Maximum type 1 error rate inflation in multiarmed clinical trials with adaptive interim sample size modifications

Alexandra C. Graf[1,2], **Peter Bauer**[1], **Ekkehard Glimm**[3], and **Franz Koenig**[*,1]

[1] Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria
[2] Competence Center for Clinical Trials, University of Bremen, Linzer Strasse 4, 28359 Bremen, Germany
[3] Novartis Pharma AG, Novartis Campus, 4056 Basel, Switzerland

Received 1 August 2013; revised 20 January 2014; accepted 22 January 2014

Sample size modifications in the interim analyses of an adaptive design can inflate the type 1 error rate, if test statistics and critical boundaries are used in the final analysis as if no modification had been made. While this is already true for designs with an overall change of the sample size in a balanced treatment-control comparison, the inflation can be much larger if in addition a modification of allocation ratios is allowed as well. In this paper, we investigate adaptive designs with several treatment arms compared to a single common control group. Regarding modifications, we consider treatment arm selection as well as modifications of overall sample size and allocation ratios. The inflation is quantified for two approaches: a naive procedure that ignores not only all modifications, but also the multiplicity issue arising from the many-to-one comparison, and a Dunnett procedure that ignores modifications, but adjusts for the initially started multiple treatments. The maximum inflation of the type 1 error rate for such types of design can be calculated by searching for the "worst case" scenarios, that are sample size adaptation rules in the interim analysis that lead to the largest conditional type 1 error rate in any point of the sample space. To show the most extreme inflation, we initially assume unconstrained second stage sample size modifications leading to a large inflation of the type 1 error rate. Furthermore, we investigate the inflation when putting constraints on the second stage sample sizes. It turns out that, for example fixing the sample size of the control group, leads to designs controlling the type 1 error rate.

Keywords: Conditional error function; Interim analysis; Maximum type 1 error; Sample size reassessment; Treatment selection.

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

In the last decade, adaptivity in clinical trials with design modifications such as sample size reassessment or treatment selection at an interim analysis has gained increasing attention. One may argue that there have always been modifications when performing clinical trials, for example simply covered by amendments to the study protocols. However, it has been shown that if, after design modifications, the critical boundaries and test statistics for the corresponding fixed sample size design are used, then the type 1 error rate is inflated. For the comparison of the means of a normally distributed outcome with known variance between a single treatment and a control in parallel groups and balanced sample sizes,

*Corresponding author: e-mail: franz.koenig@meduniwien.ac.at, Phone: +43-140400-7480, Fax: +43-140400-7477

that is equal sample size in the treatment and control group, Proschan and Hunsberger (1995) derived the maximum possible type 1 error rate inflation. They assumed that the experimenter, for any interim outcome, would choose the second stage sample sizes in such a way that the conditional type 1 error rate is maximized ("worst case scenario"). This strategy will also maximize the overall type 1 error rate. They showed that the type 1 error rate can be inflated from 0.05 to 0.11. Graf and Bauer (2011) extended these worst case arguments to the case of unbalanced sample size reassessment showing that the maximum type 1 error rate increases to 0.19 when the allocation ratio is allowed to change at interim. However, in this unbalanced case the maximum of the conditional type 1 error rate can only occur if the experimenter knows the value of the nuisance parameter, the common mean under the null hypothesis. This may at least approximately apply for the control treatment if a large number of data from previous experiments is available.

Many methods for type 1 error control in adaptive designs are available for testing a single hypothesis (Bauer, 1989; Bauer and Koehne, 1994; Proschan and Hunsberger, 1995; Lehmacher and Wassmer, 1999; Mueller and Schaefer, 2001; Brannath et al., 2002; Mueller and Schaefer, 2004; Gao et al., 2013) and have been applied in clinical trials. Multiarmed selection designs have been proposed (e.g. Thall et al., 1988, 1988) and have been extended to allowing for adaptive design modifications (Bauer and Kieser, 1999; Koenig et al., 2008; Bretz et al., 2009; Bebu et al., 2013; Sugitani et al., 2013). With the rise of adaptive methods in clinical trials, the main emphasis has been on strict control of the type 1 error rate to maintain the strictly confirmatory nature (EMA, 2007; FDA, 2010; Wang et al., 2013).

However, there are complaints that the adaptive machinery has become too complicated with "tests that resort to nonstandard adjustments and weightings appear mysterious to all but the specialist in adaptive design" (Metha and Pocock, 2012). From an operational perspective, adaptations put a burden on data analysts who have to clean data for interim decision making and on drug supply managers who have to deal with the possibility that doses may be added to or removed from the trial. Uncertainty at the planning stage about the total funds needed for the trial can also be a concern. From a statistical perspective, it has been argued by some experts that adaptive designs offer little advantage over more conventional group-sequential designs (Tsiatis and Metha, 2003; Jennison and Turnbull, 2006; Levin et al., 2013) and that they use test statistics that might violate desirable principles like sufficiency (Burman and Sonesson, 2006). However, these criticisms of adaptive designs are not uncontroversial themselves (Brannath et al., 2006). In any case, such additional burden may prevent experimenters from using adaptive design methodology and resort to either ignoring the issue or using seemingly simple adjustments like Bonferroni or Dunnett corrections. It is therefore desirable to investigate the maximum type 1 error inflation arising from such strategies. Regarding specific clinical trials, the precise quantification of the inflation can also be a guide to decide whether the implementation of the adaptive test machinery is really necessary, or whether a simpler adjustment might suffice, possibly after additional restrictions of the interim decision options, like upper and lower limits on the allowed sample size modifications.

In this work, we investigate the maximum type 1 error rate when k test treatments are compared to a single common control and when treatment selection is allowed at interim either with or without flexible sample size reassessment. Designs of multiarmed clinical trials with interim treatment selection have attracted a lot of research in the last decade (Zeymer et al., 2001; Gaydos et al., 2009; Barnes et al., 2010). Nevertheless, the number of conducted or started trials seems to be rather limited (Elsaesser et al., 2014; Morgen et al., 2014).

In Section 2, we give a motivating example of a clinical trial where the experimenters decided to use the conservative Bonferroni procedure instead of an adaptive approach. In Section 3, we introduce the hypothesis tests and the type of interim adaptations investigated to calculate the maximum type 1 error rate. In Section 4, we consider the situation when the treatment with the largest observed interim effect is always selected for the second stage. Furthermore, we investigate the maximum type 1 error rate when second stage sample sizes are restricted to range within a prefixed interval. In Section 5, we mainly focus on the case of $k = 2$ treatment arms, always proceeding with both treatments and the control to the second stage. In Section 6, we discuss our findings in the context of the motivating example and give some practical considerations. This is followed by concluding remarks in Section 7.

2 Motivating example

Barnes et al. (2010) give a recent case study for a two-stage clinical trial on the drug indacaterol to treat chronic obstructive pulmonary disease (COPD). This study comprised a first stage for dose-finding with dose selection after 14 days of treatment, and a second stage evaluating efficacy and safety during 26 weeks of treatment. The dose-finding stage included seven randomized treatment arms, four doses of the study drug, placebo and two further treatment groups with active comparators. At an interim analysis after the first stage the indacaterol doses were selected using preset efficacy and safety data (Lawrence et al., 2014). A multiplicity correction using a Bonferroni adjustment with $\alpha/4$ was applied, despite the fact that in the final efficacy analysis only the two selected indacaterol doses should have been compared individually against placebo based on the pooled data of both stages with prefixed sample sizes. This approach controls the type 1 error rate if the sample size, as in the given example, is prefixed. However, due to the overcorrection, this approach is conservative. The authors themselves acknowledge that the approach "is statistically somewhat conservative, but it has the merit of simplicity". The question arises whether for such a design sample size reassessment strategies would have been possible without inflating the type 1 error rate.

3 Trial design

In the following, we assume that a clinical trial is designed for k treatment and one control arm where a two-stage design should be applied. In a first stage the observed outcome measures $x_{j,i}^{(1)}$ from patients $j = 1, \ldots, n_i^{(1)}$, randomized to one of $k+1$ groups, that is to the control, denoted by index $i = 0$, or to one of the treatment groups, $i = 1, \ldots, k$ are investigated. The outcome is assumed to be normally distributed with common known variance, $X_{j,i} \sim N(\mu_i, \sigma)$. Without loss of generality we set $\sigma = 1$. Having obtained at the end of the first stage $n_0^{(1)}$ observations in the control and $n_i^{(1)} = a_i n_0^{(1)}$, $i = 1, \ldots, k$ observations in the treatment groups, the sample means $\bar{x}_i^{(1)} = \frac{1}{n_i^{(1)}} \sum_{j=1}^{n_i^{(1)}} x_{j,i}^{(1)}$ for $i = 0, \ldots, k$ are calculated. The $a_i > 0$ denote the prefixed first-stage-allocation-ratios between treatment group i and control. The experimenter may set the second stage sample sizes to $n_i^{(2)} = r_i n_i^{(1)} = r_i a_i n_0^{(1)}$ in the treatment groups and to $n_0^{(2)} = r_0 n_0^{(1)}$ in the control group with the second-to-first-stage-ratios $0 \leq r_i \leq \infty$, $i = 0, \ldots, k$ based on the interim sample means.

In the final analysis, after the second stage, we test the hypotheses

$$H_{0i} : \mu_i = \mu_0 \text{ vs. } H_{Ai} : \mu_i > \mu_0 \quad \text{for} \quad i = 1, \ldots, k$$

using the standardized mean difference T_i pooling the data of both stages and comparing it to the critical boundary $c_{1-\alpha}$ as used for the fixed sample size design. This means that adaptivity is not accounted for, neither in the test statistics nor in the critical boundary. The test statistics is defined as:

$$T_i = \left(\frac{n_i^{(1)} \bar{x}_i^{(1)} + n_i^{(2)} \bar{x}_i^{(2)}}{n_i^{(1)} + n_i^{(2)}} - \frac{n_0^{(1)} \bar{x}_0^{(1)} + n_0^{(2)} \bar{x}_0^{(2)}}{n_0^{(1)} + n_0^{(2)}} \right) \Big/ \left(\sqrt{\frac{1}{n_i^{(1)} + n_i^{(2)}} + \frac{1}{n_0^{(1)} + n_0^{(2)}}} \right)$$

for $i = 1, \ldots, k$ with $\bar{x}_i^{(2)}$, $i = 0, \ldots, k$ denoting the second stage sample means.

We obtain the worst case scenarios for each possible interim outcome by searching for the second-to-first-stage ratios maximizing the conditional type 1 error rate, $\tilde{r}_i$. Generalizing the formula in Koenig et al. (2008) the conditional type 1 error rate for rejecting at least one treatment-control comparison in the final analysis, given the observed interim data is (see Appendix A1):

$$\begin{aligned} &CE_\alpha\big(a_1,\dots,a_k,Z_0^{(1)},\dots,Z_k^{(1)},r_0,\dots,r_k\big)\\ &= 1-\int_{-\infty}^{\infty}\prod_{i=1}^{k}\Phi\left[c_{1-\alpha}\sqrt{\frac{(1+r_i)(1+r_0+a_i(1+r_i))}{(1+r_0)r_i}}-\frac{Z_i^{(1)}}{\sqrt{r_i}}\right.\\ &\left.+Z_0^{(1)}\sqrt{\frac{a_i}{r_i}\frac{1+r_i}{1+r_0}}+Z_0^{(2)}\sqrt{\frac{r_0a_i}{r_i}\frac{1+r_i}{1+r_0}}\right]\phi\big(Z_0^{(2)}\big)dZ_0^{(2)} \end{aligned} \qquad (1)$$

where α is the preplanned level for the type 1 error rate and $c_{1-\alpha}$ the critical boundary of the preplanned final tests (see Remark 3.1). The $Z_i^{(j)}=(\bar{X}_i^{(j)}-\mu)\sqrt{n_i^{(j)}}$, $i=0,\dots,k$, $j=1,2$ are defined as the standardized differences between the sample mean and the common true mean μ under the global null hypothesis of stage $j=1$ (at interim) and $j=2$, respectively (without loss of generality $\mu=0$). The cumulative distribution function and density of the standard normal distribution are denoted by Φ and ϕ, respectively. Note that $Z_i^{(j)}$ follow independent standard normal distributions.

When second stage sample sizes are not constrained, the maximum type 1 error rate is given by

$$E_\alpha^*(a_1,\dots,a_k)=\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}\widetilde{CE}_\alpha\big(a_1,\dots,a_k,Z_0^{(1)},\dots,Z_k^{(1)}\big)\phi\big(Z_k^{(1)}\big)\dots\phi\big(Z_0^{(1)}\big)dZ_k^{(1)}\dots dZ_0^{(1)} \qquad (2)$$

where

$$\widetilde{CE}_\alpha\big(a_1,\dots,a_k,Z_0^{(1)},\dots,Z_k^{(1)}\big)=\max_{0\le r_0,\dots,r_k\le\infty}CE_\alpha\big(a_1,\dots,a_k,Z_0^{(1)},\dots,Z_k^{(1)},r_0,\dots,r_k\big).$$

Whereas CE_α is a function of $r_0,\dots,r_k$, $\widetilde{CE}_\alpha$ is a function of $\tilde{r}_0,\dots,\tilde{r}_k$, the second-to-first-stage-ratios leading to the maximum CE_α. The $\tilde{r}_i$ are determined for a given interim outcome $(Z_0^{(1)},\dots,Z_k^{(1)})$ and are therefore a function of the $Z_i^{(1)}$. Thus, $\widetilde{CE}_\alpha$ does not depend on $r_0,\dots,r_k$.

In the following we use a quasi Newton method provided by the R-function optim for numerical optimization and for numerical integration we used the R-function integrate (R Development Core Team, 2012). R-programs to calculate the maximum type 1 error rate are available as Supplementary Information.

Remark 3.1. The critical boundary $c_{1-\alpha}$ of the preplanned test may be defined in different ways: (i) as the $(1-\alpha)$-quantile of the standard normal distribution, $z_{1-\alpha}$, if no correction at all for multiplicity is applied or (ii) as a Dunnett critical boundary (Dunnett, 1955) based on the preplanned first-stage-allocation-ratios a_i, $i=1,\dots,k$ to adjust for multiplicity due to the treatment-control comparisons. Even strategy (ii) may not guarantee type 1 error control if additional sample size reassessment is performed at interim. Moreover, in case of sample size reassessment (and/or treatment selection) the Dunnett critical boundary would not be fixed a priori when calculated for the actual sample sizes in the final analysis. For simplicity ,we will apply the pre-fixed Dunnett boundary, $d_{1-\alpha}$, based on the preplanned first-stage-allocation-ratios a_i between treatment and control in the following. Remarks 4.1 and 4.2 discuss how results change if instead critical boundaries are based on actual (reassessed) total sample sizes in the final analysis.

Remark 3.2. For $k \geq 2$ we calculate the maximum type 1 error rate under the global null hypothesis $\mu_i = \mu_0, i = 1, \dots, k$. A proof that the maximum type 1 error is attained under the global null hypothesis is given in Appendix A2.

Remark 3.3. For $k = 1$, Graf and Bauer (2011) showed, by numerical evaluation, that the maximum type 1 error in the case of balanced first stage sample size between treatments before the interim analysis ($a_i = 1, i = 1, \dots, k$) is an upper bound. For $k \geq 1$ we will likewise set $a_i = 1$, since it is the most common scenario applied in practice. Note that for many-to-one comparisons, the scenario with $a_i = 1/\sqrt{k}$ leads to the smallest required sample size for a given power and significance level. Therefore we will also give some numerical results for this allocation ratio.

4 Selection of the most promising treatment at interim

We first consider that in the interim analysis the treatment group m with the largest observed interim effect $Z_m^{(1)} = \max_{i=1,\dots,k} Z_i^{(1)}$ is selected for the second stage, setting $r_i = 0$ for $i \notin \{0, m\}$. The second-to-first-stage-ratios r_i, $i \in \{0, m\}$ may be set based on the interim results, $0 \leq r_0, r_m \leq \infty$. In the final analysis, only the selected treatment group m is compared to the control group (using data of both stages). The corresponding null hypothesis H_{0m} is rejected, if the final test statistic T_m exceeds the critical value $c_{1-\alpha}$. Note that the maximum type 1 error rate for the case of always selecting the best treatment is an upper bound for the maximum type 1 error rate when in a particular trial another single treatment is selected, for example the treatment with the second largest observed effect at interim because of potential safety issues for the most effective treatment. Clearly, under the global null hypothesis and for balanced first stage sample sizes over the k treatments, selecting a treatment with an interim effect smaller than the largest observed interim effect will reduce the maximum type 1 error rate. Following the lines of Graf and Bauer (2011), the conditional type 1 error rate (1) for this scenario simplifies to

$$CE_\alpha\big(a_m, r_0, r_m, Z_0^{(1)}, Z_m^{(1)}\big) = 1 - \Phi\left[\frac{c_{1-\alpha}\sqrt{\dfrac{1}{a_m(1+r_m)} + \dfrac{1}{1+r_0}} - \dfrac{Z_m^{(1)}}{\sqrt{a_m}(1+r_m)} + \dfrac{Z_0^{(1)}}{1+r_0}}{\sqrt{\dfrac{r_m}{a_m(1+r_m)^2} + \dfrac{r_0}{(1+r_0)^2}}}\right]. \tag{3}$$

Note that if the treatment with the largest observed interim effect is selected, m is random and therefore also a_m is a random variable. In the following we set $a_1 = \dots = a_k$ so that a_m is no longer a random variable and the maximum type 1 error rate can be evaluated by

$$E_\alpha^* = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \widetilde{CE}_\alpha\big(a_m, Z_0^{(1)}, Z_m^{(1)}\big) k\Phi\big(Z_m^{(1)}\big)^{k-1}\phi\big(Z_m^{(1)}\big)\phi\big(Z_0^{(1)}\big)dZ_m^{(1)}dZ_0^{(1)} \tag{4}$$

where

$$\widetilde{CE}_\alpha\big(a_m, Z_0^{(1)}, Z_m^{(1)}\big) = \max_{0 \leq r_0, r_m \leq \infty} CE_\alpha\big(a_m, r_0, r_m, Z_0^{(1)}, Z_m^{(1)}\big). \tag{5}$$

and $k\Phi(x)^{(k-1)}\phi(x)$ is the probability density function of the maximum of independent standard normal distributions.

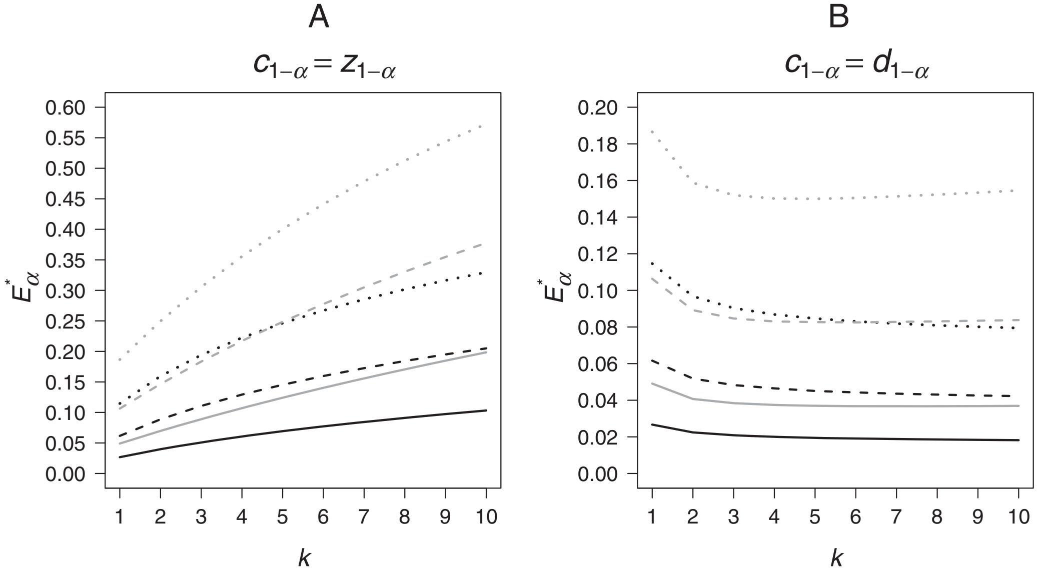

Figure 1 Maximum type 1 error rate E^*_α when always selecting the treatment with the maximum effect at interim for an increasing number of treatment groups k. Results are given when using the uncorrected critical boundary $z_{1-\alpha}$ (A) or the fixed Dunnett critical boundary $d_{1-\alpha}$ (B) for equal (black lines) and flexible (gray lines) second-to-first-stage ratios. Nominal one-sided α was set to 0.01 (solid lines), 0.025 (dashed lines), and 0.05 (dotted lines).

4.1 Equal second-to-first-stage-ratios

Let $r_0 = r_m = r$ with $0 \leq r \leq \infty$, that means only allowing for equal second-to-first-stage ratios, and let furthermore $a_1 = \cdots = a_k = 1$ indicating balanced first stage sample sizes for the treatment and the control groups. After the second stage, the selected treatment group is compared to the control group (using data of both stages) applying the critical value $c_{1-\alpha}$ of the pre-planned test. Note that for this scenario the final test is balanced between both groups. In a slight modification of Proschan and Hunsberger (1995), the conditional type 1 error rate (3) of the final treatment-control comparison for $r_0 = r_m = r$ and $a_1 = \cdots = a_k = 1$ reduces to

$$CE_\alpha\big(r, Z_0^{(1)}, Z_m^{(1)}\big) = 1 - \Phi\left[c_{1-\alpha}\sqrt{\frac{r+1}{r}} - \big(Z_m^{(1)} - Z_0^{(1)}\big)\frac{1}{\sqrt{2r}}\right]$$

$$= 1 - \Phi\left[c_{1-\alpha}\sqrt{\frac{r+1}{r}} - T_m^{(1)}\frac{1}{\sqrt{r}}\right]$$

For notational convenience, the first stage test statistics $T_m^{(1)} = (Z_m^{(1)} - Z_0^{(1)})/\sqrt{2}$ is used. The conditional type 1 error rate does not depend on the unknown nuisance parameter μ.

Calculation of E^*_α in this balanced case follows the lines of Proschan and Hunsberger (1995). The essential difference is that the density of the maximum of k independent standard normal distributions has to be used in the integration. The subspaces of the interim sample space to perform separate optimizations remain the same (see Appendix A3).

The black lines in Fig. 1 show that if no correction for multiplicity is done (Fig. 1A), the type 1 error is highly inflated and increases with k. Using Dunnett boundaries for k treatment-control

Table 1 Maximum type 1 error rate for $k = 2$ with and without treatment selection, with or without adjustment for multiplicity and with equal or flexible second-to-first stage ratios as compared to the case $k = 1$.

	$\boldsymbol{k = 1}$		$\boldsymbol{k = 2}$			
nominal α	$c_{1-\alpha} = z_{1-\alpha}$		$c_{1-\alpha} = z_{1-\alpha}$		$c_{1-\alpha} = d_{1-\alpha}$	
	treatment selection of most promising treatment					
	equal (Proschan and Hunsberger, 1995)	flexible (Graf and Bauer, 2011)	equal (Section 4.1)	flexible (Section 4.2)	equal (Section 4.1)	flexible (Section 4.2)
0.01	0.0267	0.0491	0.0398	0.0697	0.0224	0.0407
0.025	0.0616	0.1064	0.0887	0.1466	0.0518	0.0892
0.05	0.1146	0.1867	0.1594	0.2496	0.0968	0.1588
	without treatment selection					
	equal (Proschan and Hunsberger, 1995)	flexible (Graf and Bauer, 2011)	equal (Section 5.1)	flexible (Section 5.2)	equal (Section 5.1)	flexible (Section 5.2)
0.01	0.0267	0.0491	0.0478	0.0800	0.0263	0.0473
0.025	0.0616	0.1064	0.1058	0.1701	0.0610	0.1037
0.05	0.1146	0.1867	0.1897	0.2885	0.1138	0.1842

comparisons, that means adjusting for all initially planned comparisons (Fig. 1B), the overall type 1 error decreases with k, that means correcting for multiplicity of all possible individual treatment-control comparisons leads to a smaller inflation of the overall type 1 error as compared to $k = 1$. For increasing k, the correction is done for an increasing number of $k - 1$ hypotheses not tested in the final analysis. Correcting for all possible individual treatment-control comparisons would be a conservative approach if the second stage sample size would be fixed independently of the data, for example in the planning phase. Here the inflation of the maximum type 1 error rate is caused by the worst case sample size reassessment rule.

For a direct comparison with the case of no treatment selection discussed later (see Section 5), the columns "equal" in Table 1 show the maximum overall type 1 error rate for $k = 2$ with and without correction for multiplicity as well as for the case of $k = 1$ (Proschan and Hunsberger, 1995).

If the first-stage-allocation-ratios are set to $a_i = 1/\sqrt{k}$, $i = 1, \ldots, k$, a smaller maximum type 1 error rate was found. When using the Dunnett critical boundary, for $\alpha = 0.025$ the values are $E^*_{0.025} = 0.0378$ for $k = 2$ and $E^*_{0.025} = 0.0351$ and 0.0340 for $k = 3$ and 4, respectively. For comparison, setting $a_i = 1$, the values are $E^*_{0.025} = 0.0518$ (see Table 1 and Fig. 1), 0.0482 and 0.0463 (see Fig. 1) for $k = 2$, 3, and 4, respectively. Similar results can be found for $\alpha = 0.01$ and $\alpha = 0.05$.

Remark 4.1. To give an impression of how results may change if the actual final adapted sample sizes are used in the calculation of the critical Dunnett boundaries (see Remark 3.1) for $k = 2$, the values would become only slightly smaller than in Table 1: 0.0221, 0.0507, and 0.0948 for $\alpha = 0.01$, 0.025, and 0.05, respectively.

4.2 Flexible second-to-first-stage-ratios

"Flexible" second-to-first-stage ratios allow different sample size reassessments for the selected treatment and the control, for example a sample size decrease for the control, but a sample size increase for the selected treatment group. For each interim outcome, the worst case $\tilde{r}_0$ and $\tilde{r}_m$ may differ. The sample size of the final treatment-control comparison may then be unbalanced between treatment arms. If we again assume balanced first stage sample sizes, the conditional type 1 error rate is now calculated by (3) setting $a_m = 1$. We use the independence of $Z_0^{(1)}$ and $Z_m^{(1)}$ to get rid of the nuisance parameter μ. The conditional type 1 error rate cannot be written as a function of the test statistic $T_m^{(1)}$ as in Section 4.1. As in Graf and Bauer (2011), the calculation of E_α^* can be separated into several parts of the interim subspace using $Z_m^{(1)}$ instead of $Z_1^{(1)}$. To evaluate the maximum type 1 error rate we partition the interim subspace in a way analogous to Graf and Bauer (2011) (see Section 1 in the Supplemental Materials).

The gray lines in Figs. 1A and B show that allowing for flexible second-to-first-stage ratios substantially increases the possible maximum type 1 error rate. Using $d_{1-\alpha}$ (Fig. 1B) in all scenarios leads to a nonmonotonous behavior with respect to the number of treatments k. An explanation for this is that the fixed boundaries are correct for the worst case scenarios, where the overall sample size is balanced between treatment and control, whereas for the unbalanced worst case scenarios they lead to smaller critical boundaries as compared to the boundaries using the actual total sample sizes. For larger k this difference in the correlation matrices is extended to all the $k - 1$ dropped treatments at interim, so that the differences between unbalanced and balanced critical boundaries tend to increase with increasing k which in the end leads to an increase in the maximum type 1 error rate. Again, to allow a direct comparison to the other discussed scenarios, the columns "flexible" in Table 1 show the values for $k = 2$ for both choices of the critical boundary as well as $k = 1$ (Graf and Bauer, 2011).

If $a_i = 1/\sqrt{k}$, $i = 1, \ldots, k$, as in the case of equal second-to-first stage ratios, a smaller maximum type 1 error was found. When using the Dunnett critical boundary, for $\alpha = 0.025$ the values are $E_{0.025}^* = 0.0860$, 0.0792, and 0.0753 for $k = 2$, 3, and 4, respectively. For comparison, setting $a_i = 1$, the values are $E_{0.025}^* = 0.0892$ (see Table 1 and Fig. 1), 0.0846, and 0.0830 (see Fig. 1) for $k = 2$, 3, and 4, respectively. Similar results can be found for $\alpha = 0.01$ or $\alpha = 0.05$.

Remark 4.2. When using Dunnett critical boundaries as in Remark 4.1, the maximum type 1 error rate up to $k = 10$ (data not shown) is smaller than for Dunnett critical boundaries based on balanced sample sizes. The maximum type 1 error rate is decreasing in k and hence also differences between the two approaches increase with k.

4.3 Constrained second stage sample size

Unconstrained sample size reassessment of course will hardly be used in practice. We therefore put constraints on the second-to-first-stage-ratios r_i, $r_{i,lo} \leq r_i \leq r_{i,up}$, $i \in \{0, m\}$. The ranges for the maximization in formula (5) are therefore changed to $r_{0,lo} \leq r_0 \leq r_{0,up}$ and $r_{m,lo} \leq r_m \leq r_{m,up}$. Figure 2 shows the maximum type 1 error rate E_α^*, $\alpha = 0.025$ for different constraints on sample size reassessment using the Dunnett critical boundary $d_{1-\alpha}$:

I. $r_{0,lo} = r_{m,lo} = 0$, $r_{0,up} = r_{m,up}$, $r_{m,up} = 1, 2, \ldots, 10$: Setting the lower boundary to 0 means that we allow for early rejection at interim. The solid lines in Fig. 2 show that E_α^* is increasing with increasing upper boundary, flattening off for larger values. Allowing for flexible second-to-first-stage-ratios (solid lines in Fig. 2B), the increase with the upper boundary is even steeper than for equal ratios (Fig. 2A). However, the results for $k \geq 3$ are very similar.

II. $r_{0,lo} = r_{m,lo} = 1$, $r_{0,up} = r_{m,up}$, $r_{m,up} = 1, 2, \ldots, 10$: In this scenario, the second stage sample size has to be at least as large as the first stage sample size for the selected treatment and the control. The dashed lines in Fig. 2A. show that for equal ratios and $k \geq 4$, the maximum type

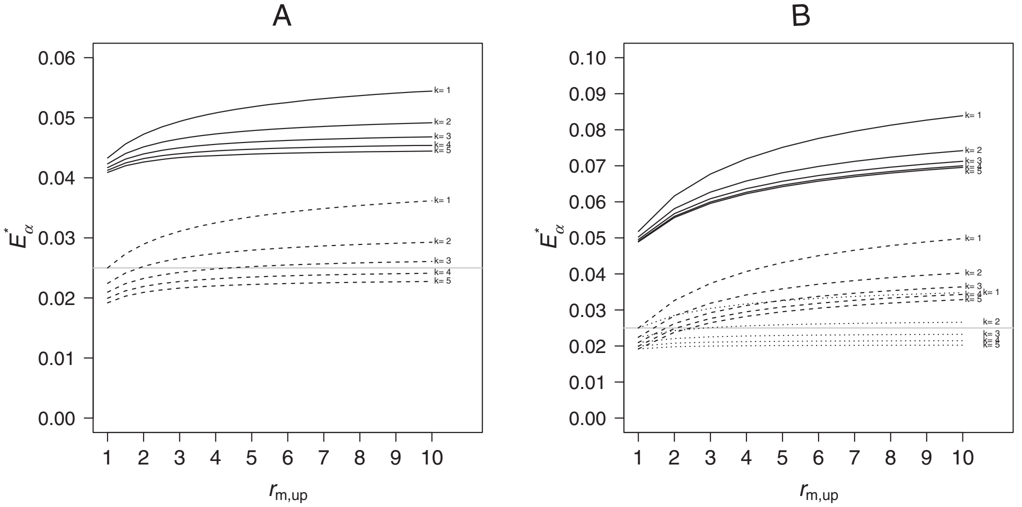

Figure 2 Maximum type 1 error rate E^*_α as a function of the upper boundary $r_{m,up}$ for the second-to-first-stage-ratio when always selecting the treatment with the maximum effect at interim for constrained second stage sample size for equal (A) and flexible (B) second-to-first-stage-ratios using Dunnett corrected critical boundaries. Solid lines: $r_{0,lo} = r_{m,lo} = 0$, $r_{0,up} = r_{m,up}$, dashed lines: $r_{0,lo} = r_{m,lo} = 1$, $r_{0,up} = r_{m,up}$ and dotted lines: $r_{0,lo} = r_{m,lo} = 1$, $r_{0,up} = 1$. Nominal one-sided α was set to 0.025.

1 error is always below the nominal $\alpha = 0.025$. Calculations including numerical integration of E^*_α for $k = 4$ and $r_{i,up} = \infty$ give a value of 0.02509. Therefore, for $k = 4$ selecting always only one treatment and the control, such type of constraints may be safely applied in practice without inflating the type 1 error rate. The reason is that there is a tradeoff (i) between the overcorrection from using Dunnett boundaries adjusting for treatment-control comparisons that are not carried over to the final test and (ii) the inflation due to data-dependent choice of the final sample size of the selected treatment (equal ratios, total sample size per selected treatment at least twice the first stage sample size per group). The smaller the prefixed range for the second stage sample sizes the smaller the impact of the latter effect.

Similar results can be found for a nominal α of 0.05 and 0.01. For $k = 4$ and $r_{i,up} = \infty$ the values are $E^*_{0.01} = 0.0106$ and $E^*_{0.05} = 0.0483$. Note that in the scenario for $k = 4$ without any interim sample size reassessment, for example: $r_{0,lo} = r_{m,lo} = r_{0,up} = r_{m,up} = 1$, the selection of one treatment and the control would happen quite late in the trial in terms of total sample over all groups (at a fraction of 5/7).

Allowing for flexible second-to-first-stage-ratios (Fig. 2B) only for smaller windows (smaller $r_{0,up}$ and $r_{m,up}$) E^*_α does not exceed α. For example for $\alpha = 0.025$ and $r_{0,up} = r_{m,up} = 2$, the number of treatments has to be larger than 3 so that E^*_α will always be below 0.025.

III. $r_{0,lo} = r_{m,lo} = 1$, $r_{0,up} = 1$, $r_{m,up} = 1, 2, \ldots, 10$: In this case, the second-to-first-stage-ratios are allowed to be flexible by definition, the only option for sample size adaptation is the choice of a second stage sample size for the selected treatment to be at least as large as in the first stage and not to exceed $r_{m,up}$ (see dotted lines in Fig. 2B). Such an adaptation may arise from a rare adverse event in the selected treatment group requiring additional information. It is interesting to note that for $k > 2$ the maximum type 1 error rate E^*_α will never exceed the nominal level, even if the upper boundary is set to ∞. For $k = 2$ no inflation occurs with $r_{m,up} = 2$. Similar results can be found for a nominal α of 0.05 and 0.01. Note that Fig. 2B shows that the type 1 error rate is not inflated when Dunnett critical boundaries are used in case of an allocation ratio to control of $1/(k+1)$ between treatment(s) and control in both stages, that is $r_{m,lo} = r_{m,up} = k$.

5 No treatment selection at interim

Since selecting only the treatment with the largest interim effect is a natural strategy often discussed in the literature (Cohen and Sackrowitz, 1989; Bowden and Glimm, 2008; Friede and Stallard, 2008; Stallard et al., 2008; Bauer et al., 2010), we first elaborated on this in Section 4. However, if all initially planned treatment arms are further investigated in the second stage, under the global null hypothesis, the maximum type 1 error rate is larger than for any other case with treatment selection. The reason is that dropping treatments at the interim analysis can be viewed as a constrained sample size reestimation problem (with $r_i = 0$ or $r_i = 1$ as the only options for treatment i), and this cannot produce a larger maximum of the conditional type 1 error rate than the unconstrained optimization problem.

For $k > 1$ we were not able to find a general closed solution for the maximum type 1 error rate (even if a single constant $c_{1-\alpha}$ is used as a critical boundary for all the k standardized treatment vs. control test statistics). To put the above optimization problem into a manageable framework, we illustrate the calculation for the case of two experimental treatment arms ($k = 2$) in the following. For the less complex scenario of equal second-to-first-stage ratios, numerical results are reported for $k > 2$.

5.1 Equal second-to-first-stage-ratios

As an extension to Proschan and Hunsberger (1995) we first investigate the case of equal second-to-first-stage-ratios setting $r_0 = r_1 = r_2 = r$. Assuming furthermore that the first stage sample sizes are balanced, that is setting $a_1 = a_2 = 1$ (and therefore also that the final stage sample sizes are balanced between treatment arms), for $k = 2$ formula (1) simplifies to

$$CE_\alpha\big(Z_0^{(1)}, Z_1^{(1)}, Z_2^{(1)}, r\big) = 1 - \int_{-\infty}^{\infty} \prod_{i=1}^{2} \Phi\left[c_{1-\alpha}\sqrt{\frac{2(1+r)}{r}} - \sqrt{\frac{2}{r}}T_i^{(1)} + Z_0^{(2)}\right] \phi\big(Z_0^{(2)}\big)dZ_0^{(2)}$$

$$= 1 - \Phi_{\mathbf{0},\Sigma}\begin{bmatrix} c_{1-\alpha}\sqrt{\dfrac{1+r}{r}} - \dfrac{1}{\sqrt{r}}T_1^{(1)} \\ c_{1-\alpha}\sqrt{\dfrac{1+r}{r}} - \dfrac{1}{\sqrt{r}}T_2^{(1)} \end{bmatrix}.$$

As in Section 4.1, for notational convenience, the first stage test statistics $T_i^{(1)}$ for comparing treatment i to the control are used. The conditional type 1 error rate does not depend on the nuisance parameter μ. The cumulative distribution function of the multivariate normal distribution with two-dimensional mean zero-vector $\mathbf{0}$ and covariance-matrix Σ with elements $\sigma_{11} = \sigma_{22} = 1$ and covariance $\sigma_{12} = 1/2$ is denoted by $\Phi_{\mathbf{0},\Sigma}(\mathbf{x})$. To calculate the worst case conditional type 1 error rate we have to partition the $(T_1^{(1)}, T_2^{(1)})$-plane.

I. If $T_1^{(1)} < 0$ and $T_2^{(1)} < 0$ the largest conditional type 1 error rate is obtained by setting $\tilde{r} = \infty$, $\tilde{r}$ denoting the worst case second-to-first-stage ratio. The second stage is now overruling the negative interim effect and therefore yielding a $\widetilde{CE}_\alpha = 1 - \Phi_{\mathbf{0},\Sigma}\left[\mathbf{c}_{1-\alpha}\right]$ that is equal to α if $\mathbf{c}_{1-\alpha} = (d_{1-\alpha}, d_{1-\alpha})'$. Since $P[(T_1^{(1)} < 0) \cap (T_2^{(1)} < 0)] = \frac{1}{3}$ for the bivariate normal distribution with $\sigma_{12} = 1/2$ (see e.g. Kotz et al., 2000), the contribution of this subspace to the overall maximum type 1 error rate E_α^* is $\frac{1}{3}\left(1 - \Phi_{\mathbf{0},\Sigma}\left[\mathbf{c}_{1-\alpha}\right]\right)$.

II. If $T_1^{(1)} > c_{1-\alpha}$ or $T_2^{(1)} > c_{1-\alpha}$ the largest conditional type 1 error rate $\widetilde{CE}_\alpha = 1$ (applying early rejection at interim and setting $\tilde{r} = 0$) is obtained. This leads to a contribution to E_α^* of $P[(T_1^{(1)} > c_{1-\alpha}) \cup (T_2^{(1)} > c_{1-\alpha})] = 1 - \Phi_{\mathbf{0},\Sigma}\left[\mathbf{c}_{1-\alpha}\right]$ that is equal to α if $\mathbf{c}_{1-\alpha} = (d_{1-\alpha}, d_{1-\alpha})'$.

In the remaining interim subspace we were not able to find a closed solution for $\widetilde{CE}_\alpha$. Therefore, we used numerical optimization of the single parameter r. The "equal"-columns of Table 1 show the results for the overall E^*_α for the case of $k = 2$, with and without correction for multiplicity. As is to be expected, applying the naive unadjusted critical boundary $z_{1-\alpha}$ may result in a further considerable type 1 error rate inflation as compared to $k = 1$. An interesting finding is that when using the Dunnett critical value, E^*_α are close to the results for $k = 1$.

For $k = 3$ and using Dunnett critical boundaries for $\alpha = 0.025$ the maximum type 1 error rate is still inflated up to 0.0545, but interestingly the inflation is smaller compared to $k = 2$. For $k = 4$ treatments, $E^*_{0.025}$ is flattening off at an inflated level of 0.0543. For $\alpha = 0.01$ and 0.05 the same tendencies can be found.

5.2 Flexible second-to-first-stage-ratios

If we allow for flexible second-to-first-stage-ratios, we again have to use the independent $Z_i^{(1)}$ (instead of the test statistics $T_i^{(1)}$) to get rid of the nuisance parameter μ. If we assume balanced first stage sample size, the conditional type 1 error rate is now calculated by (1) setting $k = 2$ and $a_1 = a_2 = 1$. To explain the worst case scenarios in more detail, we will focus on the subspaces in terms of the interim outcome of the control group $Z_0^{(1)}$.

A. Subspace ($Z_0^{(1)} \leq -c_{1-\alpha}$): $\widetilde{CE} = 1$ is obtained by setting either $\tilde{r}_1$ or $\tilde{r}_2$ to ∞ and $\tilde{r}_0 = 0$. The contribution of this subspace to E^*_α therefore is $\Phi(-c_{1-\alpha})$.

B. Subspace ($Z_0^{(1)} \geq 0$): The worst case choice is setting $\tilde{r}_0 = \infty$ in the final analysis, getting two independent tests against the asymptotically fixed mean $\mu = 0$. Hence the conditional type 1 error rate reduces to

$$1 - \Phi\left(c_{1-\alpha}\sqrt{\frac{1}{r_1} + 1} - Z_1^{(1)}\frac{1}{\sqrt{r_1}}\right)\Phi\left(c_{1-\alpha}\sqrt{\frac{1}{r_2} + 1} - Z_2^{(1)}\frac{1}{\sqrt{r_2}}\right)$$

independent of $Z_0^{(1)}$. A detailed explanation for the calculation of the maximum type 1 error rate for this subspace B is given in Section 2 in the Supplemental Materials. Summing up the results for $Z_0^{(1)} \geq 0$, the contribution to the overall maximum type 1 error rate can be calculated by

$$\frac{1}{2}\left(1 - \frac{1}{16}e^{-c^2_{1-\alpha}} + \frac{1}{2}e^{\frac{-c^2_{1-\alpha}}{2}}\Phi(c_{1-\alpha}) - \Phi(c_{1-\alpha})^2\right).$$

C. Subspace ($-c_{1-\alpha} < Z_0^{(1)} < 0$): In this region the worst case conditional type 1 error rate depends on all three interim values of control and treatment groups, respectively. If either $Z_1^{(1)}$ or $Z_2^{(1)}$ is larger than $\min(c_{1-\alpha}\sqrt{2} + Z_0^{(1)}, c_{1-\alpha})$ again a conditional type 1 error rate of 1 can be achieved. For the remaining regions we used numerical point-wise optimization and integration for calculating the contribution to the overall type 1 error rate E^*_α.

The columns "flexible" for $k = 2$ of Table 1 show the total E^*_α for flexible second-to-first-stage-ratios applying critical boundaries $z_{1-\alpha}$ or $d_{1-\alpha}$. Without any correction for multiplicity ($z_{1-\alpha}$), the maximum type 1 error is clearly increased as compared to the case $k = 1$. Interestingly, as for the results of equal second-to-first-stage ratios (see Section 5.1), when using the pre-specified Dunnett critical boundary, E^*_α is close to the results for $k = 1$.

Due to the numerical burden we did not calculate the maximum type 1 error rate for $k > 2$. However, we expect similar findings as for the case of equal second-to-first-stage ratios at least for $k = 3$ and

4, that is the maximum type 1 error rate sightly decreasing when using a Dunnett adjusted critical boundary.

6 Practical recommendations

The results presented for the case of selecting the most promising hypothesis at interim are of great practical interest, because they demonstrate that, given certain restrictions on the second stage sample size, naive strategies may even lead to an adequate control of the type 1 error rate. For example, if the sample size per treatment group in the second stage is at least as large as in the first stage and we only allow for equal second-to-first-stage-ratios, no inflation of the type 1 error rate occurs for the number of treatments $k \geq 4$ when simply using the Dunnett critical boundaries. For $k = 3$, no inflation occurs when restricting the second-stage sample size to be at maximum 4 times the first-stage sample size (see Fig. 2A). If we fix the overall sample size in the control group, allowing for any choice of the overall sample size in the selected treatment group that increases its first stage sample size more than twofold does not lead to an inflation of α for $k \geq 3$ (see Fig. 2B). Therefore, if in the case study of Barnes et al. (2010) (see Section 2) only the selection of a single treatment group and control had been pre-specified, the experimenter would have been permitted to do any balanced increase of the sample size, even when using the conventional test statistic and the less conservative Dunnett critical boundary (instead of the applied Bonferroni adjustment) for final testing. If a flexible sample size reassessment for the second stage would have been allowed for (as in Section 4.2), no type 1 error inflation would have occurred if the second stage sample size would have been constrained to be between the first-stage and twice the first stage sample size. However, it has to be noted that for realistic scenarios (as e.g. an upper bound of twice the first stage sample size) and a larger k, the obtained maximum type 1 error rate may be much smaller than α so that even using the Dunnett critical boundaries would lead to conservative procedures. Note that these results only apply when using prespecified-binding constraints on the selection rules.

Allowing for early rejection at interim, the maximum type 1 error rate will always be inflated. In such scenarios, if the use of conventional test statistics is preferred, one may adjust the critical boundary so that the maximum type 1 error rate is controlled. As an example, assume that we only allow for equal second-to-first-stage ratios setting the upper bound of the second-stage-sample size of the selected treatment and control to be twice the first stage sample size. For $k = 2$ an adjusted level of 0.013 (instead of 0.025) has to be used to control the maximum type 1 error rate. In more detail, if we assume for both treatments an effect size of 0.5 times the standard deviation, a sample size of $n = 65$ per group would be needed to achieve 80% power. Compared to a fixed sample size test with Dunnett adjusted critical boundaries, this would be a 20.4% increase of the per-group sample size. For increasing k, this is only slightly decreasing: for $k = 3$ an increase of 18.8% and for $k = 4$ an increase of 17.0% of the per-group sample size is needed to control the maximum type 1 error rate when additionally allowing for the given sample size reassessment. To achieve a power of 90%, a slightly smaller increase in the per-group sample size is needed, that means an increase of 16.4%, 16.7%, and 15.6% would be needed for $k = 2$, 3, and 4, respectively. All these examples show that adjusting for the worst case would be a rather conservative strategy and adaptive tests should be implemented instead (Koenig et al., 2008; Bretz et al., 2009).

7 Discussion

In this paper, we have investigated the maximum type 1 error rate arising from the application of a nonadaptive test used by experimenters who freely adapt their ongoing trials. This problem has been addressed by Proschan and Hunsberger (1995) for the comparison of one treatment with a control and balanced sample sizes before and after the adaptive interim analysis. They considered a restricted rule incorporating a stopping for futility criterion. This leads to procedures where the effect of adjusting the adaptation of the sample size is no longer dramatic. Graf and Bauer (2011) have extended the worst case calculations allowing for unbalanced sample sizes. In this paper, a further level of complexity has been added by considering multiple comparisons of k treatments with a single control. For the case without selection of a treatment arm at interim, we calculate the maximum type 1 error rate for $k = 2$ in the case of equal and flexible second-to-first-stage-ratios (assuming balanced first stage sample sizes). Not surprisingly, when applying uncorrected level α treatment-control comparisons, the worst case type 1 error is dramatically inflated. By using Dunnett-adjusted critical boundaries, the worst case inflation is still large. Interestingly, the inflation is very similar to the case of comparing $k = 1$ treatment to a control (Graf and Bauer, 2011). This means that when adjusting for the number of treatments for $k = 2$, no noticeable further maximum inflation of the type 1 error rate occurs as compared to $k = 1$.

The case of equal and flexible second-to-first-stage ratios was investigated for scenarios where only a single treatment and the control are selected at the interim analysis. In this scenario, there is a trade-off between inflation due to sample size reassessment and the overcorrection for the $k - 1$ treatments finally not selected and not tested in the statistical analysis. For equal ratios, the maximum type 1 error is monotonically decreasing with k with a finite limit noticeably larger than the nominal level α. As expected, the impact of flexible ratios is more severe, the maximum inflation of the actual level α, though decreasing for small k, is increasing with larger k.

There are several caveats to be mentioned here. First, for the case of flexible ratios the conditional error can only be calculated when the nuisance parameter, the common mean under the global null hypothesis, is known. Secondly, the maximum type 1 error only occurs if the experimenters apply the worst case sample size reassessment rule (maximizing the conditional type 1 error rate) at any point in the interim sample space. Thirdly, in some interim subspace, the maximum is assumed if some of the second stage sample sizes go to infinity. Although theoretically interesting, this of course means that these maximum type 1 error rates can never be reached in real clinical trials. Adjusting for these "unrestricted worst cases" would be an extremely conservative strategy and cannot be recommended for use in practice. Therefore, we also investigated maximum type 1 error rates that arise when the second stage sample sizes are constrained by upper and lower limits. Some of these results are practically interesting, because they demonstrate that in certain cases, when putting restrictions on the second-stage-sample sizes, naive strategies can control the type 1 error rate. Such calculations under constraints could replace simulations of the type 1 error rate in designs with adaptive selection rules, the latter being considered problematic by some researchers (Posch et al., 2011).

Open research problems are at present the unconstrained optimization for $k > 2$, which imposes a burden of numerical integration and optimization. For the unconstrained scenario of $k = 2$, the optimization lasts up to one half second for one grid point on an Intel(R)Core(TM)i5 CPU M540 processor with 2.53GHz and it is therefore still a time consuming numerical challenge to derive a sufficiently narrow grid over the three dimensional interim subspaces with sufficiently accurate values of the maximum conditional error functions to be integrated. Also scenarios where the selection of s, $1 < s < k$ out of k treatment groups and the control are prespecified are of high interest.

As a conclusion, we do not recommend the use of unrestricted "worst case" adjustments since they will be far too conservative for serious consideration. If limits on sample size modifications can be imposed, it is still important to compare the operating characteristics of adaptive designs with the maximum-type-1-error-based adjustments discussed here. Only then we can decide whether sample size limits can or should be imposed and how tight they might be.

Acknowledgments The work of Alexandra Graf was supported by the Austrian Science Fund (FWF), Project No.: J3344-N26. Part of her work was undertaken during a research stay at the University of

Bremen. She is grateful to Werner Brannath for hospitality at the Competence Center for Clinical Trials as well as for helpful comments. Peter Bauer's and Franz Koenig's research has received funding from the European Union Seventh Framework Programme [FP7 2007–2013] under grant agreement No.: 602552. Furthermore, we thank the unknown reviewers, the associate editor and editor Lutz Edler for helpful suggestions as well as Byron Jones for critical proofreading of the paper, which have improved the paper substantially.

Conflict of interest

The authors have declared no conflict of interest.

Appendix

A.1 Calculation of the conditional type 1 error rate

In the following we assume that the global null hypothesis applies ($\mu_i = \mu$ for $i = 0, \ldots, k$). The conditional type 1 error for rejecting at least one treatment-control comparison in the final analysis, given the interim data, can be calculated as follows:

In the final analysis after the second stage each test (comparing treatment i to the control group) is based on the following global test statistic:

$$T_i = \frac{\frac{n_i^{(1)}\bar{x}_i^{(1)}+n_i^{(2)}\bar{x}_i^{(2)}}{n_i^{(1)}+n_i^{(2)}} - \frac{n_0^{(1)}\bar{x}_0^{(1)}+n_0^{(2)}\bar{x}_0^{(2)}}{n_0^{(1)}+n_0^{(2)}}}{\sqrt{\frac{1}{n_i^{(1)}+n_i^{(2)}} + \frac{1}{n_0^{(1)}+n_0^{(2)}}}}$$

$$= \frac{\frac{(\bar{x}_i^{(1)}-\mu)n_i^{(1)}+(\bar{x}_i^{(2)}-\mu)n_i^{(2)}}{n_i^{(1)}+n_i^{(2)}} - \frac{(\bar{x}_0^{(1)}-\mu)n_0^{(1)}+(\bar{x}_0^{(2)}-\mu)n_0^{(2)}}{n_0^{(1)}+n_0^{(2)}}}{\sqrt{\frac{1}{n_i^{(1)}+n_i^{(2)}} + \frac{1}{n_0^{(1)}+n_0^{(2)}}}}$$

$$= \frac{\frac{Z_i^{(1)}+\sqrt{r_i}Z_i^{(2)}}{\sqrt{a_i}(1+r_i)} - \frac{Z_0^{(1)}+\sqrt{r_0}Z_0^{(2)}}{1+r_0}}{\sqrt{\frac{1}{a_i(1+r_i)} + \frac{1}{1+r_0}}}$$

where $Z_i^{(j)} = (\bar{x}_i^{(j)} - \mu)\sqrt{n_i^{(j)}}$, $i \geq 0$, $j = 1, 2$, have independent standard normal distributions. If the overall test statistic T_i is larger than the critical boundary $c_{1-\alpha}$ we get a false positive decision, which leads to the following inequality:

$$\frac{\sqrt{r_i}Z_i^{(2)}}{\sqrt{a_i}(1+r_i)} > c_{1-\alpha}\sqrt{\frac{1}{a_i(1+r_i)} + \frac{1}{1+r_0}} + \frac{Z_0^{(1)}}{(1+r_0)} + \frac{Z_0^{(2)}\sqrt{r_0}}{1+r_0} - \frac{Z_i^{(1)}}{\sqrt{a_i}(1+r_i)}$$

leading to

$$Z_i^{(2)} > c_{1-\alpha}\sqrt{\frac{1+r_i}{r_i} + \frac{a_i(1+r_i)^2}{(1+r_0)r_i}} + Z_0^{(1)}\sqrt{\frac{a_i}{r_i}}\frac{1+r_i}{1+r_0} + Z_0^{(2)}\sqrt{\frac{r_0 a_i}{r_i}}\frac{1+r_i}{1+r_0} - \frac{Z_i^{(1)}}{\sqrt{r_i}}$$

Since $Z_i^{(2)}$ and $Z_0^{(2)}$ have independent standard normal distributions for every set of values a_i, r_0, r_i, $Z_0^{(1)}$ and $Z_i^{(1)}$ (and hence are independent of these quantities), the conditional error can be written as in formula (1).

A.2 The maximum type 1 error rate is attained under the global null hypothesis

For $k \geq 2$ the maximum type 1 error rate is attained under the global null hypothesis $\mu_i = \mu_0, i = 1, \ldots, k$. To see this, assume (without loss of generality) that the null hypothesis is not true for H_{01}. Then the conditional error for rejecting at least one true hypothesis is calculated in the same way as $CE_\alpha(a_1, \ldots, a_k, Z_0^{(1)}, \ldots, Z_k^{(1)}, r_0, \ldots, r_k)$ from formula (1), but the product going from 2 to k rather than 1 to k, further on denoted by $CE_\alpha^*(a_2, \ldots, a_k, Z_0^{(1)}, Z_2^{(1)}, \ldots, Z_k^{(1)}, r_0, r_2, \ldots, r_k)$. Obviously, for given $r_0, r_2, \ldots, r_k$, $CE_\alpha^* \leq CE_\alpha$. This statement is true since the integrated term in formula (1), for given $r_0, r_2, \ldots, r_k$, is the same for CE_α and CE_α^* but r_1 does not appear in CE_α^*. Since the integrated function in (1) is at every point larger (or equal) in CE_α^*, the whole integral must be larger (or equal) for CE_α^* as compared to CE_α for every given constellation of $r_0, r_2, \ldots, r_k$.

Let now denote $r_0^*, r_2^*, \ldots, r_k^*$ the second-two-first-stage-ratios leading to the maximum $\widetilde{CE}_\alpha^*$. Due to the above arguments

$$\widetilde{CE}_\alpha^*\big(a_2, \ldots, a_k, Z_0^{(1)}, Z_2^{(1)}, \ldots, Z_k^{(1)}\big) \leq CE_\alpha\big(a_1, \ldots, a_k, Z_0^{(1)}, \ldots, Z_k^{(1)}, r_0^*, r_1, r_2^*, \ldots, r_k^*\big)$$

for all r_1. The ratios $r_0^*, r_2^*, \ldots, r_k^*$ may not be the ratios maximizing CE_α, but finding the ratios leading to $\widetilde{CE}_\alpha$ can only increase the conditional type 1 error CE_α and thus

$$\widetilde{CE}_\alpha^*\big(a_2, \ldots, a_k, Z_0^{(1)}, Z_2^{(1)}, \ldots, Z_k^{(1)}\big) \leq \widetilde{CE}_\alpha\big(a_1, \ldots, a_k, Z_0^{(1)}, \ldots, Z_k^{(1)}\big).$$

This domination also holds for the integral for the type 1 error $E_\alpha^*(a_1, \ldots, a_k)$ in formula (2), showing that the global null indeed gives the parameter constellation leading to the largest type 1 error inflation.

A.3 Maximum conditional type 1 error rate when selecting the most promising treatment for the scenario of equal second-to-first-stage-ratios

Following the lines of Proschan and Hunsberger (1995) the maximum conditional type 1 error rate when selecting the treatment with the largest observed interim effect for the scenario of equal second-to-first-stage-ratios (refer to Section 4.1) can be calculated by dividing the interim sample space into three subspaces:

I. If $T_m^{(1)} \leq 0$ (equivalently $Z_m^{(1)} - Z_0^{(1)} < 0$) the worst case arises from setting $\tilde{r} = \infty$ so that the second stage overrules the first stage adverse effect leading to $\widetilde{CE}_\alpha(Z_0^{(1)}, Z_m^{(1)}) = 1 - \Phi(c_{1-\alpha})$. This results in a contribution to E_α^* of $(1 - \Phi(c_{1-\alpha}))/(k+1)$ since $P[T_m^{(1)} < 0] = \Phi_{\mathbf{0},\Sigma}(\mathbf{0}) = \frac{1}{1+k}$, $\mathbf{0}$ here denoting the k-dimensional zero vector and Σ the k-dimensional covariance matrix with $\sigma_{ii} = 1$ and $\sigma_{ij} = 1/2$ for $i \neq j$.

II. Within the subspace $0 < T_m^{(1)} < c_{1-\alpha}$ (or equivalently $Z_0^{(1)} < Z_m^{(1)} < c_{1-\alpha}\sqrt{2} + Z_0^{(1)}$), Proschan and Hunsberger (1995) showed that $\tilde{r} = (\frac{c_{1-\alpha}}{T_m^{(1)}})^2$ is leading to a worst case conditional type 1 error rate $\widetilde{CE}_\alpha(T_m^{(1)}) = 1 - \Phi(\sqrt{c_{1-\alpha}^2 - (T_m^{(1)})^2})$. We found no simplification of the two-dimensional integration in this subspace.

III. If $T_m^{(1)} > c_{1-\alpha}$ (or equivalently $Z_m^{(1)} > c_{1-\alpha}\sqrt{2} + Z_0^{(1)}$) the test can be rejected already at interim ($\tilde{r} = 0$ with $\widetilde{CE}_\alpha = 1$) leading to a contribution of this subspace of $1 - \Phi_{\mathbf{0},\Sigma}[\mathbf{c}_{1-\alpha}]$ where $\mathbf{c}_{1-\alpha}$ is a k-dimensional vector with values $c_{1-\alpha}$ which is $P[T_m^{(1)} > c_{1-\alpha}]$ and reduces to α when using the multiplicity corrected Dunnett critical boundary $d_{1-\alpha}$.

References

Barnes, P. J., Pocock, S. J., Magnussen, H., Iqbal, A., Kramer, B., Higgins, M. and Lawrence, D. (2010). Integrating indacaterol dose selection in a clinical study in COPD using an adaptive seamless design. *Pulmonary Pharmacology and Therapeutics* **23**, 165–171.

Bauer, P. (1989). Multistage testing with adaptive designs. *Biometrie und Informatik in Medizin und Biologie* **20**, 130–148.

Bauer, P. and Koehne, K. (1994). Evaluations of experiments with adaptive interim analysis. *Biometrics* **50**, 1029–1041.

Bauer, P. and Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* **18**, 1833–1848.

Bauer, P., Koenig, F., Brannath, W. and Posch, M. (2010). Selection and bias - Two hostile brothers. *Statistics in Medicine* **29**, 1–13.

Bebu, I., Dragalin, V. and Luta, G. (2013). Confidence intervals for confirmatory adaptive two-stage designs with treatment selection. *Biometrical Journal* **55**, 294–309.

Bowden, J. and Glimm, E. (2008). Unbiased estimation of selected treatment means in two-stage trials. *Biometrical Journal* **50**, 515–527.

Brannath, W., Posch, M. and Bauer, P. (2002). Recursive combination tests. *JASA* **97**, 236–244.

Brannath, W., Bauer, P. and Posch, M. (2006). On the efficiency of adaptive designs for flexible interim decisions in clinical trials. *Journal of Statistical Planning and Inference* **136**, 1956–1961.

Bretz, F., Koenig, F., Brannath, W., Glimm, E. and Posch, M. (2009). Adaptive designs for confirmatory clinical trials. *Statistics in Medicine* **28**, 1181–1217.

Burman, C. F. and Sonesson, C. (2006). Are flexible designs sound? *Biometrics* **62**, 664–669.

Cohen, A. and Sackrowitz, H. (1989). Two stage conditionally unbiased estimators of the selected mean. *Statistics and Probability Letters* **8**, 273–278.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *JASA* **50**, 1096–1121.

Elsaesser, A., Regnstroem, J., Vetter, T., Koenig, F., Hemmings, R., Greco, M., Papaluca-Amati, M. and Posch, M. (2014). Adaptive designs in European Marketing Authorisation—a survey of advice letters at the European Medicines Agency, submitted.

EMA (2007). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. Doc. Ref. CHMP/EWP/2459/02 Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003616.pdf

FDA Draft Guidance (2010). Adaptive Design Clinical Trials for Drugs and Biologics. Available at: http://www.fda.gov/downloads/Drugs/.../Guidances/ucm201790.pdf

Friede, T. and Stallard, N. (2008). A comparison of methods for adaptive treatment selection. *Biometrical Journal* **50**, 767–781.

Gao, P., Liu, L. and Mehta, C. (2013). Adaptive designs for noninferiority trials. *Biometrical Journal* **55**, 310–321.

Gaydos, B. (2009). Phase 2/3 adaptive design utilizing a Bayesian decision analytic approach to dose selection. *EMA/EFPIA 2nd Workshop: Adaptive Design in Confirmatory Trials*, EMA Headquaters, UK, Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Minutes/2010/04/WC500089206.pdf

Graf, A. C. and Bauer, P. (2011). 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* **30**, 1637–1647.

Jennison, C. and Turnbull, B. W. (2006). Adaptive and non-adaptive group sequential tests. *Biometrika* **93**, 1–21.

Koenig, F., Brannath, W., Bretz, F. and Posch, M. (2008). Adaptive Dunnett tests for treatment selection. *Statistics in Medicine* **27**, 1612–1625.

Kotz, S., Balakrishnan, N. and Johnson, N. L. (2000). *Continuous Multivariate Distributions*. John Wiley and Sons, New York, NY.

Lawrence, D., Bretz, F. and Pocock, S. (2014). INHANCE: an adaptive confirmatory study with dose selection at interim In: A. Trifilieff (Ed.), *Indacaterol: The First Once-daily Long-acting Beta2 Agonist for COPD, Milestones in Drug Therapy*. Springer, Basel, CH, pp. 77–92.

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

Levin, G. P., Emerson, S. C. and Emerson, S. S. (2013). Adaptive clinical trial designs with pre-specified rules for modifying the ample size: understanding efficient types of adaptation. *Statistics in Medicine* **32**, 1259–1275.

Mehta, C. and Pocock, S. (2012). Authors' reply. *Statistics in Medicine* **31**, 99–100.

Morgan, C. C., Huyck, S., Jenkins, M., Chen, L., Bedding, A., Coffey, C. S., Gaydos, B. and Wathen, J. K. (2014). Adaptive design: results of 2012 survey on perception and use. *Therapeutic Innovation and Regulatory Science*, doi:10.1177/2168479014522468.

Mueller, H. H. and Schaefer, H. (2001). Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* **95**, 886–891.

Mueller, H. H. and Schaefer, H. (2004). A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine* **23**, 2497–2508.

Posch, M., Maurer, W. and Bretz, F. (2011). Type 1 error rate control in adaptive designs for confirmatory clinical trials with treatment selection at interim. *Pharmaceutical Statistics* **10**, 96–104.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of Studies based on conditional power. *Biometrics* **51**, 1315–1324.

R Development Core Team (2012). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria: ISBN 3-900051-07-0. Available at: http://www.R-project.org

Stallard, N., Todd, S. and Whitehead, J. (2008). Estimation following selection of the largest of two normal means. *Journal of Statistical Planning and Inference* **138**, 1629–1638.

Sugitani, T., Hamasaki, T. and Hamada, C. (2013). Partition testing in confirmatory adaptive designs with structured objectives. *Biometrical Journal* **55**, 341–359.

Thall, P. F., Simon, R. and Ellenberg, S. S. (1988). Two-stage selection and testing designs for comparative clinical trials. *Biometrika* **75**, 303–310.

Thall, P. F., Simon, R. and Ellenberg, S. S. (1988). A two-stage design for choosing among several experimental treatments and control in clinical trials. *Biometrics* **45**, 537–547.

Tsiatis, A. A. and Metha, C. R. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* **90**, 367–378.

Wang, S. J., Bretz, F., Dmitrienko, A., Hsu, J., Hung, H. M., Huque, M. and Koch, G. (2013). Panel forum on multiple comparison procedures: a commentary from a complex trial design and analysis plan. *Biometrical Journal* **55**, 275–293.

Zeymer, U., Suryapranata, H., Monassier, J. P., Opolski, G., Davies, J., Rasmanis, G., Linssen, G., Tebbe, U., Schroder, R., Tiemann, R., Machnig, T. and Neuhaus, K. L. (2001). The Na^+/H^+ exchange inhibitor Eniporide as an adjunct to early reperfusion therapy for acute myocardial infarction. *Journal of the American College of Cardiology* **38**, 1664–1651.