Statistics
in Medicine

Research Article

Received 12 January 2015, Accepted 24 September 2015 Published online 12 October 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6761

Precision of maximum likelihood estimation in adaptive designs

Alexandra Christine Graf,[a*†] Georg Gutjahr[b] and Werner Brannath[b]

There has been increasing interest in trials that allow for design adaptations like sample size reassessment or treatment selection at an interim analysis. Ignoring the adaptive and multiplicity issues in such designs leads to an inflation of the type 1 error rate, and treatment effect estimates based on the maximum likelihood principle become biased. Whereas the methodological issues concerning hypothesis testing are well understood, it is not clear how to deal with parameter estimation in designs were adaptation rules are not fixed in advanced so that, in practice, the maximum likelihood estimate (MLE) is used. It is therefore important to understand the behavior of the MLE in such designs. The investigation of Bias and mean squared error (MSE) is complicated by the fact that the adaptation rules need not be fully specified in advance and, hence, are usually unknown. To investigate Bias and MSE under such circumstances, we search for the sample size reassessment and selection rules that lead to the maximum Bias or maximum MSE. Generally, this leads to an overestimation of Bias and MSE, which can be reduced by imposing realistic constraints on the rules like, for example, a maximum sample size. We consider designs that start with k treatment groups and a common control and where selection of a single treatment and control is performed at the interim analysis with the possibility to reassess each of the sample sizes. We consider the case of unlimited sample size reassessments as well as several realistically restricted sample size reassessment rules. © 2015 The Authors. *Statistics in Medicine* Published by John Wiley & Sons Ltd.

Keywords: maximum likelihood estimation; sample size reassessment; treatment selection; adaptive designs

1. Introduction

There has been increasing interest over the last years in adaptive two-stage clinical trials where more than one treatment group are compared with one common control. These trials allow for design adaptations as, for example, sample size reassessment or treatment selection at an interim analysis. It is well known that ignoring the adaptive and multiplicity issues lead to a considerable inflation of the type 1 error rate and that effect estimates based on the maximum likelihood principle may be biased. For the comparison of a single treatment with a control and balanced sample sizes between groups, Proschan and Hunsberger [1] showed that the maximum type 1 error rate can be inflated from 0.05 to 0.11. Graf and Bauer [2] extended this arguments to allow for individual sample size reassessment rules in the treatment and control group respectively, which increases the maximum type 1 error to 0.19. However, when selecting one out of k treatments and control for a second stage, Graf *et al.* [3] showed that if using the Dunnett test to adjust for multiplicity [4], the maximum type 1 error rate may not exceed the pre-specified α-level for specific restrictions on the second stage sample size reassessment rule because of the over-correction for the treatments not tested at the end of the study. A large number of hypothesis testing methods have been developed that allow for flexible sample size adaptations (not pre-fixed in advance) without compromising the overall type 1 error rate based on the combination test approach [5–7] or the conditional error principle [8,9] and have been extended to multi-armed clinical trials allowing for treatment selection [6,7,10–12].

[a]*Medical University of Vienna, Center for Medical Statistics, Informatics and Intelligent Systems, Spitalgasse 23, 1090 Vienna, Austria*
[b]*University of Bremen, Competence Center for Clinical Trials, Linzer Strasse 4, 28359 Bremen, Germany*
[*]*Correspondence to: Alexandra Christine Graf, Medical University of Vienna, Center for Medical Statistics, Informatics and Intelligent Systems, Spitalgasse 23, 1090 Vienna, Austria.*
[†]*E-mail: alexandra.graf@meduniwien.ac.at*

Whereas the methodological issues concerning hypothesis testing are well understood, up to now, it is not clear how to deal with parameter estimation after flexible interim adaptations. Several methods have been proposed to reduce or remove the Bias [12–19]. The Bias depends on many different features as the selection procedure, the sample size reassessment rule, or the unknown parameters. The proposed methods therefore do only apply to specific adaptation rules and, hence, are not generally applicable. In particular, in designs were adaptation rules are not fixed in advance, estimation is still an unsolved issue, so that in practice the maximum likelihood estimate (MLE) is still used.

Bauer *et al*. [20] investigated the impact of treatment selection on the mean Bias and the mean squared error (MSE) when selecting those j (out of k) treatments with the largest observed effects while fixing the total per-group sample size. They further considered designs where the sample size is reshuffled equally to the selected treatment arms and control with the conclusion that due to regression of the mean, Bias decrease as compared with the scenario without reshuffling. To our knowledge, no investigations of other types of sample size and selection rules where undertaken yet. Hence, the behavior of MLE is not yet fully understood for adaptive designs.

Adaptive designs have the practically important feature that the selection and sample size reassessment rule need not be fully pre-specified. This complicates the investigation of Bias and MSE, which depend on the actually unknown sample size and selection rule. Simulations or numerically investigations under typical adaption rules are important, however, can only give partial answers. We therefore investigate the behavior of the MLE from another point of view; we search for the selection and sample size reassessment rule leading to the maximum mean Bias or maximum MSE when using the MLE at the end of the adaptive trial to estimate the treatment effect. Brannath *et al.* [16] calculated the maximum mean Bias for the case of a one-sample z-test and concluded that the maximum mean Bias in a flexible two stage design is in general and not larger than that of a conventional group sequential design. We will consider scenarios where more than one treatment groups are compared with a common control, and one treatment and the control are selected for the second stage. Moreover, we also allow for flexible choices of the second stage allocation ratio, permitting, for example, a larger increase in sample size for the selected treatment than for the control group.

The case of unlimited sample sizes provides upper bounds for Bias and MSE. We therefore consider also scenarios with restrictions on the sample size to obtain less conservative estimates for real adaptive trials. We will, for instance, investigate bounded second stage sample sizes as well as the restriction on the control group to have a smaller sample size than that of the treatment group. We will also consider designs with a fixed overall sample size for the control group and designs with a fixed total sample size permitting only a reshuffling between the selected treatment and the control, with and without the restriction of a smaller control group.

We will see in this paper that the maximum mean Bias and maximum MSE of the MLE are independent from the true means in the treatment and control groups, without and with restrictions on the second stage sample sizes. As a consequence, they are the same under the null and all alternative hypothesis. This is a very attractive property of the maximum mean Bias and maximum MSE of the MLE that simplifies its investigation and discussion considerably.

The rest of the paper is organized as follows. In Section 2, we describe the type of interim adaptations investigated to calculate the maximum mean Bias and maximum MSE. In Section 3, we investigate the maximum mean Bias and maximum MSE for the case when only $k = 1$ treatment is compared with one control. In Section 4 we generalize the arguments to the scenario of selecting one out of $k > 1$ treatments and control for the second stage. A strategy that is intensively discussed in the literature [11–13, 18, 20]. We will end with a discussion of the results in Section 5.

2. Designs with treatment selection

Assume a clinical trial with parallel groups and a two-stage design that starts at the first stage with k treatments and a control and continues in the second stage with one selected treatment and the control. We assume normally distributed outcomes, $X_{(i,j,l)} \sim N(\mu_i, \sigma^2)$, $i = 0, \dots, k$, where i represents the treatment group, with $i = 0$ for the control and $i = 1, \dots, k$ for the experimental treatments, and $j \in \{1, 2\}$ is the index for the stage. The index l stands for the individual, where $l = 1, \dots, n_{(i,1)}$ in the first stage and $l = 1, \dots, n_{(i,2)}$ in the second stage for each treatment group i. The common variance σ^2 is assumed to be known.

An interim analysis is performed after recruitment of $n_{(i,1)}$ patients in the ith experimental treatment and $n_{(0,1)}$ in the control group. For simplicity, we assume balanced sample sizes in the first stage, that is, $n_{(i,1)} = n_{(0,1)} = n$ for all $i = 1, \dots, k$, which is a common scenario. However, the second stage sample sizes can be unbalanced. Based on the data of the first stage, $X_{(i,1,l)}$, $i = 0, \dots, k$ and $l = 1, \dots, n$, we select one out of the k treatments, say treatment $s \in \{1, \dots, k\}$, and the control for the second stage. We may also reassess the second sample sizes based on the first stage data. In the second stage, $n_{(s,2)} = r_s n$, patients are recruited in the selected treatment and $n_{(0,2)} = r_0 n$ in the control group, where second-to-first-stage ratios $0 \leqslant r_i \leqslant \infty$ for $i = 0, s$, can depend on the first stage data. Note that the selected treatment (or control) can also be stopped at interim by setting $r_s = 0$ (or $r_0 = 0$). In contrast to the majority of the literature on point estimation in designs with treatment selection, we do not assume a specific selection or sample size reassessment rule and thereby consider the full flexibility permitted with adaptive designs [6, 7, 10, 11].

Treatment selection here means to decide on the treatment 'of interest' for which the effect estimate will further be investigated. For treatments, not selected in the interim analysis, we assume that the treatment effect is not of interest at the end of the trial. In the final analysis, the overall effect of the selected treatment to control is calculated using the maximum likelihood estimators, calculated over both stages:

$$\bar{x}_i = \frac{\bar{x}_{(i,1)} + r_i \bar{x}_{(i,2)}}{1 + r_i},$$

where $i = 0, s$, $\bar{x}_{(i,1)} = \frac{1}{n} \sum_{l=1}^{n} x_{(i,1,l)}$ is the sample mean of the first stage, and $\bar{x}_{(i,2)}$ is the sample mean of the second stage for group $i = 0, s$. If sample size adjustments are performed based on the first stage data, the overall sample mean $\bar{x}_i$ may be biased (see e.g. Brannath *et al.* [16]).

Our intention is to derive the worst case, meaning that we are searching for the sample size reassessment and selection rule maximizing the mean Bias (denoted in sequel as "Bias" for short) or the MSE for the selected treatment compared to the control. We prefer to consider the 'root mean squared error', $\text{RMSE} = \sqrt{\text{MSE}}$, because it is on the same scale as the mean and the Bias. In the context of designs with treatments selection, the Bias and MSE are defined as follows:

$$\text{Bias} = \mathbf{E}\left[(\bar{X}_s - \bar{X}_0) - (\mu_s - \mu_0)\right] \qquad \text{and} \qquad \text{MSE} = \mathbf{E}\left[((\bar{X}_s - \bar{X}_0) - (\mu_s - \mu_0))^2\right].$$

These quantities have also been denoted by 'selection Bias' and 'selection MSE' (cf. Bauer *et al.* [20]).

The general idea of this paper is to determine the maximum Bias or maximum MSE by maximizing at each interim sample point the conditional Bias or conditional MSE given the interim data. By searching for the treatment selection and sample size adaptation rules that maximizes the conditional Bias (or MSE), we obtain the treatment selection and sample size rules that maximizes the overall Bias (or MSE). This idea has been used in Brannath *et al.* [16] to obtain the maximum Bias in the one-sample case and, thereby, also in the balanced two-sample case. A similar idea has earlier (and later) been used to determine the maximum type 1 error rate of the naive z-test or Dunnett-test [1–3].

3. Two-arm trials with sample size reassessment

For illustrative purposes, we start with the scenario where only one treatment group ($k = s = 1$) is compared with a control. The results will be generalized to $k > 1$ in Section 4. We start with a discussion of the maximum Bias and then proceed with a similar investigation of the maximum RMSE.

3.1. Maximum Bias

Brannath *et al.* [16] calculated the maximum Bias of the one-sample mean in a two-stage design with data-driven sample size reassessments. Their result easily generalizes to the treatment effect estimate in a two-arm parallel group design with balanced first and second stage sample sizes, because the treatment effect estimate in a balanced two-arm trial is formally equal to the one-sample mean of observations with variance $2\sigma^2$. According to the result in [16], the maximum Bias in an adaptive two-stage trial with two-arms and the restriction $n_{(0,2)} = n_{(1,2)}$ becomes

$$\text{B}^*(n, \sigma, r_{\min}, r_{\max}) = \frac{\sqrt{2}\sigma}{\sqrt{n}} \phi(0) \left(\frac{1}{1 + r_{\min}} - \frac{1}{1 + r_{\max}} \right) \approx \frac{0.6\,\sigma}{\sqrt{n}} \left(\frac{1}{1 + r_{\min}} - \frac{1}{1 + r_{\max}} \right) \qquad (1)$$

where ϕ denotes the standard normal density and $r_{\min}$ and $r_{\max}$ are pre-specified lower and upper bounds for the data driven second-to-first-stage ratio $r = n_{(0,2)}/n = n_{(1,2)}/n$. Note that the maximum Bias is independent from the true means μ_0 and μ_1. We can set $r_{\min} = 0$ and $r_{\max} = \infty$ if no such bounds exist. In this case, the maximum Bias becomes $\mathrm{B}^*(n, \sigma, 0, \infty) = \phi(0)\sqrt{2}\,\sigma/\sqrt{n} \approx 0.6\,\sigma/\sqrt{n}$.

3.1.1. Flexible second-to-first-stage ratios. The restriction to $n_{(0,2)} = n_{(1,2)}$ may be too strong for applications because it does not permit an unequal increase or decrease of the sample sizes in the two arms. For instance, if the control is a placebo, ethical reasons may advise us to increase the sample size only in the treatment group or even decrease it in the control arm. A reduction in the placebo allocation ratio will usually also increase the willingness to participate in the trial. We, therefore, also consider the Bias under unequal sample size adaptations, which can be determined by maximizing the conditional Bias with regard to $n_{(0,2)}$ and $n_{(1,2)}$ without the constraint $n_{(0,2)} = n_{(1,2)}$. We will see in subparagraph 3.1.5 (where we describe our calculations) that the maximum Bias remains independent from μ_1 and μ_0 for flexible second stage sample sizes $n_{(0,2)}$ and $n_{(1,2)}$. Note that $r_0 = n_{(0,2)}/n$ and $r_1 = n_{(1,2)}/n$ are the individual second-to-first-stage ratios for the control and treatment group with $r_{\min} \leqslant \min(r_0, r_1) \leqslant \max(r_0, r_1) \leqslant r_{\max}$.

Figure 1 (A) for $k = 1$ shows the maximum Bias, B^*, standardized by the standard error $\sqrt{2\sigma^2/n}$ of the first-stage mean. The shown results do therefore also not depend on the first stage sample size n or the common known variance σ. The solid lines in Figure 1 (A) show $\mathrm{B}^*/\sqrt{2\sigma^2/n}$ for $r_{\min} = 0,\ 0.5,$ and 1 and $r_{\max}$ varying from 0 to 3. As expected, the maximum Bias is increasing with decreasing $r_{\min}$ and increasing $r_{\max}$, showing that more flexibility leads to a larger maximum Bias. For example, for $r_{\min} = 1$, meaning that the second-stage sample size has to be as least as large as the first-stage sample size and $r_{\max} = 2$ allowing a doubling of the second-stage sample size as compared with the first stage, the maximum Bias is 0.09 times the first-stage standard error, increasing to 0.19 and 0.38 for $r_{\min} = 0.5$ and 0, respectively. This shows that the option for sample size reductions (including early stopping) can largely increases B^*.

The maximum Bias appears to be large for some of the scenarios in Figure 1 (A). However, recall that we have plotted B^* in units of $\sqrt{2\sigma^2/n}$ and that $\sqrt{2\sigma^2/n}$ decreases with increasing per group first-stage sample size n. Assume, for instance, that n is half the sample size required for a z-test with power 90% at $\delta = \mu_i - \mu_0$ in a classical two-armed parallel group design at one-sided level $\alpha = 0.025$. Then $\sqrt{2\sigma^2/n} = \delta\{\Phi^{-1}(0.975) + \Phi^{-1}(0.9)\}^{-1} = 0.31\,\delta$, and if $(r_{\min}, r_{\max}) = (1, 2)$, then $\mathrm{B}^* = 0.09\,\sqrt{2\sigma^2/n}$ is only 3% of the effect size δ assumed in the sample size calculation. Allowing for more flexibility, as, for example, $(r_{\min}, r_{\max}) = (0, 2)$, the bias is substantially increasing to 12% of the effect size.

3.1.2. Restriction to $r_1 \geqslant r_0$. A reasonable constraint to reduce the maximum Bias is to require that the experimental treatment group is never smaller than the control group. In this case, r_0 can vary between $(r_{\min}, r_{\max})$, while r_1 is restricted to $(r_0, r_{\max})$. The dot-dashed lines in Figure 1 (A) show the standardized maximum Bias for this type of restriction. As expected, the maximum Bias is always smaller than the maximum Bias with flexible ratios but larger than the one with balanced second-stage sample sizes. Its line is right in the middle of the two other lines. The dotted lines in Figure 1 (A) shows the standardized maximum Bias if we restrict the second-stage sample sizes to be balanced. If $r_{\max} = 2$, the standardized maximum Bias under the given constraint becomes 0.08, 0.16, and 0.32 for $r_{\min} = 1,\ 0.5,$ and 0, respectively. We can see that for $r_{\min} \geqslant 0.5$ the difference in Bias between the constraints $r_1 = r_0$ and $r_1 \geqslant r_0$ is small.

3.1.3. Fixing r_0. A stronger restriction is to fix the total sample size of the control group (that is, fixing r_0) while in the experimental treatment group sample sizes are reassessed within the window $(r_0, r_{\max})$. The MLE of the control group is then unbiased. The maximum Bias, therefore, does not depend on the interim outcome of the control group. However, it depends on the fixed r_0. The dashed lines in Figure 1 (A) give the standardized maximum Bias for $r_0 = 0, 0.5,$ or 1 while $r_{\max}$ varies from r_0 to 3. For example, if $r_0 = 1$ and $r_{\max} = 2$, then the standardized maximum Bias is 0.05, that is, a little more than half of the Bias with flexible sample size reallocations. We are aware that fixing $r_0 = 0$ may be an unrealistic scenario always resulting in a second stage without control. However, for complete presentation of the results, the Figure also shows the line for $r_0 = 0$.

3.1.4. The effect of $r_{\min}$. Figure 1 (A) indicates that the minimum $r_{\min}$ for the second-stage sample sizes has quite some impact on the maximum Bias. To further elaborate the impact of $r_{\min}$, we have calculated

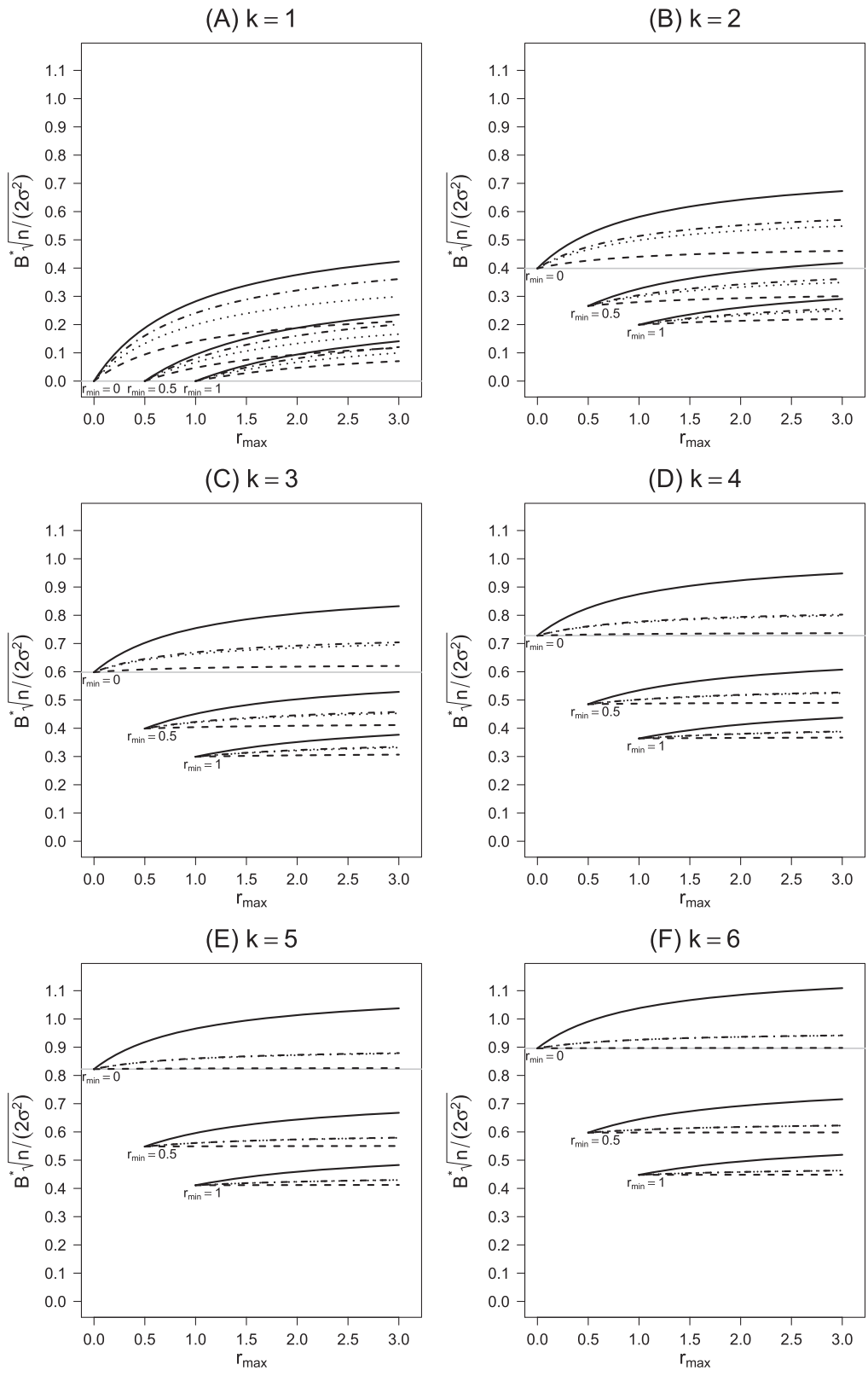

Figure 1. Standardized maximum Bias as a function of r_{max} for $r_{min} = 0, 0.5$, and 1. Values are given for a number of $k = 1$ to 6 treatments in panels (A) to (F). Within one panel the standardized maximum Bias is shown for different restrictions on the sample size reassessment rule: flexible second-to-first-stage-ratios (solid lines), equal second-to-first-stage ratios (dotted lines), restricting $r_1 \geqslant r_0$ (dot-dashed lines), and fixing the control (dashed lines). The gray horizontal line shows the standardized Bias for a fixed-size-sample test with post-trial selection.

the maximum Bias for $r_{\max} = \infty$ and $r_{\min} = 0, 0.5, 1$. Table I shows the results for the standardized maximum Bias. The row 'flexible' gives the maximum Bias with flexible second-stage allocation ratios; in the row '$r_1 \geqslant r_0$', the sample size of the experimental treatment group is constraint to be at least as large as in the control group. '$r_0 = r_1$' means that the second-stage sample sizes are restricted to be balanced, and 'fix r_0' that the sample size of the control group is fixed. Here, we are interested in the case $k = 1$ (one experimental treatment only). We observe that the maximum Bias is halved by letting $r_{\min} = 1$ (i.e. forcing the second-stage sample sizes to be at least as large as the first-stage ones) compared with $r_{\min} = 0$. With $r_{\min} = 0.5$, the maximum Bias is reduced by about 33%. These factors for the maximum Bias seem to be completely independent from the additional restrictions on r_1 and r_0 which is a remarkable finding. A possible explanation for this finding is that the maximum bias is dominated by the minimum sample size $r_{\min}$.

3.1.5. Determination of maximum Bias. As mentioned in the introduction, the sample size reassessment rule, which maximizes the Bias, is obtained by maximizing the conditional Bias, that is, the deviation of the conditional mean of the treatment effect estimate (given the interim data) from the true parameter value.

For the calculation of the conditional Bias, we standardize the individual stage-wise means $Z_{(i,j)} = (\bar{X}_{(i,j)} - \mu_i)\sqrt{n_{(i,j)}/\sigma^2}$, $i = 0, 1, j = 1, 2$. Recall that $n_{(1,1)} = n_{(0,1)} = n$ and our definition of $r_i = n_{(i,2)}/n$, $i = 0, 1$, with $r_{\min} \leqslant \min(r_0, r_1) \leqslant \max(r_0, r_1) \leqslant r_{\max}$. To simplify the notation, we will omit the index j for the first stage data and summaries, for example, denoting the first-stage standardized means by z_i, $i = 0, 1$. Similar calculations as in [16] give the conditional Bias:

$$\begin{aligned} \mathrm{CB}(z_0, z_1, r_0, r_1, n, \sigma) &= \mathbf{E}\left[(\bar{X}_1 - \bar{X}_0) - (\mu_1 - \mu_0) \mid Z_{(i,1)} = z_i, i = 0, 1\right] \\ &= \frac{\sigma}{\sqrt{n}}\left[\frac{z_1}{1 + r_1} - \frac{z_0}{1 + r_0}\right]. \end{aligned} \tag{2}$$

To evaluate the worst case, the second-to-first-stage ratios r_1 and r_0 are searched to maximize (2):

$$\widetilde{\mathrm{CB}}(z_0, z_1, n, \sigma, r_{\min}, r_{\max}) = \max_{r_{\min} \leqslant r_0, r_1 \leqslant r_{\max}} \mathrm{CB}(z_0, z_1, r_0, r_1, n, \sigma) \tag{3}$$

Note again that we assume in the following the same lower and upper bounds $r_{\min}, r_{\max}$ for r_0 and r_1 and that (3) corresponds to the fully flexible case without additional restrictions on r_0 and r_1 (like e.g. $r_1 \geqslant r_0$). The generalization to different boundaries and additional restrictions on (r_1, r_0) are formally straightforward. Clearly, $\widetilde{\mathrm{CB}}$ depends on the restrictions made for the ratios r_i.

To assess the worst case reassessment rule for a given interim result, the true means of treatment and control group have to be known. However, our intention is to evaluate an upper bound for the overall Bias. The maximum Bias B^* is evaluated by integrating the maximum conditional Bias over all interim outcomes:

$$\mathrm{B}^*(n, \sigma, r_{\min}, r_{\max}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \widetilde{\mathrm{CB}}(z_0, z_1, n, \sigma, r_{\min}, r_{\max})\phi(z_1)\phi(z_0)dz_1 dz_0 \,. \tag{4}$$

Obviously, the maximum Bias $\mathrm{B}^*(n, \sigma, r_{\min}, r_{\max})$ does not depend on the unknown μ_0 and μ_1.

Fortunately, (2) is the sum of two terms depending only on r_1 or r_0. Hence, in the fully flexible case, we can maximize each term separately in r_1 or r_0, respectively. Denoting the worst case sample size fractions by $\tilde{r}_i$, $i \in \{0, 1\}$, we obtain $\tilde{r}_1 = r_{\min}$ for $z_1 > 0$ and $\tilde{r}_1 = r_{\max}$ for $z_1 < 0$. Similarly, $\tilde{r}_0 = r_{\max}$ for $z_0 > 0$ and $\tilde{r}_0 = r_{\min}$ for $z_0 < 0$. Figure (A) in the Appendix shows the four subsets of the interim outcome space corresponding the four values of the tuple $(\tilde{r}_0, \tilde{r}_1)$. The maximum Bias, B^*, is obtained by integrating the maximum conditional Bias in each subset and summing up the four integrals. This leads to

$$\mathrm{B}^*(n, \sigma, r_{\min}, r_{\max}) = \frac{2\sigma}{\sqrt{n}}\phi(0)\left[\frac{1}{1 + r_{\min}} - \frac{1}{1 + r_{\max}}\right] \approx \frac{0.8\sigma}{\sqrt{n}}\left[\frac{1}{1 + r_{\min}} - \frac{1}{1 + r_{\max}}\right]. \tag{5}$$

Without any restrictions on the second-stage sample size reassessment rule, that is, setting $(r_{\min}, r_{\max}) = (0, \infty)$, the maximum Bias simplifies to $\mathrm{B}^*(n, \sigma, 0, \infty) = 2\phi(0)\sigma/\sqrt{n} \approx 0.8\sigma/\sqrt{n}$. Comparison of (1)

Table I. The standardized maximum Bias, $B_k^* \sqrt{n/(2\sigma^2)}$, as well as the standardized maximum root mean squared error, $\sqrt{MSE_k^*(n/(2\sigma^2))}$, for different restrictions for the sample size reassessment rules: flexible r_s and r_0, balanced second-stage sample size ($r_s = r_0$), a larger second stage sample size in the treatment group ($r_s \geqslant r_0$) and fixing the sample size in the control (fix r_0). Values are given for $r_{min} = 0$, 0.5 and 1 setting $r_{max} = \infty$. For comparison, the fixed design with $r_{min} = r_{max}$ is given showing the maximum selection Bias and mean squared error, respectively.

		r_{min}					
		$B_k^* \sqrt{n/(2\sigma^2)}$			$\sqrt{MSE_k^* n/(2\sigma^2)}$		
k	type	0	0.5	1	0	0.5	1
1	$r_{min} = r_{max}$	0.000	0.000	0.000	1.000	0.817	0.707
	flexible	0.564	0.376	0.282	1.129	0.859	0.723
	$r_1 \geqslant r_0$	0.482	0.321	0.241	1.092	0.843	0.717
	$r_1 = r_0$	0.399	0.266	0.199	1.039	0.820	0.707
	fix r_0	0.282	0.188	0.141	1.080	0.842	0.717
2	$r_{min} = r_{max}$	0.399	0.266	0.199	1.246	0.955	0.799
	flexible	0.764	0.509	0.382	1.320	0.980	0.809
	$r_s \geqslant r_0$	0.628	0.419	0.314	1.276	0.963	0.801
	$r_s = r_0$	0.598	0.399	0.299	1.258	0.956	0.799
	fix r_0	0.482	0.321	0.241	1.271	0.962	0.801
3	$r_{min} = r_{max}$	0.598	0.399	0.299	1.389	1.040	0.856
	flexible	0.910	0.607	0.455	1.446	1.059	0.864
	$r_s \geqslant r_0$	0.739	0.493	0.370	1.402	1.042	0.857
	$r_s = r_0$	0.728	0.485	0.364	1.395	1.040	0.856
	fix r_0	0.628	0.419	0.314	1.399	1.042	0.857
4	$r_{min} = r_{max}$	0.728	0.485	0.364	1.489	1.099	0.897
	flexible	1.022	0.681	0.511	1.537	1.117	0.904
	$r_s \geqslant r_0$	0.827	0.551	0.414	1.495	1.100	0.897
	$r_s = r_0$	0.882	0.548	0.411	1.492	1.099	0.897
	fix r_0	0.739	0.493	0.370	1.493	1.100	0.897
5	$r_{min} = r_{max}$	0.822	0.548	0.411	1.565	1.145	0.929
	flexible	1.109	0.739	0.555	1.608	1.161	0.935
	$r_s \geqslant r_0$	0.898	0.599	0.449	1.567	1.146	0.929
	$r_s = r_0$	0.896	0.597	0.488	1.566	1.145	0.929
	fix r_0	0.827	0.551	0.414	1.567	1.145	0.926
6	$r_{min} = r_{max}$	0.895	0.597	0.448	1.625	1.181	0.954
	flexible	1.180	0.787	0.590	1.666	1.197	0.960
	$r_s \geqslant r_0$	0.957	0.638	0.479	1.627	1.182	0.954
	$r_s = r_0$	0.956	0.637	0.478	1.627	1.181	0.954
	fix r_0	0.898	0.599	0.449	1.626	1.182	0.954

and (5) reveals, that when dropping the constraint of equal second-stage sample sizes, the maximum Bias is increased by the factor $\sqrt{2}$, that is, by about 41%.

We finally note how to account for constraints like $r_1 \geqslant r_0$. To account for $r_1 \geqslant r_0$, we need to rule out that $r_1 = r_{min}$ and $r_0 = r_{max}$. For $z_0 > 0$ and $z_1 > 0$, we therefore maximize $CB(z_0, z_1, r_0, r_1, n, \sigma)$ under the assumption $r_1 = r_0$. In this case, the maximum depends on $z_1 - z_0$: it is attained for $r_1 = r_0 = r_{min}$ if $z_1 - z_0 > 0$ and otherwise for $r_1 = r_0 = r_{max}$. The maximization of $CB(z_0, z_1, r_0, r_1, n, \sigma)$ under the constraint of a fixed r_0 follows similar lines as in the fully flexible case (leading to a rule that depends on z_1 only).

3.1.6. Reshuffling. Assume now that a sample size of n_g patients per group is pre-planned over both stages, resulting in a total of $2n_g$ patients in the trial. This overall patient number is kept fix. The interim analysis is performed after recruitment of tn_g patients per group where $t \in (0,1)$. The ratio t can be interpreted as the timing of the interim analysis. To keep the overall sample size, the second stage needs to consist of $2(1-t)n_g$ patients in total. This number is allocated to the experimental treatment and control group in a data dependent manner. This means that in the interim analysis, a second-stage sample size allocation rate v, $0 \leqslant v \leqslant 1$, is chosen based on the interim results, such that in the second stage a number of $v2(1-t)n_g$ patients is allocated to the control and $(1-v)2(1-t)n_g$ to the experimental treatment group. The conditional Bias (2) can be rewritten as follows:

$$\mathrm{CB}(z_0, z_1, v, n_g, t, \sigma) = \frac{\sigma}{\sqrt{tn_g}} \left[\frac{z_1}{1+(1-v)w_t} - \frac{z_0}{1+vw_t} \right] \tag{6}$$

where we use the notation $w_t = 2(\frac{1}{t} - 1)$ for mathematical convenience. The allocation ratio $0 \leqslant v \leqslant 1$ is now searched to maximize the conditional Bias (6). By setting the first derivative of the conditional Bias to zero, we obtain a quadratic equation with the two roots

$$v^{(1),(2)} = \frac{-z_1 + z_0(w_t+1) \pm \sqrt{-z_0 z_1}(w_t+2)}{(z_0+z_1)w_t}$$

that are candidates for the $\tilde{v}$ maximizing the conditional Bias. The candidates $v^{(1)}$ and $v^{(2)}$ do only exist if z_0 and z_1 have different signs and $z_0 \neq -z_1$. If $z_0 = -z_1 > 0$, one can see from (6) that the conditional Bias is maximized for $v^{(3)} = 1/2$. Furthermore, $v^{(1)}$ and $v^{(2)}$ are ineligible if larger than 1 or smaller than 0. Whether $v^{(1)}$ or $v^{(2)}$ is actually the maximizer depends on z_0 and z_1. To assess the global maximum, the candidates $v^{(4)} = 0$ and $v^{(5)} = 1$ also have to be investigated. Note that for $z_0 = -z_1 < 0$, candidates $v^{(4)}$ and $v^{(5)}$ coincide and show the maximizer of the conditional Bias. The worst case conditional Bias is the maximum of the five candidates.

$$\widetilde{\mathrm{CB}}(z_0, z_1, n_g, t, \sigma) = \max_{i=1,\ldots,5 : 0 \leqslant v^{(i)} \leqslant 1} \left[\mathrm{CB}(z_0, z_1, v^{(i)}, n_g, t, \sigma) \right]. \tag{7}$$

Figure (B) in the Appendix shows the subspaces of the interim outcome in terms of the standardized means in the treatment and control groups corresponding to the different maximizer $v^{(i)}$ for $t = 0.5$. The white area gives the subspace where either $v^{(1)}$ or $v^{(2)}$ are the global maximum. The dashed line gives the subspace, where $v^{(3)}$ is the global maximum. It can be seen that $v^{(1)}$ is no global optimum for $t = 0.5$. Numerical integration can be used to compute the overall Bias:

$$\mathrm{B}^*(n_g, t, \sigma) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \widetilde{\mathrm{CB}}(z_0, z_1, n_g, t, \sigma) \phi(z_1) \phi(z_0) dz_1 dz_0. \tag{8}$$

For numerical integration, we used the R-package `R2Cuba` [21]. In the following, we also show results under the restriction $0 \leqslant v \leqslant 0.5$, which guarantees that the second-stage sample size of the experimental treatment group is never smaller than that of the control group. The solid black line, marked with 1 in Figure 3 (A), shows the standardized maximum Bias as a function of the timing of the interim analysis t for $k = 1$ and $0 \leqslant v \leqslant 1$. The maximum Bias is now standardized by the standard error of a fixed-size-sample test with per group sample size n_g, that is, $\sqrt{2\sigma^2/n_g}$. We are not standardizing with the standard error of the interim estimate because it depends on t. The dashed line (marked with 1) gives the standardized maximum Bias under the restriction $0 \leqslant v \leqslant 0.5$. One can see that (for $k = 1$) the standardized maximum Bias is decreasing with increasing t, that is, the later interim analysis, the smaller the maximum Bias. This is due to the larger first and smaller second-stage sample sizes. For $t = 0.5$, that is, planning the interim analysis half way, the standardized maximum Bias is 0.40 if $0 \leqslant v \leqslant 1$ and decreases to 0.21 if $0 \leqslant v \leqslant 0.5$.

3.2. Maximum mean squared error

To maximize the MSE, we proceed similar to calculating the maximum Bias. For each interim outcome, the sample size reassessment rule is searched that maximize the conditional MSE (worst case). The conditional MSE, given the interim outcome, can be calculated as follows (see Appendix A.1):

$$\begin{aligned}\mathrm{CMSE}(z_0, z_1, r_0, r_1, n, \sigma) &= \mathbf{E}\left[\left((\bar{X}_1 - \bar{X}_0) - (\mu_1 - \mu_0)\right)^2 \mid Z_{(i,1)} = z_i, i = 0, 1\right] \\ &= \frac{\sigma^2}{n}\left[\left(\frac{z_1}{1+r_1} - \frac{z_0}{1+r_0}\right)^2 + \frac{r_1}{(1+r_1)^2} + \frac{r_0}{(1+r_0)^2}\right]\end{aligned} \tag{9}$$

For each z_0 and z_1, r_0 and r_1 are searched to maximize the CMSE:

$$\widetilde{\mathrm{CMSE}}\left(z_0, z_1, n, \sigma, r_{\min}, r_{\max}\right) = \max_{r_{\min} \leqslant r_0, r_1 \leqslant r_{\max}} \mathrm{CMSE}\left(z_0, z_1, n, \sigma, r_0, r_1\right) \tag{10}$$

where $r_{\min}$ and $r_{\max}$ again denote the lower and upper bounds for the second-to-first-stage ratios r_i, $i = 0, 1$. Additional constraints on (r_0, r_1) like $r_1 \geqslant r_0$ need to be accounted in the maximum (10). Integrating over all interim outcomes gives the maximum MSE, denoted by MSE^* in the sequel,

$$\mathrm{MSE}^*(n, \sigma, r_{\min}, r_{\max}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \widetilde{\mathrm{CMSE}}(z_0, z_1, n, \sigma, r_{\min}, r_{\max})\phi(z_0)\phi(z_1)dz_0dz_1\ . \tag{11}$$

Note that MSE^* is also independent from the group means μ_i, $i = 0, 1$.

3.2.1. Flexible second-to-first-stage ratios. We start investigating the case of completely flexible r_1 and r_0 within the boundaries $(r_{\min}, r_{\max})$. Note again that we assume equal bounds for the treatment and the control group; however, the sample size reassessment rule for the treatment and control group can be different. To maximize the CMSE in (9), for given z_0 and z_1 at interim, a total of nine candidates have to be investigated, and the global maximum is the maximum over these nine candidates. Integrating over all interim outcomes gives the MSE^*. Details can be found in Appendix A.2.

The solid lines in Figure 2 (A) show the maximum RMSE, say RMSE^*, divided by the standard error of the first-stage mean difference, that is, $\sqrt{\mathrm{MSE}^*}/\sqrt{2\sigma^2/n}$. Note that we use here the same standardization as for the maximum Bias and that the standardized RMSE^* does not depend on n or σ. As for the Bias, for increasing $r_{\min}$ and decreasing $r_{\max}$, the standardized RMSE^* is decreasing. Setting $r_{\max} = 2$ and $r_{\min} = 0$, RMSE^* is 1.10 times first-stage standard error. Increasing $r_{\min}$ to 0.5 or 1, the values are decreasing to 0.84 and 0.71.

The gray horizontal line through 1 represents the standardized RMSE of the first-stage mean difference. For the investigated $r_{\min} \geqslant 0.5$, the standardized RMSE^* is always smaller than 1, meaning that we gain in precision from the second stage, independently of the sample-size reassessment rule. If $r_{\min} = 0$, RMSE^* is larger than the first-stage RMSE, indicating that we can lose in precision if sample sizes are reassessed and the trial can be stopped at interim (compared with a trial that consist of the first stage only). The latter is because of the Bias that is possible under sample-size reductions and early stopping (Figure 1 (A)).

Setting $r_{\min} = r_{\max} > 0$ gives the RMSE to a fixed-size sample test with a sample size larger than the first stage. For example, $r_{\min} = r_{\max} = 0.5$, the standardized RMSE is 0.82 decreasing to 0.71 for $r_{\min} = r_{\max} = 1$. It is interesting to see that for $r_{\min} \geqslant 0.5$, RMSE^* under flexible sample-size reassessments is only slightly increasing in $r_{\max}$ and remains close to the RMSE of the fixed-size-sample test with second stage per group sample size $r_{\min}n$. Hence, for sufficiently large $r_{\min}$, the Bias from any adaptive sample increase will not have a substantial negative effect on the precision of the overall maximum likelihood estimate.

The rows 'flexible' in Table I shows the standardized RMSE^* when setting $r_{\max} = \infty$ for $r_{\min} = 0$, 0.5, 1. Without any restrictions on the reassessment rule, setting $(r_{\min}, r_{\max}) = (0, \infty)$, the standardized RMSE^* is 1.13. Setting $r_{\min} = 1$ and $r_{\max} = \infty$, it is 0.72 as compared with 0.71 for the corresponding fixed-size-sample test.

3.2.2. Balanced second-stage sample sizes. Restricting the second-stage sample size to be balanced between the treatment groups $(r = r_1 = r_0)$ reduces the CMSE (9) to

$$\mathrm{CMSE}(y, n, \sigma, r) = \frac{2\sigma^2}{n(1+r)^2}\left(y^2 + r\right), \tag{12}$$

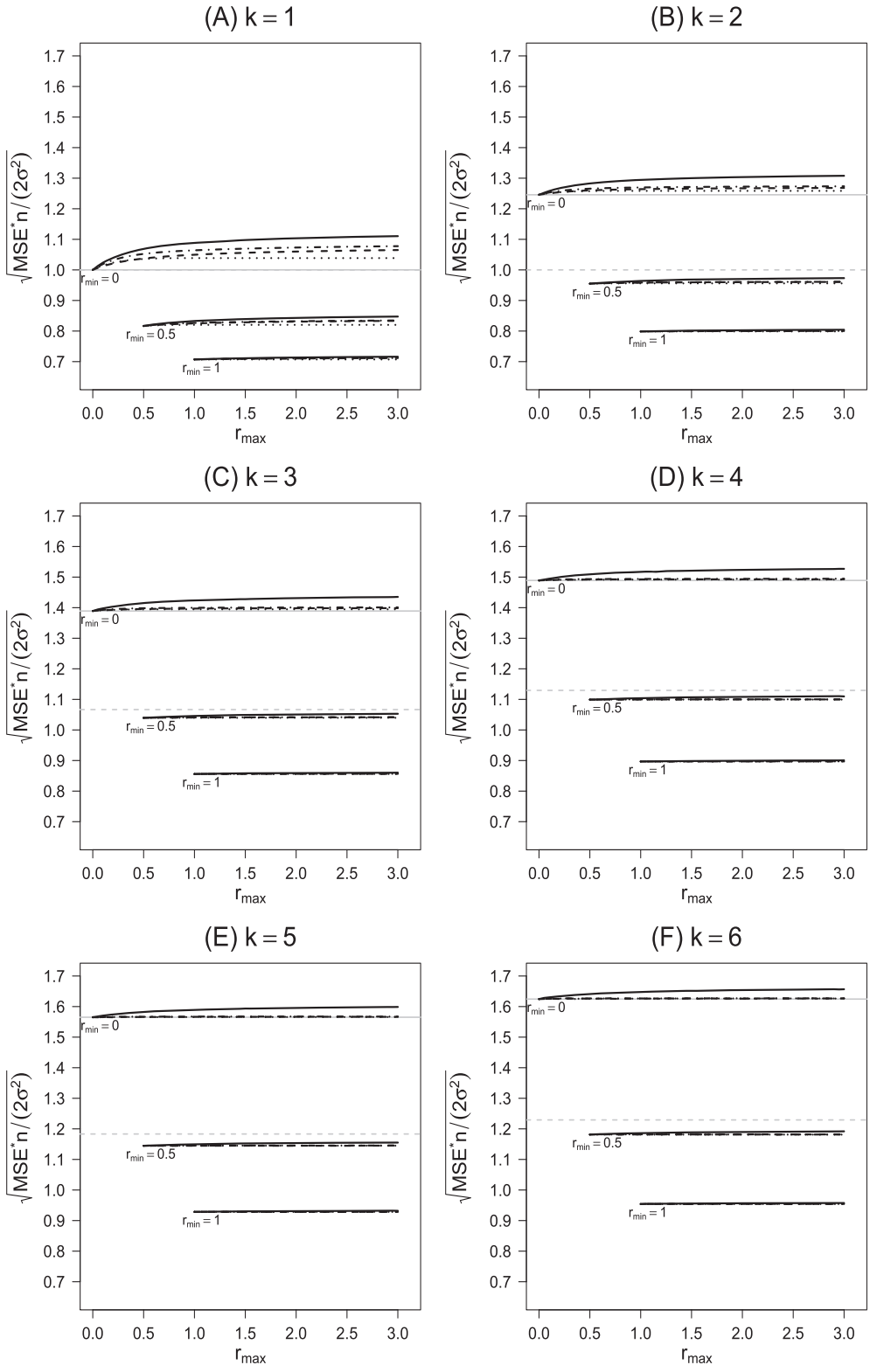

Figure 2. Standardized maximum root mean squared error (RMSE) as a function of r_{max} for $r_{min} = 0, 0.5$, and 1. Values are given for a number of $k = 1$ to 6 treatments in panels (A) to (F). Within one panel, the standardized maximum RMSE is shown for different restrictions on the sample-size reassessment rule: flexible second-to-first-stage ratios (solid lines), equal second-to-first-stage ratios (dotted lines), restricting $r_1 \geqslant r_0$ (dot-dashed lines), and fixing the control (dashed lines). The solid gray horizontal line shows the standardized maximum RMSE for a fixed-size-sample test with post-trial selection. The dashed gray horizontal line shows the standardized RMSE of a fixed-sample-size test when selecting the treatment with the maximum effect at the end.

where $y = (z_1 - z_0)/\sqrt{2}$ is standard and normally distributed. Setting the first derivative to 0, a candidate for the global maximum can be evaluated by $r^{(1)} = 1 - (z_1 - z_0)^2$. By calculating the second derivative at the point $r^{(1)}$, it can be shown, that $r^{(1)}$ is a maximum if $|\ (z_1 - z_0)\ | \leqslant \sqrt{2}$. This candidate is the global maximum if $r_{\min} \leqslant r^{(1)} \leqslant r_{\max}$; otherwise, the global maximum is achieved for $r^{(2)} = r_{\min}$ or $r^{(3)} = r_{\max}$. The worst case CMSE can be evaluated as the maximum over all three candidates.

$$\widetilde{\mathrm{CMSE}}(y, n, \sigma, r_{\min}, r_{\max}) = \max_{i=1,2,3 : r_{\min} \leqslant r^{(i)} \leqslant r_{\max}} \mathrm{CMSE}(y, n, \sigma, r^{(i)}), \tag{13}$$

and the maximum MSE is obtained by numerical integration

$$\mathrm{MSE}^*(n, \sigma, r_{\min}, r_{\max}) = \int_{-\infty}^{\infty} \widetilde{\mathrm{CMSE}}(y, n, \sigma, r_{\min}, r_{\max}) \phi(y) dy$$

The dotted lines in Figure 2 (A) show the standardized RMSE* for the case of equal second-stage sample sizes in the groups. The restriction to balanced sample sizes decreases the RMSE*, the decrease being smaller for larger $r_{\min}$. Setting $r_{\max} = 2$, the standardized RMSE* is 1.04, 0.82, and 0.71 for $r_{\min} = 0$, 0.5, or 1, respectively. Note that, for $r_{\min} \geqslant 0.5$, the lines corresponding to the different restrictions are indistinguishable. The rows '$r_1 = r_0$' for $k = 1$ in Table I show the standardized RMSE* for $r_{\max} = \infty$. Without any restrictions ($r_{\min} = 0$), the standardized RMSE* is 1.04 and, hence, can still be larger than the RMSE of the first stage. For $r_{\min} = 0.5$ and 1, the standardized RMSE* is more or less equal to the standardized RMSE of the corresponding fixed-size-sample test with per group sample size $r_{\min} n$. This shows that, for sufficiently large $r_{\min}$, the worst case Bias from data driven, balanced second-stage sample size increases has a more or less negligible effect on the precision of the overall maximum likelihood estimate.

3.2.3. Restricting the treatment to $r_1 \geqslant r_0$. The dot-dashed lines in Figure 2 (A) show the standardized RMSE* when restricting the sample size of the treatment group to be as least as large as the sample size of the control group. Setting $r_{\max} = 2$, the standardized RMSE* is 1.07, 0.83 and 0.71 for $r_{\min} = 0$, 0.5 or 1, respectively, which is only slightly larger than RMSE* under balanced second-stage sample sizes. The rows '$r_1 \geqslant r_0$' for $k = 1$ in Table I shows the maximum for $r_{\max} = \infty$. Without any restrictions ($r_{\min} = 0$), the standardized RMSE* is 1.09. Here, we see some inflation of the RMSE* compared with the one under the constraint $r_1 = r_0$.

3.2.4. Fixing r_0. Note that, when fixing r_0, the MLE of the control group is unbiased. In the maximization of the CMSE only three of the nine candidates remain. Two candidates are derived by setting $r_1 = r_0$ ($= r_{\min}$) or $r_1 = r_{\max}$. The third candidate can be calculated as candidate $r^{(6)}$ in the maximization of the CMSE with flexible ratios in Appendix A.2. with $r_{\min}$ replaced by r_0.

The dashed lines in Figure 2 (A) show the standardized RMSE*, assuming $r_0 = 0, 0.5$, or 1 and $r_0 \leqslant r_1 \leqslant r_{\max}$. As expected, the standardized RMSE* is smaller than the ones with flexible reassessment in the control; however, the difference decreases with increasing $r_{\min}$. Setting $r_{\max} = 2$, the standardized RMSE* is 1.06, 0.83, and 0.71 for $r_0 = 0$, 0.5, and 1, respectively. We can see from Table I and Figure 2 (A) that, for sufficiently large $r_{\min}$ or large $r_{\max}$, the differences in RMSE* between the rule with fixed r_0 and the one with $r_1 \geqslant r_0$ are only small, so that there is no substantial gain in (minimum) precision from fixing r_0 (the lines in Figure (A) are indistinguishable).

3.2.5. Reshuffling. In case of reshuffling, the CMSE can be rewritten as follows:

$$\mathrm{CMSE}(z_0, z_1, v, n_g, t, \sigma) = \frac{\sigma^2}{t n_g} \left[\left(\frac{z_1}{1 + (1-v) w_t} - \frac{z_0}{1 + v w_t} \right)^2 + \frac{(1-v) w_t}{(1 + (1-v) w_t)^2} + \frac{v w_t}{(1 + v w_t)^2} \right], \tag{14}$$

where, as before, $w_t = 2(\frac{1}{t} - 1)$. Recall that the per-group first-stage sample size is $t n_g$, and the total overall sample size is fixed at $2 n_g$. At the second stage, $v 2(1-t) n_g$ patients are allocated to the control and $(1-v) 2 (1-t) n_g$ patients to the treatment group.

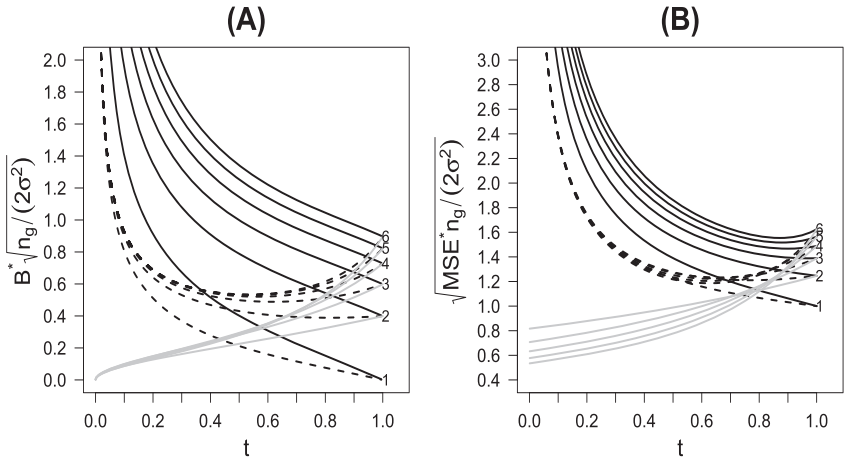

Figure 3. Standardized maximum Bias, panel (A), and the standardized maximum root mean squared error, panel (B), as a function of the timing t of the interim analysis for $k = 1$ to 6 treatments compared with one common control for the case of full (solid black lines) and restricted reshuffling (dashed black-lines). For comparison, the standardized Bias and root mean squared error are given for an adaptive design with treatment selection and a sample size of $(1-t)n_g(k+1)/2$ in the second stage (gray lines).

By setting the first derivative of (14) to zero, candidates for the global maximum are found. This problem can be reduced to finding the roots of a third-degree polynomial, and therefore, at maximum three candidates ($v^{(1)}, v^{(2)}, v^{(3)}$) must be assessed. Note that we did not derive these candidates analytically. Instead, we used the R-function polyroot [22] for the numerical root finding. Considering furthermore $v^{(4)} = 0$ and $v^{(5)} = 1$, the worst case CMSE is the maximum over five candidates. Integrating over all interim outcomes gives the maximum MSE, denoted as before by MSE*. Figure (D) in the Appendix gives the subspaces of the interim outcome of treatment and control to evaluate the worst case CMSE. In the white area either $v^{(1)}$, $v^{(2)}$ or $v^{(3)}$ are the global maximizer. As for the maximum Bias, we will, furthermore, also give results when restricting $0 \leqslant v \leqslant 0.5$, which means that a larger sample size has to be allocated to the treatment group.

The solid line marked with 1 (the case $k = 1$) in Figure 3 (B) shows RMSE* for $0 \leqslant v \leqslant 1$ divided by the standard error of a fixed-size-sample test with per-group sample size n_g. The standardized RMSE* is plotted as a function of the timing of the interim analysis t. The dashed line marked with 1 gives the corresponding standardized RMSE* if $0 \leqslant v \leqslant 0.5$. Like the Bias, the standardized RMSE* is decreasing with increasing t. For $t = 0.5$, the standardized RMSE* is 1.39 if $0 \leqslant v \leqslant 1$ and decreases to 1.23 if $0 \leqslant v \leqslant 0.5$. Note that the standardized RMSE* is always larger than 1, that is, the RMSE with sample reshuffling (between the experimental and control group) is always larger, and for small t substantially larger, than the RMSE of the reference fixed-sample design with the same overall sample size.

4. Multi-arm trials with interim treatment selection

In this section, we consider two-stage designs, which start with a control and $k > 1$ experimental treatment groups and where one experimental treatment, say treatment $s \in \{1, \ldots, k\}$, and the control are selected for the second stage. The second stage sample sizes are then set based on the interim results. Again, we assume balanced sample sizes in the first stage, while in some of our rules the second-stage sample sizes are permitted to be unbalanced.

4.1. Maximum Bias

To evaluate the maximum Bias, we search for the selection and sample size adaptation rule that maximize Bias. These are obtained by first maximizing for each MLE, $\bar{X}_i - \bar{X}_0$, the conditional Bias (see formula (2) for $k = 1$) with respect to the sample size fractions r_i and r_0 and then selecting the treatment s with largest maximized conditional Bias;

$$s = \arg\max_{i=1,\ldots,k} \widetilde{\mathrm{CB}}\left(z_0, z_i, n, \sigma, r_{\min}, r_{\max}\right) \tag{15}$$

Integrating over all interim outcomes gives the worst case Bias;

$$\mathrm{B}_k^*\left(n, \sigma, r_{\min}, r_{\max}\right) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \widetilde{\mathrm{CB}}\left(z_0, z_s, n, \sigma, r_{\min}, r_{\max}\right) \phi(z_k) \ldots \phi(z_0) dz_k \cdots dz_0, \tag{16}$$

where s is data dependently determined as in (15) and z_s, z_0 denote the observed interim outcome of the selected treatment and control group, respectively. Note that, for the given z_i, each $\widetilde{\mathrm{CB}}(z_0, z_i, n, \sigma, r_{\min}, r_{\max})$ can be calculated according to the case of $k = 1$ (Section 3.1).

4.1.1. Flexible second-to-first-stage ratios. In case of flexible second-to-first-stage ratios, r_0 is maximized independently from r_s (see formula (2) with r_1 replaced by r_s). Because the conditional Bias and thereby also its maximum $\widetilde{\mathrm{CB}}$ are increasing in z_s for fixed z_0, we have

$$\max_{i=1,\ldots,k} \widetilde{\mathrm{CB}}\left(z_0, z_i, n, \sigma, r_{\min}, r_{\max}\right) = \widetilde{\mathrm{CB}}\left(z_0, \max_{i=1,\ldots,k} z_i, n, \sigma, r_{\min}, r_{\max}\right). \tag{17}$$

This means that the treatment with the largest worst case conditional Bias at interim is the treatment with the largest observed z_i, that is, $z_s = \max_{i=1,\ldots,k} z_i$ and (15) reduces to the selection of treatment

$$s = \arg\max_{i=1,\ldots,k} \widetilde{\mathrm{CB}}\left(z_0, z_i, n, \sigma, r_{\min}, r_{\max}\right) = \arg\max_{i=1,\ldots,k} z_i$$

The maximum Bias (16) can therefore be reduced to

$$\mathrm{B}_k^*\left(n, \sigma, r_{\min}, r_{\max}\right) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \widetilde{\mathrm{CB}}\left(z_0, z_s, n, \sigma, r_{\min}, r_{\max}\right) k\Phi(z_s)^{k-1} \phi(z_s)\phi(z_0) dz_s dz_0, \tag{18}$$

where Φ denotes the cumulative distribution function of the standard normal distribution. Note, the probability density function of the maximum of independent standard normal distributions is $k\Phi(x)^{k-1}\phi(x)$. Like for $k = 1$, the maximum Bias is independent from μ_i for all $i = 0, 1, \ldots, k$.

The solid lines in Figure 1 (B) to (F) show the standardized maximum Bias for $k = 2$ to 6 as a function of $r_{\max}$ for $r_{\min} = 0$, 0.5, and 1. The maximum Bias is standardized by the first-stage standard error $\sqrt{2\sigma^2/n}$ of one treatment-to-control comparison. Because of this standardization, the shown Biases are also independent of the first stage sample size n and the common variance σ^2.

The gray horizontal line shows the standardized Bias for a fixed sample size of n patients per treatment group and post-trial selection [20], which results here from setting $r_{\min} = r_{\max} = 0$. Setting $r_{\min} = r_{\max} > 0$ gives an adaptive design where in an interim analysis, one treatment and the control are selected for the second stage and a second stage with a fixed sample size is performed. This means that $r_{\min} = r_{\max}$ gives the selection Bias without any additional Bias because of sample size reassessment.

As expected, the standardized maximum Bias is increasing with increasing k. For flexible reassessment rules, the difference of the maximum Bias to the pure selection Bias (the case $r_{\min} = r_{\max}$) is large for all shown k; however, it is not increasing (rather decreasing) in k. Again, the maximum Bias is effectively decreased by increasing $r_{\min}$.

4.1.2. Maximum Bias under restrictions on the second-stage sample-size ratios. Obviously, equality (17) holds also under restrictions like $r_0 = r_s$, $r_0 \leqslant r_s$ or a fixed r_0. Hence, we can also utilize the mathematical results from Section 3.1 when restricting the second-to-first-stage ratios.

The dotted and the dashed-dotted lines in Figure 1 (B) to (F) show the standardized maximum Biases for $k = 2$ to 6 with balanced second-stage sample sizes ($r_s = r_0$) and under the restriction $r_s \geqslant r_0$, respectively. Both restrictions substantially reduce the maximum Bias. For $k \geqslant 2$, we see only a small difference between the maximum Bias with balanced sample sizes and the one with the restriction $r_s \geqslant r_0$.

The difference of the maximum Bias to the selection Bias ($r_{\min} = r_{\max}$), that is, the additional Bias due to sample size reassessment is still rather large for $r_{\min} = 0$ but becomes substantially smaller for $r_{\min} \geqslant 0.5$ and is decreasing with increasing number of treatments k. This means that with a larger number of treatments the selection Bias dominates the Bias from data-driven sample-size reassessments. Fixing r_0 (dashed lines) leads to a further reduction in the maximum Bias, which is then close to the selection Bias.

4.1.3. Maximum Bias for $r_{\max} = \infty$. To investigate the influence of $r_{\min}$ more carefully, we give in Table I the standardized maximum Biases for different $r_{\min}$, setting $r_{\max} = \infty$ also for $k \geqslant 2$. For comparison, the rows $r_{\min} = r_{\max}$ contain the selection Bias without any sample size reassessments. The maximum Bias is decreasing in $r_{\min}$ also in this fixed sample size case because of the increasing second stage sample sizes. Like for the case $k = 1$, the reduction in the maximum Bias due to an increase in $r_{\min}$ is more or less independent from the further restrictions on r_s and r_0, and it seems independent from k: the maximum Bias always decreases by about 33% when setting $r_{\min} = 0.5$ (compared with $r_{\min} = 0$) and by about 50% for $r_{\min} = 1$.

The table confirms the finding that the restriction $r_s \geqslant r_0$ leads to a substantial reduction in the maximum possible Bias, while the restriction to balanced second-stage sample sizes does not lead to a substantial further reduction. For $k \geqslant 2$ and $r_{\min} \geqslant 0.5$, fixing r_0 has some (but not a large) additional effect on the maximum Bias and brings the Bias close to the pure selection Bias ($r_{\min} = r_{\max}$). We may deduce from this findings that a data driven increase in the sample size for the selected experimental treatment group will (when initially large enough) not lead to a substantially additional Bias.

4.1.4. Reshuffling. Like in Section 3.2.5, we assume now that a total sample size of n_g patients per-group is pre-planned for the two stages, whereby tn_g per-group are used in the first stage, t denoting the timing of the interim analysis. As a consequence, the overall pre-planned second-stage sample size is $(1-t)n_g(k+1)$. Now, in the interim analysis, one treatment is selected and the second-stage sample size $(1-t)n_g(k+1)$ is reshuffled between the selected treatment and control that means that for some $v \in (0, 1)$, $(1-v)(1-t)n_g(k+1)$ patients are allocated to the selected experimental treatment and $v(1-t)n_g(k+1)$ patients to the control group. The conditional Bias can be calculated to be (6) with $w_t = (k+1)/t-(k+1)$. Note that the sample size over both stages is $tn_g + (1-v)(1-t)n_g(k+1)$ in the selected treatment and $tn_g + v(1-t)n_g(k+1)$ in the control group. It can be shown that equality (17) holds also in the case of reshuffling (Appendix A.3), so that in the calculation of the maximum Bias B_k^*, the $(k+1)$ dimensional integral can be reduced to a two-dimensional integral.

Figure 3 (A) shows values of B_k^* standardized by the standard error of a two-group fixed-size-sample test with sample size n_g, that is, $\sqrt{2\sigma^2/n_g}$. The standardized maximum Bias is shown as function of t for $k = 1$ to 6. The solid lines show the values for $0 \leqslant v \leqslant 1$ and the dashed lines for $0 \leqslant v \leqslant 0.5$. Recall that $v \leqslant 0.5$ corresponds to the constraint that the control group is smaller or as large as the selected experimental treatment group. For comparison, the gray solid lines give the maximum Bias for an adaptive design with interim selection of one treatment and control at time point t, the second-stage sample size being $(1-t)n_g(k+1)/2$ per-group. This is the selection Bias without additional sample size reassessment. The selection Bias is 0 for $t = 0$ because we then perform a fixed-sample-size test with only one treatment and control. It is increasing with increasing t, being the selection Bias for a trial with post-trial selection for $t = 1$. This is equivalent to setting $r_{\min} = r_{\max} = 0$ in Figure 1. The standardized maximum Bias (including selection Bias and the Bias due to sample size reassessment) is decreasing with increasing t for $v \leqslant 1$; however, there is a non-monotonous behavior for the standardized maximum Bias if restricting v to be smaller than 0.5. The maximum Bias is depending on both the selection Bias and the Bias for additional sample-size reassessment. The selection Bias is increasing with t, and the Bias due to sample size reassessment is decreasing with t. This leads to a tradeoff between both types of Bias for $k \geqslant 1$.

For $t = 0.5$ and $k = 2, 3, 4$, the standardized maximum Bias is 0.80, 1.00, 1.14 if $v \leqslant 1$ and 0.43, 0.50, 0.52 if $v \leqslant 0.5$, respectively. For comparison, the selection Bias is 0.23, 0.28, and 0.29 for $k = 2, 3$, and 4. In summary, a sample-size reshuffling between the selected treatment and the control group can lead to a substantial Bias. The maximum Bias is halved by the constraint that the control group is never larger than the selected experimental treatment group.

4.2. Maximum mean squared error

To evaluate the maximum MSE, we proceed similar to evaluating the maximum Bias. The selection rule to maximize the MSE is to select the treatment with the maximum worst case CMSE based on the interim result:

$$s = \arg\max_{i=1,\ldots,k} \widetilde{\mathrm{CMSE}}\left(z_0, z_i, n, \sigma, r_{\min}, r_{\max}\right) \tag{19}$$

Note that the treatment with the maximum worst case CMSE is not necessarily the treatment with the maximum observed z_i at interim or the treatment with the maximum absolute difference to the control $|z_i - z_0|$, because the conditional error (9) cannot be written as a function of $z_i - z_0$. The maximum MSE is a $(k+1)$ dimensional integral over all interim outcomes:

$$\mathrm{MSE}_k^*\left(n, \sigma, r_{\min}, r_{\max}\right) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \widetilde{\mathrm{CMSE}}\left(z_0, z_s, n, \sigma, r_{\min}, r_{\max}\right) \phi(z_k) \ldots \phi(z_0) dz_k \ldots dz_0, \tag{20}$$

where $\widetilde{\mathrm{CMSE}}$ is calculated as discussed in Section 3.2 and s is chosen as in (19). To evaluate the $(k+1)$ dimensional integral, numerical integration was performed using the R-package `R2Cuba` [21].

4.2.1. Results. Figure 2 (B) to (F) shows the standardized RMSE_k^* for $k = 2$ to 6. As for $k = 1$, the maximum RMSE was standardized by the standard error of the first stage, that is, $\sqrt{MSE_k^*}/\sqrt{2\sigma^2/n}$. The solid lines show the scenario with full flexibility on the reassessment rules within the boundary $(r_{\min}, r_{\max})$, the dashed lines when fixing the sample size of the control group. The dot-dashed lines show the values when restricting the sample size of the treatment to be larger as the sample size of the control $(r_0 \geqslant r_s)$ and the dotted lines when restricting the second stage sample size to be balanced $(r_0 = r_s)$. For comparison, the dashed gray horizontal line shows the standardized RMSE of a fixed-sample-size test when selecting the treatment with the maximum effect at the end. The solid gray horizontal line represents the case $r_{\min} = r_{\max} = 0$ where we select after n observations per group the treatment with maximum CMSE. By definition (see also formula (9)), for $r_{\min} = r_{\max} = 0$, the CMSE is simply the square of the difference between the estimated and true effect.

Note also that, if we restrict the second stage sample size to be balanced between groups, the treatment with the maximum CMSE at interim is the treatment with the maximum observed $|\, z_i - z_0 \,|$. Again, the values for $r_{\min} = r_{\max}$ give the selection RMSE without additional sample size reassessment. We can see from the figures that even though the adaptive sample size reassessment may increase Bias substantially, it has only a small effect on the RMSE, that is, sample-size reassessments do not increase the RMSE much over the RMSE under treatment selection, at least when sample-size reductions are limited to $r_{\min} \geqslant 0.5$. Especially for $r_{\min} \geqslant 0.5$ lines for the different restrictions are indistinguishable because of the small difference between the results.

Table I give the standardized RMSE_k^* for several scenarios setting $r_{\max} = \infty$. Recall that, for a comparison, the rows $r_{\min} = r_{\max}$ show the RMSE under treatment selection only. Like the maximum Bias, RMSE_k^* is increasing with increasing k; however, for $r_{\min} > 0$, the additional increase in RMSE_k^* due to sample size reassessment is small in particular under the additional restrictions on r_s and r_0. The difference becomes smaller, the larger $r_{\min}$ and k are. The difference between the fixed and adaptive sample size case is particularly small with balanced second stage sample sizes. This may be due to the fact that balanced sample sizes are optimal with regard to the variance of the second-stage effect estimate. Moreover, aiming on the reduction of RMSE_k^*, we find for $k \geqslant 1$ that fixing r_0 is not more and can even be less effective than the restriction to balanced sample sizes (in contrast to what we find for the maximum Bias). Again, there is no large difference between the constraints $r_s \geqslant r_0$ and $r_s = r_0$, in particular, for larger k.

4.2.2. Reshuffling. For the case of a sample-size shuffling between the selected treatment and the control group, the maximum conditional mean squared error, $\widetilde{\mathrm{CMSE}}$, can be calculated as in Section 3.2.5. Figure 3 (B) shows the resulting standardized RMSE_k^* for $k = 1$ to 6 for $0 \leqslant v \leqslant 1$ (solid black lines) and for the restriction $0 \leqslant v \leqslant 0.5$ (dashed black lines). The gray solid lines gives the maximum selection RMSE for an adaptive design, selecting one treatment and control at interim time point t, allocating in the

second stage $(1-t)n_g(k+1)/2$ patients to each of the two groups. In this balanced case, the treatment with the maximum CMSE at interim is the treatment with the maximum absolute observed difference $|z_i - z_0|$ at interim. Note again that this is not necessarily true if we allow for reshuffling leading to unbalanced second stages.

The selection RMSE (gray lines) is increasing with increasing t. This is similar to the results of [20] where the selection Bias was calculated for the case of selecting the treatment with maximum effect at interim. We note again that selecting the treatment with the maximum treatment effect is not the same as selecting the treatment with the maximum CMSE at interim. As for the Bias, the standardized RMSE_k^* shows a non-monotonous behavior. There is a trade-off between the variance, which is increasing with t for selection and the Bias due to sample size reassessment, which is decreasing with t. For $k \geqslant 2$ and with the constraint $v \leqslant 0.5$, there is a t for which the RMSE is minimal. The minimum is achieved at t-values close to 0.5.

The later the interim analysis the smaller the difference between RMSE_k^* and selection RMSE. For $t = 0.5$ and $k = 2, 3, 4$, the standardized maximum Bias is 1.56, 1.67, 1.74 if $v \leqslant 1$ and 1.28, 1.27, 1.25 if $v \leqslant 0.5$. For a comparison, the selection RMSE is 0.99, 0.93, 0.88 for $k = 2, 3$, and 4, respectively. Note that the case $t = 1$ gives the worst case RMSE of a classical fixed-sample-size parallel group design where the single treatment is selected post-trial (in a fully flexible manner), and that the RMSE_k^* of an adaptive design with mid-trial treatment selection and sample size reshuffling is smaller.

5. Discussion

We investigated in this paper the maximum effect of data-driven sample-size reassessments and treatment selection on Bias and precision of maximum likelihood estimators in multi-armed adaptive designs. We assumed that in an interim analysis, one out of k treatments and the control are selected for a second stage and sample sizes are reassessed in a fully flexible manner with and without restrictions. To best of our knowledge, we are the first who consider Bias and MSE under flexible selection and sample size reassessment rules. In [20], for instance, selection Bias and MSE were considered without sample size reassessment and only for some specific selection rules.

To cope with flexible decision rules, we calculated the maximum Bias and maximum MSE searching at each possible interim outcome for the worst case treatment selection and sample size assignments, which maximize the conditional Bias or conditional MSE. We are aware of the fact that the determination of maximum Bias and MSE will lead to an overestimation and that Bias and MSE may in reality be (substantially) smaller. To bound the conservatism of our approach, we considered several restrictions on the sample-size rules, like balanced second-stage sample sizes or to rules for which the selected experimental treatment group is as least as large as the control group. We saw that these restrictions substantially reduce the maximum Bias and maximum MSE and that in some cases (e.g. when $k = 1$ and $r_{\min} = 1$) the maximum Bias and maximum inflation of the MSE is small enough to justify the use of the MLE.

In spite of the conservatism of our approach, we have been able to draw several important conclusions. One important conclusion is that a lower bound for the second stage sample sizes may effectively reduce Bias and inflations of the MSE. We saw, for instance, that under the constraint that the second stage sample sizes are at least as large as the first stage n (i.e. the case $r_{\min} = 1$); Bias is in general limited and not much larger than the pure selection Bias. This is particularly the case under the restriction that in the second stage sample the treatment group is as least as large as the control group. Moreover, we found that the maximum Bias is not much further decreased by forcing the treatment groups to be balanced at the second stage or the size of the control group to be fixed. Constraining the second stage sample sizes to be at least as large as the first stage n has an even more pronounced effect on the maximum MSE, which is more or less independent from the maximum sample size ($r_{\max}$) and the additional restriction on the second-stage allocation ratios. We can, therefore, conclude that with a sufficiently large minimal second-stage sample size a further increase of the sample size in the selected treatment group has only a limited negative effect on Bias and MSE.

We also learned that when fixing the total sample size and reshuffling the (fixed) second-stage sample size between the control and selected treatment group the additional Bias and MSE due to the sample-size reassessments may be substantial even under the (realistic) constraint that the control group is not

larger than the experimental treatment group. This is particularly the case when the interim analysis is done early. Note that the results with fixed and flexible overall sample sizes are not easy to compare because we had to use different standardizations for the reshuffling and the other cases, and because in the other cases, the total sample size is not fixed but data dependent and of an less determined magnitude.

Our paper necessarily leaves important questions open. It is known that the selection Bias can be severe even without sample-size reassessments if selection is done late. Early selection in general will reduce the Bias as compared with 'post-trial' selection [20]. Our findings confirm these results. Hence, an important question, that goes beyond the scope of this paper, is the performance of adjusted estimates to account for the selection Bias under flexible selection and sample-size reassessments. To this end, it is important to note that Bias adjusted estimates have only been suggested and considered for designs with fixed (known) selection rules, namely selecting the seemingly most efficient treatment. We consider shrinkage estimates as one of the most interesting candidates as they are known to perform well in terms of the MSE under the common treatment selection process (compared with [19]) but other estimates may be considered as well. Another interesting and important extension of our work would be the consideration of selections rules with more than one selected experimental treatment and with realistic constraints on the selection process. Selection of more than one treatment for the play-the-winner rule without additional sample-size reassessment was investigated in [20]. Calculation of the maximum Bias or MSE for further selection rules would be an interesting contribution.

Appendix

A.1 Calculation of the CMSE

For the calculation of the conditional mean squared error, we set $Z_{(i,j)} = (\bar{X}_{(i,j)} - \mu_i)\sqrt{n/\sigma^2}$ for $i = 0, 1$ and $j = 1, 2$.

$$
\begin{aligned}
&\mathrm{CMSE}(z_0, z_1, r_0, r_1, n) = \\
&= \mathbf{E}\left[\left((\bar{X}_1 - \bar{X}_0) - (\mu_1 - \mu_0)\right)^2 \,\middle|\, Z_{(i,1)} = z_i, i = 0, 1\right] \\
&= \mathbf{E}\left[\left(\frac{\frac{\sigma}{\sqrt{n}}Z_{(1,1)} + \frac{r_1\sigma}{\sqrt{n_{(1,2)}}}Z_{(1,2)}}{1 + r_1} - \frac{\frac{\sigma}{\sqrt{n}}Z_{(0,1)} + \frac{r_0\sigma}{\sqrt{n_{(0,2)}}}Z_{(0,2)}}{1 + r_0}\right)^2 \,\middle|\, Z_{(i,1)} = z_i, i = 0, 1\right] \\
&= \mathbf{E}\left[\left(\frac{\frac{\sigma}{\sqrt{n}}Z_{(1,1)}}{1 + r_1} - \frac{\frac{\sigma}{\sqrt{n}}Z_{(0,1)}}{1 + r_0}\right)^2 + 2\left(\frac{\frac{\sigma}{\sqrt{n}}Z_{(1,1)}}{1 + r_1} - \frac{\frac{\sigma}{\sqrt{n}}Z_{(0,1)}}{1 + r_0}\right)\left(\frac{\frac{r_1\sigma}{\sqrt{nr_1}}Z_{(1,2)}}{1 + r_1} - \frac{\frac{r_0\sigma}{\sqrt{nr_0}}Z_{(0,2)}}{1 + r_0}\right)\right. \\
&\quad \left. + \left(\frac{\frac{r_1\sigma}{\sqrt{nr_1}}Z_{(1,2)}}{1 + r_1} - \frac{\frac{r_0\sigma}{\sqrt{nr_0}}Z_{(0,2)}}{1 + r_0}\right)^2 \,\middle|\, Z_{(i,1)} = z_i, i = 0, 1\right] \\
&= \frac{\sigma^2}{n}\left[\left(\frac{z_1}{1 + r_1} - \frac{z_0}{1 + r_0}\right)^2 + \frac{r_1}{(1 + r_1)^2} + \frac{r_0}{(1 + r_0)^2}\right]
\end{aligned}
$$

because

$$
\mathbf{E}\left[\left(\frac{\frac{r_1\sigma}{\sqrt{nr_1}}Z_{(1,2)}}{1 + r_1} - \frac{\frac{r_0\sigma}{\sqrt{nr_0}}Z_{(0,2)}}{1 + r_0}\right)^2 \,\middle|\, Z_{(i,1)} = z_i, i = 0, 1\right] = \frac{\sigma^2}{n}\left(\frac{r_1}{(1 + r_1)^2} + \frac{r_0}{(1 + r_0)^2}\right)
$$

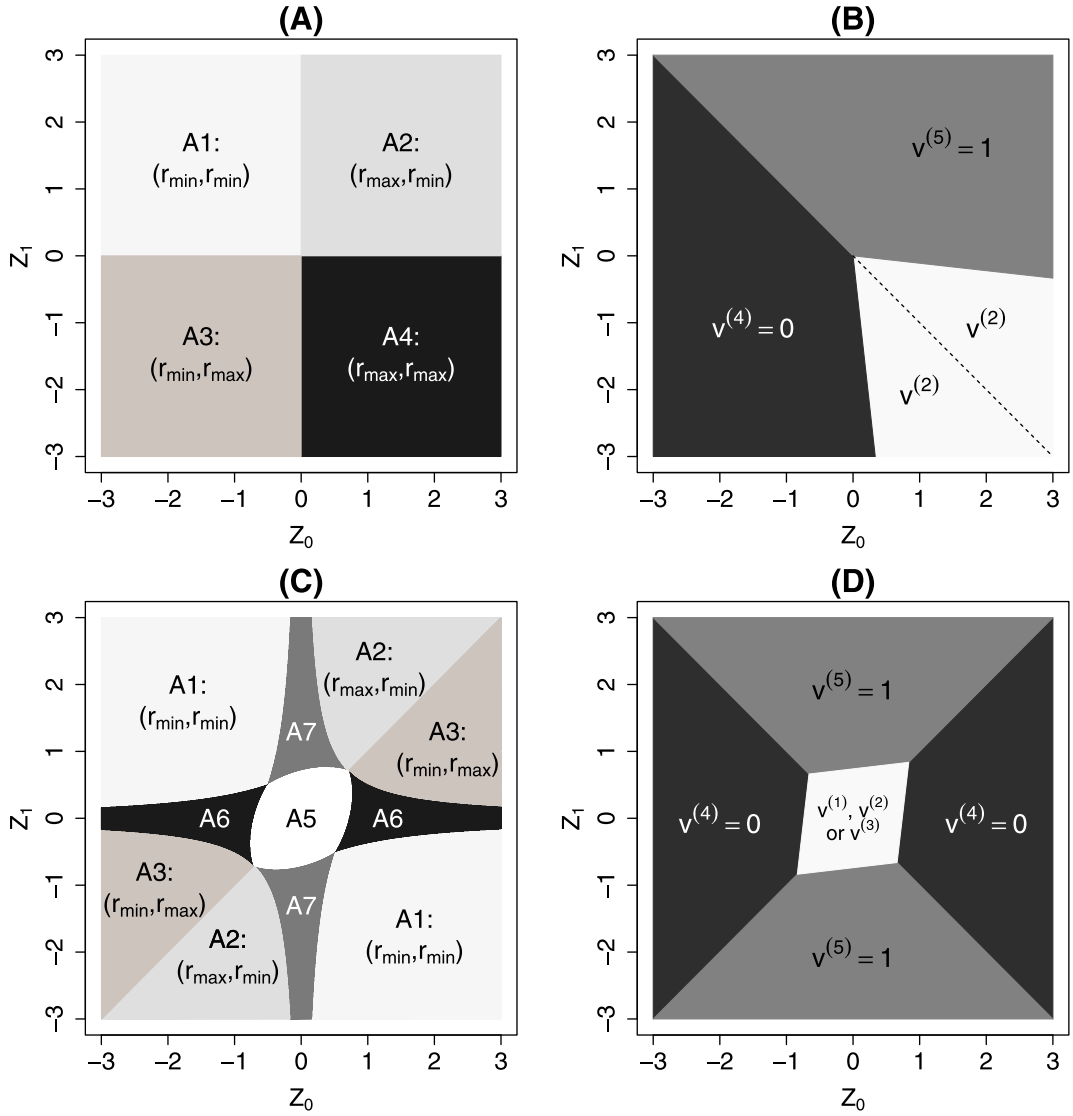

Figure A1. Subsets of the interim outcome of treatment and control (z_0, z_1) to be used for evaluating the worst case conditional Bias (first row) and the worst case conditional MSE (second row). Subsets are given for flexible second-to-first-stage ratios (first column) and the case of reshuffling (second column).

and

$$\mathbf{E}\left[\left.\frac{\frac{r_1\sigma}{\sqrt{nr_1}}Z_{(1,2)}}{1+r_1}-\frac{\frac{r_0\sigma}{\sqrt{nr_0}}Z_{(0,2)}}{1+r_0}\right| Z_{(i,1)}=z_i, i=0,1\right]=0$$

A.2 Maximizing the CMSE

To maximize the CMSE (9) for given z_0 and z_1 at interim, nine candidates have to be investigated. Apparently, candidates $(r_0^{(1)}, r_1^{(1)}) = (r_{\min}, r_{\min})$, $(r_0^{(2)}, r_1^{(2)}) = (r_{\max}, r_{\min})$, $(r_0^{(3)}, r_1^{(3)}) = (r_{\min}, r_{\max})$ and $(r_0^{(4)}, r_1^{(4)}) = (r_{\max}, r_{\max})$ have to be investigated. By setting the first derivative (with respect to r_0 and r_1) to 0 and solving the corresponding system of two equations, candidate (5) can be assessed with

$$\left(r_0^{(5)}, r_1^{(5)}\right) = \left(\frac{-1+2z_0^2-z_0z_1+z_1^2}{-1+z_0z_1+z_1^2}, \frac{-1+2z_1^2-z_0z_1+z_0^2}{-1+z_0z_1+z_0^2}\right).$$

Depending on the given z_1 and z_0, this candidate is either a minimum or a maximum. If a maximum, this candidate is ineligible if either $r_0^{(5)}$ or $r_1^{(5)}$ is larger than $r_{\max}$ or smaller than $r_{\min}$. Setting $r_0^{(6)} = r_{\min}$, the corresponding worst case reassessment rule for the treatment group can be calculated by setting the first derivative with respect to r_1 (assuming r_0 fixed) to zero and results in

$$\left(r_0^{(6)}, r_1^{(6)}\right) = \left(r_{\min}, \frac{1 + r_{\min} + 2z_0z_1 - 2z_1^2 - 2r_{\min}z_1^2}{1 + r_{\min} - 2z_0z_1}\right)$$

Candidates $(r_0^{(7)}, r_{\min})$, $(r_{\max}, r_1^{(8)})$ and $(r_0^{(9)}, r_{\max})$ can be assessed similarly. The global maximum for given z_0 and z_1 is the maximum over all nine candidates and formula (10) can be rewritten as.

$$\widetilde{\text{CMSE}}\left(z_0, z_1, n, \sigma, r_{\min}, r_{\max}\right) = \max_{i=1,\ldots,9 : r_{\min} \leqslant r_0^{(i)}, r_1^{(i)} \leqslant r_{\max}} \left(\text{CMSE}\left(z_0, z_1, n, \sigma, r_0^{(i)}, r_1^{(i)}\right)\right) \tag{21}$$

Evaluating $\widetilde{\text{CMSE}}$ for all possible (z_0, z_1) and integrating over all interim outcomes gives MSE*.

Figure (C) in the Appendix shows the subsets (corresponding to the several candidates) when setting $r_{\min} = 0$ and $r_{\max} = \infty$. The subset A5 is the area where candidate 5 is the global maximum. See also the subsets for candidates 1 (area A1), 2 (area A2), 3 (area A3), 6 (area A6), and 7 (area A7). It can be seen that candidates 4, 8, and 9 are no global maxima. Some of the regions are similar to the regions maximizing the conditional Bias (Figure A). For z_1 or z_0 close to 0, the worst-case reassessment rule to maximize the CMSE is different from setting r_0 or r_1 to $r_{\min}$ or $r_{\max}$.

A.3 Maximum CB under reshuffling

The following equality holds in the case of reshuffling:

$$\max_{i=1,\ldots,k} \widetilde{\text{CB}}\left(z_0, z_i, n_g, t, \sigma\right) = \widetilde{\text{CB}}\left(z_0, \max_{i=1,\ldots,k} z_i, n_g, t, \sigma\right). \tag{22}$$

For fix v, n_g, t, and z_0, the conditional Bias (6) is monotonous in z_i. Assume now that, for some observed z_s and z_0, the optimal second-stage allocation rate is $\tilde{v}$. For a $z_s^* > z_s$, because of monotonicity, $\text{CB}(z_0, z_s^*, \tilde{v}, n_g, t, \sigma) > \text{CB}(z_0, z_s, \tilde{v}, n_g, t, \sigma)$. $\tilde{v}$ may not be the allocation rate maximizing the conditional Bias for z_s^*, but finding the actual optimum $\tilde{v}^*$ can only increase the Bias and therefore $\text{CB}(z_0, z_s^*, \tilde{v}^*, n_g, t, \sigma) \geqslant \text{CB}(z_0, z_s^*, \tilde{v}, n_g, t, \sigma)$ concluding that $\widetilde{\text{CB}}$ is monotonous in z_i (for fixed z_0). Therefore, equality (22) holds.

Acknowledgements

Part of the work of Alexandra Graf was supported by the Austrian Science Fund (FWF), Project No. J3344-N26. Additional funding has been provided by the UK Medical Research Council, Project No. MR/M005755/1. Georg Gutjahr received funding from the DFG project BR 373/1-1.

References

1. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* 1995; **51**:1315–1324.
2. Graf AC, Bauer P. Maximum inflation of the type 1 error rate when sample size and allocation rate are adapted in a pre-planned interim look. *Statistics in Medicine* 2011; **30**:1637–1647.
3. Graf AC, Bauer P, Glimm E, Koenig F. Maximum type 1 error rate inflation in multi-armed clinical trials with interim sample size modifications. *Biometrical Journal* 2014; **56**:614–630.
4. Dunnett C. A multiple comparison procedure for comparing several treatments with a control. *JASA* 1955; **50**:1096–1121.
5. Bauer P, Koehne K. Evaluations of experiments with adaptive interim analysis. *Biometrics* 1994; **50**:1029–1041.
6. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 1999; **18**:1833–1848.
7. Bretz F, Koenig F, Brannath W, Glimm E, Posch M. Adaptive designs for confirmatory clinical trials. *Statistics in Medicine* 2009; **28**:1181–1217.
8. Mueller HH, Schaefer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* 2001; **95**:886–891.

9. Mueller HH, Schaefer H. A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine* 2004; **23**:2497–2508.
10. Koenig F, Brannath W, Bretz F, Posch M. Adaptive Dunnett tests for treatment selection. *Statistics in Medicine* 2008; **27**:1612–1625.
11. Friede T, Stallard N. A comparison of methods for adaptive treatment selection. *Biometrical Journal* 2008; **50**:767–781.
12. Stallard N, Todd S, Whitehead J. Estimation following selection of the largest of two normal means. *Journal of Statistical Planning and Inference* 2008; **138**:1629–1634.
13. Cohen A, Sackrowitz HB. Two stage conditionally unbiased estimators of the selected mean. *Statistics and Probability Letters* 1989; **8**:273–278.
14. Shen L. An improved method of evaluating drug effect in a multiple dose clinical trial. *Statistics in Medicine* 2001; **20**: 1913–1929.
15. Posch M, Koenig F, Branson M, Brannath W, Dunger-Baldauf C, Bauer P. Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine* 2006; **24**:3697–3714.
16. Brannath W, Koenig F, Bauer P. Estimation in flexible two stage designs. *Statistics in Medicine* 2006; **25**:3366–3381.
17. Stallard N, Todd S. Point estimates and confidence regions for sequential trialsinvolving selection. *Journal of Statistical Planning and Inference* 2005; **135**:402–419.
18. Bowden J, Glimm E. Unbiased estimation of selected treatment means in two-stage trials. *Biometrical Journal* 2008; **50**:515–527.
19. Carreras M, Brannath W. Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection. *Statistics in Medicine* 2013; **32**:1977–1690.
20. Bauer P, Koenig F, Brannath W, Posch M. Selection and Bias - Two hostile brothers. *Statistics in Medicine* 2010; **29**:1–13.
21. Hahn T. CUBA-a library for multidimensional numerical integration. *Computer Physics Communication* 2005; **168**:78–95.
22. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2013. http://www.R-project.org.

