 1.93$ | $\bar{y}_{14} = 1.45$ | $\bar{y}_{21} = 3$ | $\bar{y}_{22} = 2.7$ | $\hat{\theta}_{2,L_1}$ | 2.164 |
| | | | | | | | | $\hat{\theta}_{2,L_2}$ | 2.194 |

Here, we only give explicit details of computing $\hat{\theta}_{2,UMV}$ and $\hat{\theta}_{2,U}$ as they are easier to compute, and since based on the simulations in the next section, we recommend $\hat{\theta}_{2,UMV}$.

For the UMVCUE ($\hat{\theta}_{2,UMV}$) given by expression (3), the first term $\hat{\theta}_{s,N} = \hat{\theta}_{2,N} = 2.614$. Furthermore, $\sigma_{1s}^2 = \sigma_{12}^2 = 1.089$ and $\sigma_{2s}^2 = \sigma_{22}^2 = 0.817$ so that $\frac{\sigma_{2s}^2}{\sqrt{\sigma_{1s}^2+\sigma_{2s}^2}} = \frac{\sigma_{22}^2}{\sqrt{\sigma_{12}^2+\sigma_{22}^2}} = 0.592$ and $\frac{\sqrt{\sigma_{1s}^2+\sigma_{2s}^2}}{\sigma_{1s}^2} = \frac{\sqrt{\sigma_{12}^2+\sigma_{22}^2}}{\sigma_{12}^2} = 1.268$. Since $(p_i - p_{i-1}) = 0.25$ for all $i = 1, \ldots, 4$, then $u$ (the observed value for $U$) is given by

$$\begin{aligned} u &= \min\left\{\frac{p_3 b - (p_3 - p_2)\bar{x}_{13}}{p_2}, \frac{p_4 b - \sum_{i=3}^{4}(p_i - p_{i-1})\bar{x}_{1i}}{p_2}\right\} \\ &= \min\left\{\frac{(0.75\times 2) - (0.25\times 0.8)}{0.5}, \frac{2 - [0.25\times(0.8+0)]}{0.5}\right\} = 2.6. \end{aligned}$$

Note that $f(b) = 1.268 \times (2.614 - 2) = 0.779$ and $f(u) = 1.268 \times (2.614 - 2.6) = 0.018$, so that substituting into expression (3), $\hat{\theta}_{2,UMV} = 2.839$.

For the unbiased estimator ($\hat{\theta}_{2,U}$) given by expression (4), we make the following calculations. The naive estimates for partitions 1 and 2 are $\hat{\delta}_{1,N} = [(90 \times 3) + (120 \times 3)]/210 = 3$ and $\hat{\delta}_{2,N} = [(90 \times 2) + (120 \times 2.4)]/210 = 2.229$, respectively. Note that $\frac{\tau_{21}^2}{\sqrt{\tau_{11}^2+\tau_{21}^2}} = \frac{\tau_{22}^2}{\sqrt{\tau_{21}^2+\tau_{22}^2}} = 0.837$ and $\frac{\sqrt{\tau_{11}^2+\tau_{21}^2}}{\tau_{11}^2} = \frac{\sqrt{\tau_{21}^2+\tau_{22}^2}}{\tau_{21}^2} = 0.896$. Since $p_i - p_{i-1} = 0.25\,(i = 1, \ldots, 4)$, $v_1 = \frac{1}{0.25}\left(p_2 b - 0.25\sum_{\substack{i=1\\i\neq 1}}^{2}\bar{x}_{1i}\right) = 4 \times [(0.5 \times 0.2) - (0.25 \times 2)] = 2$ and $v_2 = \frac{1}{0.25}\left(p_2 b - 0.25\sum_{\substack{i=1\\i\neq 2}}^{2}\bar{x}_{1i}\right) = 4 \times [(0.5 \times 0.2) - (0.25 \times 3)] = 1$, respectively. For partition 1,

$$\begin{aligned} w_1 &= \min\left\{\frac{p_3 b - 0.25\sum_{\substack{i=1\\i\neq 1}}^{3}\bar{x}_{1i}}{0.25}, \frac{p_4 b - 0.25\sum_{\substack{i=1\\i\neq 1}}^{4}\bar{x}_{1i}}{0.25}\right\} \\ &= \min\left\{\frac{(0.75\times 2) - 0.25\times(2+0.8)}{0.25}, \frac{2 - 0.25\times(2+0.8+0)}{0.25}\right\} = 3.2. \end{aligned}$$

Similarly, for partition 2, $w_2 = 2.2$. Then, for partition 1, $f(v_1) = 0.896 \times (3 - 2) = 0.896$ and $f(w_1) = 0.896 \times (3 - 3.2) = -0.179$, and for partition 2, $f(w_2) = 0.896 \times (2.229 - 1) = 1.101$ and $f(w_2) = 0.896 \times (2.229 - 2.2) = 0.026$. Now, we have all components required to obtain UMVCUEs for the effects in partitions 1 and 2, which give $\hat{\delta}_{1,UMVCUE} = 3.272$ and $\hat{\delta}_{2,UMVCUE} = 2.657$, respectively. The unbiased estimate is the weighted sum of the UMVCUEs in the partitions giving $\hat{\theta}_{2,U} = 2.965$.

The estimates $\hat{\theta}_{2,UMV}$, $\hat{\theta}_{2,U}$, $\hat{\theta}_{2,SI}$, and $\hat{\theta}_{2,MI}$ are greater than $\hat{\theta}_{2,N}$ (see Table 2). This may be explained by the observation in Section 5.2 that, in some scenarios, the naive estimator is negatively biased. The estimate $\hat{\theta}_{2,SI}$ is slightly smaller than $\hat{\theta}_{2,MI}$. Again, this may be explained by an observation in Section 5.2 that, for all scenarios in the simulation study, on average, the single-iteration bias-adjusted estimator gives a smaller estimate than the multiple-iteration estimator.

# 5 | SIMULATIONS TO COMPARE THE VARIOUS ESTIMATORS

## 5.1 | Simulations setting

To evaluate the properties of the various estimators, we conducted simulations with $\sigma^2 = 1$ and $b = 0$. We initially consider the case of $K = 4$ and $p_i - p_{i-1} = 0.25$ ($i = 1, \dots, 4$). In all simulations, if the trial continues to stage 2, the combined stages 1 and 2 sample size is set to be 800. For example, if the stage 1 sample size is 400 patients, the stage 2 sample size is 400. The available patients in stage 1 are equally split among the four partitions and treatment arms. For example, with 400 patients in stage 1, in each partition, 50 patients are randomly allocated to each of the control and experimental treatment. Similarly, the patients available for testing in stage 2 are equally split among the partitions that continue to stage 2 and among the treatment arms. Hence, with 400 patients available in stage 2, if $F$ is selected, the patient allocation in stage 2 is as in stage 1 with 400 patients. If $S_2$ is selected so that two partitions are tested in stage 2, in each partition, 100 patients are randomly allocated to each of the control and experimental treatment. We perform simulations for three cases of stage 1 sample size (200, 400, and 600 patients). Taking the combined stages 1 and 2 to be 800 patients is justified in the supplementary material.

We consider seven scenarios with true treatment effects as summarized in Table 3. The selection rule and estimators developed are aimed at identifying predictive effects, but since we are estimating mean differences, the methods are valid with or without prognostic effects. If the biomarker has no predictive effect but has a prognostic effect, we are in a scenario of equal treatment effects in all partitions. Scenarios 1, 3, and 7 could be such cases. If there are prognostic and predictive effects, we are in a scenario of unequal treatment effects in partitions. Scenarios 2, 4, 5, and 6 could be such cases. In Scenarios 1 to 3, the right decision is to continue to stage 2 with $F$, but with decreasing probability of selecting $F$. The right decisions for Scenarios 4 to 6 are to continue with $S_3$, $S_2$, and $S_1$, respectively. The ideal decision for Scenario 7 is to stop at stage 1. The probabilities for various decisions for different scenarios when stage 1 includes 200 patients (25 in each treatment arm in each partition) are also given in Table 3. These have been calculated using expressions in Section 2.4 and in the supplementary material. As expected, the probability of stopping the trial at stage 1 (last column) increases as the treatment effects in partitions become less than $b$ in more partitions (from 0.007 for Scenario 1 to 0.482 for Scenario 7). In each of Scenarios 4 to 6, the probability of continuing with $F$ is substantially larger than the probability of making the right decision, demonstrating that, in some configurations, decision making is challenging. In Section 5.2.1, simulations show that incorrect decisions tend to be made when observed means are substantially different from the true means and hence lead to bias.

Table 4 gives probabilities of various decisions when the stage 1 sample sizes are 400 and 600. For scenario 3, where treatment effects are equal in all partitions and equal to the futility boundary, the probabilities of various decisions are approximately equal for different stage 1 sample sizes. For the other scenarios, by comparing the probabilities in bold, the probability of making a correct decision increases with stage 1 sample size.

For each of the seven scenarios and three different stage 1 sample sizes, we simulated stage 1 data for $N = 1\,000\,000$ trials. For each trial, the subpopulation with the largest simulated sample mean difference $\geq 0$ continues to stage 2. If no subpopulation fulfills this, the trial stops. We consider estimation conditional on continuing to stage 2 and so bias and MSE for each estimator are evaluated based on simulated trials that continue to stage 2. Using $\hat{\theta}_{s,SI}$ for illustration, for each $s (s = 1, \dots, 4)$, bias and MSE are calculated as $\text{bias}(\hat{\theta}_{s,SI}) = \sum_{i=1}^{N} (\hat{\theta}_{i,SI} - \theta_i) \mathbf{1}_{[i=s]} / \sum_{i=1}^{N} \mathbf{1}_{[i=s]}$ and $\text{MSE}(\hat{\theta}_{s,SI}) = \sum_{i=1}^{N} (\hat{\theta}_{i,SI} - \theta_i)^2 \mathbf{1}_{[i=s]} / \sum_{i=1}^{N} \mathbf{1}_{[i=s]}$.

**TABLE 3** Treatment effects and probabilities of different decisions for the various scenarios in the simulation study (probabilities of correct decisions are in bold)

| | Treatment Effect | | | | | Probability of a Decision ($n_1 = 200$) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | $\delta_1$ | $\delta_2$ | $\delta_3$ | $\delta_4$ | Ideal Selection | $F$ | $S_3$ | $S_2$ | $S_1$ | Stop |
| 1 | 0.3 | 0.3 | 0.3 | 0.3 | $F$ | **0.983** | 0.005 | 0.003 | 0.002 | 0.007 |
| 2 | 0.2 | 0.1 | 0.1 | 0.1 | $F$ | **0.812** | 0.049 | 0.035 | 0.034 | 0.070 |
| 3 | 0.0 | 0.0 | 0.0 | 0.0 | $F$ | **0.500** | 0.083 | 0.070 | 0.073 | 0.274 |
| 4 | 0.1 | 0.0 | 0.0 | −0.2 | $S_3$ | 0.430 | **0.179** | 0.093 | 0.093 | 0.205 |
| 5 | 0.1 | 0.0 | −0.2 | −0.1 | $S_2$ | 0.362 | 0.112 | **0.179** | 0.115 | 0.232 |
| 6 | 0.1 | −0.2 | −0.1 | −0.1 | $S_1$ | 0.298 | 0.098 | 0.104 | **0.214** | 0.286 |
| 7 | −0.1 | −0.1 | −0.1 | −0.1 | Stop | 0.240 | 0.083 | 0.087 | 0.108 | **0.482** |

**TABLE 4** Probabilities of different decisions for different stage 1 sample sizes for various scenarios in the simulation study (probabilities of correct decisions are in bold)

|  | Ideal | Probability of a Decision ($n_1 = 400$) |  |  |  |  | Probability of a Decision ($n_1 = 600$) |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Scenario** | **Selection** | $F$ | $S_3$ | $S_2$ | $S_1$ | **Stop** | $F$ | $S_3$ | $S_2$ | $S_1$ | **Stop** |
| 1 | $F$ | **0.9987** | 0.0004 | 0.0002 | 0.0002 | 0.0005 | **0.99988** | 0.00004 | 0.00002 | 0.00002 | 0.00004 |
| 2 | $F$ | **0.8944** | 0.0312 | 0.0212 | 0.0200 | 0.0332 | **0.93711** | 0.02033 | 0.01326 | 0.01213 | 0.01717 |
| 3 | $F$ | **0.5000** | 0.0833 | 0.0698 | 0.0734 | 0.2735 | **0.50000** | 0.08333 | 0.06981 | 0.07342 | 0.27344 |
| 4 | $S_3$ | 0.4013 | **0.2286** | 0.0983 | 0.0971 | 0.1747 | 0.37973 | **0.26859** | 0.10095 | 0.09838 | 0.15235 |
| 5 | $S_2$ | 0.3085 | 0.1220 | **0.2386** | 0.1261 | 0.2048 | 0.27015 | 0.12853 | **0.28802** | 0.13147 | 0.18183 |
| 6 | $S_1$ | 0.2266 | 0.0977 | 0.1156 | **0.2939** | 0.2662 | 0.17916 | 0.09454 | 0.12250 | **0.35893** | 0.24487 |
| 7 | Stop | 0.1587 | 0.0724 | 0.0842 | 0.1157 | **0.5690** | 0.11034 | 0.06193 | 0.07895 | 0.11756 | **0.63122** |

## 5.2 | Simulation results

### 5.2.1 | Comparing biases for the various estimators

Figure 2 summarizes biases when the stage 1 sample size is 200 (top plots) and 600 patients (bottom plots). Plots for the case where the stage 1 sample size is 400 patients are provided in the supplementary material. Plots in Columns 1 to 4 correspond to the cases of selecting $F$, $S_3$, $S_2$, and $S_1$, respectively. The $y$-axes correspond to biases divided by approximate standard errors (SEs). The approximate SE $= \sqrt{4/(m_{1s} + m_{2s})}$ and so SEs are only equal when $F$ is selected (Column 1). Although SEs are not equal, we will later observe from the boxplots of the estimates that the trend for bias is the same when bias is not divided by SE. The $x$-axes correspond to the seven scenarios. As per the legend, biases for different estimators are distinguished by different line types. Estimators $\hat{\theta}_{s,UMV}$ and $\hat{\theta}_{s,U}$ are not included in Figure 2 because they are mean unbiased. For Scenario 1, the probabilities for selecting $S_3$, $S_2$, and $S_1$ are low and so simulations results are highly variable when $S_3$, $S_2$, or $S_1$ is selected but this does not change the general findings in this paper.

We first describe the results for the case where the stage 1 sample size is 200 (top row). When $F$ is selected, the naive estimator ($\hat{\theta}_{s,N}$) and the first shrinkage estimator ($\hat{\theta}_{s,L_1}$) are the same and correspond to the line showing the largest biases. Focusing on the naive estimator, the bias when $F$ is selected (Column 1) is positive in all scenarios. For scenarios where the right decision is to continue with $F$ (Scenarios 1 to 3, see Table 3), bias when $F$ is selected is attributable to the futility rule with the bias negligible when the effect in $F$ is substantially larger than the futility boundary (Scenario 1). When the right decision is not to continue with $F$ (Scenarios 4 to 7) but $F$ is selected, the impact of selection and futility on bias would increase and consequently give a larger bias. Still focusing on the top row, when $S_3$ is selected (Column 2), the naive estimator for $\theta_3$ is negatively biased for some scenarios and positively biased for other scenarios. The explanation for this pattern is given in the supplementary material. Comparing the bias when $F$, $S_3$, $S_2$, and $S_1$ are selected (Columns 1 to 4), the bias is smallest when $S_1$ is selected. This can be attributed partly to the enrichment, where the stage 2 sample size is fixed regardless of the size of the population selected so that when $S_1$ is selected, proportionally, there are more unbiased stage 2 data to estimate $\theta_1$ compared to when $F$, $S_3$, or $S_2$ is selected. In summary, note that, in some scenarios, the bias of the naive estimator is substantial and so it is essential to use an estimator that corrects for subpopulation selection.

Still focusing on the top row, when $F$ is selected, practically, the single-iteration bias corrected estimator $\hat{\theta}_{s,SI}$ is mean unbiased, especially for Scenarios 1 to 3 where the correct decision is to select $F$. When $S_3$ is selected, $\hat{\theta}_{s,SI}$ almost eradicates bias in Scenarios 3 to 7 and is better than the naive estimator in Scenarios 1 and 2. When $S_2$ or $S_1$ is selected, $\hat{\theta}_{s,SI}$ eradicates almost all bias in Scenarios 2 to 7 but does not do so in Scenario 1. In all scenarios, the line for the multiple-iteration bias-adjusted estimator ($\hat{\theta}_{s,MI}$) is always slightly above that of $\hat{\theta}_{s,SI}$. Hence, comparing $\hat{\theta}_{s,SI}$ and $\hat{\theta}_{s,MI}$, when $\hat{\theta}_{s,SI}$ is negatively biased, $\hat{\theta}_{s,MI}$ is preferable, whereas $\hat{\theta}_{s,SI}$ is preferable when it is positively biased.

Comparing biases for the naive estimator for different stage 1 sample sizes (top versus bottom plots), as also indicated by expression (2), the bias increases with the proportion of stage 1 data. Increase in bias is also seen for both the single-iteration ($\hat{\theta}_{s,SI}$) and multiple-iteration ($\hat{\theta}_{s,MI}$) bias-adjusted estimators. From the bottom row, $\hat{\theta}_{s,SI}$ and $\hat{\theta}_{s,MI}$ perform worst when some partitions that should be dropped at stage 1 continue to stage 2 or when some partitions that should continue to stage 2 are dropped. As before, the line for $\hat{\theta}_{s,MI}$ is above that of $\hat{\theta}_{s,SI}$ with the distances between the lines increasing with stage 1 sample size.

The pattern of the shrinkage estimators is best understood by considering all results in Figure 2. In all cases, the line for the first shrinkage estimator ($\hat{\theta}_{s,L_1}$) overlaps or is above that of the second shrinkage estimator ($\hat{\theta}_{s,L_2}$). Estimator $\hat{\theta}_{s,L_1}$

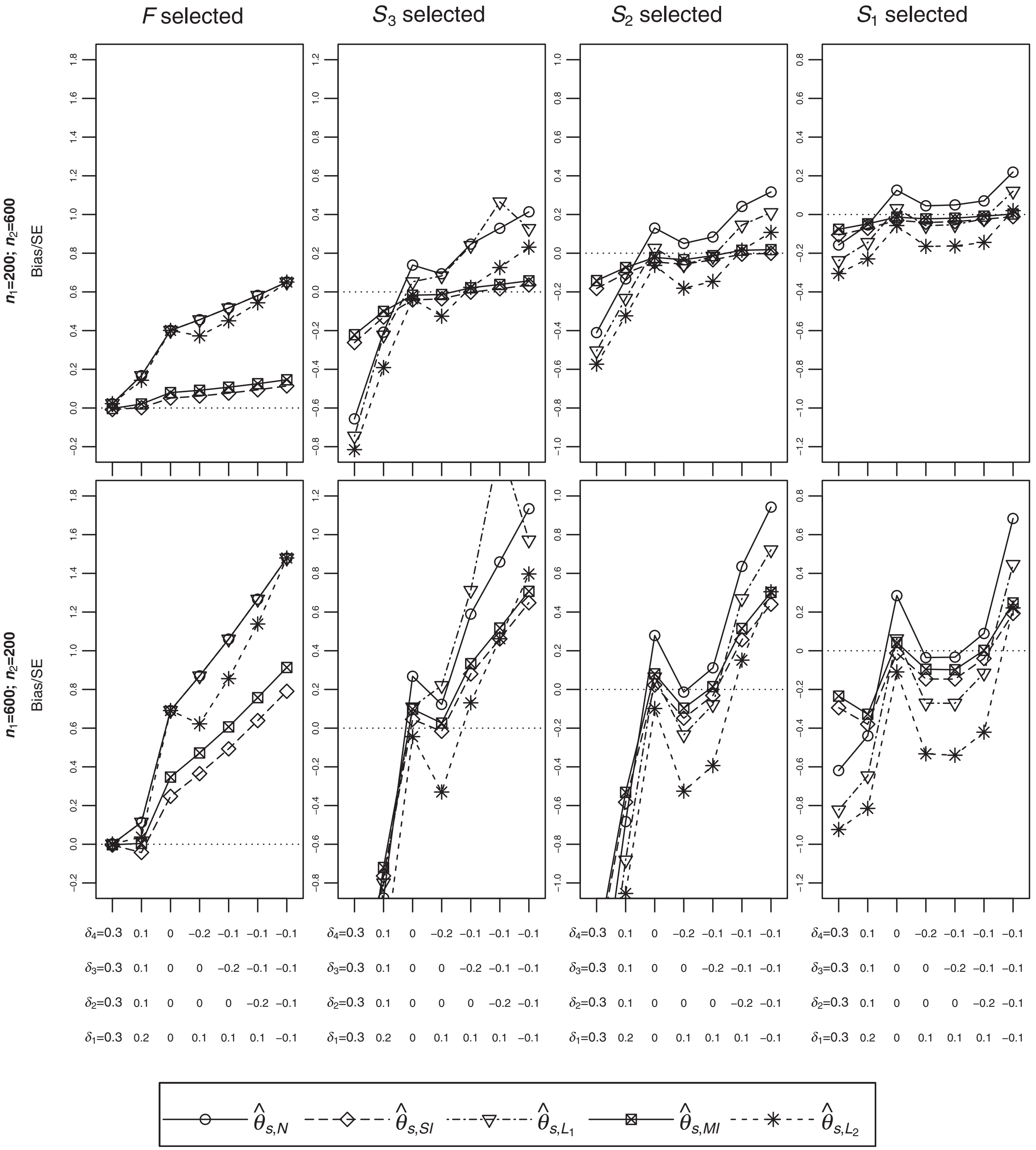

**FIGURE 2** Biases in units of approximate standard error for different configurations. The dotted line is the point of no bias. Other line types correspond to different estimators. SE, standard error

performs similar to or better than $\hat{\theta}_{s,L_2}$ when the selected subpopulation consists of partitions that should continue to stage 2 such as when $F$ is selected in Scenarios 1 to 3 and such as when $S_3$ is selected in Scenarios 1 to 4. Estimator $\hat{\theta}_{s,L_2}$ performs better than $\hat{\theta}_{s,L_1}$ when the selected subpopulation consists of partitions that should not continue to stage 2 such as when $F$ is selected in Scenarios 4 to 7 and such as when $S_3$ is selected in Scenarios 5 to 7.

In almost all scenarios, the two shrinkage estimators perform worse than the other estimators that account for adaptation. One reason for this may be the fact that the shrinkage estimators do not account for stopping for futility. When

$F$ is selected, the naive estimator is the same as the first shrinkage estimator. This is because the stage 1 estimate in partition $i$ is $\hat{C}_{+}\bar{X}_{1i} + (1 - \hat{C}_{+})\bar{Y}_{1K}$ so that the shrinkage estimator shrinks to the effect in the full population, that is, to $\bar{Y}_{1K} = (\bar{X}_{11} + \cdots + \bar{X}_{1K})/K$. A reasonable alternative would be to use a weighted mean of $\bar{Y}_{11}, \bar{Y}_{12}, \ldots, \bar{Y}_{1K}$. For example, if we shrink to $(\bar{Y}_{11} + \cdots + \bar{Y}_{1K})/K$, in terms of sample means in partitions, we are shrinking to a weighted sum such that for $i < i'$, $\bar{X}_{1i}$ has more weight than $\bar{X}_{1i'}$. In such a case, shrinkage estimators will be closer to the naive estimators when fewer partitions are selected (see additional simulations in the supplementary material).

### 5.2.2 | Comparing MSEs for the various estimators

Mean squared errors for the various estimators are given in Figure 3. The $y$-axes are root mean squares ($\text{RMSE} = \sqrt{\text{MSE}}$) divided by approximate SEs. The best shrinkage estimator in terms of bias (either $\hat{\theta}_{s,L_1}$ or $\hat{\theta}_{s,L_2}$ depending on the scenario) has smaller or practically the same MSEs as the naive estimator. Hence, the best shrinkage estimators may be considered to be better than the naive estimator in terms of MSE. The challenge, however, is determining the best shrinkage estimator since the true treatment means are unknown.

Since estimators that extend the works of Kimani et al ($\hat{\theta}_{s,U}$) and of Robertson et al ($\hat{\theta}_{s,UMV}$) are mean unbiased, their MSEs are variances. When $S_1$ is selected, by derivation, the two estimators are the same and, hence, have equal MSE. For any other selection, as expected, $\hat{\theta}_{s,UMV}$ has smaller MSE than $\hat{\theta}_{s,U}$. The differences increase with stage 1 sample size (top versus bottom plots) and the size of the selected subpopulation (right to left panels). The MSEs of $\hat{\theta}_{s,U}$ and $\hat{\theta}_{s,UMV}$ are mostly larger than the MSEs for all the other estimators with the differences substantial when selection is performed later in the trial.

In general, the MSEs for the single-iteration ($\hat{\theta}_{s,SI}$) and multiple-iteration ($\hat{\theta}_{s,MI}$) bias-adjusted estimators are practically the same. Hence, since their biases are also similar, the two estimators are approximately equivalent and so it is sufficient to compare one of them to the other estimators. The MSE for $\hat{\theta}_{s,SI}$ is larger than that of the naive estimator ($\hat{\theta}_{s,N}$) in most cases while it is always smaller than the MSEs for the unbiased estimators ($\hat{\theta}_{s,U}$ and $\hat{\theta}_{s,UMV}$).

### 5.2.3 | Comparing the estimators using both bias and MSE

Comparing the shrinkage estimators ($\hat{\theta}_{s,L_1}$ and $\hat{\theta}_{s,L_2}$) to the naive estimator ($\hat{\theta}_{s,N}$), we prefer $\hat{\theta}_{s,N}$. This is because although a shrinkage estimator sometimes has a smaller MSE, it can have substantially higher bias than $\hat{\theta}_{s,N}$ (for example, compare Columns 4 in Figures 2 and 3).

Comparing the single-iteration bias-adjusted estimator ($\hat{\theta}_{s,SI}$) and the naive estimator ($\hat{\theta}_{s,N}$), when $F$ is selected, $\hat{\theta}_{s,SI}$ is preferable as it reduces bias substantially and has smaller MSE. However, when $S_1$ is selected, $\hat{\theta}_{s,N}$ is better as it has smaller MSE and it does not differ from $\hat{\theta}_{s,SI}$ in terms of bias. When $S_3$ or $S_2$ is selected, $\hat{\theta}_{s,N}$ is better when bias is not substantial (Scenarios 3 and 4), whereas for Scenarios 5 to 7, $\hat{\theta}_{s,SI}$ is better as it reduces bias and its MSE is better or only slightly higher than that of $\hat{\theta}_{s,N}$. Overall, we consider $\hat{\theta}_{s,SI}$ as a better estimator than $\hat{\theta}_{s,N}$ as it performs better in cases with substantial bias.

When $F$ is selected, the bias of the naive estimator ($\hat{\theta}_{s,N}$) is substantial and compared to the UMVCUE ($\hat{\theta}_{s,UMV}$), we prefer the latter since the difference in RMSE between the two estimators is smaller than the bias eradicated. When $S_1$ is selected, we would also recommend $\hat{\theta}_{s,UMV}$ over $\hat{\theta}_{s,N}$ as the former is mean unbiased in all scenarios, with the only case where it is not clearly superior due to high RMSE being when $n_1 = 600$. The conclusion when $S_3$ or $S_2$ is selected is the same as when $S_1$ is selected, that is, $\hat{\theta}_{s,UMV}$ is better than $\hat{\theta}_{s,N}$.

Comparing the single-iteration bias-adjusted estimator $\hat{\theta}_{s,SI}$ to the UMVCUE $\hat{\theta}_{s,UMV}$, we recommend the latter since, when $F$ is selected, $\hat{\theta}_{s,SI}$ has substantial bias that is larger than the difference in RMSE between it and $\hat{\theta}_{s,UMV}$. In addition, when $S_1$ is selected, the difference in RMSE between the two estimators is smaller than the bias of $\hat{\theta}_{s,SI}$. Consequently, based on the performance across the scenarios in the simulation study, we recommend $\hat{\theta}_{s,UMV}$ when an adaptive threshold enrichment design is used.

For a more detailed comparison of the estimators, Figures 4 and 5 give boxplots of simulated estimates for Scenarios 1 (top plots), 4 (middle plots), and 6 (bottom plots) described in Table 3 when $F$ and $S_3$ are selected. The boxplots emphasize the findings summarized above. As an example, when $n_1 = 600$ (Figure 5), for Scenario 6 (bottom left panel), almost all naive estimates are above the true value and $\hat{\theta}_{s,UMV}$ performs well in that case. From the left panels, we note the unbiased estimators ($\hat{\theta}_{s,UMV}$ and $\hat{\theta}_{s,U}$) have substantially higher variances compared to the other estimators.

### 5.2.4 | Summary findings and recommendations from the simulation study

The bias of the naive estimator can be substantial, and so it is essential to use an estimator that corrects for the decision made using stage 1 data. We recommend the estimator that follows the work of Robertson et al ($\hat{\theta}_{s,UMV}$) since it is mean unbiased. Although it has larger MSE than some estimators, the bias eradicated in most cases was larger than the difference in RMSEs. Although the simulation study was based on four partitions and specific treatment effect

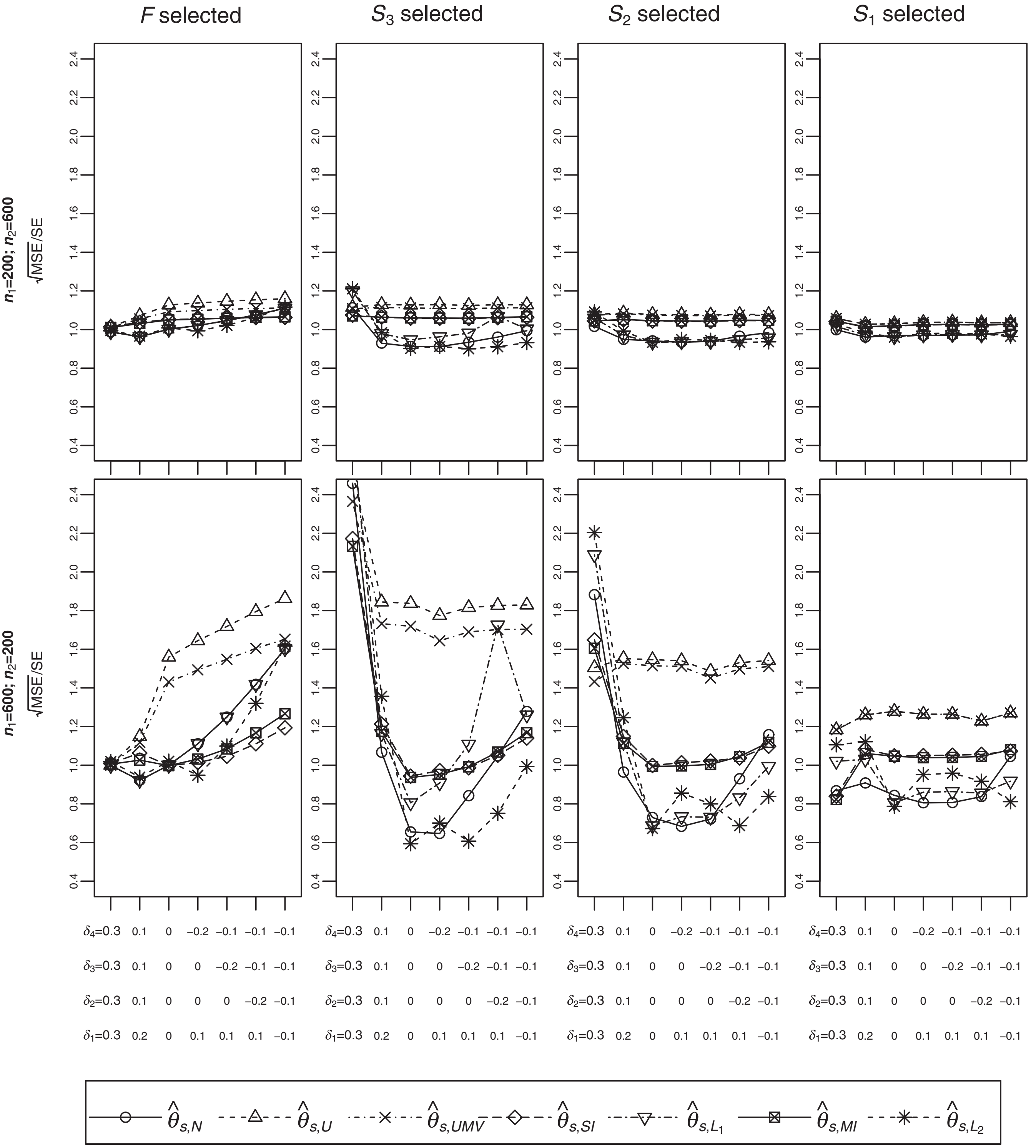

**FIGURE 3** Root mean squares in units of approximate standard error for different configurations. Different line types correspond to different estimators. MSE, mean squared error; SE, standard error

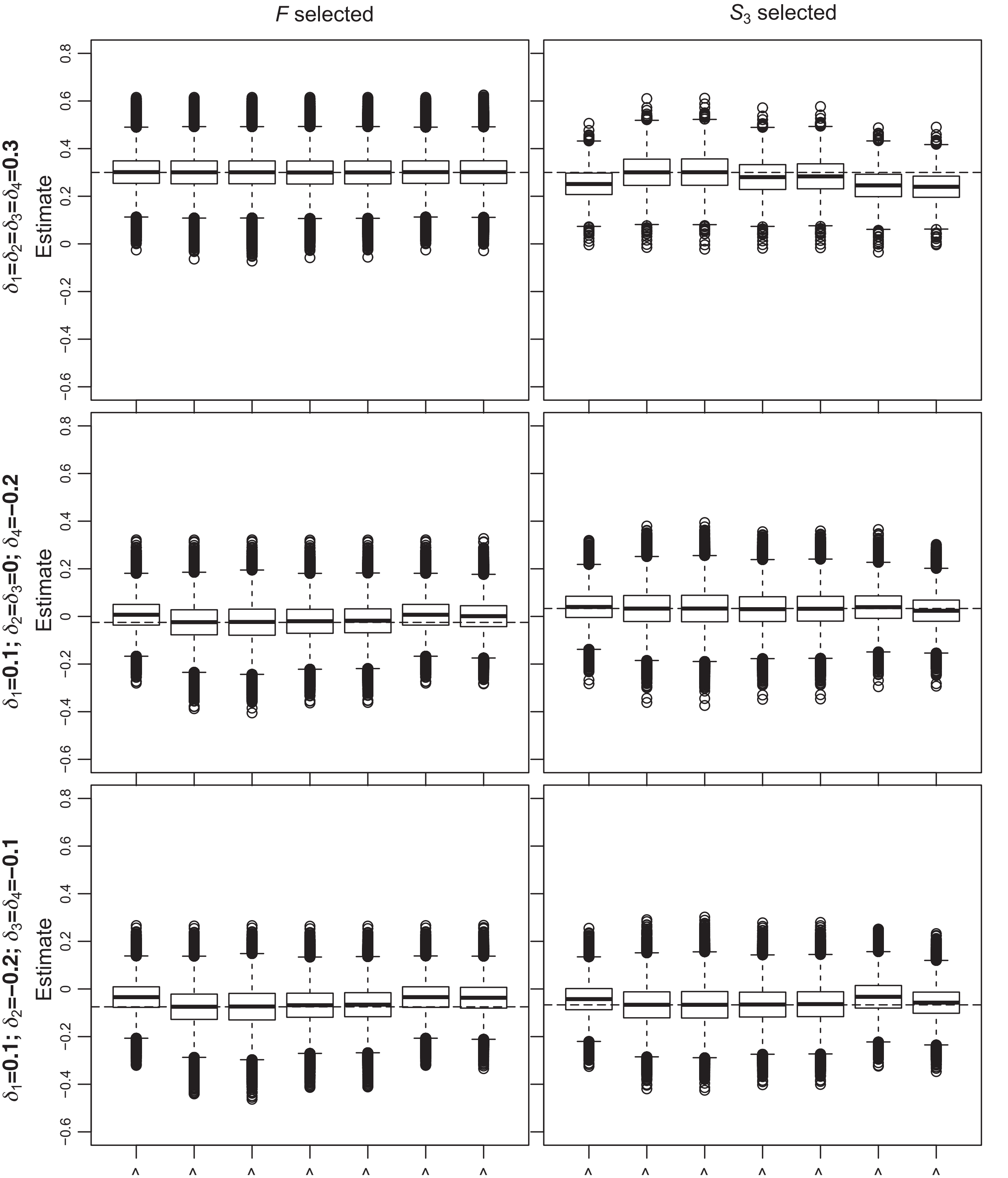

**FIGURE 4** Boxplots of estimates for different estimators when $n_1 = 200$. Results have been chosen when $F$ and $S_3$ were selected and for Scenarios 1 (top panels), 4 (middle panels), and 6 (bottom panels). The dashed lines correspond to the true means in the selected subpopulation

scenarios, we expect similar findings for other configurations (that is, more candidate partitions and/or different effect sizes). Simulations for the case of 8 partitions are in the supplementary material.

We have recommended one estimator for all scenarios. An alternative is a hybrid estimator where the recommended estimator ($\hat{\theta}_{s,SI}$ or $\hat{\theta}_{s,UMV}$) depends on the subpopulation selected. This is suitable if investigators are willing to sacrifice unbiasedness for more precision. In this case, before the trial, a simulation study based on plausible scenarios would be required to compare bias and MSE conditional on the selected subpopulation.

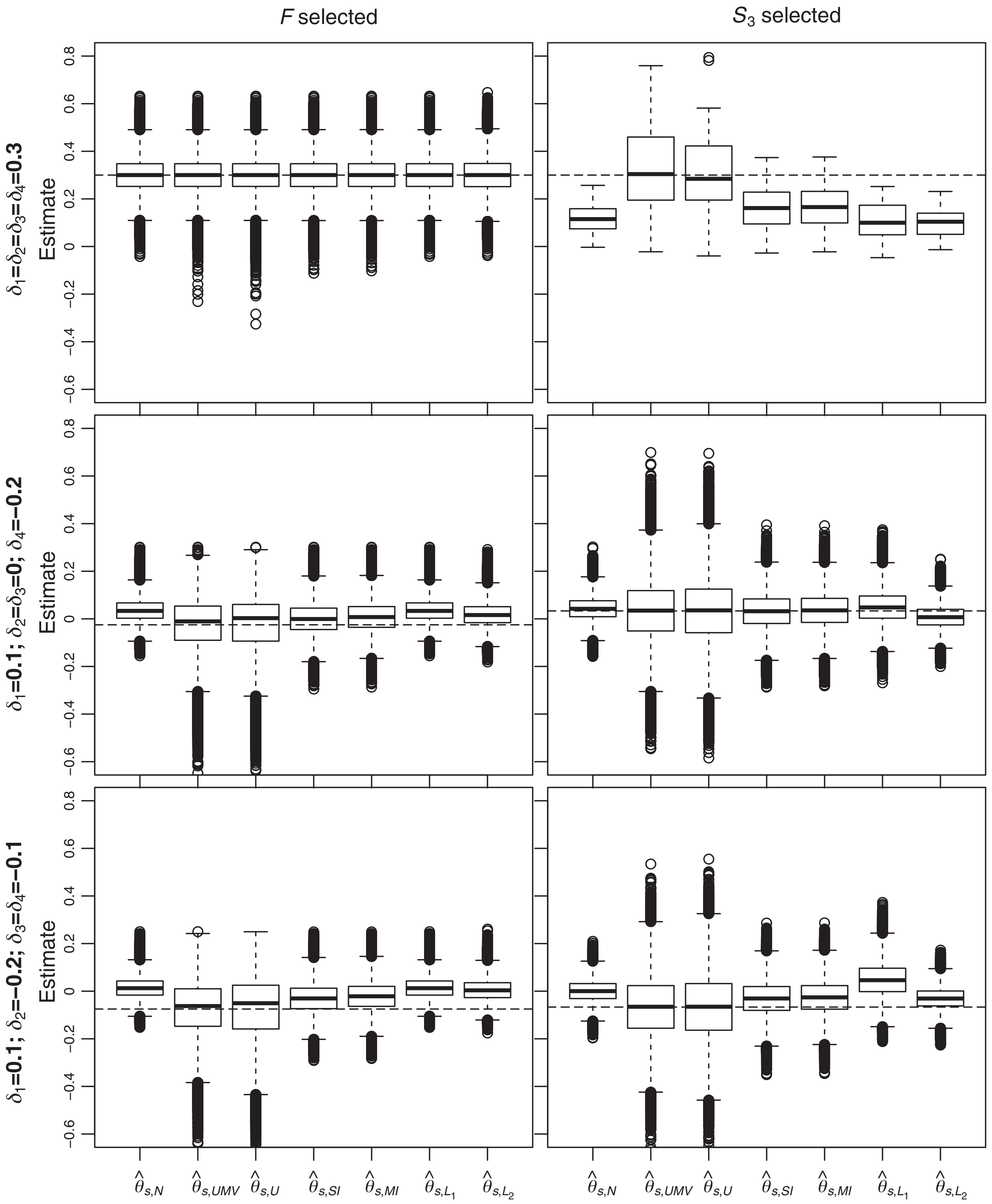

**FIGURE 5** Boxplots of estimates for different estimators when $n_1 = 600$. Results have been chosen when $F$ and $S_3$ were selected and for Scenario 1 (top panels), 4 (middle panels), and 6 (bottom panels). The dashed lines correspond to the true means in the selected subpopulation

## 6 | DISCUSSION

Acknowledging that different patients may require different care has led to trial designs that incorporate assessment of treatment effects in different subsets of the population. Most statistical methodologies for such designs focus on hypothesis testing.[2,5,24,26,29-33] In this paper, we have considered point estimation following an adaptive threshold enrichment clinical trial. We have assessed bias for the naive estimator when different subpopulations are selected. Depending on the scenario, the bias of the naive estimator of the treatment effect in the selected subpopulation is substantial and can be negative or positive. There is thus a need for new estimators. Building on estimators that have been proposed for treatment selection, we have derived several estimators that account for subpopulation selection. By derivation, two estimators are mean unbiased. In this paper, we have recommended the best among these two, that is, the UMVCUE. An alternative is a hybrid estimator where different estimators are recommended based on the selected subpopulation. This would require a simulation study before the trial and is suitable if investigators can accept some unbiasedness for a more precise estimator.

We have considered a specific selection rule but the proposed estimators can be modified for other selection rules. For example, it may be desired that different subpopulations have different futility boundaries. Futility boundaries may be based on factors such as subpopulation prevalence, and sponsor and public health gains.[34] Another factor is safety where the futility boundary may be chosen to reflect investigators' willingness to accept moderate efficacy if the new treatment is substantially safer than the control. The selection rule we have used specifies that a higher biomarker value leads to a smaller treatment effect. If this is a misspecification of the relationship between the biomarker and treatment effect, the unbiased estimators will remain so because we condition on the selection rule. However, the probability of making the right decision will be low and we anticipate that the naive estimator will have more bias and that the unbiased estimators will have higher MSE.

In the derivations, we have not required the prevalences in different partitions to be equal. If the biomarker values are approximately continuous, then it is reasonable to subdivide the full population into equal partitions as we have done in the example and the simulations. Other numerical biomarker values may be discrete with few possible values, leading to partitions with varying sizes.

We have assumed the number of patients in each partition, and hence prevalence, is known. For the case of two partitions and a fixed cut-off value, taking the stage 1 number of patients in a partition to have a binomial distribution, Kimani et al[7] showed that using stage 1 prevalence estimates in the expressions for the unbiased estimators provides unbiased estimates for the treatment effects. This extends to the case of more than two partitions, where numbers of patients in partitions are taken to have a multinomial distribution. The proof is based on the fact that the estimator in a partition is unbiased conditional on the number of patients in an interval and that the proportion of patients in a partition is unbiased for the prevalence in the partition. The proof for the case of estimating the cut-off values using stage 1 data is similar.

Conditional on continuing to stage 2, we have derived estimators for the effect in the selected subpopulation. Continuing to stage 2 is necessary for the unbiased estimators. This is not the case for the other estimators as they involve obtaining stage 1 estimates in all partitions that correct for the subpopulation selection and then combine them with the stage 2 unbiased estimates. Hence, estimates for effects in the dropped partitions that correct for subpopulation selection can be obtained using the shrinkage and bias-adjusted estimators. However, they are not necessarily mean unbiased.

Methods developed for normally distributed data following treatment selection have been adapted for time-to-event data.[28] Even after assuming asymptotic normality of the log hazard ratio, some of the estimators we have derived such as the UMVCUE may not be valid for time-to-event data. For example, if there is a quantitative interaction with hazard ratios in different partitions being unequal, a model that accounts for this is required. In this case, obtaining separate estimates for each partition is the valid approach.

Finally, since in all simulations, the combined stages 1 and 2 sample size was 800, for the different stage 1 sample sizes considered, there would be no savings or losses in terms of the cost of treating patients. The saving/loss is only made in terms of costs associated with biomarker testing. Hence, the case for performing subpopulation selection with a small proportion of patients can be justified if the biomarker is expensive, leading to savings if $F$ is selected. The case for performing subpopulation selection with a large proportion of patients is justifiable if the biomarker is not expensive. In this case, the resources loss is not substantial if $F$ is selected and yet, if only a part of the population will benefit, there is a higher probability of making the right decision that may improve power. The setting of fixed total sample size is sometimes referred to as enrichment because if some partitions are dropped in stage 2, the number of patients recruited from partitions in stage 2 is higher than if more partitions were selected. To save money on treatment costs or reduce the total sample size, subpopulation selection could be performed early, with no enrichment in stage 2. With no enrichment, the number of patients in a partition in stage 2 is fixed. The statistical properties of the estimators for the setting with no enrichment can be evaluated as in the case of enrichment.

## ACKNOWLEDGEMENTS

This work was funded by the UK Medical Research Council (grant MR/N028309/1). We thank the editor, the associate editor, and two reviewers for comments that greatly improved this paper.

## ORCID

*Peter K. Kimani* http://orcid.org/0000-0001-8200-3173

## REFERENCES

1. Ballman KV. Biomarker: predictive or prognostic? *J Clin Oncol*. 2015;33:3968-3971.
2. Simon N, Simon R. Adaptive enrichment designs for clinical trials. *Biostatistics*. 2013;14(4):613-625.
3. Mok TS, Wu Y-L, Thongprasert S, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med*. 2009;361(10):947-957.
4. Tran HT, Liu Y, Zurita AJ, et al. Prognostic or predictive plasma cytokines and angiogenic factors for patients treated with pazopanib for metastatic renal-cell cancer: a retrospective analysis of phase 2 and phase 3 trials. *Lancet Oncol*. 2012;13(8):827-837.
5. Renfro LA, Coughlin CM, Grothey AM, Sargent DJ. Adaptive randomized phase II design for biomarker threshold selection and independent evaluation. *Chin Clin Oncol*. 2014;3(1):3.
6. Li W, Chen C, Lia X, Beckmanb RA. Estimation of treatment effect in two-stage confirmatory oncology trials of personalized medicines. *Statist Med*. 2017;36:1843-1861.
7. Kimani PK, Todd S, Stallard N. Estimation after subpopulation selection in adaptive seamless trials. *Statist Med*. 2015;34:2581-2601.
8. Kunzmann K, Benner L, Kieser M. Point estimation in adaptive enrichment designs. *Statist Med*. 2017;36(25):3935-3947.
9. Cohen A, Sackrowitz HB. Two stage conditionally unbiased estimators of the selected mean. *Stat Probab Lett*. 1989;8:273-278.
10. Stallard N, Todd S. Point estimates and confidence regions for sequential trials involving selection. *J Stat Plan Infer*. 2005;135:402-419.
11. Bowden J, Glimm E. Unbiased estimation of selected treatment means in two-stage trials. *Biom J*. 2008;50(4):515-527.
12. Bauer P, Koenig F, Brannath W, Posch M. Selection and bias—two hostile brothers. *Statist Med*. 2010;29:1-13.
13. Kimani PK, Todd S, Stallard N. Conditionally unbiased estimation in phase II/III clinical trials with early stopping for futility. *Statist Med*. 2013;32(17):2893-2910.
14. Carreras M, Brannath W. Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection. *Statist Med*. 2013;32(10):1677-1690.
15. Robertson DS, Prevost AT, Bowden J. Unbiased estimation in seamless phase II/III trials with unequal treatment effect variances and hypothesis-driven selection rules. *Statist Med*. 2016a;35(22):3907-3922.
16. Stallard N, Kimani PK. Uniformly minimum variance conditionally unbiased estimation in multi-arm multi-stage clinical trials. *Biometrika*. 2018;105:495-501.
17. Robertson DS, Prevost AT, Bowden J. Accounting for selection and correlation in the analysis of two-stage genome-wide association studies. *Biostatistics*. 2016b;17(4):634-649.
18. Whitehead J. On the bias of maximum likelihood estimation following a sequential test. *Biometrika*. 1986;73(3):573-581.
19. Hwang JT. Empirical Bayes estimation for the means of the selected populations. *Indian J Stat A*. 1993;55:285-311.
20. Uher R, Tansey KE, Malki K, Perlis RH. Biomarkers predicting treatment outcome in depression: what is clinically significant? *Pharmacogenomics*. 2012;13(2):233-240.
21. Rush AJ, Fava M, Wisniewski SR, et al. Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. *Control Clin Trials*. 2004;25(1):119-142.
22. Uher R, Maier W, Hauser J, et al. Differential efficacy of escitalopram and nortriptyline on dimensional measures of depression. *Br J Psychiatry*. 2009;194:252-259.
23. Tsao AS, Liu S, Lee JJ, et al. Clinical outcomes and biomarker profiles of elderly pretreated NSCLC patients from the BATTLE trial. *J Thorac Oncol*. 2012;7(11):1645-1652.
24. Jiang W, Freidlin B, Simon R. Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *J Natl Cancer Inst*. 2007;99(13):1036-1043.
25. Grothey A, Strosberg JR, Renfro LA. A randomized, double-blind, placebo-controlled phase II study of the efficacy and safety of Monotherapy Ontuxizumab (MORAb-004) plus best supportive care in patients with chemorefractory metastatic colorectal cancer. *Clin Cancer Res*. 2018;24(2):316-325.
26. Joshi A, Zhang J, Fang L. Statistical design for a confirmatory trial with a continuous predictive biomarker: a case study. *Contemp Clin Trials*. 2017;63:19-29.

27. Morris CN. Parametric empirical Bayes inference: theory and applications. *J Am Stat Assoc*. 1983;78:47-55.

28. Brückner M, Titman A, Jaki T. Estimation in multi-arm two-stage trials with treatment selection and time-to-event endpoint. *Statist Med*. 2017;36:3137-3153.

29. Liu A, Liu C, Li Q, Yu KF, Yuan VW. A threshold sample-enrichment approach in a clinical trial with heterogeneous subpopulations. *Clin Trials*. 2010;7:537-545.

30. Wason J, Marshall A, Dunn J, Stein RC, Stallard N. Adaptive designs for clinical trials assessing biomarker-guided treatment strategies. *Br J Cancer*. 2014;110:1950-1957.

31. Spencer AV, Harbron C, Mander A, Wason J, Peers I. An adaptive design for updating the threshold value of a continuous biomarker. *Statist Med*. 2016;35:4909-4923.

32. Antoniou M, Jorgensen AL, Kolamunnage-Dona R. Biomarker-guided adaptive trial designs in phase II and phase III: a methodological review. *PLOS ONE*. 2016;11(2):e0149803.

33. Simon N, Simon R. Using Bayesian modeling in frequentist adaptive enrichment designs. *Biostatistics*. 2018;19(1):27-41.

34. Ondra T, Jobjörnsson S, Beckman RA, et al. Optimizing trial designs for targeted therapies. *PLOS ONE*. 2016;11(9):e0163726.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**How to cite this article:** Kimani PK, Todd S, Renfro LA, Stallard N. Point estimation following two-stage adaptive threshold enrichment clinical trials. *Statistics in Medicine*. 2018;37:3179–3196. https://doi.org/10.1002/sim.7831