STATISTICS IN MEDICINE
Statist. Med. (2009)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3538

TUTORIAL IN BIOSTATISTICS

Adaptive designs for confirmatory clinical trials

Frank Bretz[1,2,‡], Franz Koenig[3,*,†,‡], Werner Brannath[3],
Ekkehard Glimm[1] and Martin Posch[3]

[1] *Novartis Pharma AG, Lichtstrasse 35, 4002 Basel, Switzerland*
[2] *Department of Biometry, Medical University of Hannover, 30623 Hannover, Germany*
[3] *Section of Medical Statistics, Medical University of Vienna, Spitalgasse 23, 1090 Wien, Austria*

SUMMARY

Adaptive designs play an increasingly important role in clinical drug development. Such designs use accumulating data of an ongoing trial to decide how to modify design aspects without undermining the validity and integrity of the trial. Adaptive designs thus allow for a number of possible adaptations at midterm: Early stopping either for futility or success, sample size reassessment, change of population, etc. A particularly appealing application is the use of adaptive designs in combined phase II/III studies with treatment selection at interim. The expectation has arisen that carefully planned and conducted studies based on adaptive designs increase the efficiency of the drug development process by making better use of the observed data, thus leading to a higher information value per patient.

In this paper we focus on adaptive designs for confirmatory clinical trials. We review the adaptive design methodology for a single null hypothesis and how to perform adaptive designs with multiple hypotheses using closed test procedures. We report the results of an extensive simulation study to evaluate the operational characteristics of the various methods. A case study and related numerical examples are used to illustrate the key results. In addition we provide a detailed discussion of current methods to calculate point estimates and confidence intervals for relevant parameters. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: adaptive seamless design; design modification; flexible design; combination test; conditional error rate; interim analysis; many-to-one comparisons; treatment selection

[*] Correspondence to: Franz Koenig, Section of Medical Statistics, Core Unit for Medical Statistics and Informatics, Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria.

[†] E-mail: franz.koenig@meduniwien.ac.at

[‡] The first two authors (in alphabetic order) have made equal contributions to this paper.

Contract/grant sponsor: FWF; contract/grant number: P18698-N15

Received 4 May 2008
Accepted 8 December 2008

1. INTRODUCTION

Interim analyses are often conducted in clinical trials because of ethical and economical reasons. On the one hand, clinical trials should not be continued (and decisions postponed) if a clear tendency favoring a particular treatment evolves so that patients in need can benefit quickly from the medical progress. On the other hand, patients should not be treated with a new therapy (for which in such situations only limited knowledge about the risks is available) if the ongoing trial gives no indication for a potential benefit. Moreover, clinical trial designs that allow for early decisions during the conduct of an ongoing study may reduce the overall costs and timelines of the development program for the new therapy.

Repeatedly looking at the data with the possibility for interim decision making, however, may inflate the overall type I error rate since the primary null hypotheses of interest are tested anew at each interim analysis. Special analysis methods are therefore required to maintain the validity of a clinical trial, if confirmatory conclusions at the final analysis are required, such as in late phase II or phase III trials. Group sequential designs are commonly used to account for the repeated data analyses [1–4]. Such designs allow one to stop a trial at any interim analysis for either futility or superiority while controlling the overall type I error rate. However, standard design aspects, like the number of interim analyses, the group sample sizes and the decision boundaries, have to be specified in the planning phase and cannot be changed during the ongoing trial. Once the design has been fixed and the trial has been started, it has to be conducted according to the pre-specified decision rules. Group sequential designs are thus characterized by a *pre-specified adaptivity* [5], that is, trial design modifications based on the information collected up to an interim analysis are not possible. One possibility to introduce more flexibility is to consider error spending functions [6], which allow for a flexible number of stages and group sample sizes. Such methods, however, require that the adaptations are independent from the interim test statistics. Extensions thereof to multi-armed clinical trials were investigated, among others, by Follmann *et al.* [7], Hellmich [8] and Stallard and Friede [9].

To overcome the inherent limitations of such designs, *confirmatory adaptive designs* have been proposed instead, which enable the user to perform design modifications during an ongoing clinical trial while maintaining the overall type I error rate [10, 11]. This new class of trial designs is characterized by an *unscheduled adaptivity* [5], that is, adapting design parameters can be done without a complete pre-specification of the adaptation rules. Applying such a clinical trial design with flexible interim decisions allows the user to learn from the observed data during an ongoing trial and to react quickly to emerging unexpected results. Important examples of mid-term adaptations include sample size reestimation or reallocation based on the observed nuisance parameters [12] or dropping of treatment arms in combined phase II/III trials [13, 14].

To control the overall type I error rate, confirmatory adaptive designs satisfy a common invariance principle [5]: Separate test statistics are calculated from the samples at the different stages and combined in a pre-specified way for the final test decisions. Hence, any design modification that preserves the distributional properties of the separate stagewise test statistics under a given null hypothesis H of interest does not lead to an inflation of the overall type I error rate [11, 15]. Approaches based on the *combination test principle* combine the stagewise p-values using a pre-specified combination function [10, 11] and by construction satisfy the invariance principle. A closely related approach is based on the *conditional error principle*, which computes the type I error under the null hypothesis H conditional on the observed data at interim and use this quantity for the final analysis [16–18]. Such confirmatory adaptive designs thus enable the user to perform mid-trial design modifications based on the complete observed information from an ongoing study, possibly including information external to the trial (historical data, parallel study results, etc.). In particular, Bayesian interim decision tools can be applied to guide the interim decision process without undermining the frequentist character of the final analysis [19, 20].

The investigation of novel adaptive trial designs has recently attracted much attention because of the recognized need to improve the efficiency of current drug development processes [21]. The Pharmaceutical Research and Manufacturers of America (PhRMA), for example, has initiated a working group to facilitate a wider usage and regulatory acceptance of these designs [22]. Adaptive designs are mentioned explicitly in the Critical Path Opportunities List published by the U.S. Food and Drug Administration as one example for creating innovative and efficient clinical trials [23]. The European Medicines Agency has recently published a Reflection Paper, which gives regulatory guidance on methodological issues in confirmatory clinical trials planned with an adaptive design [24]. Continual discussions between regulatory agencies, the pharmaceutical industry and academia have helped to foster a better mutual understanding of the issues and opportunities related to adaptive designs [25, 26]. In view of these ongoing discussions and activities, it is the aim of this paper to review the current methodology for confirmatory adaptive trial designs controlling strictly the overall type I error rate. Out of scope for this review are adaptive designs applied in early drug development, which often take place under different constraints and use different methodologies, such as Bayesian adaptive designs [27, 28] or adaptive dose ranging studies [29]. An overview of opportunities for adaptivity in the entire drug discovery and development process is described in [30].

Accordingly, the paper is organized as follows. In Section 2, we introduce a case study, which will be used in the subsequent sections to illustrate the methods. In Section 3, we describe the core methodology for the analysis of confirmatory adaptive designs. This section includes an overview of the adaptive design methodology for a single null hypothesis and how to perform adaptive designs with multiple hypotheses. In Section 4, we discuss important design considerations, such as power calculation, efficient interim decision rules, and sample size reestimation considerations. In Section 5, we present the results of an extensive simulation study to evaluate the operational characteristics of the various methods. In Section 6, we review the current methods for calculating point estimates and confidence intervals for relevant parameters. Section 7 is devoted to practical considerations when implementing confirmatory adaptive designs. Concluding remarks are given in Section 8.

2. A CASE STUDY

This case study refers to the late development phase of a drug for the indication of generalized anxiety disorder. The primary objective is (i) to select the most promising dose levels out of three different dose levels under investigation and (ii) to demonstrate subsequently that the selected dose levels lead to a statistically significant and clinically relevant efficacy as compared with placebo. The primary endpoint of this study is the change from baseline at week 8 of treatment in the total score on the Hamilton Rating Scale for Anxiety (HAM-A, [31]). This psychiatric scale was developed to quantify the anxiety symptomatology. It consists of 14 items, each defined by a series of symptoms. Each item is rated on a 5-point scale, ranging from 0 (not present) to 4 (severe). It is reasonable to assume that the total HAM-A score is normally distributed. Furthermore, it is assumed based on the outcome of previous studies that the common standard deviation across the

DOI: 10.1002/sim

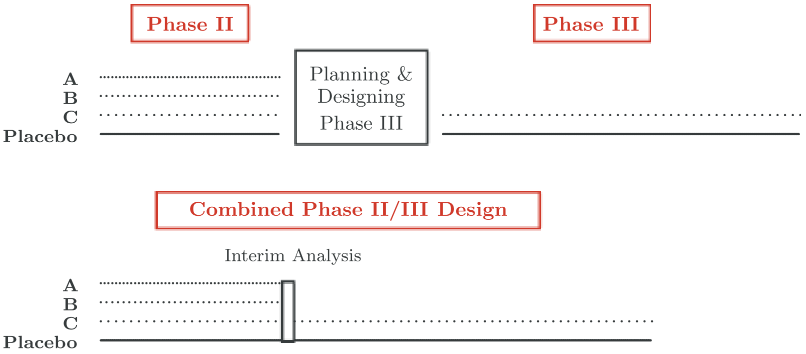

Figure 1. Top: classical development with two sperate studies. Bottom: combined phase II/III design with treatment selection at interim.

dose groups is 6 points of the HAM-A scale. The sample size of the study should ensure that the power for the individual pairwise comparisons is at least $1-\beta=0.8$, assuming a clinically relevant benefit of 2 points over placebo and an overall significance level of $\alpha=0.025$ (assuming one-sided null hypotheses).

The clinical team was requested to investigate the potential application of a two-stage adaptive design with dose selection at interim as compared with a traditional development program with two independent studies. In the following we describe these two development options in more detail.

Following the classical drug development paradigm, two separate studies are required. The three dose levels are first compared with placebo in a phase II study. Based on the results from that study, it is decided, whether to continue the drug development and which dose levels to carry forward to phase III (Figure 1, top). If the decision is to carry out the phase III study, previously available information is typically only taken into account when discussing the parameter estimates necessary for the sample size determination. Other than that, the phase III study is run independently from the phase II trial. In particular, the analysis of the confirmatory phase III study does not use the data from phase II.

In a two-stage confirmatory adaptive design one or more dose levels are selected at the interim analysis (after the first stage) and carried forward to the second stage together with placebo (Figure 1, bottom). The final comparisons of the selected dose levels with placebo include the patients of both stages and are performed such that the overall type I error rate is controlled at a pre-specified significance level α. The decision rules for the interim analysis should be set up in such a way that dose levels unlikely to show a clinically relevant difference to placebo are stopped early for futility. The expectation is that the combined phase II/III design leads to a reduction of the intermediate decision time ('white space') and increases the information value by combining the evidence across different development phases. Such designs are by construction *inferentially seamless*, as opposed to *operationally seamless* designs [32], where data from from different phases or stages are not combined. Note that in the present case study the same patient population and the same primary endpoint are investigated in both phases. This ensures that the stages are as homogeneous as possible, which is an essential requirement of combined phase II/III designs.

In the following we will use this case study to illustrate the design and analysis methods for adaptive designs in confirmatory studies as compared to a traditional development program.

DOI: 10.1002/sim

The numerical calculations are based on a set of SAS/IML macros, which are available at *www.meduniwien.ac.at/user/franz.koenig*. This web site also contains a script file, which includes the calls for the simulations conducted in Section 4.

3. ANALYSIS OF CONFIRMATORY ADAPTIVE DESIGNS

In this section, we describe the core methodology for the analysis of confirmatory adaptive designs. In Section 3.1, we explain the underlying conditional invariance principle. In Section 3.2, we describe adaptive tests for a single null hypothesis, reviewing both the combination test and the conditional error principle. These methods are extended in Section 3.3 to test adaptively multiple hypotheses based on the closure principle. The resulting test procedures are illustrated in Section 3.4 with a numerical example. For the sake of simplicity we consider only two-stage designs with a single interim analysis. The generalization of the subsequent methods and results to more than two stages is mostly straightforward.

3.1. Conditional invariance principle

Adaptive designs follow a common principle called *conditional invariance principle* [5]. Assume a trial with two sequential stages, where design characteristics of the second stage are chosen based on the data from the first stage as well as external information. We consider here the behavior of the trial under a specific elementary null hypothesis H. Let T_2 denote the test statistic for H applied to the second stage data. Due to the data-driven choice of the design characteristics, T_2 will in general depend on the interim data. However, we often can transform T_2 in such a way that the *conditional* null distribution of T_2 given the interim data and the second stage design equals a fixed pre-specified null distribution, and hence is *invariant* with respect to the interim data and mid-trial design adaptations. An invariant conditional distribution is typically achieved by transforming T_2 to a p-value p_2, which is uniformly distributed under H (conditionally on the interim data and the second stage design). Usually, the invariance of the conditional null distribution of p_2 implies that p_2 is stochastically independent of the first stage data. Since the joint distribution of the interim data and p_2 is known and invariant with respect to the unknown mid-trial adaptation rule, we can specify an α-level rejection region in terms of the interim data and p_2. This gives a test that controls the overall type I error rate at level α independently of the interim adaptation. The current most rigorous discussion of this can be found in [33].

3.2. Adaptive tests for a single null hypothesis

3.2.1. Combination test approach. Consider a null hypothesis H which is tested in two stages. Let p_1 and p_2 denote the stagewise p-values for H, such that p_1 is based only on the first stage and p_2 only on the second stage data. A two-stage combination test [10, 11] is defined by a combination function $C(p_1, p_2)$ which is monotonically increasing in both arguments, some boundaries α_1 and α_0 for early stopping as well as a critical value c for the final analysis. The trial is stopped in the interim analysis if $p_1 \leqslant \alpha_1$ (early rejection of the null hypothesis) or $p_1 > \alpha_0$ (early acceptance due to futility). If the trial proceeds to the second stage, i.e. $\alpha_1 < p_1 \leqslant \alpha_0$, the null hypothesis H is rejected at the final analysis if $C(p_1, p_2) \leqslant c$, where c solves

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{[C(x,y) \leqslant c]}\, \mathrm{d}y\, \mathrm{d}x = \alpha \tag{1}$$

Here, the indicator function $\mathbf{1}_{[\cdot]}$ equals 1 if $C(p_1, p_2) \leqslant c$ and 0 otherwise. By definition of c, the combination test $C(p_1, p_2)$ is an α-level test. This remains true if the design of the second stage (for example, the sample size or the test statistic) is based on the interim data. The only requirement is that under H the distribution of the second stage p-value p_2 conditioned on p_1 is stochastically larger than or equal to the uniform distribution [15]. This is the case, for example, when new patients are recruited in the second stage and conservative tests are used at each stage.

We define the decision function of a combination test through

$$\varphi_C(p_1, p_2) = \begin{cases} 1 & \text{if } p_1 \leqslant \alpha_1 \text{ or both } p_1 \leqslant \alpha_0 \text{ and } C(p_1, p_2) \leqslant c \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Thus, $\varphi_C = 1$ ($\varphi_C = 0$) corresponds to the rejection (non-rejection) of H, respectively.

At the planning stage of a clinical trial one has to specify the combination function and the design of the first stage, including the sample size and the test statistic for the first stage. The second stage design does not have to be specified in advance, but it must be specified latest at the interim analysis. Note that this allows in principle to use different test statistics for the two stages, for example, analyzing the first stage data with a non-parametric test and the second stage data with a parametric test. In Section 7, we give recommendations about the amount of pre-specification required in a study protocol.

Well-known examples of combination functions include Fisher's product criterion $C(p_1, p_2) = p_1 p_2$ [10, 11] and the weighted inverse normal combination function [34]

$$C(p_1, p_2) = 1 - \Phi[w_1 \Phi^{-1}(1-p_1) + w_2 \Phi^{-1}(1-p_2)] \tag{3}$$

where w_1 and w_2 denote pre-specified weights such that $w_1^2 + w_2^2 = 1$ and Φ denotes the cumulative distribution function of the standard normal distribution. Cui *et al.* [35] proposed a method that in effect is equivalent to the inverse normal method (3), although the latter is more general. Note that for the one-sided test of the mean of normally distributed observations with known variance, the inverse normal combination test with pre-planned stagewise sample sizes n_1, n_2 and weights $w_1^2 = n_1/(n_1+n_2)$, $w_2^2 = n_2/(n_1+n_2)$ is equivalent to a classical two-stage group sequential test if no adaptations are performed (the term in squared brackets is simply the standardized total mean). Thus, the quantities α_1, α_0 and c required for the inverse normal method can be computed with standard software for group sequential trials [36].

3.2.2. Binding versus non-binding stopping rules for futility. As seen from (1), the critical value c depends on the boundaries α_1 and α_0. For a fixed value of α_1, decreasing values of α_0 result in larger values of c. This has two contradictory implications on the power. On the one hand, it will cause a decrease in power because decreasing α_0 increases the probability of stopping for futility after the first stage. On the other hand, with a larger c one has a higher chance to reject after the second stage. It depends on the distribution of the p-value, which of these two implications has a larger overall effect on the power. In general one can argue that decreasing values of α_0 indeed leads to larger power for sufficiently powered studies. However, choosing $\alpha_0 < 1$ implies a *binding* futility boundary: If the observed first stage p-value p_1 is larger than the futility threshold α_0, the null hypothesis H must be retained, irrespective of the second stage results. Consequently, $p_1 > \alpha_0$ implies in practice a futility stop of a study at the interim analysis. Ignoring the binding futility rule by continuing the study with the possibility to reject H in the final analysis may inflate the overall type I rate. Alternatively, setting $\alpha_0 = 1$ guarantees the full flexibility to continue the study whenever it is appropriate. Such flexibility is often important in order to react quickly to unpredicted results or trends. It should be noted that even if $\alpha_0 = 1$, the study can be stopped at any time for futility without inflating the overall type I error rate, implying that the null hypothesis H is not rejected. This is sometimes called *non-binding* stopping for futility. Note that designs with binding stopping rules have larger power than those with non-binding stopping rules at the cost of reduced flexibility. Ultimately, the selection of an adequate threshold value α_0 is application specific and needs to be decided on a case-by-case basis, possibly taking simulations into account.

3.2.3. Conditional error function for adaptive combination tests. An alternative test procedure to the combination function approach is based the conditional error function described below [16]. Let X_1 denote the first stage sample and let $A(X_1)$ denote a measurable function from the first stage sample space to the unit interval [0,1] such that

$$E_H(A) \leqslant \alpha \tag{4}$$

The function A is referred to as the *conditional error function*. The sample size and the test statistic for the second stage can be modified based on the interim data, thus resulting in a second stage p-value p_2 (based on the data of the second stage, only). Finally, H is rejected if

$$p_2 \leqslant A(X_1) \tag{5}$$

Note that if $A = 0$ (early acceptance) or $A = 1$ (early rejection), no second stage needs to be performed for the test decision. This procedure controls the overall type I error rate at level α as long as the conditional distribution of the second stage p-value p_2, given the first stage data, is stochastically larger or equal to the uniform distribution under H.

One can also define the conditional error function $A(X_1)$ via a (single- or multi-stage) test φ [17]. As before, let $\varphi = 1$ ($\varphi = 0$) denote the rejection (non-rejection) of the null hypothesis H, respectively. Then the corresponding conditional error function of φ conditioning on the first stage data X_1 is given by

$$A(X_1) = E_H(\varphi | X_1)$$

Note that with this choice of the conditional error function the original test is applied, if no adaptations are performed. At the interim analysis, one thus has the option to complete the trial as initially planned, or to choose any other test for H (with new observations) at the level of the conditional error function. If adaptations are performed, the null hypothesis H is rejected based on the second stage p-value p_2 whenever (5) is satisfied. It follows that the conditional error function is the conditional probability of rejecting H under the assumption that H is true, given the first stage p-value p_1. We refer to [37] for a graphical illustration of the conditional error function in the plane for the p-values of both stages. Note that the conditional error rate may depend on nuisance parameters, if present. In some situations the conditional error can then be approximated, such as, for example, in the t test setting [38].

The conditional error function corresponding to any combination test defined through (2) is given by

$$A(p_1)=\begin{cases}1, & p_1\leqslant\alpha_1\\ \max\limits_{x\in[0,1]}\{x\cdot\mathbf{1}_{[C(p_1,x)\leqslant c]}\}, & \alpha_1<p_1\leqslant\alpha_0\\ 0, & p_1>\alpha_0\end{cases}$$

Note that due to (1) the level condition (4) is always satisfied. Obviously, φ_C rejects H if and only if $p_2\leqslant A(p_1)$. Consider as an example Fisher's product combination test $C(p_1,p_2)=p_1p_2$. The corresponding conditional error function is then given by

$$A(p_1)=\begin{cases}1, & p_1\leqslant\alpha_1\\ c/p_1, & \alpha_1<p_1\leqslant\alpha_0\\ 0, & p_1>\alpha_0\end{cases}$$

Similarly, the conditional error function for the weighted inverse normal method defined in (3) is given by

$$A(p_1)=\begin{cases}1, & p_1\leqslant\alpha_1\\ 1-\Phi\left(\dfrac{\Phi^{-1}(1-c)-w_1\Phi^{-1}(1-p_1)}{w_2}\right), & \alpha_1<p_1\leqslant\alpha_0\\ 0, & p_1>\alpha_0\end{cases}$$

3.3. Multiple testing in confirmatory adaptive designs

We now consider the case of testing k elementary null hypotheses $H_1,\ldots,H_k$. In Section 3.3.1, we review the closure principle to the extent required in Section 3.3.2, where we describe the methodology for confirmatory adaptive designs with multiple hypotheses. To illustrate the ideas, we assume the comparison of k treatments with a control. Accordingly, we denote by $H_i:\theta_i\leqslant\theta_0, i\in\mathscr{T}_1=\{1,\ldots,k\}$, the related k one-sided null hypotheses of comparing treatment i with the control 0, where θ_i denotes the mean effect of treatment i. It should be noted that the methodology described in the following is more general and covers many other applications. We refer to [20, 39–41] for clinical examples addressing other types of design modifications than treatment selection or sample size reassessment.

3.3.1. The closure principle. Performing an α-level test for each of k hypotheses H_i may lead to a substantial inflation of the overall type I error rate. That is, the probability to reject at least one true null hypothesis may be larger than the pre-specified significance level α. However, it is mandatory for confirmatory clinical trials that the probability to declare at least one ineffective treatment as effective is bounded by α. Hence the need of multiple test procedures, which strongly control the familywise error rate (FWER) at level α, where strong FWER is defined as the maximum probability of rejecting at least one true null hypothesis irrespective of the configuration of true and false null hypotheses [42].

The closure principle [43] is a general methodology to construct multiple test procedures controlling the FWER in the strong sense. The closure principle considers all intersection

DOI: 10.1002/sim

hypotheses that are constructed from the initial set of elementary null hypotheses. To control the FWER, an elementary null hypothesis H_i can only be rejected if all intersection hypotheses implying H_i are rejected, too. Then the resulting closed test procedure is operationally defined as follows:

1. Let $H_i, i \in \mathcal{T}_1 = \{1, \dots, k\}$, be the set of elementary null hypotheses.
2. Construct all intersection hypotheses $H_{\mathcal{S}} = \bigcap_{i \in \mathcal{S}} H_i, \mathcal{S} \subseteq \mathcal{T}_1$.
3. Define suitable α-level tests $\varphi_{\mathcal{S}}$ for each intersection hypothesis $H_{\mathcal{S}}, \mathcal{S} \subseteq \mathcal{T}_1$.
4. Reject an elementary null hypothesis H_i, $i \in \mathcal{T}_1$, if all intersection hypotheses $H_{\mathcal{S}}$ with $i \in \mathcal{S} \subseteq \mathcal{T}_1$ are rejected by their α-level tests $\varphi_{\mathcal{S}}$.

It is shown that the above procedure controls the FWER strongly at level α [43]. Note that we can specify any α-level test $\varphi_{\mathcal{S}}$ for $H_{\mathcal{S}}$ and in particular different tests can be used for different hypotheses. This property will be exploited when constructing adaptive tests based on the closure principle in Section 3.3.2.

Consider as an example the case of comparing $k = 2$ active treatments with a control. It follows from the closure principle that $H_i, i = 1, 2$, is rejected at level α if and only if both the test for the intersection hypothesis $H_1 \cap H_2$ and the test for the elementary hypothesis H_i are significant. For further examples illustrating the closure principle to the extent required in the context of adaptive designs, we refer to [14].

3.3.2. Multiple testing in confirmatory adaptive designs. In this section, we describe how to test adaptively multiple hypotheses by combining the methodology for adaptive tests (Section 3.2) with the closure principle (Section 3.3.1) [44–46]. We describe the methodology for the case that hypotheses are dropped at the interim analysis (for example, because treatments are dropped at interim for futility reasons and the related treatment-control comparisons become irrelevant). Extensions to other situations are mostly straightforward [46].

According to the closure principle we define for each intersection hypothesis $H_{\mathcal{S}}, \mathcal{S} \subseteq \mathcal{T}_1$, an adaptive combination test $\varphi_{\mathcal{S}}$ with combination function $C(p_1, p_2)$, stopping boundaries α_1, α_0 and critical value c. Note that different combination functions as well as different stopping boundaries can be used for different intersection hypotheses. Let $p_{1,\mathcal{S}}$ and $p_{2,\mathcal{S}}$ denote the stagewise p-values for the intersection hypothesis $H_{\mathcal{S}}, \mathcal{S} \subseteq \mathcal{T}_1$. One key characteristic of adaptive closed test procedures is that the multiplicity adjustment for $H_{\mathcal{S}}$ is addressed within each stage separately and the information is combined only afterwards using the pre-specified combination function $C(p_1, p_2)$. To address the multiplicity within one stage, we can use any standard fixed-sample multiple test procedure, as reviewed below. This means that within a stage we can test $H_{\mathcal{S}}$ with any (pre-specified) single stage multiple test procedure and plug the associated stagewise multiplicity adjusted p-values $p_{1,\mathcal{S}}$ and $p_{2,\mathcal{S}}$ into the combination function $C(p_1, p_2)$.

Assume now that at the interim analysis some hypotheses are dropped. Let $\mathcal{T}_2 \subseteq \mathcal{T}_1$ denote the index set of those hypotheses that are tested in the final analysis after the second stage. In the example of dropping treatments at interim, $\mathcal{T}_2$ would denote the index set of the remaining treatment-control hypotheses to be tested at the end of the study. In this example, no data of the discontinued treatment arms would be available for the second stage. When testing $H_{\mathcal{S}}$ for those intersections $\mathcal{S}$ that include treatments that were dropped at interim, the second stage multiplicity adjusted p-value $p_{2,\mathcal{S}}$ accounts only for the active treatments continued to the second stage. That is, for all hypotheses $H_{\mathcal{S}}, \mathcal{S} \subseteq \mathcal{T}_2$, the associated p-value $p_{2,\mathcal{S}}$ based on the second stage data

DOI: 10.1002/sim

X_2 can be calculated using any suitable multiple test procedure. For all other $\mathscr{S}\subseteq\mathscr{T}_1$ we set

$$p_{2,\mathscr{S}}=p_{2,\mathscr{S}\cap\mathscr{T}_2} \tag{6}$$

where $p_{2,\emptyset}=1$. For each $\mathscr{S}\subseteq\mathscr{T}_1$ the intersection hypothesis $H_{\mathscr{S}}$ is then rejected if $\varphi_C(p_{1,\mathscr{S}}, p_{2,\mathscr{S}})=1$. An individual hypothesis $i\in\mathscr{T}_1$ is rejected if all $H_{\mathscr{S}}$, $\mathscr{S}\subseteq\mathscr{T}_1$, with $i\in\mathscr{S}$, are rejected by their combination tests. We will illustrate this with an example in Section 3.4.

As mentioned before, the closure principle allows the use of different tests for different intersection hypotheses. Let s denote the number of treatments in $\mathscr{S}\subseteq\mathscr{T}_1$. Let further $p_{j,i}$ denote the individual stagewise p-value for stage $j=1,2$ and hypothesis H_i and let $p_{j,(i)}$ denote the ordered p-values within stage j. If the common Bonferroni procedure [42] is used, the stagewise adjusted p-value for $H_{\mathscr{S}}$, $\mathscr{S}\subseteq\mathscr{T}_j$, at stage $j=1,2$ is $p_{j,\mathscr{S}}=\min\{1, s\min_{i\in\mathscr{S}} p_{j,i}\}$. Similarly, one can use the Šidak test [42] with $p_{j,\mathscr{S}}=1-[1-\min_{i\in\mathscr{S}} p_{j,i}]^s$, or the Simes test [47] with $p_{j,\mathscr{S}}= s\min_{i\in\mathscr{S}} p_{j,(i)}/i$. If normality can be assumed, the Dunnett test [48] or other parametric tests can alternatively be used. Note that for the second stage the adjusted p-values for $H_{\mathscr{S}}$, $\mathscr{S}\subseteq\mathscr{T}_2$, are defined as above. For all other $\mathscr{S}\subseteq\mathscr{T}_1$ (such that $\mathscr{S}\nsubseteq\mathscr{T}_2$) the p-values for the intersection hypotheses are given in (6). Finally we note that closed test procedures based on the conditional error function approach are also available, see [49] for adaptive Dunnett tests with treatment selection at interim.

3.4. Numerical example

We now illustrate some of the methods from the previous sections with a numerical example. Motivated by the case study from Section 2 we assume a two-stage design with dose selection at interim. We assume that the outcomes are approximately normal with a common standard deviation of $\sigma=6$ points. Suppose that the planned total group sample size is $n=142$ and that the interim analysis is made halfway after $n_1=71$ observations per group. We use an adaptive combination test with the inverse normal combination (3) and weights proportional to the stagewise sample sizes, i.e. $w_1=w_2=\sqrt{1/2}$. In order to define reasonable early stopping boundaries, we use an O'Brien–Fleming design [2] with $\alpha_1=0.0054$, $\alpha_0=0.1$ and $c=0.0359$ for $\alpha=0.025$. The quantities α_1 and c can be computed with standard software packages for given values of α and α_0 [36]. Note that the small value of α_0 indicates a rather aggressive *binding* futility stopping rule and is used here only for illustration purposes. Finally, we assume that the study involves the comparison of three active dose groups with placebo, resulting in $k=3$ null hypotheses $H_i:\theta_i\leqslant\theta_0, i\in\mathscr{T}_1=\{1,2,3\}$. Let $\bar{X}_{j,i}$ denote the observed mean values in dose group $i=0,\ldots,3$, at stage $j=1,2$.

Assume that after the first stage we have observed the mean values $\bar{X}_{1,i}$ shown in Table I, which also summarizes the resulting standardized z-statistics and unadjusted p-values $p_{1,i}$. If we would not correct for multiplicity, we could already reject the null hypothesis H_3 at the interim analysis, since $p_{1,3}=0.0049<0.0054=\alpha_1$. But since this study is supposed to be confirmatory, we apply the closure principle and use the Bonferroni test for each intersection hypothesis, see Figure 2. The Bonferroni adjusted p-value for the global null hypothesis $H_{\{1,2,3\}}$ at the first stage is $p_{1,\{1,2,3\}}=3\min(p_{1,1}, p_{1,2}, p_{1,3})=3\times 0.0049=0.0147$. Since $\alpha_1<p_{1,\{1,2,3\}}$ it follows from the closure principle that we cannot reject any of the three elementary null hypothesis at interim. Moreover, for the global intersection hypothesis $H_{\{1,2,3\}}$ we have $p_{1,\{1,2,3\}}<\alpha_0$, the binding futility threshold is not crossed and we can proceed to the second stage. However, we have to consider also the other intersection hypotheses $H_{\{1,2\}}$, $H_{\{1,3\}}$ and $H_{\{2,3\}}$ with Bonferroni adjusted p-values $p_{1,\{1,2\}}=0.1364$, $p_{1,\{1,3\}}=0.0098$ and $p_{1,\{2,3\}}=0.0098$. Since $p_{1,\{1,2\}}>\alpha_0$, the binding futility

DOI: 10.1002/sim

Table I. Summary statistics from the first stage of an adaptive design with dose selection at interim.

i	$\bar{X}_{1,i}$	z-Statistic	$p_{1,i}$
0	0	—	—
1	0.8	0.794	0.2135
2	1.5	1.490	0.0682
3	2.6	2.582	0.0049

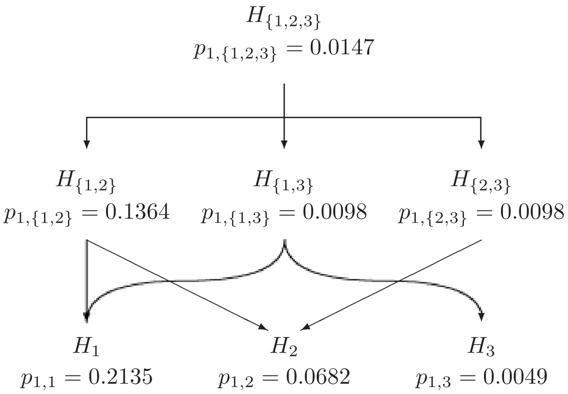

Figure 2. Closed test procedure for the first stage data. Bonferroni adjusted p-values are reported for each intersection hypothesis.

threshold is crossed and we have to accept H_{12}, and consequently the elementary hypotheses H_1 and H_2 as well, irrespective of the second stage results. In practice, this would imply a discontinuation of the dose groups $i=1,2$ and only the high dose group $i=3$ and placebo $i=0$ would be continued in the second stage. Note that even if the binding futility threshold had not been crossed, the adaptive design methodology employed here would allow the sponsor to discontinue any dose group (e.g. because of unexpected safety results).

To finalize the numerical example, assume the second stage means $\bar{X}_{2,0}=0$ and $\bar{X}_{2,3}=1.9$, resulting in the p-value $p_{2,3}=0.0296$. Since all other treatments have been dropped at interim, we plug $p_{2,3}$ as the second stage p-value into the adaptive combination tests for $H_{\{1,2,3\}}$, $H_{\{1,3\}}$, $H_{\{2,3\}}$ and H_3 (all other intersection hypotheses cannot be rejected anyway because of the interim results). Since $C(p_{1,\{1,2,3\}}, p_{2,3})<c$, $C(p_{1,\{1,3\}}, p_{2,3})<c$, $C(p_{1,\{2,3\}}, p_{2,3})<c$ and $C(p_{1,3}, p_{2,3})<c$, we finally can reject H_3 and conclude that dose $i=3$ is superior to placebo.

4. DESIGN CONSIDERATIONS

In this section we describe some important design considerations when planning a confirmatory adaptive design. We start with discussing different measures to establish the success of a trial in Section 4.1, resulting in different power concepts. In Section 4.2, we describe the concept of

DOI: 10.1002/sim

conditional power and briefly review the related problem of sample size reestimation. Finally, we discuss statistical considerations for interim decisions in Section 4.3.

4.1. Power concepts

An important decision to be made at the design stage of a clinical trial is to define a suitable metric to measure the success according to the study objectives, which is not always an easy task. Clinical trials employing an adaptive design to address multiple hypotheses typically intend to reach a decision about a number of null hypotheses H_i, $i=1,\ldots,k$, and sometimes intersections of these, such as the global null hypothesis $H_{\{1,\ldots,k\}}=\bigcap_{i=1}^{k} H_i$, for example. Power concepts to measure the success of a study are then associated with the probabilities of rejecting H_i, $i=1,\ldots,k$, if they are in fact wrong. Let K_i denote the alternative hypotheses associated with $H_i, i=1,\ldots,k$. The problem is that the individual events

$$H_i \text{ is rejected in favor of } K_i$$

can be combined in many different ways to obtain a measure of success, such as the probability to reject *at least one* false null hypothesis (disjunctive power) or the probability to reject *all* false null hypotheses (conjunctive power) [50].

To illustrate the issue of defining an appropriate measure of trial success, assume, for example, that the null hypotheses are given by $H_i:\theta_i\leqslant 0$, where θ_i denotes the mean effect of treatment i. The individual power to reject H_i at a given point $\tilde{\theta}_i$ in the parameter space is $p_{\tilde{\theta}}(\tilde{\theta}_i)=P_{\tilde{\theta}}(H_i \text{ is rejected}|\theta_i=\tilde{\theta}_i)$, where $\tilde{\theta}=(\tilde{\theta}_1,\ldots,\tilde{\theta}_k)$. Two appealing power concepts are then

$$P_{\tilde{\theta}}(\text{reject } H_i \text{ for at least one treatment with } \tilde{\theta}_i>0) \tag{7}$$

which is related to the disjunctive power introduced before, and

$$p\left(\max_{i=1,\ldots,k} \tilde{\theta}_i\right) \tag{8}$$

i.e. the probability of rejecting the treatment that is truly best in terms of the efficacy parameter. In adaptive designs involving treatment arm selection at an interim stage, there is a positive probability of discontinuing the treatment arm b with $\theta_b=\max_{i=1,\ldots,k}(\theta_i)$ at interim. If no early stopping for success is foreseen, this discontinuation is associated with a non-rejection of H_b. Of course, this does not preclude the possibility of a successful trial in the sense of (7). However, if the trial aim is to identify the best treatment, the interpretation of (7) as the probability of success is inappropriate and (8) should be preferred. If only one treatment is selected at interim and 'best treatment' is entirely determined by the parameters θ_i,

$$P(\text{treatment } b \text{ is selected and } H_b \text{ is rejected}) \tag{9}$$

is an appropriate probability to measure a trial success.

However, the reality of clinical trials is more complex than that. A typical situation is one where θ_i is an efficacy parameter, but the 'quality' of a treatment also depends on safety considerations (e.g. the frequency of adverse events associated with different treatment regimens or different doses). Such safety concerns are often hard to quantify and even if they are, it is not always clear how to combine safety and efficacy parameters into a single parameter θ_i. In such situations, the investigation of the operating characteristics of a planned clinical trial should include the

DOI: 10.1002/sim

calculation and/or simulation of multiple power concepts, including (9) as well as (7) and (8) in the context of adaptive designs for confirmatory trials. Ultimately, the totality of information needs to be discussed with the clinical team to justify a reasonable sample size and to ensure a successful trial according to the trial objectives.

4.2. Conditional power and sample size reestimation

In the interim analysis it is tempting to determine the conditional power to reach a rejection in an on-going trial given the observed results. In multi-armed clinical trials the precise meaning of the conditional power depends on the precise definition of the overall power. For example, the conditional rejection probability to reject an elementary null hypothesis H_i is given by $\mathrm{CRP}_i = P_{\tilde{\theta}_i}(H_i \text{ is rejected}|X_1)$, where X_1 denotes the first stage sample and $\tilde{\theta}_i$ denotes the true effect size.

For simplicity, the conditional power concept will be illustrated for the case of comparing one treatment versus control using a two-stage combination test $C(p_1, p_2)$. Hence the conditional power to reach a rejection in the final analysis given the first stage p-value p_1 is

$$CRP = \begin{cases} 1, & p_1 \leqslant \alpha_1 \\ \displaystyle\int_0^1 \mathbf{1}_{[C(p_1,y)\leqslant c]} dF_{p_2}^{\tilde{\theta}}(y), & \alpha_1 < p_1 \leqslant \alpha_0 \\ 0, & p_2 > \alpha_0 \end{cases} \tag{10}$$

where $F_{p_2}^{\tilde{\theta}}$ denotes the distribution of the second stage p-value p_2 which depends on the true effect size $\tilde{\theta}$. Note that for $\tilde{\theta}=0$ the conditional rejection probability in (10) reduces to the conditional error function in Section 3.2.3. Since the true effect size $\tilde{\theta}$ is unknown, the conditional power CRP may be calculated by using the effect size for which the study has been powered in the planning phase, or by using an interim estimate of the true size (or a combination of both). The estimates of the conditional power based on the interim effect size estimate are highly variable, depending on the first stage sample size [51]. Their use as a decision-making tool for possible adaptations should therefore always be supported by computations or simulations of the operating characteristics with respect to a success probability. Several authors have proposed to reassess the second stage sample size such that the conditional power is controlled at a pre-specified level. However, such sample size rules can be very inefficient and may lead to large average and maximal sample sizes unless appropriate early stopping rules and upper sample size limits are applied [52, 53]. Alternatively, prediction intervals conditional on the interim results can be computed [54] or Bayesian predictive power probabilities can be used [55, 56] to better quantify the uncertainty of the interim results. For a discussion of different sample size reassessment rules, we refer to [12] and the references therein for further details.

Numerical example (*continued*). We come back to the numerical example from Section 3.4 to illustrate the calculation of the conditional power, assuming that only the high dose group $i=3$ and placebo $i=0$ are continued to the second stage. To this end, we plug in the desired effect size from the planning phase ($\tilde{\theta}=2$, see Section 2) for the true unknown effect sizes in (10). In our specific example of selecting the best treatment and due to properties of the selected closed test procedure, we know that a rejection of the intersection hypothesis $H_{\{1,2,3\}}$ implies a rejection of all other relevant (intersection) hypotheses $H_{\{1,3\}}$, $H_{\{2,3\}}$ and H_3. Thus, it is sufficient to calculate the conditional power for $H_{\{1,2,3\}}$ only. Based on the interim data shown in Table I the conditional

DOI: 10.1002/sim

power is then given by

$$\begin{aligned}
CRP_{H_3} &= P_{\tilde{\theta}=2}(H_3 \text{ is rejected with the adaptive closed test}|X_1)\\
&= P_{\tilde{\theta}=2}(C(3p_{1,3}, p_{2,3})<c|X_1)\\
&= P_{\tilde{\theta}=2}(C(0.0147, p_{2,3})<c|X_1)\\
&= 1-\Phi\left(\frac{\Phi^{-1}(1-c)-w_1\Phi^{-1}(1-3p_{1,3})}{w_2}-\frac{\tilde{\theta}}{\sigma}\sqrt{\frac{n_2}{2}}\right)\\
&= 0.96
\end{aligned}$$

Since the conditional power is persuasive enough we retain the pre-planned second stage sample size $n_2=71$. Note that in general the closed test can imply a more complicated rejection region, in which case the calculation of the conditional power requires either multidimensional integration or simulations.

4.3. Statistical considerations for interim decisions

It is a difficult, if not unsolvable, problem to completely foresee at the design stage of a clinical trial the decision processes at an interim analysis. At this stage we often do not know, for example, how many and in which way treatments will be selected at interim, since other considerations than the observed efficacy results may influence the decision. For example, safety concerns may arise at interim and suggest to continue with the treatment having the smaller observed interim effect size. Thus, scientific expert knowledge not available at the planning phase and unknown random event processes in the background will in general influence the interim decision process. Nevertheless, it is essential to understand the operating characteristics of an adaptive design before the start of an actual trial. To evaluate them properly, one typically has to rely on simulations, which of course depend on the study design under investigation and on the (unknown) interim decision rules. In the following we propose a general framework that helps formalizing such simulations, so that the impact of different interim decision rules can be quantified and evaluated. The proposed framework consists of three steps and for the sake of illustration we use again the standard application of comparing several treatments with a common control.

First, a ranking scheme has to be implemented, which ranks the experimental treatment arms. Different ranking schemes are possible, the most natural being the observed (standardized) treatment-control mean differences at interim. Alternatively, one could compute the probabilities $P(\text{treatment } i \text{ has rank } j)$, where the one with largest probability of being best is assigned rank 1, from the remaining treatments, the one with largest probability of being best is assigned rank 2, and so on. Predictive probabilities of success can help quantify the success probabilities at the end of the study given the observed data at interim, see [19, 20] for applications in the context of confirmatory adaptive designs. More complex ranking schemes taking safety into account or assessing cost–benefit ratios are also possible. If the treatments correspond to increasing dose levels of an experimental drug, modeling approaches can be used to better predict the expected treatment means [57].

Second, it is proposed to define the number of retained treatments once their ranking is available. One possibility is to define a probability vector $r=(r_1,\ldots,r_k)$, where $0\leqslant r_i \leqslant 1$ denotes the probability of retaining i treatments and $\sum r_i = 1$. If, for example, $r_1 = 1$, exactly one treatment is selected at interim; if $r_k = 1$, all treatments are selected for the next stage. Note that r can also depend on the distance of the measures from the first step. For example, if two observed standardized differences are similar, it is likely to continue with both arms. In addition, clinical relevance thresholds can be included as necessary conditions for a treatment arm to be selected.

Third, after having *ranked* the treatments and decided on the *number* of treatments to be carried forward (through the r vector), one needs to formalize *which* treatments actually to select at interim. One possibility is to specify a probability vector $s = (s_1, \ldots, s_k)$, where $0 \leqslant s_i \leqslant 1$ denotes the probability of retaining the ith best treatment arm according to the selected ranking scheme. Consider, for example, a study with $k = 3$ experimental arms. If $r = (1, 0, 0)$ and $s = (1, 0, 0)$, we select the best ranked treatment at interim. Suppose otherwise that the clinical team is reluctant because of possible unexpected safety problems and they assume a probability 0.3 for this to happen. Thus, they may consider two arms to be retained at the interim analysis, i.e. $r = (0, 1, 0)$, and $s = (0.7, 0.2, 0.1)$. We would then have a probability 0.7 of picking the best performing study arm, 0.2 of picking the second best and 0.1 of picking the third best. Having picked the first arm to be retained we pick from the remaining two arms with their probabilities renormalized. If treatment 1 is picked as the first arm, s is renormalized to $(0, 0.67, 0.33)$.

5. A SIMULATION STUDY

In this section, we report the results of an extensive simulation to illustrate some of the methods described previously, motivated by the case study introduced in Section 2 and covering a wide range of practical scenarios. In Section 5.1, we describe the design of this simulation study, including its assumptions and scenarios as well as the performance metrics used to evaluate different statistical operational characteristics of the various methods. The results of the simulation study are summarized in Section 5.2. Because of the large number of scenarios and performance metrics, only a subset of the possible plots is included here to illustrate the key findings. We conclude the simulation study with a few remarks in Section 5.3.

5.1. Design of simulation study

Motivated by the case study from Section 2, consider the comparison of $k = 2$ and 3 treatments with a control in the homoscedastic normal model with known common variance σ^2. Let n denote the pre-specified total group sample size. For simplicity we choose n such that the individual treatment-control comparisons have a power of $1 - \beta = 0.80$ for a one-sided z-test with $\alpha = 0.025$. That is, we calculate $n = 2(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2 / \tilde{\theta}^2$, where z_c denotes the c-quantile of the standard normal distribution and $\tilde{\theta}$ is the treatment effect to be detected. For the simulation study we assume $\sigma = 6$ and $\tilde{\theta} = 2$, resulting in a total sample size of $n = 142$ per treatment group. We further assume a two-stage design with one interim analysis. No early efficacy testing in the interim analysis is foreseen. We do, however, investigate the impact of non-binding early futility stopping on power. The subsequent results are obtained by simulating 100 000 trials for each scenario using SAS/IML.

Based on the considerations in Section 4.1, we compute (i) the probability to reject correctly at least one of the hypotheses under investigation at the final analysis (disjunctive power) and (ii) the

DOI: 10.1002/sim

probability to reject correctly a specific elementary null hypothesis (individual power). We also considered other power measures, but the results are not reported here. We consider the following decision rules to be adopted in the interim analysis:

(I) Continue with all treatments in the second stage.
(II) Select the best treatment based on the observed first stage mean values.
(III) Select the best treatment only if the mean difference to control is above a threshold Δ.
(IV) Select all treatments where the mean difference to control is above Δ.

The threshold Δ reflects a (non-binding) early futility stopping rule. It may happen that none of the treatments achieves the threshold, in which case the study would be stopped for futility. In reality, we typically do not know the value for Δ since it is not a formalized threshold pre-specified at the design stage of a trial, as explained in Section 4.3. Nevertheless, one should investigate the impact of Δ via simulation to better understand the operating characteristics of the employed adaptive design.

The following test procedures are investigated in the simulation study:

(A) Adaptive combination test using Dunnett adjusted p-values for the intersection hypotheses at each stage and combining the stagewise p-values using the inverse normal method with weights $w_i = \sqrt{n_i/(n_1+n_2)}$ corresponding to the pre-planned sample size fractions. Note that the Dunnett test for the second stage reduces to the z-test if only one treatment is selected.
(B) The single stage Dunnett test with treatment selection at interim, which uses Dunnett critical boundaries accounting for the pre-specified number k of treatment-control comparisons. Note that dropping a treatment in a single stage design means that one accepts the corresponding null hypothesis. That is, the associated test statistic is set to $-\infty$.
(C) Separate phase II/III design, where the information of the first stage (phase II) is only used to address the question of how many and which treatments to investigate in the second stage (phase III). For the final analysis only the second stage data are used. We use the Dunnett test in the final analysis to adjust for the number of active treatments actually continued to the second stage. For example, if two active treatments are selected at interim, the final analysis uses the Dunnett test for two treatment-control comparisons. If only one active treatment is selected, the second stage data are tested using a z-test.

For methods (A) and (C) a sample size reassessment can be performed in the interim analysis without compromising the overall type I error rate. In the simulations we investigated a sample size reallocation strategy in case that treatments are dropped in the interim analysis. In such cases we reallocate the subjects originally assigned to the discontinued treatments at stage 2 evenly to the selected treatments (including the control group). It is thus ensured that the total number of observations in the trial is constant $N = kn$.

5.2. Simulation results

From the many different scenarios investigated in the simulation study (corresponding to different combinations of methods, performance metrics, interim decision rules, etc.) we decided to report in detail the results for four scenarios (corresponding to Figures 3–6) to illustrate the key findings. Let θ_i denote the mean of treatment group $i = 0, 1, 2, (3)$, where $i = 0$ denotes the control group.

DOI: 10.1002/sim

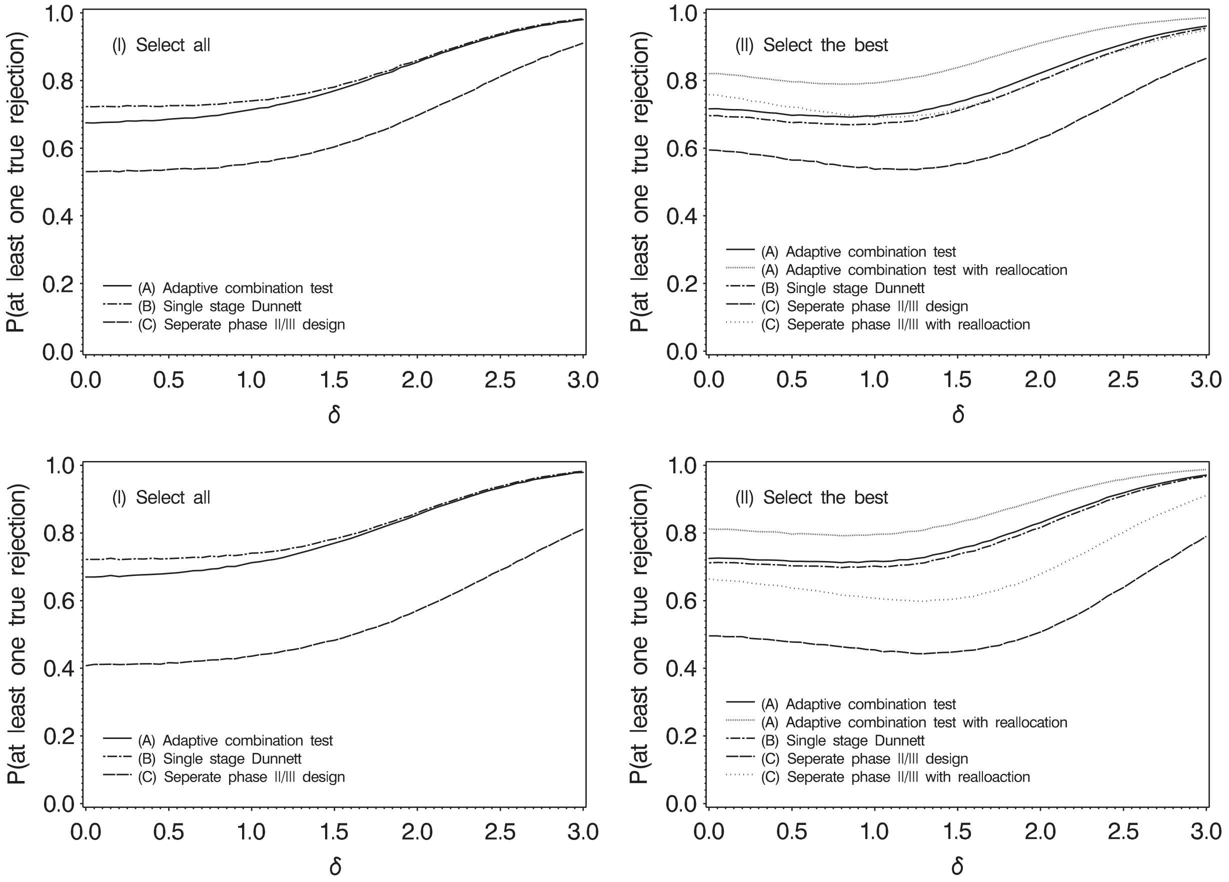

Figure 3. Disjunctive power for comparing $k=2$ treatments with a control using the information fractions $n_1/n=\frac{1}{3}$ (first row) and $n_1/n=\frac{1}{2}$ (second row). The power is the probability to reject correctly at least one of the false null hypotheses at the final analysis for the efficacy profile $(\theta_0,\theta_1,\theta_2)=(0,\delta,2)$ where $\delta\in[0,3]$ is plotted on the abscissa. Left column—decision rule (I): Continue with all treatments in the second stage. Right column—decision rule (II): select the treatment with the larger observed first stage mean value.

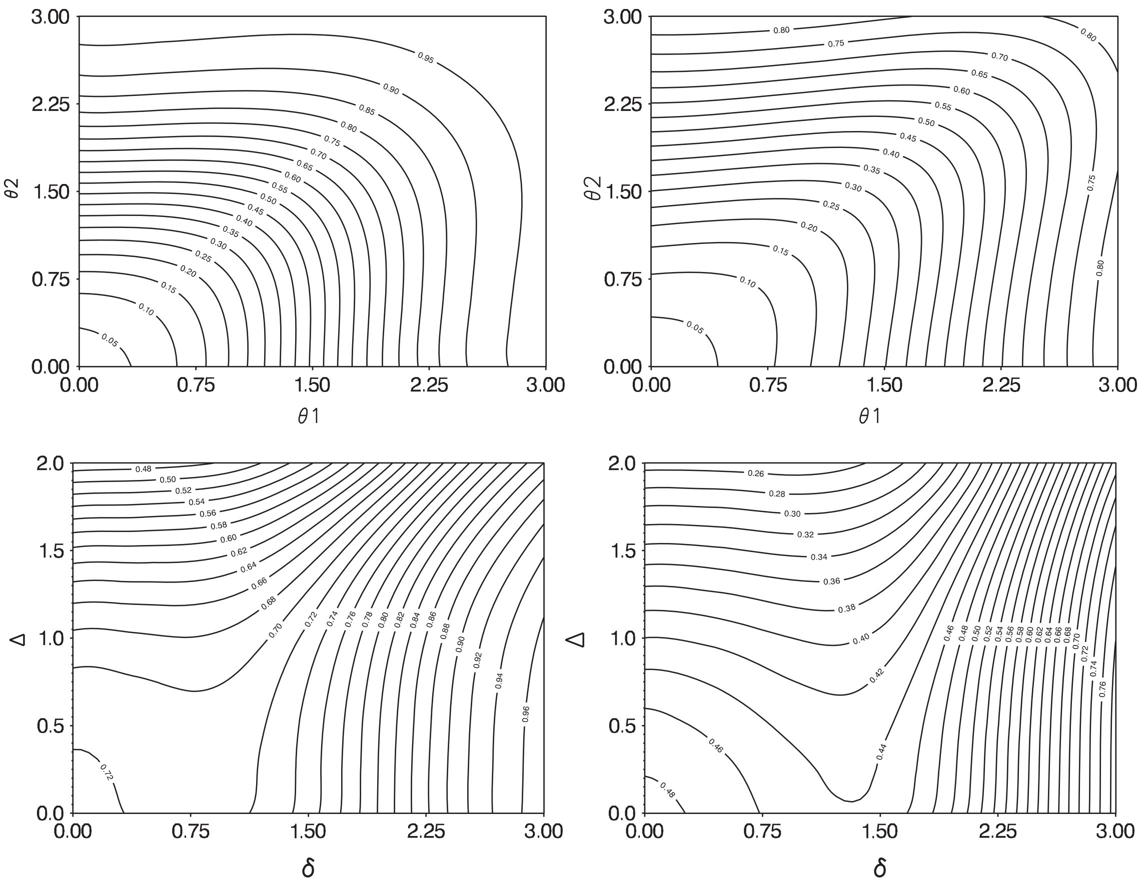

Figure 4. Disjunctive power for comparing $k=2$ treatments with a control. First row—decision rule (II) using the information fraction $n_1/n=\frac{1}{2}$ for different efficacy profiles. Second row—decision rule (III): Select the better treatment only if the interim mean difference to control is above a threshold Δ for the efficacy profile $(\theta_0, \theta_1, \theta_2)=(0, \delta, 2)$, where $\delta \in [0, 3]$. Left column—method (A): Adaptive combination test using Dunnett adjusted p-values. Right column—method (C): separate phase II/III design.

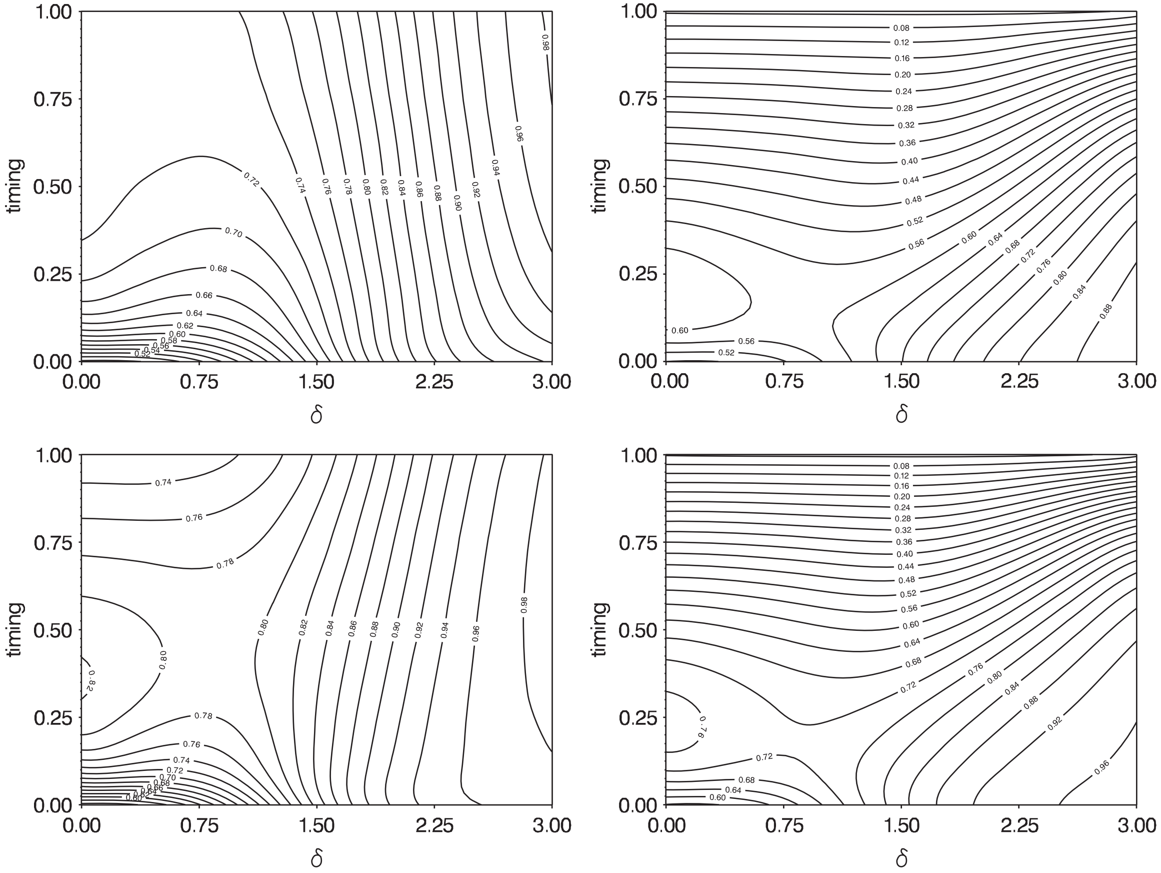

Figure 5. Disjunctive power for comparing $k=2$ treatments with a control using decision rule (II): select the better treatment based on the observed first stage mean values. The power is computed for the efficacy profile $(\theta_0, \theta_1, \theta_2) = (0, \delta, 2)$, where $\delta \in [0, 3]$ is plotted on the abscissa. The information fraction n_1/n is plotted on the y-axis. Left column—method (A): adaptive combination test using Dunnett adjusted p-values. Right column—method (C): separate phase II/III design. Both methods without (first row) and with sample size reallocation (second row).

DOI: 10.1002/sim

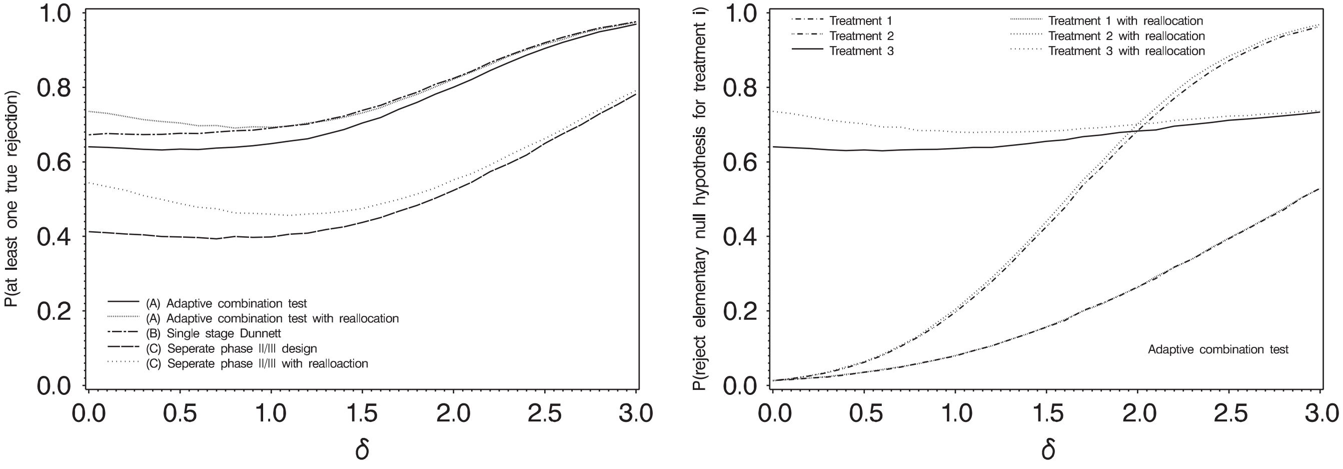

Figure 6. Power for comparing $k=3$ treatments with a control using the information fraction $n_1/n=\frac{1}{2}$ and decision rule (IV), where only those treatments are selected which have a positive treatment-control difference. In the left figure the disjunctive power is shown for methods (A), (B) and (C) and in the right figure the individual power for method (A). The power depends on the efficacy profile $(\theta_0, \theta_1, \theta_2, \theta_3)=(0, \delta/2, \delta, 2)$, where $\delta \in [0, 3]$ is plotted on the abscissa.

We first compare methods (A)–(C) with and without reallocation for decision rules (I) and (II), two different information fractions $n_1/n=\frac{1}{3},\frac{1}{2}$ for the interim analysis, and $(\theta_0,\theta_1,\theta_2)=(0,\delta,2)$, where $\delta\in[0,3]$ and $k=2$ (Figure 3). When selecting both treatments, the single stage procedure (B) has a higher power than the competing methods. The difference between methods (A) and (B) decreases for increasing values of δ. The slight reduction in power for method (A) in comparison with (B) for small values of δ might be acceptable due to the gain in flexibility. When selecting the most promising treatment, method (A) is more powerful than the single stage Dunnett test (B), since for method (A) no multiplicity adjustment for the second stage data has to be applied. In contrast, for method (B) we have to account for the pre-specified number of k treatment-control comparisons regardless of whether and how many treatments are dropped at the interim analysis. Comparing the timing of the interim analysis at $n_1/n=\frac{1}{3}$ (first row) with $n_1/n=\frac{1}{2}$ (second row) shows a large impact on the power for the separate phase II/III design (C), where only the data of the second stage are used for the final analysis. Note that even a sample size reallocation strategy can hardly compensate for the loss in power as compared with methods (A) and (B). Performing sample size reallocation with method (A) leads to a largely increased power, which is even larger than the power for method (B) when selecting all treatments and using the same total sample size. Because of the relatively small power differences between methods (A) and (B) and in order to illustrate here only the key findings, we omit method (B) in the subsequent summary of the simulations for $k=2$. Instead, we refer to [49, 58] for further results.

In Figure 4 we compare the disjunctive power (i.e. the probability to reject at least one false null hypothesis) of methods (A) and (C) for decision rules (II) and (III) and different efficacy profiles, $k=2$, where the interim analysis is performed after $n_1=n/2$ patients per treatment group. If decision rule (II) is applied (Figure 4; top row) we observe that the power increases symmetrically in both θ_1 and θ_2, as expected. Note that the power for the combination test is substantially larger than for the separate phase II/III design. In the bottom row of Figure 4 we plot the disjunctive power for methods (A) and (C) for different values of the non-binding futility threshold Δ when applying the decision rule (III). Increasing the value of Δ makes it more difficult to proceed to the second stage, thus leading to smaller power. If $\Delta=0$ we proceed to the second stage with the most promising treatment only if at least one of the observed treatment-control difference at interim is positive. Note that for small fixed value of Δ the power is not monotonous in θ_1: The power first decreases in θ_1 and then increases for large values of θ_1. The treatment selection probabilities (not reported here) explain this behavior. For values around $\theta_1=0$ the more efficient treatment $i=2$ (with $\theta_2=2$) is almost always selected in the interim analysis. With increasing values of θ_1, the probability of selecting treatment $i=1$ (which has a lower efficacy) increases. This reduces the disjunctive power of rejecting at least one null hypothesis for small values of θ_1. Note that this phenomenon can be observed in Figure 3.

In Figure 5 we investigate the impact of the timing for the interim analysis on the disjunctive power for $k=2$ when selecting the best treatment at interim (decision rule (II)). We start considering method (A) without sample size reallocation (top left graph). For a fixed value of θ_1 the power increases with increasing information fraction n_1/n. From a power perspective these results suggest to not perform an interim analysis at all. This is to be expected, since we select the better of the two treatments at interim and dropping a treatment arm leads to a reduction of the total sample size, thus leading to a smaller power. Note however that the decrease in power is negligible for $n_1/n>\frac{1}{2}$ while the total sample size is substantially decreased for smaller information fractions. When including sample size reallocation, however, the total sample size of the trial is constant $(=3n)$. Now an adaptive design leads to substantial power gain, as depicted in the bottom left

DOI: 10.1002/sim

graph of Figure 5. Consider a fixed value for θ_1, say, $\theta_1 = 1$, and compare the power values for different information fractions. One notes that for increasing values of n_1/n the power first increases and then decreases. For $\theta_1 = 1$ (and recalling that $\theta_2 = 2$ throughout), the maximum power is approximately 80 per cent, which is achieved at about $n_1/n = 0.4$. Note that the maximum power and the optimal interim time point depend on the (unknown) efficacy profile. For example, for $\theta_1 = 0$, the maximum power is approximately 82 per cent, which is achieved after observing about one-third of the patients. For the separate phase II/III design (C) (right column in Figure 5), we observe similarly that the power is not monotone in n_1/n for a fixed value of θ_1, irrespective of whether sample size reallocation is applied or not. Note however that for large fractions n_1/n the sample size of the second stage decreases to 0. Since the final analysis uses only the second stage data, the power thus quickly goes to 0. Finally we note that method (A) depends much less on the information fraction n_1/n than method (C), thus being more robust. This can be concluded from the almost vertical contour lines observed in the left column of Figure 5, as opposed to the almost horizontal contour lines in the right column.

In Figure 6 we summarize some results for the case of comparing $k = 3$ treatments with a control when using decision rule (IV) with $\Delta = 0$, selecting at interim those treatments for the second stage which have a positive treatment-control difference. Thus, either 0, 1, 2, or 3 treatments can be selected for the second stage, depending on the observed interim results. The left plot displays the disjunctive power for methods (A)–(C). As for the case $k = 2$ in Figure 3, both methods (A) and (B) lead to substantially larger power than method (C). Note, however, that here the impact of sample size reallocation is smaller since in most cases we select at least two of the active treatments. For example, if $\delta = 0$ and thus $\theta_1 = \theta_2 = 0$, the probability is 50 per cent to select treatment $i = 1, 2$ for the second stage. Since the probability to select treatment $i = 3$ ($\theta_3 = 2$) is almost 1, the resulting probability to select at least two treatments is 75 per cent in this case. So far we considered only the disjunctive power. Alternatively, one might be interested in the individual power as displayed in the right plot of Figure 6 for method (A). Clearly, the individual power curves follow the relative efficacy profiles. For example, the individual power for treatment $i = 1, 2$ ($\theta_1 = \delta/2, \theta_2 = \delta$) increases monotonically in $\delta \in [0, 3]$. Note that although $\theta_3 = 2$, the individual power for treatment 3 is not constant, but depends on δ: Applying the closed test procedure for method (A) induces dependencies among the decisions for the elementary hypotheses by testing the intersection hypotheses. Note also that sample size reallocation only leads to a noticeable power increase for treatment 3 and small values for δ, since the probability to select all three treatments for the second stage quickly goes to 1 for large values of δ. Similarly, treatment 1 never benefits from a sample size reallocation, since if treatment 1 is selected, the other two treatments are selected as well.

5.3. *Concluding remarks*

In the comparison of strategies (A), (B) and (C) we allow sample size reallocation for (A) and (C), but not for (B). The reason for doing so is that a strategy like (B) may result in a substantial inflation in size when allowing for other adaptations than considered here, including sample size reassessment. For a single treatment-control comparison, Proschan and Hunsberger [16] have shown that the maximum type I error rate can exceed 2α.

It is worthwhile pointing out that the simulation study is based on assumptions, which in practice are not always satisfied. Motivated by the case study from Section 2, we considered combined phase II/III studies that would be most applicable when the same patient population and the same

DOI: 10.1002/sim

primary endpoint are investigated in both phases. In practice, however, the distinction between phase II and phase III is unlikely to be just a sample size determination after phase II results are obtained. Often, phase II has narrower patient population likely to have a larger treatment effect size, uses surrogate endpoints instead of clinical endpoints that take much longer time to observe, in addition to selecting dose regimens. For a discussion of such issues and related simulations, we refer to [59].

Also, in current clinical practice substantially fewer patients are investigated in phase II than in phase III. As seen from the simulation study, combining the data from both phases leads to negligible power increase for small information fractions (say, $n_1/n \approx 0.1$ or less). One may argue that larger information fractions are not relevant in pharmaceutical practice and the comparisons involving method (C) thus become obsolete. But given a total sample size $N = kn$ for phase II and III, the simulations also suggest that there is an optimal trade-off between power gain and sample size savings somewhere around $n_1/n \approx 0.5$, depending on the effect sizes, the interim decision rules, whether sample size reallocation and/or early stopping (for futility or success) is foreseen, etc. In essence, from a power perspective it is thus advantageous to have the interim analysis not too early for given N and consequently the size of the first stage (phase II) relative to the second stage (phase III) should be larger than customary in current practice. Coupled with that, a more precise treatment or dose selection at interim is to be expected since a larger body of evidence is available when choosing a later time point for the interim analysis.

When it comes to the decision, whether or not to apply an adaptive design, it is important to ensure that the finally chosen design adequately addresses the primary study objectives. The conduct of extensive simulations, similar to the ones reported here, are then of critical importance to understand and compare the operating characteristics of different design options and interim decision strategies. In the end, the final decision depends on many considerations and not all of them are of statistical nature (see also Section 7). But well-planned and extensive clinical trial scenario evaluations are likely to support the interdisciplinary discussions. Substantial lead time for their conduct should be devoted at the planning stage.

6. ESTIMATION

An open question in adaptive designs is the construction of adequate point estimates and confidence intervals for the treatment effects of interest. These will be reviewed below with the understanding that research is still ongoing and the results may soon be outdated. This section is somewhat more technical than the previous ones because no established methods for point estimates and confidence intervals exist and some of the reviewed methods require a more involved notation. We refer to the original papers for simulation studies comparing the different methods.

6.1. Point estimates

Interim adaptations in clinical trials will have an impact on the estimates of treatment effects. This is obvious in the case of treatment arm selection based on the observed effect size at interim, but it is also true for other types of adaptations. Consider as an example a clinical trial that aims at inference on the treatment effects θ_i for which (asymptotically) normally distributed sample estimates $\bar{X}_{j,i}$ are available. To be more precise, consider

$$\bar{X}_{j,i} \sim \mathrm{N}(\theta_i, \sigma^2/n_j) \tag{11}$$

DOI: 10.1002/sim

where n_j denotes the (balanced) group samples sizes at stage j and n_j/σ^2 is the information on θ_i from stage $j=1,2$. In the context of the case study from Section 2, the $\bar{X}_{j,i}$ would simply denote the estimated means per treatment arm i and stage j and θ_i would denote the true mean of treatment i. Equation (11) gives the marginal distributions of the $\bar{X}_{j,i}$. These may be correlated (for example, if the parameters θ_i denote differences between several treatments and a common control). In addition, not all combinations of i and j are necessarily available if treatment arms are dropped or if the trial is stopped early.

Note that the estimates $\bar{X}_{j,i}$ are unbiased. Thus, if unbiasedness is the only issue, then for any given treatment i, both $\bar{X}_{1,i}$ and $\bar{X}_{2,i}$ are unbiased estimates of θ_i, irrespective of the decisions made at an interim analysis. Unfortunately, of course, the stagewise means $\bar{X}_{j,i}$ are not an efficient estimate of θ_i, if treatment i is investigated in both stages. In the following, we discuss the impact of different types of modifications on the estimation of treatment effects.

6.2. *Sample size reestimation*

For simplicity, we will discuss in this section the case of just one treatment (or a treatment versus control difference) whose expected value is denoted by θ_1. Assume a two-stage trial where, based on the results of the interim analysis, the originally intended second stage sample size n_2 is modified to $\tilde{n}_2$. Irrespective of the decision rule used to modify the sample size, the maximum likelihood estimate (MLE) of θ_1 is given by $\hat{\theta}_1=(n_1\bar{X}_{1,1}+\tilde{n}_2\bar{X}_{2,1})/(n_1+\tilde{n}_2)$ [60]. This estimate will only be unbiased (i.e. the expected value of $\hat{\theta}_1$ is θ_1) if either $\tilde{n}_2$ is identical n_2 (i.e. the originally planned second stage sample size remains always unmodified), or the sample size reassessment is made independent of $\bar{X}_{1,1}$. The latter, for example, is (approximately) the case if the sample size reassessment is based on the usual (pooled) variance estimate for the common variance σ^2 of the treatment group(s). If the sample size reestimation is based on $\bar{X}_{1,1}$ directly, $\hat{\theta}_1$ is typically biased. A bias is also observed if the reestimation depends on a measure that is correlated with $\bar{X}_{1,1}$. One practically relevant example of the latter is the case of blinded sample size review where the total variance ignoring group differences is used to reassess the sample size in a trial which compares a treatment with a control [61].

An upper bound for the bias of the MLE is given by

$$\frac{\sigma}{\sqrt{2\pi\cdot n_1}}\left(\frac{n_1}{n_1+n_{2,\min}}-\frac{n_1}{n_1+n_{2,\max}}\right)$$

where $n_{2,\min}\leqslant\tilde{n}_2\leqslant n_{2,\max}$ are pre-specified lower and upper bounds on the second stage sample size $\tilde{n}_2$ [60]. It is shown by simulations that the MLE has good properties with respect to its mean squared error (MSE). The variance and bias of the MLE are unknown for unspecified sample size reestimation rules, but simulations indicate that the bias is small relative to the variance.

We now consider two alternative estimates to the MLE. The *mean-unbiased estimate*

$$\hat{\theta}_{1u}=u\bar{X}_{1,1}+(1-u)\bar{X}_{2,1},\quad 0\leqslant u\leqslant 1$$

uses fixed weights u and $1-u$ for the stagewise means. For example, if the inverse normal combination function (3) is used for testing with weights w_1 and $w_2=\sqrt{1-w_1^2}$, it is natural to

DOI: 10.1002/sim

use $u=w_1^2$ for the mean-unbiased estimate. However, it has been shown that $\hat{\theta}_u$ does not behave well in terms of the MSE [60]. The *median-unbiased estimate* (MEUE) is given by

$$\hat{\theta}_{1m}=\frac{w_1\sqrt{n_1}\bar{X}_{1,1}+w_2\sqrt{\tilde{n}_2}\bar{X}_{2,1}}{w_1\sqrt{n_1}+w_2\sqrt{\tilde{n}_2}},\quad w_1^2+w_2^2=1,\quad w_1,w_2\geqslant 0 \tag{12}$$

Simulations indicate that this estimate seems to perform reasonably well with respect to the MSE [60]. Note that $\hat{\theta}_{1m}$ is a compromise that 'shrinks' $\hat{\theta}_{1u}$ toward $\hat{\theta}_1$. To see this, suppose that there is an 'intended' second stage sample size n_2 which is specified in the planning phase of the trial. We would then use the weight $u=n_1/(n_1+n_2)$ for the mean-unbiased estimate. For the MEUE, we would use $w_1=\sqrt{n_1/(n_1+n_2)}$, resulting in the weight

$$\frac{w_1\sqrt{n_1}}{w_1\sqrt{n_1}+w_2\sqrt{\tilde{n}_2}}=\frac{n_1}{n_1+\sqrt{n_2\tilde{n}_2}}$$

for $\hat{\theta}_{1m}$. Recalling that the weight for the MLE is given by $n_1/(n_1+\tilde{n}_2)$, we see that the relations

$$\frac{n_1}{n_1+\tilde{n}_2}<\frac{n_1}{n_1+\sqrt{n_2\tilde{n}_2}}<\frac{n_1}{n_1+n_2}\quad \text{if } \tilde{n}_2>n_2$$

and

$$\frac{n_1}{n_1+\tilde{n}_2}>\frac{n_1}{n_1+\sqrt{n_2\tilde{n}_2}}>\frac{n_1}{n_1+n_2}\quad \text{if } \tilde{n}_2<n_2$$

hold for the weights of $\hat{\theta}_1$, $\hat{\theta}_{1m}$, and $\hat{\theta}_{1u}$, respectively.

If $n_{2,\min}=0$, there is a positive probability that the trial is stopped after the first interim analysis. In this case the mean-unbiased estimate $\hat{\theta}_{1u}$ cannot be calculated if $\bar{X}_{2,1}$ is not available due to $\tilde{n}_2=0$. The median-unbiased estimate $\hat{\theta}_{1m}$ is formally defined for $\tilde{n}_2=0$ and equals $\bar{X}_{1,1}$ in this case, although it will no longer be 'median unbiased'. Note that, however, with only one additional observation the estimate $\hat{\theta}_{1m}$ would become median unbiased. Since the influence of one additional observation on $\hat{\theta}_{1m}$ is typically small (for reasonably large n_1), the estimate $\hat{\theta}_{1m}$ will not exhibit a large median bias even if $\tilde{n}_2=0$ is possible.

Several suggestions for bias-corrected estimates are available [62–64], which either attempt to correct the bias approximately or deliberately over-correct for it (i.e. deliberately underestimate the true treatment effect). If the stopping rules are made explicit and declared as binding in advance, some improvements like explicit analytical expressions for the bias of the MLE [65], and some improved estimators exploiting the stopping boundaries [60, 66] are available. However, these estimators are somewhat complicated to derive, have no closed-form representation and usually only address one requirement (e.g. median unbiasedness or truncation adaptable unbiasedness) with other properties unknown or unfavorable.

The inclusion of early stopping further complicates the derivation of adequate point estimates and no fully satisfactory solution seems to exist. We thus recommend to report the MLE or the MEUE, keeping in mind that these can be biased. Note that if stopping occurs after the first stage, the MLE and the MEUE coincide.

Numerical example (*continued*). To illustrate the proposed estimates we revisit the numerical example from Section 3.4. For simplicity, we consider only the largest dose level (treatment $i=3$ in Section 3.4), ignoring the two intermediate dose levels $i=1,2$. Recall that the observed first

Table II. Impact of different sample size reassessment strategies on various point estimates.

	Pre-planned sample size	Sample size increase	Sample size decrease
	$n_2=71$	$\tilde{n}_2=2n_2=142$	$\tilde{n}_2=35$
Maximum likelihood ($\hat{\theta}_1$)	2.25	2.13	2.37
Mean-unbiased ($\hat{\theta}_{1u}$)	2.25	2.25	2.25
Median-unbiased ($\hat{\theta}_{1m}$)	2.25	2.19	2.31

stage mean value is 2.6 and the second stage mean is 1.9. To illustrate the impact of different sample size reassessment strategies on the various point estimates, we consider the pre-planned sample size for the second stage ($n_2=71$), an increased second stage sample size $\tilde{n}_2=2n_2=142$ and a decreased second stage sample size $\tilde{n}_2=35$. The weights for the mean-unbiased estimate are given by the pre-planned stagewise sample sizes with $u=n_1/(n_1+n_2)$. The results are shown in Table II for three different point estimates. If no sample size reassessment is performed, all three estimates are identical. If the sample size is changed at the interim analysis, the median unbiased estimate is a compromise between the two others, as discussed above.

6.3. *Treatment arm selection*

In multi-armed clinical trials, where treatments can be dropped at an interim analysis based on the unblinded interim data, the second stage sample size $\tilde{n}_{i2}$ of a treatment arm i depends on the interim data. For instance, dropping treatment arm i implies $\tilde{n}_{i2}=0$. It is also common to reassess sample sizes for the selected treatments (e.g. to reallocate the pre-planned sample sizes of the dropped treatment arms to the selected treatment groups). Hence, estimates in trials with treatment selection are similarly affected as in trials with sample size reassessment and exhibit similar marginal statistical properties. For instance, the MLE and MEUE of a specific treatment i perform well in terms of the marginal MSE, and the marginal median of the MEUE is close to the true treatment effect, also in trials with treatment selection.

The statistical properties of the estimates for the *selected* treatment(s) are another concern in clinical trials with treatment selection: when selecting the treatment with the largest observed effect at the interim analysis then the final MLE for the selected treatment will exhibit a positive mean bias. This bias is due to the selection process and the use of the interim data in the MLE. Hence, we call it *selection bias*. The selection bias is closely related to the publication or *reporting bias* caused by not reporting (or under-reporting) statistically non-significant results. Since the selection of treatments is based on the interim data whereas statistical significance is based on all data, the selection bias is smaller than the reporting bias. The selection bias is most relevant if effects are reported only for the selected treatments and omitted for the dropped treatments. The problem may be mitigated if estimates are reported for all treatments whereby the effect of the dropped treatments is estimated from the interim data. (Although the bias from the sample size reassessment remains of relevance also in this case.) If, however, effects of several treatments are similar, the selection bias can be substantial and confidence in the MLE might be limited.

Regarding the MLE, the selection bias is largest if treatments have identical effect. It goes to 0 as differences between treatments go to infinity. The selection bias is avoided by the uniformly

DOI: 10.1002/sim

minimum variance conditional unbiased estimate (UMVCUE). The UMVCUE is defined as an estimator of the effect of the selected treatment which is unbiased conditional on the order of treatments with respect to the mean effects from the first stage. Among all estimators unbiased in this sense, it has minimum variance. Since conditional unbiasedness is a stronger requirement than unconditional unbiasedness, the UMVCUE is also unconditionally unbiased. It is given by

$$\hat{\theta}_{(i)} = \frac{\sigma_2^2 \bar{X}_{1,(i)} + \sigma_1^2 Y}{\sigma_1^2 + \sigma_2^2} - \frac{\sigma_2^2}{\sqrt{\sigma_1^2+\sigma_2^2}} \frac{\phi(W_{i,i+1}) - \phi(W_{i,i-1})}{\Phi(W_{i,i+1}) - \Phi(W_{i,i-1})} \tag{13}$$

where

$$W_{s,i} = \frac{1}{\sigma_1^2}\left(\frac{\sigma_2^2 \bar{X}_{1,(s)} + \sigma_1^2 Y}{\sqrt{\sigma_1^2+\sigma_2^2}} - \sqrt{\sigma_1^2+\sigma_2^2}\bar{X}_{1,(i)} \right), \quad \bar{X}_{1,(0)} := \infty \quad \text{and} \quad \bar{X}_{1,(k+1)} := -\infty$$

k denotes the number of treatments at the start of the trial and ϕ, Φ denote the standard normal density and distribution function, respectively. Here, $\bar{X}_{1,(i)}$ denotes the effect estimate of the treatment with the ith largest observed effect at the interim analysis and $Y \sim \mathrm{N}(\theta_{(i)}, \sigma_2^2)$ denotes the second stage effect estimate of the selected treatment. Formula (13) gives a conditionally unbiased estimate for all explicitly selected treatments. That is, it can be applied for interim decision rules that pre-specify in advance to continue the best, the best two, etc. treatments to the second stage. It can also be applied if the treatment with the largest effect is not among those continued to the second stage.

If equal variance σ^2 is assumed for every treatment, then $\sigma_1^2 = \sigma^2/n_1$ and $\sigma_2^2 = \sigma^2/n_2$, where n_j is the common sample size per group in stage j. The UMVCUE is due to Cohen and Sackrowitz [67], who derived it for selecting the best of k treatments and under the assumption that $n_1 = n_2$. Their proof can be extended to yield the stated results. It can also be extended to the case of unequal sample sizes per treatment arm in the first stage [68]. Note that if more than one treatment arm is selected at the interim analysis, the UMVCUE, while still being unbiased, displays some undesirable properties; in particular, it is not the minimum variance estimate anymore.

In terms of MSE, the UMVCUE is inferior to the MLE. On the other hand, its MSE conditional on the selected treatment is much less sensitive to the treatment selection than the conditional MSE of the MLE. If the wrong treatment is selected, the conditional MSE of the MLE given this choice quickly becomes very large. This tendency is also present, but much weaker for the UMVCUE. Closed-form expressions for the variance of the UMVCUE are not available, but approximations can be derived [68].

To illustrate the previous discussion, we calculated the bias and the MSE for both the UMVCUE and the MLE in case of two treatments with selection of the better treatment at interim. For simplicity we assume that the two treatment estimates $\bar{X}_{1,1}$ and $\bar{X}_{1,2}$ from the first stage are distributed according to $\mathrm{N}(0,1)$ and $\mathrm{N}(\delta,1)$, respectively. We further assume that the second stage treatment effect estimate is given by Y as explained above, with $\sigma_2^2 = 1$. This corresponds to the situation where the interim analysis is performed after 50 per cent of observations in the selected treatment arm. The results are displayed in Figure 7. In this situation, the bias of the MLE is given by $(1/\sqrt{2})\phi(\delta/\sqrt{2})$. The bias of the UMVCUE is 0 by construction. The two curves for MSE are the 'conditional' MSEs, that is, the mean-squared deviations of the estimate from the selected treatment mean, conditional on the selection (see [13] for a formal

DOI: 10.1002/sim

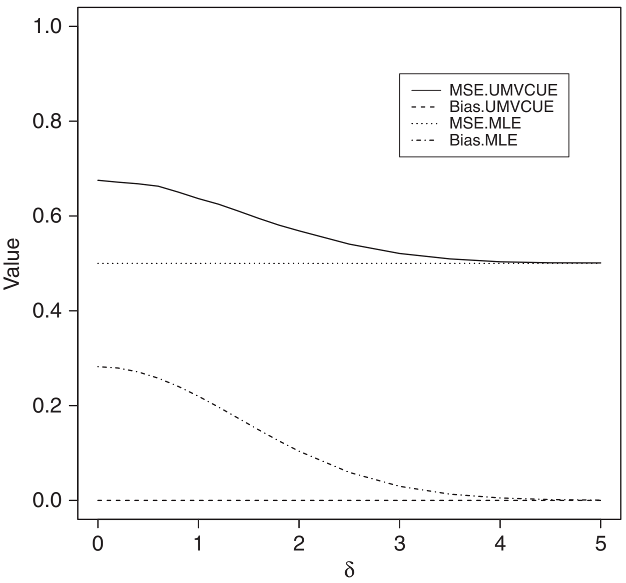

Figure 7. MSE and bias of effect estimates after treatment selection at interim.

definition). For the MLE, this is 0.5 in our situation. As seen from Figure 7, the MLE has a larger bias as compared with the UMVCUE, but a smaller MSE and vice versa. This underpins the difficulties in estimating the treatment effects. Which estimates to consider needs to be decided on a case-by-case basis, where at least bias and MSE should be balanced against each other.

6.4. Repeated confidence intervals

Just like point estimation, interval estimation is also affected by trial modifications made after an interim analysis. Several suggestions have been made. *Repeated confidence intervals* (RCIs) for a parameter θ of interest are defined as a sequence of confidence intervals I_j, $j=1,2$, which fulfill

$$P(\theta\in I_j \text{ for both stages } j=1,2)\geqslant 1-\alpha$$

for all θ [69]. Repeated confidence intervals can be obtained based on the duality of confidence intervals and hypothesis tests. In the following we first consider the univariate case for a single parameter of interest and then consider the problem of deriving simultaneous confidence intervals for multiple parameters.

6.4.1. Univariate confidence intervals. Assume that a combination test for

$$H:\theta_1-\theta_0\leqslant 0 \quad \text{against } K:\theta_1-\theta_0>0$$

has been performed at level α. To construct simultaneous confidence intervals for the treatment-control difference $\theta=\theta_1-\theta_0$, we define for all parameters μ the hypotheses pair

$$H(\mu):\theta\leqslant\mu \quad \text{against } K(\mu):\theta>\mu$$

For each μ let $p_1(\mu)$ and $p_2(\mu)$ denote the first and second stage p-values for $H(\mu)$. For example, in a treatment-control comparison with a normally distributed endpoint and balanced group sizes the first stage z-test p-value is

$$p_1(\mu) = 1 - \Phi[\sqrt{n_1}(\bar{x}_1 - \bar{x}_0 - \mu)/\sqrt{2}\sigma]$$

and $p_2(\mu)$ is defined by analogy. Here, $p_2(\mu) = 1$ if no second stage data are available. Now, given the stagewise p-values are monotone in μ, a 100 per cent $(1-\alpha)$ confidence interval is given by all μ such that the combination test does not reject the corresponding null hypothesis, i.e.

$$I = \{\mu | p_1(\mu) > \alpha_0 \text{ or } p_1(\mu) > \alpha_1, C(p_1(\mu), p_2(\mu)) > c\}$$

As an example, consider a two-stage design for a normally distributed endpoint where the sample size is reassessed after the interim analysis. If the inverse normal combination function with $\alpha_0 = 1$ (no binding stopping for futility) and critical values α_1 and c is used, the repeated $(1-\alpha)$-confidence intervals are given by

$$\left[\bar{X}_1 - \frac{\sqrt{2}\sigma z_{1-\alpha_1}}{\sqrt{n_1}}, \bar{X}_1 + \frac{\sqrt{2}\sigma z_{1-\alpha_1}}{\sqrt{n_1}}\right]$$

if the trial is stopped at interim, and

$$\left[\hat{\theta}_{1m} - \frac{\sqrt{2}\sigma z_{1-c}}{w_1\sqrt{n_1} + w_2\sqrt{n_2}}, \hat{\theta}_{1m} + \frac{\sqrt{2}\sigma z_{1-c}}{w_1\sqrt{n_1} + w_2\sqrt{n_2}}\right]$$

if the trial continues to the second stage [34]. Here, $\hat{\theta}_{1m}$ is the median-unbiased estimate of the difference in effect sizes between treatment and control given by (12), n_1 and n_2 are the (potentially modified) sample sizes per stage and w_1, w_2 are the pre-specified weights. If no modification of sample size is performed, these intervals simplify to the ordinary asymptotic confidence intervals for μ at a local level α_1 and c. Note that RCIs are conservative in the sense that the true probability $\mathrm{P}(\theta \in I_j$ for both stages $j = 1, 2)$ will usually be larger than $1-\alpha$.

6.4.2. Other confidence bounds. For group sequential tests there is no unique ordering of sample points across stages and thus there are many ways to define confidence bounds [3]. Besides repeated confidence intervals, the *monotone confidence interval* approach [70] has been applied to adaptive designs [15]. Repeated confidence intervals are typically strictly conservative but give valid confidence intervals even if one does not adhere to the pre-specified stopping rules. The monotone confidence bounds in contrast are exact but can only be computed at stages where a stopping boundary has been crossed or in the final analysis. The RCIs as introduced in the previous section based on $\hat{\theta}_m$ may not contain the MLE $\hat{\theta}$. Brannath *et al.* [60] suggest to extend them such that $\hat{\theta}$ is contained. They also discuss other (conservative) methods of constructing CIs based on the MLE $\bar{X}$ (e.g. [16]). However, simulations performed by these authors imply that these ML-based RCIs are wider than $\hat{\theta}_m$-based CIs for most practically relevant situations (i.e. if sample size modification that are extremely different from what was originally planned are avoided). Hence, we recommend the use of $\hat{\theta}_m$-based RCIs.

DOI: 10.1002/sim

6.4.3. Simultaneous confidence bounds. Posch *et al.* [13] constructed simultaneous confidence bounds as follows. Assume an adaptive test (based on the closure of combination tests) of the one sided hypotheses

$$H_i : \theta_i - \theta_0 \leqslant 0 \quad \text{against } K_i : \theta_i - \theta_0 > 0,\ i = 1, \ldots, k$$

at multiple level α has been performed. For all parameter vectors $\mu = (\mu_1, \ldots, \mu_k)$ define

$$H_i(\mu_i) : \theta_i - \theta_0 \leqslant \mu_i \quad \text{against } K_i(\mu_i) : \theta_i - \theta_0 > \mu_i,\ i = 1, \ldots, k \tag{14}$$

and let $H_{\mathscr{T}_1}(\mu) = \bigcap_{i \in \mathscr{T}_1} H_i(\mu_i)$ denote the global intersection hypotheses. By the duality of confidence intervals and hypothesis tests we can obtain a simultaneous confidence region by testing $H_{\mathscr{T}_1}(\mu)$ with an α-level combination test for all possible vectors μ. The set of all μ for which $H_{\mathscr{T}_1}(\mu)$ cannot be rejected gives a $(1-\alpha)100$ per cent confidence region. More formally, define for each μ the stage-wise p-values $p_{1,\mathscr{T}_1}(\mu)$ and $p_{2,\mathscr{T}_2}(\mu)$ for the global intersection hypothesis $H_{\mathscr{T}_1}(\mu)$ (if $\mathscr{T}_2 = \emptyset$ we set $p_{2,\mathscr{T}_2}(\mu) = 1$). Now, the confidence region for $\theta_i - \theta_0$, $i = 1, \ldots, k$, is given by all vectors μ such that $\varphi_C[p_{1,T_1}(\mu), p_{2,T_2}(\mu)] = 0$. To obtain simultaneous confidence intervals (instead of a general confidence region) we enlarge the confidence region to a rectangle. To this end we define at each stage j and for all treatments $i \in \mathscr{T}_1$ an adjusted p-value

$$p_{j,i}^{\text{adj}}(\mu_i) = \sup_{\xi \in \mathbb{R}^k,\ \xi_i \leqslant \mu_i} p_{j,\mathscr{T}_j}(\xi) \tag{15}$$

where $\xi = (\xi_1, \ldots, \xi_k)$. Now, by [13] the simultaneous confidence intervals for $\theta_i - \theta_0$ are given by all μ_i such that $\varphi_C[p_{1,i}^{\text{adj}}(\mu_i), p_{2,i}^{\text{adj}}(\mu_i)] = 0$.

Many p-values of tests for intersection hypotheses as, for example, the Bonferroni, Šidak or Dunnett test, can be written as a function of the unadjusted individual p-values, such that for a suitable monotonic function f we have $p_{j,\mathscr{T}_j}(\mu) = f(p_{j,1}(\mu_1), \ldots, p_{j,k}(\mu_k))$, where $p_{j,i}(\mu_i)$ are unadjusted stage-wise p-values for hypothesis $H_i(\mu_i)$ (see Section 3.3.2). Then, $p_{j,i}^{\text{adj}}(\mu_i) = f(1, \ldots, 1, p_{j,i}(\mu_i), 1, \ldots, 1)$ (assuming that $\lim_{\mu_l \to \infty} p_{j,l}(\mu_l) = 1$ and that the $p_{j,l}(\mu_l)$ are increasing in μ_l for $l \in \mathscr{T}_1$). For example, for the Bonferroni and the Simes test $p_{j,i}^{\text{adj}}(\mu_i) = \min(1, |\mathscr{T}_j| \, p_{j,i}(\mu_i))$, where $|\mathscr{T}_1| = k$ and $|\mathscr{T}_2|$ denotes the number of treatments (besides the control group) in the second stage. For the Šidak test $p_{j,i}^{\text{adj}}(\mu_i) = 1 - [1 - p_{j,i}(\mu_i)]^{|\mathscr{T}_j|}$.

The practical computation of the simultaneous confidence intervals involves only one-dimensional numeric root finding. The confidence intervals are computed separately for each treatment i. First, the μ_i are determined that are rejected at the first stage. Assuming monotonicity of $p_{1,i}^{\text{adj}}(\mu_i)$ in μ_i this is done by solving the equation $p_{1,i}^{\text{adj}}(\mu_i) = \alpha_1$ in μ_i. Denoting the solution by $\mu_{a,i}$, all $\mu_i \leqslant \mu_{a,i}$ can be excluded from the confidence interval. If a binding futility bound is specified, additionally we need to determine the solution of $p_{1,i}^{\text{adj}}(\mu_i) = \alpha_0$, denoted by $\mu_{b,i}$. All $\mu_i \geqslant \mu_{b,i}$ have to be included in the confidence interval, since the respective null hypothesis is accepted already at the first stage. If the trial stops in the interim analysis, or no second stage data for treatment i are available, the confidence interval for treatment i is given by $(\mu_{a,i}, \infty)$. If the trial continues and second stage data for treatment i are available, we compute the solution of $C(p_{1,i}^{\text{adj}}(\mu_i), p_{2,i}^{\text{adj}}(\mu_i)) = c$, denoted by $\mu_{c,i}$. Thus, at the second stage all $\mu_i \leqslant \mu_{c,i}$ that have not been accepted at the first stage (i.e. where $\mu_i \leqslant \mu_{b,i}$ holds) can be excluded from the confidence interval. Thus, the final interval is given by $(\mu_{lb,i}, \infty)$, where $\mu_{lb,i} = \min(\max(\mu_{a,i}, \mu_{c,i}), \mu_{b,i})$.

Note that by enlarging the confidence region to a rectangle the resulting confidence intervals may be incompatible with the test decision (the confidence interval may not exclude the parameter values that were rejected in the original multiple test procedure). However, if, for example, Bonferroni or Šidak tests are used to test the intersection hypotheses and in the interim analysis only the dose with the smallest interim p-value is selected, then the corresponding confidence interval will be compatible in the above sense.

Numerical example (*continued*). For the numerical example of Section 3.4 the simultaneous lower confidence bounds can be computed as follows. Let $p_{1,3}^{\text{adj}}(\mu_3)=\min\{3\,\Phi[-\sqrt{71/2}(X_{1,3}-X_{1,0}-\mu_3)/6],1\}$ and $p_{2,3}^{\text{adj}}(\mu_3)=\Phi[-\sqrt{71/2}(X_{2,3}-X_{2,0}-\mu_3)/6]$. Now, let $\mu_{a,3}$ denote the solution of $p_{1,3}^{\text{adj}}(\mu_3)=\alpha_1$, $\mu_{b,3}$ the solution of $p_{1,3}^{\text{adj}}(\mu_3)=\alpha_0$ and $\mu_{c,3}$ the solution of $1-\Phi\{w_1\Phi^{-1}[1-p_{1,3}^{\text{adj}}(\mu_3)]+w_2\Phi^{-1}[1-p_{2,3}^{\text{adj}}(\mu_3)]\}=c$. After the first stage, all $\mu_3<\mu_{a,3}$ are early rejected and all $\mu_3>\mu_b$ early accepted. Note that μ_a, μ_b are just the standard fixed sample confidence bounds of the first stage data at level $\alpha_1/3$ and $\alpha_0/3$, respectively. The final lower confidence bound is given by $\mu_{lb,3}=\min(\max(\mu_{a,3},\mu_{c,3}),\mu_{b,3})$. In the example $\mu_{a,3}=-0.332$, $\mu_{b,3}=0.753$, $\mu_{c,3}=0.697$ such that $\mu_{lb,3}=0.697$. Additionally, for the dropped doses 1 and 2, the simultaneous confidence bounds are given by the standard fixed sample confidence bounds of the first stage data at level $\alpha_1/3$ given by -2.13 and -1.43, respectively.

7. PRACTICAL CONSIDERATIONS

Many basic aspects for planning and conducting a clinical trial with a confirmatory adaptive design are the same as in other, more traditional monitoring settings, such as in group sequential trials [3, 71]. This is in particular true for the need to set up independent data monitoring committees (DMC), the need to restrict access to the interim analysis results in order to protect the integrity of the trial and to facilitate quick access and analyses of validated data at interim. Other aspects of designing and executing an adaptive design, however, may be different, both on the trial level as well as on the level of a whole drug development program. In the following we discuss some practical considerations for the implementation of a confirmatory adaptive design.

In order to retain the validity of a confirmatory trial, a strict type I error rate control is mandatory for an acceptable adaptive design. Proper statistical methodology should thus be applied, such as the methods reviewed in this paper. This requirement also implies that the experimental questions and hypotheses to be investigated are well specified upfront in the study protocol. The number of potential adaptations should generally be kept to a minimum; explorative, hypotheses generating adaptive designs in a confirmatory drug development phase are not acceptable to regulatory agencies.

Because of their complexity, both in methodology and logistics, adaptive designs require a careful preplanning. This includes the amount of pre-specification required in the study protocol as well as the conduct of extensive simulations for a good understanding of the operational characteristics. To strictly control the overall type I error rate, it is mandatory to pre-specify the design of the first stage and how the information across the stages is combined. In addition, it is essential to concisely describe the study objectives, the response variables and the type of adaptive designs (e.g. treatment selection, selection of a pre-specified sub-population, etc.) beside other standard information required in the study protocol. The decision rules allowing early stopping for success have to be stated explicitly in the protocol. For other adaptations, like treatment selection or dropping a dose, this is not the case, since the type I error rate is not affected by the nature of these rules if appropriate statistical methods are used. This implies that full flexibility is guaranteed on how to perform the treatment selection at interim, which can thus be based on any evidence available up to the interim decision point, including information from outside the ongoing trial if necessary. However, in order to assess the impact of the interim decision rules on treatment effect estimates and the operating characteristic of the trial, and as a provision of guidance for the decision makers at interim analyses, potential interim decision rules need to be investigated in the context of clinical scenario evaluations and requires extensive simulations.

An issue which is still under some debate is the degree of sponsor involvement in the interim decision process. Some adaptive designs may involve complex decisions that lie in a domain which is traditionally a sponsor responsibility and in which important sponsor interests may be involved in arriving at the best decision. For example, currently the evidence available at the end of phase II will be used to select a combination of dose, application mode, etc. for the confirmatory phase III studies and thus potentially for the later marketed drug, once it has been approved. These considerations at the end of phase II are thus business critical and go through senior decision boards of a pharmaceutical company. With an adaptive design combining phases II and III some of these decisions have to be taken before starting the trial (and considerably earlier than in a conventional program), with the notable exception of the remaining uncertainty, which the interim analyses is supposed to resolve. Current regulatory guidance in more traditional monitoring settings, such as in group sequential designs, holds that sponsors should not have access to interim data while trials are ongoing. One concern in the context of adaptive designs is that unanticipated complexities might not fit a pre-specified algorithm that can be implemented without sponsor participation. One model on how to implement flexible interim decisions in practice is to include a proposal of potential interim decision rules in the DMC charter with the understanding that the DMC has the discretion to deviate from them as necessary and involve the sponsor based on the following principles: a clear rationale for a sponsor involvement; sponsor representatives properly distanced from trial operations; clear understanding by all parties involved of the issues and potential risks; and documentation of the processes followed with restrictive firewalls put in place. The general aim should be a minimal sponsor exposure sufficient to make decisions, meaning that the smallest possible number of sponsor representatives should only get involved at the adaptation point with the minimally relevant information. Such an approach would minimize the sponsor's involvement and associated information leakage but still guarantee the sponsor's interest in case of unexpected emerging results. We refer to [72] for further details.

Another important aspect when planning a confirmatory adaptive design is to keep the focus on the project level. Adaptive designs may only be of benefit if sufficient evidence is expected from the combined phase II/III study as compared to the strategy with a phase II trial that is followed by a separate phase III clinical trial. Thus, before embarking on a phase II/III study it needs to be ensured that the totality of information is sufficient to support a submission at the end of phase III. Accordingly, if an adaptive design is applied to a confirmatory study, a second pivotal trial is typically needed to replicate the findings of an independent first pivotal trial. An adaptive phase II/III study thus does not replace the full phase III program. Also, the necessary information on safety, regimen, mode of application, endpoint, etc., which is needed for a successful confirmatory phase III, must be collected before the start of the combined phase II/III study.

The decision in favor of an adaptive design or a traditional approach is of course also influenced by many non-statistical considerations, especially, but not only, operational issues. Among the

DOI: 10.1002/sim

factors that impinge on the decision are, for example, recruitment speed, type of endpoint, the cost and time consumption of interim analyses (including, for example, requirements such as central laboratory reviews or extensive data cleaning). Adaptive trials are more beneficial if recruitment is slow, tedious or expensive, but treatment effect can be measured immediately, than if recruitment is quick, easy and inexpensive, but it takes time for the treatment to take effect or the endpoint to be observable (like, for example, in some time-to-event trials). If recruitment is very quick, it may be necessary to stop it for the sake of an interim analysis. This is considered unattractive by many clinical trial managers, because closing down and reopening centers creates operational friction. Slowing down recruitment during this period of time may be a compromise.

In an adaptive trial involving treatment arm selection, there will inevitably be patients who were randomized to a discontinued treatment, but have not reached the endpoint at the time of interim analysis. If the time to endpoint is quick, there will of course be fewer such patients. Usually, such 'overrunners' would be allowed to finish their respective treatment and be monitored for efficacy and safety, unless ethical considerations require that they are switched to a treatment that has emerged as clearly superior. The same is of course true for a trial that was stopped early for efficacy.

On an operational level, it is good practice to clearly define the interim safety and efficacy analyses upfront, limiting them to ones which are really required for decision making. It is also recommended to synchronise interim analyses for efficacy with regular safety updates, if the latter are required in an ongoing drug development program.

For manufacturing and drug supply management, adaptive designs pose a challenge because it is more difficult to predict the amount of drug required in a trial, both overall and by formulation, treatment regimen etc. Hence, manufacturing and drug supply management should be involved early in the planning phase of an adaptive trial such that the required resources for a smoothly executed trial are allocated—and then eventually freed. Of course, the decision on whether to use an adaptive or a more traditional design also depends on the magnitude of difficulties associated with drug supply. However, awareness of these issues has recently increased and progress has been made towards an adaptive drug supply management [73].

Likewise, participating centers need to be informed about the planned trial and made aware of the impact adaptations might have on their role in the different stages of the trial, especially with regard to flexible sample sizes. Where the risk/benefit is different in subsequent stages of a trial, it may also be necessary to ask study participants to renew their informed consent form.

In summary, properly conducted adaptive designs pose a number of operational challenges with a potential risk of extending the timelines due to the more complex planning, logistics, and regulatory interactions. Therefore, the rationale for an adaptive design needs to be well justified as compared to an independent phase II/phase III development program and it needs to be ensured that the evidence base for regulatory decision making is not diminished when conducting an adaptive design.

8. DISCUSSION

Pharmaceutical companies, regulatory agencies, ethical committees, the medical community and, last but not least, the patients are all interested in clinical trials that convincingly establish the efficacy and safety of a new treatment. All these stakeholders are also interested in an efficient drug development process, such that an effective new treatment can be made available as early as

DOI: 10.1002/sim

possible to the patients in need. The expectation has arisen that carefully planned and conducted studies based on adaptive designs are an important tool which can help to fulfil these requirements. However, apart from the benefits, there are also limitations to the potential use of confirmatory adaptive designs. We mentioned the mandatory, strict type I error rate control, the pre-specification of the experimental questions and hypotheses to be investigated upfront in the study protocol and that adaptive designs in the confirmatory context should not have an explorative, hypotheses-generating theme as the primary objective. Likewise, logistical hurdles (such as, for example, long time to assessment of response, fast recruitment, trial integrity issues, drug supply management) may hinder the application of an adaptive design and need to be considered in detail at the planning stage. Each adaptive design has its own peculiarities, depending on the specific application. Although adaptive designs provide more flexibility than traditional designs, they should always be conducted keeping international guidelines on good clinical practice in mind. In particular, any adaptive design should ensure the validity and integrity of a clinical study. The best way to minimise pitfalls and the difficulties associated with adaptive designs is to allocate enough lead time in order to carefully think through and possibly simulate different scenarios. It is certainly advantageous to involve statistical and logistical experts for this multidisciplinary effort and borrow from the experience of those teams which have previously undertaken the planning and execution of such a trial.

There are many open issues in adaptive designs. One criticism on adaptive designs is that in case of performing design modification one has to use non-standard test statistics instead of the common sufficient test statistics [74]. Another concern about confirmatory adaptive designs is that discrepancies in the results across stages may render the overall results hard to interpret [75], in particular if it cannot be ruled out that this difference is due to intentional or unintentional leakage of interim results. To address these concerns the sponsor may be asked to pre-plan methods which ensure that results from different stages can be justifiably combined. As seen in Section 6, estimation remains an open topic for research. However, many of the criticisms of adaptive designs also apply to classically used designs and methods, which in turn have received increased attention more recently because of the novelty of adaptive designs.

It should be emphasized that the methods reviewed in this paper all control the familywise error rate strongly at a pre-specified significance level. In the recent past, simulation-based approaches have been described, which approximate the critical value by simulating a large number of clinical trials based on pre-specified assumptions. When using such methods, caution is advisable, if the underlying assumptions are violated or cannot be verified, which is often the case in clinical practice. For these methods, a strong type I error rate control is usually difficult to assess, even when performing large-scale simulations. We refer to [76] for a critical discussion of this topic.

Finally, we believe that adaptive designs can lead a higher scientific standard for the design and conduct of clinical trials. As pointed out by Bauer [77], the current clinical practice is faced with an irritating large number of protocol amendments, which—as one may argue—have a larger impact on operational bias, type I error rate control, etc. than a well planned adaptive design following sound statistical principles and where potential adaptations are considered beforehand.

ACKNOWLEDGEMENTS

We thank Harald Pohlmann for providing us the case study example. We are particularly indebted to Peter Bauer and Willi Maurer for their continuous support and constant discussions throughout the preparation

DOI: 10.1002/sim

of this manuscript. We thank five reviewers and an associate editor for their comments, which led to a considerably improved version of the manuscript.

REFERENCES

1. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977; **64**:191–199.
2. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **5**:549–556.
3. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC: Boca Raton, 2000.
4. Proschan MA, Lan KKG, Wittes JT. *Monitoring of Clinical Trials: A Unified Approach*. Springer: New York, 2006.
5. Brannath W, Koenig F, Bauer P. Multiplicity and flexibility in clinical trials. *Pharmaceutical Statistics* 2007; **6**:205–216.
6. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663.
7. Follmann DA, Proschan MA, Geller NL. Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics* 1994; **50**:325–336.
8. Hellmich M. Monitoring clinical trials with multiple arms. *Biometrics* 2001; **57**:892–898.
9. Stallard N, Friede T. A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine* 2008; **27**:6209–6227.
10. Bauer P. Multistage testing with adaptive designs (with Discussion). *Biometrie und Informatik in Medizin und Biologie* 1989; **20**:130–148.
11. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994; **50**:1029–1041. Correction: *Biometrics* 1996; **52**:380.
12. Chuang-Stein C, Anderson K, Gallo P, Collins S. Sample size re-estimation: A review and recommendations. *Drug Information Journal* 2006; **40**:475–484.
13. 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* 2005; **24**:3697–3714.
14. Bretz F, Schmidli H, Koenig F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts (with Discussion). *Biometrical Journal* 2006; **48**:623–634.
15. Brannath W, Posch M, Bauer P. Recursive combination tests. *Journal of the American Statistical Association* 2002; **97**:236–244.
16. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* 1995; **51**:1315–1324.
17. Müller HH, Schäfer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* 2001; **57**:886–891.
18. Müller HH, Schäfer H. A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine* 2004; **23**:2497–2508.
19. Schmidli H, Bretz F, Racine-Poon A. Bayesian predictive power for interim adaptation in seamless phase II/III trials where the endpoint is survival up to some specified timepoint. *Statistics in Medicine* 2007; **26**:4925–4938.
20. Brannath W, Zuber E, Branson M, Bretz F, Gallo P, Posch M, Racine-Poon A. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine* 2009; in press.
21. FDA. Innovation/Stagnation: challenge and opportunity on the critical path to new medical products. *FDA Report*, March 2004. Available at http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html.
22. Gallo P, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, Pinheiro J. Adaptive designs in clinical drug development—an executive summary of the PhRMA Working Group. *Journal of Biopharmaceutical Statistics* 2006; **16**:275–283.
23. FDA. Innovation/Stagnation: critical path opportunities report. *FDA Report*, March 2006. Available at http://www.fda.gov/oc/initiatives/criticalpath/reports/opp_report.pdf.
24. EMEA. Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. *EMEA Doc. Ref. CHMP/EWP/2459/02*, October 2007. Available at http://www.emea.europa.eu/pdfs/human/ewp/245902 enadopted.pdf.
25. Krams M, Burman CF, Dragalin V, Gaydos B, Grieve AP, Pinheiro J, Maurer W. Adaptive designs in clinical drug development: opportunities, challenges, and scope. Reflections following PhRMA's November 2006 workshop. *Journal of Biopharmaceutical Statistics* 2006; **17**:957–964.

DOI: 10.1002/sim

26. Laurie D, Branson M, Bretz F, Maurer W, Thomas P. Designing and conducting confirmatory adaptive clinical trials. *The Regulatory Affairs Journal Pharma* 2008; **19**:85–91.
27. Inoue LYT, Thall PF, Berry DA. Seamlessly expanding a randomized phase II trial to phase III. *Biometrics* 2002; **58**:823–831.
28. Berry DA, Müller P, Grieve AP, Smith M, Parke T, Blazek R, Mitchard N, Krams M. Adaptive Bayesian designs for dose-ranging drug trials. *Case Studies in Bayesian Statistics*, Lecture Notes in Statistics, vol. 162. Springer: New York, 2002; 99–181.
29. Bornkamp B, Bretz F, Dmitrienko A, Enas G, Gaydos B, Hsu CH, Koenig F, Krams M, Liu Q, Neuenschwander B, Parke T, Pinheiro J, Roy A, Sax R, Shen F. Innovative approaches for designing and analyzing adaptive dose-ranging trials (with Discussion). *Journal of Biopharmaceutical Statistics* 2007; **17**:965–995.
30. Bretz F, Branson M, Burman CF, Chuang-Stein C, Coffey C. Adaptivity in drug discovery and development. *Drug Development Research* 2009; DOI: 10.1002/ddr.20285.
31. Hamilton M. Development of a rating scale for primary depressive illness. *British Journal of the Society for Clinical Psychology* 1967; **6**:278–296.
32. Maca J, Bhattacharya S, Dragalin V, Gallo P, Krams M. Adaptive seamless phase II/III designs—background, operational aspects, and examples. *Drug Information Journal* 2006; **40**:463–473.
33. Liu Q, Proschan MA, Pledger GW. A unified theory of two-stage adaptive designs. *Journal of the American Statistical Association* 2002; **97**:1034–1041.
34. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics* 1999; **55**:1286–1290.
35. Cui L, Hung HMJ, Wang SJ. Modification of sample size in group sequential clinical trials. *Biometrics* 1999; **55**:853–857.
36. Wassmer G, Vandemeulebroecke M. A brief review on software developments for group sequential and adaptive designs. *Biometrical Journal* 2006; **48**:732–737.
37. Vandemeulebroecke M. Group sequential and adaptive designs—A review of basic concepts and points of discussion. *Biometrical Journal* 2008; **50**(4):541–557.
38. Posch M, Timmesfeld N, Koenig F, Müller HH. Conditional rejection probabilities of Student's t-test and design adaptations. *Biometrical Journal* 2004; **46**:389–403.
39. Schmidli H, Bretz F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: applications and practical considerations. *Biometrical Journal* 2006; **48**:635–643.
40. Wang SJ, O'Neill RT, Hung JHM. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics* 2007; **6**:227–244.
41. Maurer W, Branson M, Posch M. Adaptive designs and confirmatory hypothesis testing. In *Multiple Testing Problems in Pharmaceutical Statistics*, Dmitrienko A, Tamhane A, Bretz F (eds). Taylor & Francis: New York, 2009; in press.
42. Hochberg Y, Tamhane AC. *Multiple Comparison Procedures*. Wiley: New York, 1987.
43. Marcus R, Peritz E, Gabriel KR. On closed testing procedure with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
44. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 1999; **18**:1833–1848.
45. Kieser M, Bauer P, Lehmacher W. Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biometrical Journal* 1999; **41**:261–277.
46. Hommel G. Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal* 2001; **43**:581–589.
47. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **73**:751–754.
48. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* 1955; **50**:1096–1121.
49. Koenig F, Brannath W, Bretz F, Posch M. Adaptive Dunnett tests for treatment selection. *Statistics in Medicine* 2008; **27**:1612–1625.
50. Senn S, Bretz F. Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics* 2007; **6**:161–170.
51. Bauer P, Koenig F. The reassessment of trial perspectives from interim data—a critical view. *Statistics in Medicine* 2006; **25**:23–36.
52. Jennison C, Turnbull BW. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* 2003; **22**(6):971–993.
53. Posch M, Bauer P, Brannath W. Issues in designing flexible trials. *Statistics in Medicine* 2003; **22**:953–969.

54. Evans SR, Li L, Wei LJ. Data monitoring in clinical trials using prediction. *Drug Information Journal* 2007; **41**:733–742.
55. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials (with Discussion). *Journal of the Royal Statistical Society*, *Series A* 1994; **157**:357–416.
56. Dmitrienko A, Wang MD. Bayesian predictive approach to interim monitoring in clinical trials. *Statistics in Medicine* 2006; **25**:2178–2195.
57. Bretz F, Pinheiro JC, Branson M. Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics* 2005; **61**:738–748.
58. Friede T, Stallard N. A comparison of methods for adaptive treatment selection. *Biometrical Journal* 2008; **50**(5):767–781.
59. Wang SJ, Hung HMJ, O'Neill RT. Adapting the sample size planning of a phase III trial based on phase II data. *Pharmaceutical Statistics* 2007; **5**(2):81–97.
60. Brannath W, Koenig F, Bauer P. Estimation in flexible two stage designs. *Statistics in Medicine* 2006; **25**: 3366–3381.
61. Kieser M, Friede T. Simple procedures for blinded sample size adjustment that do not affect the type I error rate. *Statistics in Medicine* 2003; **22**:3571–3581.
62. Whitehead J. On the bias of maximum likelihood estimation following a sequential test. *Biometrika* 1986; **73**:573–581.
63. Shen L. An improved method of evaluating drug effect in a multiple dose clinical trial. *Statistics in Medicine* 2001; **20**:1913–1929.
64. Stallard N, Todd S, Whitehead J. Estimation following selection of the largest of two normal means. *Journal of Planning and Statistical Inference* 2008; **138**:1629–1638.
65. Emerson SS, Fleming TR. Parameter estimation following group sequential hypothesis testing. *Biometrika* 1990; **77**:875–892.
66. Liu A, Hall WJ. Unbiased estimation following a group sequential test. *Biometrika* 1999; **86**:71–78.
67. Cohen A, Sackrowitz H. Two stage conditionally unbiased estimators of the selected mean. *Statistics and Probability Letters* 1989; **8**:273–278.
68. Bowden J, Glimm E. Unbiased estimation of selected treatment means in two-stage trials. *Biometrical Journal* 2008; **50**(4):515–527.
69. Jennison C, Turnbull BW. Repeated confidence intervals for group sequential trials. *Controlled Clinical Trials* 1984; **5**:33–45.
70. Tsiatis AA, Rosner GL, Mehta CR. Exact confidence intervals following a group sequential test. *Biometrics* 1984; **40**:797–804.
71. Emerson SS. Frequentist evaluation of group sequential clinical trial designs. *Statistics in Medicine* 2007; **26**:5047–5080.
72. Gallo P. Operational challenges in adaptive design implementation. *Pharmaceutical Statistics* 2006; **5**:119–124.
73. Quinlan JA, Krams M. Implementing adaptive designs: logistical and operational considerations. *Drug Information Journal* 2006; **40**:437–444.
74. Burman CF, Sonesson Are flexible designs sound? (with Discussion). *Biometrics* 2006; **62**:664–669.
75. Gallo P, Chuang-Stein C. What should be the role of homogeneity testing in adaptive trials? *Pharmaceutical Statistics* 2009; DOI: 10.1002/pst.342.
76. Posch M, Maurer W, Bretz F. Type I error rate control in adaptive designs for confirmatory clinical trials with treatment selection at interim 2009; submitted.
77. Bauer P. Adaptive designs: looking for a needle in the haystack—A new challenge in medical research. *Statistics in Medicine* 2008; **27**:1565–1580.