A comparison of estimation methods adjusting for selection bias in adaptive enrichment designs with time-to-event endpoints

A comparison of estimation methods adjusting for selection bias in adaptive enrichment designs with time-to-event endpoints

Fulvio Di Stefano[1], Matthieu Pannaux[2], Anne Correges[2], Stephanie Galtier[2], Veronique Robert[2], and Gaelle Saint-Hilary[1,2]

[1]Dipartimento di Scienze Matematiche (DISMA) Giuseppe Luigi Lagrange, Politecnico di Torino, Torino, Italy
[2]Department of Clinical Statistics, Institut de Recherches Internationales Servier, Suresnes, France

Abstract

Adaptive enrichment designs in clinical trials have been developed to enhance drug developments. They permit, at interim analyses during the trial, to select the sub-populations that benefits the most from the treatment. Because of this selection, the naive maximum likelihood estimation of the treatment effect, commonly used in classical randomized controlled trials, is biased. In the literature, several methods have been proposed to obtain a better estimation of the treatments' effects in such contexts. To date, most of the works have focused on normally distributed endpoints, and some estimators have been proposed for time-to-event endpoints but they have not all been compared side-by-side. In this work, we conduct an extensive simulation study, inspired by a real case-study in Heart Failure, to compare the maximum-likelihood estimator (MLE) with an unbiased estimator, shrinkage estimators and bias-adjusted estimators for the estimation of the treatment effect with time-to-event data. The performances of the estimators are evaluated in terms of bias, variance and mean squared error. Based on the results, along with the MLE, we recommend to provide the unbiased estimator and the single-iteration bias-adjusted estimator: the former completely eradicates the selection bias, but is highly variable with respect to a naive estimator; the latter is less biased than the MLE estimator and only slightly more variable.

Keywords: Subpopulation selection, selection bias, interim analysis, point estimation, adaptive design, enrichment designs, survival data

1 Introduction

In recent years, adaptive designs (ADs) clinical trials have been developed to enhance drug developments. They permit, at interim analyses during the trial, to make use of pre-planned modifications that include: refining sample size; stopping the whole trial or single doses for lack of efficacy; stopping the whole trial for success; reshuffling patients among treatment arms; selecting population that would be more likely to benefit from the treatment [1]. In particular, enrichment ADs trials consist of trials with the possibility to select, at an interim analysis, the specific sub-populations that benefits the most from the treatment, optimising the resources on the most promising groups of patients. In such trials, patients are divided in sub-groups according to some biomarker or covariate values (size of tumor, baseline heart rate, etc.) and the efficacy and safety of the treatment is assessed, at interim analyses, in each group and overall. If, according to pre-defined criteria, patients from one or several sub-groups appear to benefit more from the experimental drug than others, the recruitment is then restricted to these patients for the remainder of the trial.

Although using ADs seems a very promising option to follow, this flexibility comes with a cost. It is well known that the selection rule applied in ADs causes a biased estimation of the treatment effect [1, 2, 3]. In Figure 1, obtained similarly to Pallmann et al [1], we see that if the lowest treatment effects are excluded and the highest are not, the treatment effect is overestimated: an estimator that does not take the selection process into account produces positively biased results. Analogously, the estimation obtained from the data which are excluded is biased, but negatively. As a matter of fact, in two-stage

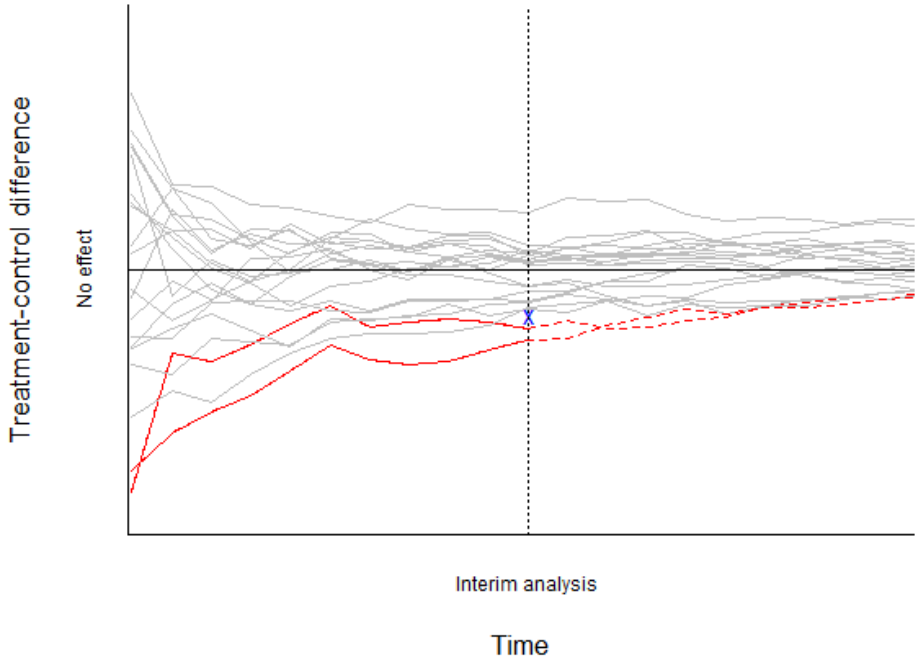

Figure 1: Illustration of bias introduced by early stopping for futility. Obtained similarly to Pallmann et al [1]. Two (in red) of 20 simulated two-arm trials with no true treatment effect are excluded because of the futility threshold (blue cross), resulting in optimistic estimation of the treatment effect.

ADs, the stage 1 naive estimators (obtained using only data before the interim analysis) are biased because of the selection rule applied, while the stage 2 naive estimators (obtained using only data after the interim analysis) are unbiased estimations of the true treatment effect for the selected treatments [3, 4], since no selection rule is applied on these.

In their Guidance for Industry on Adaptive Designs for Clinical Trials of Drugs and Biologics issued in 2019 [5], the FDA outlines that the use of AD trials can have important advantages from an ethical point of view, but that some key principles need to be satisfied in order to meet regulatory approval. In particular, the sponsor is required to evaluate the extent of bias in the estimates and, if available, to pre-specify methods for adjusting estimates to reduce or remove this bias. This is needed to avoid an optimistic estimation of the treatment effect and to control the type I error inflation which derives from the inclusion of the data used for the selection in the final analysis. This regulatory requirement motivated the work presented in this paper.

In the following, we focus on two-stage adaptive designs with sub-population selection (enrichment designs) and time-to-event data. This design permits, at one interim analysis during the trial, to select the sub-population that benefits the most from the treatment. Following the work of Kimani et al [6] on adaptive threshold enrichment clinical trials with normally distributed endpoints, we compare different approaches extended to time-to-event data:

- Unbiased estimators, so called uniformly minimum variance conditional unbiased estimator (UMVCUE), which are developed to find estimations of the true treatment effect with no bias. A first unbiased estimator developed for treatment selection was proposed by Cohen and Sackrowitz in 1989 [7]. Later, their work has been extended by Bowden and Glimm [8]. A version adapted for sub-population selection with time-to-event data has been proposed by Kimani et al [9].
- Shrinkage estimators, which attempt to reduce, but not to eradicate, the bias. They rely on the idea to shrink stage 1 estimates towards the stage 1 overall mean, reducing the bias. We compare two approaches proposed by Carreras and Brannath [4] and Brückner et al [10].
- Bias-Adjusted Estimators, which are mainly proposed by Whitehead [11] and Stallard and Todd

[12]. The main idea of these procedures is to find an estimation of the bias via an iterative procedure and then subtract it from the original naive estimator.

Brückner et al [10] compared some of these estimators in the context of multi-arm two-stage trials with treatment selection and time-to-event endpoint. Kunzmann et al[13] compared 6 other estimators in the context of adaptive enrichment designs with normally distributed endpoints, recommending a hybrid estimator which combines the UMVCUE with a conditional moment estimator as a general rule. Kimani et al [6] compared these estimators in the context of enrichment AD clinical trials with normally distributed endpoints, recommending the use of the unbiased estimator as a general rule. In a subsequent work [9], they derived expressions for an unbiased estimator in a two-stage adaptive design with time-to-event data and focused on the construction of confidence intervals. The goal of this paper is to extend these works and to compare side-by-side the performances of the presented six estimators of the treatment effects in the case of two-stage enrichment adaptive designs with time-to-event data, and to give recommendations on the estimator(s) that, in our opinion, best meet the regulatory requirements and best support internal decision-making. It is worth noting that since the primary interest of this paper is for the correct estimation of the treatment effect in the selected sub-populations, the estimators are given conditionally on the selection made. Unconditional estimators may also be of interest in other settings, but the reduction of the unconditional bias does not guarantee the reduction of the bias conditioned on a specific selection [14], which is our primary interest.

The rest of this paper is organized as follows. The methodology to obtain the six different treatment effect estimators is described in Section 2. In Section 3, we apply the methods to a case-study in cardiology assessing the effect of an experimental drug versus placebo in patients with moderate to severe chronic heart failure. In Section 4, a comprehensive simulation study comparing the performances of the six estimators in terms of bias, variance and mean squared error (MSE) is presented. We conclude with a discussion in Section 5.

2 Methods

We investigate the setting of adaptive clinical trials with two treatment arms (an experimental drug and a control), sub-population selection at one interim analysis and time-to-event data. The idea is to split the patient population according to some biomarker value and analyse the different sub-populations, indexed by $i = 1, ..., K$, separately. We define sub-populations such that the patients inside a sub-population have a biomarker value between predefined upper and lower thresholds: the sub-populations are disjoint. We identify the data before the interim analysis as stage 1 data and the data after the interim analysis as stage 2 data. Let d_{ji} denote the number of events in sub-population $i = 1, ..., K$ at stage $j = 1, 2$. In our setting, d_{2i} does not contain the events in d_{1i}.

Dealing with time-to-event data, we consider the log hazard ratio (HR) between the two treatment arms in each sub-population, defined as $\delta_i = \log\left(\frac{h_{ti}(t)}{h_{ci}(t)}\right)$ for $i = 1, ..., K$, where $h_{ti}(t)$ and $h_{ci}(t)$ are the hazard functions of the treatment and the control in sub-population i, respectively. We use a Cox proportional hazard model to obtain the estimates. We consider here that a negative value of the log HR corresponds to a reduction of risk of event with the treatment, i.e. that the treatment is effective with respect to the control; if the log HR in one sub-population is lower than in another, it means the treatment is more effective in that sub-population. The log HR is usually assumed to be normally distributed, with stage 1 and stage 2 estimators $\hat{\delta}_{1i} \sim N(\delta_i, \tau_{1i}^2)$ and $\hat{\delta}_{2i} \sim N(\delta_i, \tau_{2i}^2)$ for $i = 1, ..., K$; also $\hat{\tau}_{1i}^2$ and $\hat{\tau}_{2i}^2$ are estimated from the Cox model.

We specify a selection rule at the interim analysis as follows: given a threshold value b of the log HR between treatment arms, each sub-population $i \in (1, ..., K)$ does not continue to stage 2 if its stage 1 estimation is not lower than b ($\hat{\delta}_{1i} \geq b$); if stage 1 estimates for all sub-populations are greater or equal than b, the trial is stopped for futility. We define as $\mathcal{S}$ the set of indices corresponding to the selected sub-populations continuing in stage 2.

In this setting, one particularity of adaptive clinical trials with time-to-event data is that some stage 1 patients may not have had the event of interest yet at the time of the interim analysis. If we continue the analysis carrying these patients to stage 2, the stage 1 and stage 2 test statistics will be correlated, inducing some bias in the estimation. As a matter of fact, the independent increment structure, i.e. the independence between the test statistic of stage 1 and stage 2, is assumed for the calculation of the upcoming estimators. If the same patients continue from stage 1 in stage 2, the independent increment structure holds only approximately, even if the time of the interim and final analysis are independent of each other [9, 15]. To avoid any correlation, patients from stage 1 would have to stop the study at the

interim analysis. However, it is obviously not ethical and not applicable in practice that patients stop the study before a minimum treatment period. Instead, we use an intermediate rule following Kimani et al [9] and Jenkins et al [15]. Consider T_1 the time of the interim analysis and T_2 the time of the final analysis, defined here when a certain number of patients had experienced the event. We define $\tilde{T}_1$ (such that $T_1 \leq \tilde{T}_1 \leq T_2$) the time until which the stage 1 patients are followed up. This way, we improve the independent increment structure and we can obtain more accurate estimations. Moreover, it is realistic as, in many therapeutic areas, a maximum follow-up time could be pre-specified for all patients in the protocol. Also, note that since T_2 is fixed in terms of events, the number of events in stage 2 for each selected population d_{2i} for $i \in \mathcal{S}$ depends both on the hazard ratios and the number of partitions selected; however, $\sum_{i\in\mathcal{S}} d_{2i}$ is fixed.

In accordance with Kimani et al [9] and Bruckner et al[10], the following methodology is used to obtain the estimators. From the data from stage 1 only, using the survival times up to the time of the interim analysis (T_1), the estimators $\hat{\delta}_{1i}$ and $\hat{\tau}^2_{1i}$ are estimated directly via the score process of a Cox proportional hazard model. At the end of the trial, using all available evidence, i.e. using the survival times up to the time of the final analysis (T_2), the naive estimators $\hat{\delta}_{i,N}$ and $\hat{\tau}^2_{i,N}$ for the selected sub-populations are also estimated directly with a Cox proportional hazard model. At this point, the stage 2 estimators for the selected sub-populations are calculated as $\hat{\tau}^2_{2i} = \left(\frac{1}{\hat{\tau}^2_{i,N}} - \frac{1}{\hat{\tau}^2_{1i}}\right)^{-1}$ and $\hat{\delta}_{2i} = \hat{\tau}^2_{2i}\left(\frac{\hat{\delta}_{i,N}}{\hat{\tau}^2_{i,N}} - \frac{\hat{\delta}_{1i}}{\hat{\tau}^2_{1i}}\right)$ $\forall i \in \mathcal{S}$.

2.1 Naive estimator

The naive estimator is calculated using the estimator of the Cox proportional hazard model. It is equal to:

$$\hat{\delta}_{i,N} = \frac{\hat{\tau}^2_{2i}\ \hat{\delta}_{1i} + \hat{\tau}^2_{1i}\ \hat{\delta}_{2i}}{\hat{\tau}^2_{1i} + \hat{\tau}^2_{2i}} \qquad \forall i \in \mathcal{S}$$

and $\hat{\delta}_{i,N} = \hat{\delta}_{1i}\ \forall i \notin \mathcal{S}$.

This estimation is biased because of the selection process [1, 2, 3]. In the following, we focus only on handling the bias which arises from the selection. This estimator is calculated assuming that the independent increment structure holds and by pooling all the data available at the end of the study. Therefore, there is also a component of the bias which comes from the use of stage 1 patients also in stage 2, inducing a correlation between the stages. This correlation bias is mostly (but not fully) eradicated by choosing T_1 and T_2 independently of each other and fixing $\tilde{T}_1$ in advance and it is not adjusted further in the following. It results that this estimator is not completely naive, as as correlation between stages has been reduced.

2.2 UMVCUE

Following the work of Kimani et al [9], we calculate the uniformly minimum variance conditional unbiased estimator (UMVCUE) which corrects for the selection bias. However, because of the correlation bias between stages 1 and 2, this estimator would not be perfectly unbiased. For the selected sub-populations, it is:

$$\hat{\delta}_{i,U} = \hat{\delta}_{i,N} - \frac{\hat{\tau}^2_{2i}}{\sqrt{\hat{\tau}^2_{1i} + \hat{\tau}^2_{2i}}} \frac{-\phi(g(b))}{-\Phi(g(b))}, \qquad \forall i \in \mathcal{S}$$

where ϕ and Φ denote the density and cumulative distribution functions of a standard normal distribution, respectively, and $g(x) = \frac{\sqrt{\hat{\tau}^2_{1i}+\hat{\tau}^2_{2i}}}{\hat{\tau}^2_{1i}}\left(\hat{\delta}_{i,N} - x\right)$. This estimation, calculated only for the selected treatments and not available for the dropped ones, eradicates the bias induced by the selection process at the price of an increase in the variance.

2.3 Shrinkage estimators

The aim of this type of estimators is to have a reduction in the bias, with respect to a naive estimator, not increasing the variance. While the stage 2 estimates provide an unbiased estimation of the effects of the treatments, stage 1 estimations are biased because of selection [3, 4]. The idea behind these estimators is to shrink the stage 1 estimates towards the stage 1 overall average log HR to reduce the bias. We consider two shrinkage estimators. The first, denoted by $S1$, was developed by Carreras and Brannath [4]. We define $\hat{\delta}_{1\cdot} = \frac{1}{K}\sum_{i=1}^{K}\hat{\delta}_{1i}$ as the overall stage 1 average log HR and $t_i = \frac{d_{1i}}{d_{1i}+d_{2i}}$ as the information

fraction at the time of the interim analysis. For the selected sub-populations, the shrinkage estimator $S1$ is calculated as:

$$\hat{\delta}_{i,S1} = t_i \left[\hat{C}_i^+ \hat{\delta}_{1i} + (1 - \hat{C}_i^+)\hat{\delta}_{1\cdot}\right] + (1 - t_i)\hat{\delta}_{2i} \qquad \forall i \in \mathcal{S}$$

while they are $\hat{\delta}_{i,S1} = [\hat{C}_i^+ \hat{\delta}_{1i} + (1 - \hat{C}_i^+)\hat{\delta}_{1\cdot}]$ $\forall i \notin \mathcal{S}$, with $\hat{C}_i^+$ defined as follows. If $K \geq 4$:

$$\hat{C}_i^+ = max(0, \hat{C}_i), \qquad \hat{C}_i = 1 - \frac{(K-3)\hat{\tau}_{1i}^2}{\sum_{j=1}^K \left(\hat{\delta}_{1j} - \hat{\delta}_{1\cdot}\right)^2},$$

while if $K = 2, 3$:

$$\hat{C}_i^+ = max(0, \hat{C}_i), \qquad \hat{C}_i = 1 - \frac{(K-1)\hat{\tau}_{1i}^2}{\sum_{j=1}^K \left(\hat{\delta}_{1j} - \hat{\delta}_{1\cdot}\right)^2}.$$

The second shrinkage estimator we consider, denoted by $S2$, was proposed by Brückner et al [10] and is derived in a Bayesian framework. We suppose to have a prior distribution of the vector of true log HRs $\boldsymbol{\delta} = (\delta_1, ..., \delta_K)$, which is a multivariate normal $MVN(\boldsymbol{\mu}, \nu^2\mathbf{I}_K)$ ($\mathbf{I}_K$ is the KxK identity matrix), that is updated with the data ($\hat{\boldsymbol{\delta}}^{Stage1} \sim MVN(\boldsymbol{\delta}, \boldsymbol{\Sigma})$), to get a posterior estimation for $\boldsymbol{\delta}$. The posterior log HR of $\boldsymbol{\delta}$ is $\boldsymbol{C}\hat{\boldsymbol{\delta}}^{Stage1} + (\mathbf{I}_K - \boldsymbol{C})\boldsymbol{\mu}$, where $\boldsymbol{C} = \mathbf{I}_K - \boldsymbol{\Sigma}(\nu^2\mathbf{I}_K + \boldsymbol{\Sigma})^{-1}$. Because sub-populations are disjoint, $\boldsymbol{\Sigma}$ is a diagonal matrix containing the τ_{1i}^2 on the diagonal. Since it is unknown, we use its estimator $\hat{\boldsymbol{\Sigma}}$ with $\hat{\tau}_{1i}^2$ on the diagonal. We define the prior log HR $\boldsymbol{\mu}$ as a vector of length K containing $\hat{\delta}_{1\cdot}$, the overall average stage 1 log HR. An estimate $\hat{\nu}^2$ of ν^2 is obtained iteratively :

- Step 1: Define an initial guess of $\hat{\nu}^2$.
- Step 2: Define weights $w_i = (\hat{\nu}^2 + \hat{\boldsymbol{\Sigma}}_{ii}^2)^{-1}$ for $i = 1, ..., K$.
- Step 3: Update the estimate calculating

$$\hat{\nu}^2 = \frac{\sum_{i=1}^K w_i \left[(\hat{\delta}_{1i} - \hat{\delta}_{1\cdot})^2 - \hat{\boldsymbol{\Sigma}}_{ii}^2\right]}{\sum_{i=1}^K w_i}.$$

- Step 4: If $\hat{\nu}^2 < 0$, set $\hat{\nu}^2 = 0$.
- Step 5: Go back to step 2 using the updated $\hat{\nu}^2$, until convergence.

When the iterative approach converges, we get an estimate $\hat{\nu}^2$ used to calculate $\hat{\boldsymbol{C}} = \mathbf{I}_K - \hat{\boldsymbol{\Sigma}}(\hat{\nu}^2\mathbf{I} + \hat{\boldsymbol{\Sigma}})^{-1}$. Then the stage 1 estimator is:

$$\hat{\boldsymbol{\delta}}_{S2}^{Stage1} = \hat{\boldsymbol{C}}\hat{\boldsymbol{\delta}}^{Stage1} + (\mathbf{I}_K - \hat{\boldsymbol{C}})\mathbb{1}\hat{\delta}_{1\cdot}$$

where $\mathbb{1}$ is the vector with all entries equal to 1. Eventually, the shrinkage estimator is calculated as:

$$\hat{\delta}_{i,S2} = t_i\, \hat{\delta}_{i,S2}^{Stage1} + (1 - t_i)\hat{\delta}_{2i} \qquad \forall i \in \mathcal{S},$$

and

$$\hat{\delta}_{i,S2} = \hat{\delta}_{i,S2}^{Stage1} \qquad \forall i \notin \mathcal{S}.$$

2.4 Bias-Adjusted Estimators

Lastly, let consider bias-adjusted estimators following the work of Whitehead [11] and Stallard and Todd [12]. The main idea is to estimate the bias of the naive estimator and subtract it from the naive estimator itself. In this work, we compare two approaches: the single-iteration estimator (SI) and the multi-iteration estimator (MI). The single-iteration estimator is calculated as follows:

$$\hat{\delta}_{i,SI} = \hat{\delta}_{i,N} - \hat{b}_i(\hat{\boldsymbol{\delta}}_{i,N})$$

where $\hat{b}_i(\hat{\boldsymbol{\delta}}_{i,N})$ is an estimator of the bias of the naive estimator, and where true log hazard ratios in the expression for the bias are replaced with naive estimators. In the multiple-iteration procedure, values that replace true log hazard ratios in the expression for bias are obtained iteratively. In step 1

of the iterative procedure, naive estimates are used, i.e. the single-iteration bias-adjusted estimator is a special case of the multiple-iterations bias-adjusted estimator. In step 2, the single-iteration bias-adjusted estimator is used to replace true log hazard ratios in the estimation of the bias. Then, we calculate a new estimator subtracting the new estimated bias from the naive estimator and repeat the procedure, until convergence. Since the single-iteration approach is just a special case of the multi-iteration, in the following we show the calculation of the bias at a generic iteration with estimator $\widetilde{\boldsymbol{\delta}}$:

$$\hat{b}_i(\widetilde{\boldsymbol{\delta}}_i) = t_i(E[\hat{\delta}_{1i}|S, \widetilde{\delta}_i] - \hat{\delta}_{i,N}) \qquad \forall i \in \mathcal{S},$$

and

$$\hat{b}_i(\widetilde{\boldsymbol{\delta}}_i) = (E[\hat{\delta}_{1i}|S, \widetilde{\delta}_i] - \hat{\delta}_{i,N}) \qquad \forall i \notin \mathcal{S}.$$

The $E[\hat{\delta}_{1i}|S, \widetilde{\delta}_i]$ $i \in (1, ..., K)$ is calculated as follows:

$$E[\hat{\delta}_{1i}|S, \widetilde{\delta}_i] = \int_{-\infty}^{b} x \ \phi\left(\frac{x - \widetilde{\delta}_i}{\hat{\tau}_{1i}}\right) dx \qquad \forall i \in \mathcal{S},$$

and

$$E[\hat{\delta}_{1i}|S, \widetilde{\delta}_i] = \int_{b}^{\infty} x \ \phi\left(\frac{x - \widetilde{\delta}_i}{\hat{\tau}_{1i}}\right) dx \qquad \forall i \notin \mathcal{S}.$$

where ϕ is the probability density function of a normal distribution. This formula is constructed recalling that a sub-population is selected if the treatment has a log HR lower than b, while it is dropped if not.

3 Case-study

The analyses in this work have been inspired by a real case-study in heart failure. The initial study was a group sequential design without population selection. After the analysis, some subgroups with different efficacy were identified and it was considered retrospectively that the design could have been conducted as an adaptive enrichment design. For confidentiality reasons, the data used in this section are simulated data. Here, an experimental treatment is compared to placebo, and the initial patient population is partitioned in $K = 3$ sub-populations according to the baseline heart rate: low heart rate (below 75 bpm); medium heart rate (between 75 and 81 bpm); high heart rate (above 81 bpm). The primary endpoint is the time from randomisation to cardiovascular death or hospital admission for worsening of heart failure. The main analysis is a Cox proportional hazard model adjusted for previous beta-blocker intake at randomisation, and the treatment effect is estimated using the hazard ratio between the two treatment arms. The interim analysis is done after 630 events occurred, and the stage 1 patients are followed up at most for 18 months after the analysis; the final analysis is conducted when 1260 events have occurred, permitting to detect a hazard ratio of 0.85 with a power of 82% with a one-sided type 1 error of 2.5%. The futility threshold on the log-hazard ratio scale is set to $b = -0.1$, corresponding to a hazard-ratio of 0.9.

In Table 1, the stage 1 and stage 2 estimations from the Cox proportional hazard model are presented. The low heart rate sub-population is dropped at the interim analysis because it is below the futility threshold, while the others continue to stage 2. The medium heart rate sub-population has a large treatment effect in stage 1, but a smaller effect in stage 2 (due to sampling variations). The high heart rate sub-population has large treatment effect both in stage 1 and stage 2. Note that since the stage 1 estimates are far from b in both cases, we expect the various estimates to be close. At the end of the study, we proceed with the estimation of the treatment effect in the selected sub-populations, which results are shown in Table 2:

- For the sub-population with medium heart rate, the UMVCUE, single-iteration, multiple-iteration bias-adjusted and second shrinkage estimator estimator provide similar and slightly more conservative estimations with respect to the naive one. There is not much difference between them. On the other hand, the first shrinkage estimator provides a more conservative estimations.
- For the high heart rate sub-population, the naive estimator is also the most optimistic, but the UMVCUE, single-iteration and multiple-iteration bias-adjusted estimators remain similar to each other and to the naive one. The shrinkage estimators provide a slightly more conservative estimation.

Log HR	Low Heart Rate	Medium Heart Rate	High Heart Rate
$\hat{\delta}_{1i}(\hat{\tau}_{1i})$	-0.075 (0.155)	-0.397 (0.150)	-0.358 (0.121)
$\hat{\delta}_{2i}(\hat{\tau}_{2i})$	-	-0.109 (0.109)	-0.313 (0.097)

Table 1: Case-study in heart failure: stage 1 and stage 2 MLE estimators of the log HR.

Log HR Estimator	Medium Heart Rate	High Heart Rate
Naive estimator (N)	-0.209	-0.330
UMVCUE (U)	-0.188	-0.329
Shrinkage 1 (S1)	-0.180	-0.315
Shrinkage 2 (S2)	-0.189	-0.317
Single-iteration (SI)	-0.187	-0.327
Multiple-iteration (MI)	-0.191	-0.328

Table 2: Case-study in heart failure: comparison of the estimators of the log HR.

95% confidence intervals	Medium Heart Rate	High Heart Rate
Sidak	[-0.396;-0.021]	[-0.491;-0.170]
Bonferroni	[-0.421; 0.003]	[-0.511;-0.149]
Selection-adjusted	[-0.391; 0.045]	[-0.509;-0.140]

Table 3: Case-study in heart failure: Sidak [16], Bonferroni[17] and selection-adjusted[9] confidence intervals.

For completeness, Table 3 presents 95% confidence intervals for the selected sub-populations. Alongside the Sidak [16] and Bonferroni [17] confidence intervals, we also present the selection-adjusted confidence intervals from Kimani et al. [9]. We notice that the Bonferroni confidence intervals are wider than the Sidak's ones, as expected, with the one for the medium heart rate sub-population containing zero. However, these rely on the naive assumption that the overall estimate is normally distributed and do not correct for the selection process. Actually, we see that the selection-adjusted confidence intervals are more conservative: in the medium heart rate sub-population the lower bound is higher with respect to the other two and also the upper bound is above zero, suggesting that there could also be no difference between the treatment and the placebo; in the high heart rate sub-population the lower bound is close to the Bonferroni's, while the upper bound is higher, resulting in a wider confidence interval. Note that the selection adjusted confidence intervals are based on the stage 1 and stage 2 estimates, thus the confidence interval for the medium heart rate sub-population incorporates zero because of its smaller effect in stage 2.

4 Simulation study

In this section, we compare via simulations the performances of the six estimators presented earlier in terms of bias, variance and mean squared error (MSE). The bias is the metric we are the most interested in, because the primary objective of these estimators is to reduce or to eliminate it. However, the variance is also of interest, as an unbiased but highly variable estimator may not be preferred with respect to a reasonably biased but more precise alternative, as the latter may better support decision-making. The MSE combines the two information.

4.1 Setting

Consider a setting inspired by the case study with three sub-populations and two treatment arms: an experimental treatment and a control. Patients are recruited evenly from the 3 sub-populations during a

Log HR	$\boldsymbol{\delta_i = 0}$	$\boldsymbol{\delta_i = -0.1}$	$\boldsymbol{\delta_i = -0.2}$	$\boldsymbol{\delta_i = -0.3}$
Probability of selection	30%	49%	69%	84%

Table 4: Empirical probability of selection for the different sub-populations in the simulation study according to their log HR.

period of maximum 3 years and equally assigned to the arms (randomisation ratio 1:1), and the maximum follow-up time is 9 months. We suppose that the hazard function is constant for all the treatments and equal to $h_c = 0.0005$ for the control. A total number of 632 events is needed to detect a hazard ratio of 0.8 with a power of 80% and a one-sided type 1 error of 2.5%, and an interim analysis is conducted when half of the events are observed. Thus, we set that: the time of the interim analysis T_1 is after 316 events; the time until stage 1 patients are followed in stage 2 $\tilde{T}_1$ is set to 6 months after the interim analysis; the ending time of the trial T_2 is set when all 632 events are observed. Note that T_2 occurs when a total of 632 events are observed, also taking into account stage 1 patients followed up until $\tilde{T}_1$. Moreover, while the number of events is fixed regardless of the number of sub-populations selected, the number of patients needed to reach these number of events varies with the scenarios and the selected sub-populations, but a maximum of 3000 patients is assigned to each sub-population.

Three cases of log HR are considered: treatment ineffective in all sub-populations $\boldsymbol{\delta} = (0, 0, 0)$; treatment effective in only one sub-population $\boldsymbol{\delta} = (0\ , 0\ , -0.3)$; linear effect on the sub-populations $\boldsymbol{\delta} = (-0.1, -0.2, -0.3)$. For all the cases, the threshold is set to $b = -0.1$. We ran simulations for each scenario until we obtained 10000 simulated clinical trials with a stage 2, i.e. not stopped for futility at the interim analysis. Table 4 shows empirical selection probabilities.

For completeness, in Supplementary Material, additional simulation scenarios are presented, where: we set the threshold to $b = 0$; we set $\tilde{T}_1$ to 3 months after the interim analysis; we compare 4 sub-populations while keeping the other parameters as in this setting.

4.2 Results

In this Section, we present the estimates for the bias, variance and MSE of the estimators in our simulation study. To derive these, we define $\mathbb{S}_i$ as the set of simulations where sub-population i is selected, and $\hat{\delta}^s_{i,\cdot}$ as a generic estimator of δ_i in a simulation $s \in \mathbb{S}_i$. The bias in each sub-population is estimated via $\frac{1}{|\mathbb{S}_i|}\sum_{s\in\mathbb{S}_i}(\hat{\delta}^s_{i,\cdot} - \delta_i)$ and the MSE via $\frac{1}{|\mathbb{S}_i|}\sum_{s\in\mathbb{S}_i}(\hat{\delta}^s_{i,\cdot} - \delta_i)^2$; the variance is calculated as (MSE - bias^2). Figure 2 displays these results. Each row corresponds to the three setting we analysed: treatment ineffective in all sub-populations (top), treatment effective only in one sub-population (center), linear effect on the sub-populations (bottom). In columns are presented the metrics of interest: bias (left), variance (center) and MSE (right). In this Figure, if more than one sub-population is selected, the results for bias, variance and MSE are averaged over all sub-populations.

We see that the bias of the estimators is generally higher in the top row. The optimal decision in this case would be to stop the whole trial for futility, because none of the underlying log HR is larger than the threshold. Therefore, a sub-population is selected only when the estimated effect is substantially better than the true effect in that sub-population. The naive estimator has the greatest bias, followed by the shrinkage estimators, with the S1 slightly outperforming S2; then, the bias-adjusted estimators come next, with the single-iteration performing better than the multiple-iteration; the UMVCUE provides an almost unbiased estimation in this case, as expected, but tends to over-correct leading to a positive bias which, however, is of lower magnitude with respect to the other estimators' bias and could be attributable to the correlation bias left. In terms of variance, the naive estimator is the best performing alongside the shrinkage ones, followed by the bias-adjusted estimators performing equally, and finally the UMVCUE, which is the least precise. In terms of MSE the best performing are the S1 and S2 (in this order), followed by the single-iteration bias-adjusted estimator, multi-iteration bias-adjusted estimator and the naive estimator, which is slightly worse; the worst performing in terms of MSE is the UMVCUE.

The second row corresponds to clinical trials where, at the interim analysis, the only sub-population on which the treatment is effective should be selected, and it notably differs from the other ones. Therefore, the extent of the bias is lower than in the first row. In this case, the best performing estimator in terms of absolute bias is the multiple-iteration bias-adjusted estimator, followed in order by S1, SI, S2, UMVCUE and naive estimator. Still, we notice that the UMVCUE again over-corrects the bias, leading to a positive bias of similar magnitude with respect to the previous scenario. In terms of variance, the best estimator

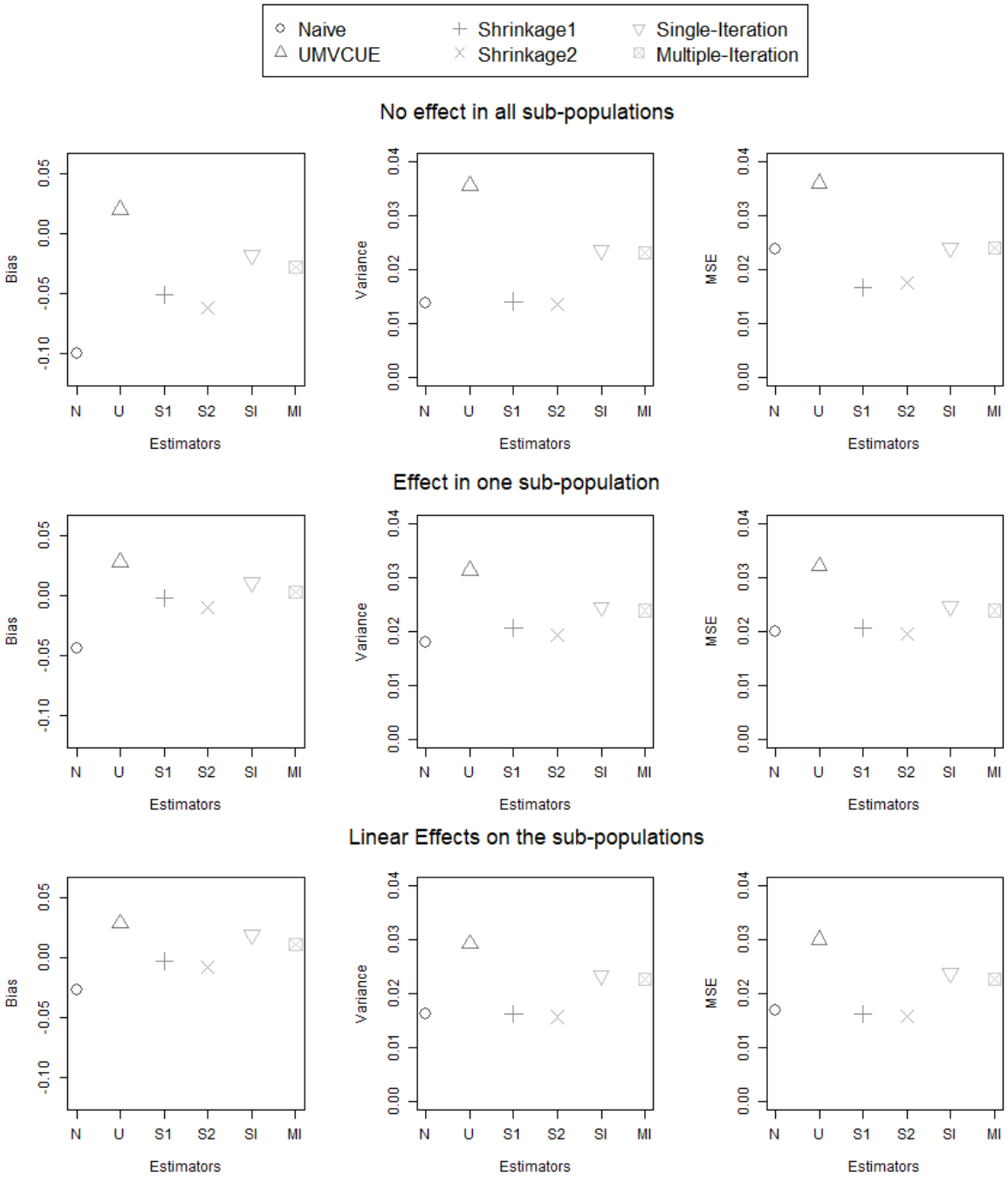

Figure 2: Estimators' performances. Top row: treatment ineffective in all sub-populations $\boldsymbol{\delta} = (0, 0, 0)$; Middle row: treatment effective only in one sub-population $\boldsymbol{\delta} = (0\ , 0\ , -0.3)$; Bottom row: linear effect on the sub-populations $\boldsymbol{\delta} = (-0.1, -0.2, -0.3)$. Left column: Bias; Centre column: Variance; Right column: Mean Squared Error.

is the naive one, followed by the shrinkage estimators, the bias-adjusted estimators, with the UMVCUE coming last. In terms of MSE, the shrinkage estimators perform best alongside the naive estimator, followed by the bias-adjusted estimators performing equally, and finally the UMVCUE.

In the third row, the correct choice is to select the two sub-populations where the treatment is more effective. Again, the naive estimator has the highest absolute bias, followed in order by UMVCUE, SI, MI with S1 and S2 that performs best at equal merit. However, the single-iteration bias-adjusted estimator, multi-iteration bias-adjusted estimator and the UMVCUE tend to over-correct the bias, leading to a positive one in their case. In terms of variance, the UMVCUE has the highest variance followed by SI, MI, naive estimator, and finally the most precise are the shrinkage estimators. Eventually, the unbiased estimator has the highest MSE, followed by the SI and MI, the naive estimator, and the shrinkage estimators in order.

At this point, we would like to have a more precise idea of the behaviour of the estimators in each sub-population. In Figure 3, we analyse the specific case of $\boldsymbol{\delta} = (-0.1, -0.2, -0.3)$, showing the results in each sub-population. We first notice that the variance and MSE of the estimators do not vary much from one sub-population to another, keeping the same order noticed in the bottom row of Figure 2, with only a slight variation of their magnitude from one row to the other. On the other hand, the bias differs substantially from one sub-population to another. The bias of the naive estimator is higher in the top row and decreases with increasing effect, as expected while moving from the threshold, approaching zero in the case of an effect equal to -0.3. The bias of the UMVCUE is constant over the three sub-populations and positive, indicating an over-correction that may be attributable to the correlation bias left. The bias of the shrinkage estimators, which seems to be approximately zero when averaged over the sub-populations in the bottom row of Figure 2, is in truth substantially variable: it is equal to the bias of the naive estimator when the effect is equal to -0.1; it is almost zero when the effect is equal to -0.2 with S1 outperforming S2; it is positive when the effect is equal to -0.3, with S1 having a higher bias with respect to the UMVCUE and S2 having a similar performance with respect to SI. Finally, regarding the bias of the bias-adjusted estimators: on the top row,it is approximately zero for both, with the SI outperforming MI; on the middle row, it is positive but lower than the UMVCUE's for the SI, with the MI outperforming the SI in this case; on the bottom row, the results are similar to the middle row.

Additional simulation scenarios' results are presented in Supplementary Material. If we set $b = 0$, the bias is lower in general for all the estimators since they are further from the threshold, but the overall order for the estimators' bias, variance and MSE is the same. When $\tilde{T}_1$ is set to 3 months after the interim analysis, the performance metrics are very similar to those presented here. When comparing 4 sub-populations, we obtain a higher bias overall, and a better performance of the S1 and UMVCUE, with the latter being almost unbiased. This may be attributable to the fact that we make the decision when we have the same number of events in stage 1 as in the previous case, but we have one more sub-population. Therefore, the decision is made with a smaller information fraction, and the stage 2 data have more impact in the overall estimates' derivation [4, 9]. In this last case, we also focus on the results in each sub-population in case of linear effects in the sub-populations, and the same pattern as seen in Figure 3 is obtained.

We can summarize the results as follows:

- The naive estimator (N) has the highest bias but very low variance, resulting in a moderate MSE compared to the other estimators.
- The UMVCUE (U) has a small constant positive bias (which is not zero because of the correlation bias, but is reduced increasing the number of sub-populations) and the highest variance, resulting in the highest MSE.
- The single-iteration (SI) bias-adjusted estimator has small bias, but more variance with respect to naive estimator, resulting in comparable MSE with respect to the naive estimator.
- The multiple-iteration (MI) bias-adjusted estimator tends to provide a less conservative estimation with respect to the single-iteration, and has similar variance and MSE with respect to the SI.
- The first shrinkage estimator (S1) has a noticeable bias which also varies substantially from one sub-population to another, but the very low variance, resulting in the very low MSE.
- The second shrinkage estimator (S2) performs similarly to S1, returning better performances in some cases and worse performances in other ones.

These results are observed in all scenarios we have considered, and they are also consistent with the results previously published in different settings [4, 6, 8, 10].

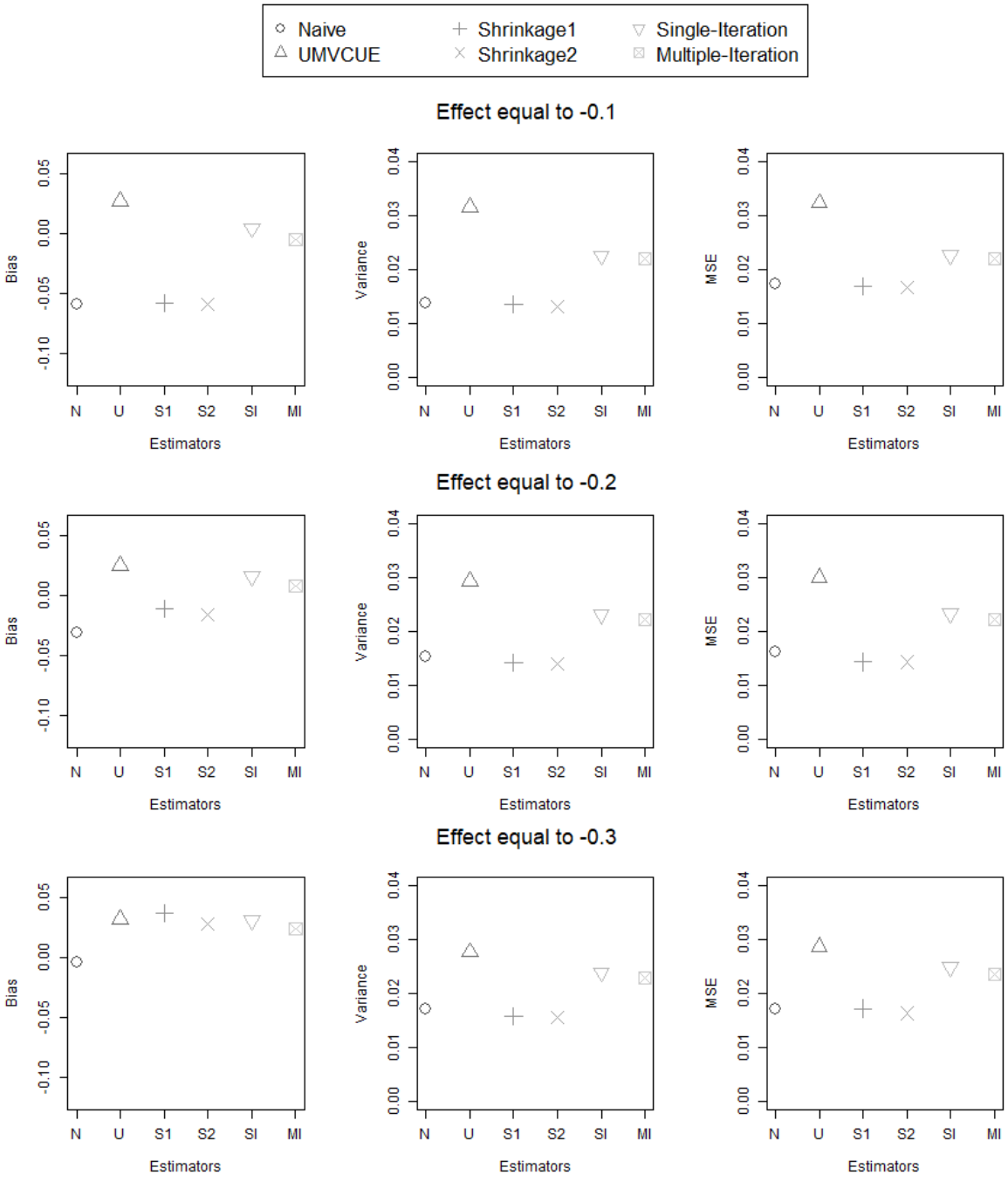

Figure 3: Estimators' performances in each sub-population in case of linear effects on the sub-populations. Top row: effect equal to -0.1; Middle row: effect equal to -0.2; Bottom row: effect equal to -0.3. Left column: Bias; Centre column: Variance; Right column: Mean Squared Error.

5 Discussion

In this paper, we investigated the handling for the selection bias [3] in two-stage enrichment adaptive designs with time-to-event data, using several estimators found in the literature [4, 7, 8, 9, 10, 11, 12]. This work was motivated by the FDA Guidance for Industry on Adaptive Design Clinical Trials for Drugs and Biologics [5], where the sponsor is required to assess the reliability, and in particular the extent of bias, of the treatment effects estimation in adaptive design trials with appropriate methods. We applied the different methods to a case-study in cardiology, and compared the performances of the estimators in several scenarios with a selection rule based on a futility threshold. The other comparative studies available in the literature were conducted in the context of adaptive designs with treatment selection [4, 10] or sub-population selection with normally distributed endpoints [6, 13].

We conducted a simulation study in order to compare the performances of the estimators in terms of bias, variance and mean squared error. Although the bias is of primary interest, the variability may also influence the choice of an estimator over another. Alongside the MLE in primary analysis, in order to address the regulatory requirements, our recommendation is to present the unbiased estimator and the single-iteration bias-adjusted estimator in sensitivity analysis: the former completely eradicates the selection bias, but is highly variable with respect to a naive estimator; the latter is less biased than a naive estimator, but only slightly more variable. Thus, the unbiased estimator can give an idea of the extent of the bias in the naive estimator, while the single-iteration bias-adjusted estimator can give a more precise and less biased estimation for decision-making. For completeness, a simulation study may be added as a complement to outline the estimator that best fits the context and aims of the trial, as suggested in literature[14].

The present study has several limitations. In particular, it focuses on clinical trials with one interim analysis, and two arms (one experimental treatment and one control), therefore a future extension could be to add more interim analyses or more arms. Moreover, we consider disjoint sub-populations, while other works do not [6, 9]. Also, we have chosen a selection rule based on comparing the HR to a pre-specified threshold. Other selection rules based on conditional power or predictive power could have been chosen, however similar conclusions are expected [18]. Besides, some works in literature have also set up the ground for further exploration, for example of the calculation of appropriate confidence intervals via simultaneous inference [19], bootstrap resampling [10], confidence regions based on orderings [12] or simultaneous inference based on the duality between hypothesis testing and confidence intervals [9]. All of these methods have advantages and limitations, and we refer to literature for further information.

Despite its limitations, the current work provide insightful information regarding the performances of the different estimators to reduce or remove the bias in enrichment adaptive designs with time-to-event data, permitting to better support decision-making and to comply with the regulatory requirements.

acknowledgements

We would like to thank Dr. Peter K. Kimani and Dr. Maximo Carreras for providing the code from their works [6, 4] that helped developing this work in the early phases. We would like to thank also two anonymous reviewers, which greatly helped us to improve the quality of the manuscript. This research is supported by Institut de Recherches Internationales Servier.

conflict of interest

The authors declare no potential conflicts of interest.

References

[1] Philip Pallmann, Alun W. Bedding, Babak Choodari-Oskooei, Munyaradzi Dimairo, Laura Flight, Lisa V. Hampson, Jane Holmes, Adrian P. Mander, Lang'o Odondi, Matthew R. Sydes, Sofía S. Villar, James M. S. Wason, Christopher J. Weir, Graham M. Wheeler, Christina Yap, and Thomas Jaki. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, 16(1), February 2018.

[2] Dirk Bassler. Stopping randomized trials early for benefit and estimation of treatment effects: Systematic review and meta-regression analysis. *JAMA*, 303(12):1180, March 2010.

[3] Peter Bauer, Franz Koenig, Werner Brannath, and Martin Posch. Selection and bias-two hostile brothers. *Statistics in Medicine*, 29:1–13, 2009.

[4] Maximo Carreras and Werner Brannath. Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection. *Statistics in Medicine*, 32(10):1677–1690, June 2012.

[5] Food and Drug Administration. Adaptive design clinical trials for drugs and biologics guidance for industry. December 2019.

[6] Peter K. Kimani, Susan Todd, Lindsay A. Renfro, and Nigel Stallard. Point estimation following two-stage adaptive threshold enrichment clinical trials. *Statistics in Medicine*, 37(22):3179–3196, May 2018.

[7] Arthur Cohen and Harold B. Sackrowitz. Two stage conditionally unbiased estimators of the selected mean. *Statistics & Probability Letters*, 8(3):273–278, August 1989.

[8] Jack Bowden and Ekkehard Glimm. Unbiased estimation of selected treatment means in two-stage trials. *Biometrical Journal*, 50(4):515–527, August 2008.

[9] Peter K. Kimani, Susan Todd, Lindsay A. Renfro, Ekkehard Glimm, Josephine N. Khan, John A. Kairalla, and Nigel Stallard. Point and interval estimation in two-stage adaptive designs with time to event data and biomarker-driven subpopulation selection. *Statistics in Medicine*, 39(19):2568–2586, May 2020.

[10] Matthias Brückner, Andrew Titman, and Thomas Jaki. Estimation in multi-arm two-stage trials with treatment selection and time-to-event endpoint. *Statistics in Medicine*, 36(20):3137–3153, June 2017.

[11] John Whitehead. On the bias of maximum likelihood estimation following a sequential test. *Biometrika*, 73(3):573–581, 1986.

[12] Nigel Stallard and Susan Todd. Point estimates and confidence regions for sequential trials involving selection. *Journal of Statistical Planning and Inference*, 135(2):402–419, December 2005.

[13] Kevin Kunzmann, Laura Benner, and Meinhard Kieser. Point estimation in adaptive enrichment designs. *Statistics in Medicine*, 36(25):3935–3947, August 2017.

[14] David S. Robertson, Munya Dimairo Babak Choodari-Oskooei, Laura Flight, Philip Pallmann, and Thomas Jaki. Point estimation for adaptive trial designs. May. https://arxiv.org/abs/2105.08836v3. Published 2021. Accessed August 30, 2021.

[15] Martin Jenkins, Andrew Stone, and Christopher Jennison. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics*, 10(4):347–356, December 2010.

[16] Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, June 1967.

[17] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, March 1961.

[18] Paul Gallo, Lu Mao, and Vivian H. Shih. Alternative views on setting clinical trial futility criteria. *Journal of Biopharmaceutical Statistics*, 24(5):976–993, August 2014.

[19] Ekkehard Glimm. Adjusting for selection bias in assessing treatment effect estimates from multiple subgroups. *Biometrical Journal*, 61(1):216–229, November 2018.