Two-Stage Adaptive Designs With Interim Treatment Selection

Dem Fachbereich 03 (Mathematik/Informatik)
der Universität Bremen
zur Erlangung des Grades eines
Dr. rer. nat.

eingereichte Dissertation

von

Máximo Carreras, M.Sc.

Datum der Einreichung: 27.02.2015

Datum des Kolloquiums: 30.04.2015

1. Gutachter: Prof. Dr. Werner Brannath, University of Bremen, Germany

2. Gutachter: Prof. Dr. Frank Bretz, University of Hanover, Germany

Dissertation Summary

This dissertation is about two-stage adaptive designs with interim treatment selection. It includes two articles entitled (1) "Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection" and (2) "Adaptive seamless designs with interim treatment selection: a case study in oncology". Both articles are published in the journal Statistics in Medicine.

Adaptive designs for clinical trials allow interim data-driven design modifications while maintaining the rigor and validity of the statistical inference [1]. Design adaptations may include early stopping for efficacy, futility or safety, reassessment of the overall sample size, adjustments to the study population (e.g. restriction to a subpopulation), changes to endpoints or hypothesis to be tested as well as dropping or adding treatment arms. Adaptive designs have gained considerable popularity in recent years among clinical trialists because of their potential to improve efficiency in drug development. There are also ethical considerations that support the use of adaptive designs; for example, adaptive designs can reduce the number of patients within the trial who are treated with non-effective treatments. Overall, adaptive designs provide the same scientific rigor that is required in more traditional study designs while potentially utilizing less resources. However, while the increase in flexibility of adaptive designs offers great opportunities, it also brings limitations and pitfalls which should be carefully assessed when adaptive designs are intended to be used in confirmatory trials [2], [3].

In the traditional drug development process, a phase II study typically compares several treatments (e.g. different doses of a new compound) to a control. The objective of the study is to determine whether the development of the compound should be continued and, if so, which treatment(s) or dose(s) should be further investigated. The phase II study also provides initial estimates of treatment effect, which are used to power the subsequent phase III study. The phase III study is then conducted as a stand-alone confirmatory study, disregarding all data collected on the phase II study.

One of the most appealing applications of adaptive designs is to combine phase II and phase III studies of the traditional drug development process into a single seamless phase II/III confirmatory study (see [4] and [5] for comprehensive summaries). Bauer and Kieser [6] proposed two-stage adaptive designs that allow the integration of the treatment selection and the confirmatory testing of efficacy for the selected treatments within a single study. An important feature of these designs is the fact that the treatment selection rule does not need to be pre-specified, which gives considerable flexibility to the inherently complex interim decision process. Hommel [7] extended Bauer and Kieser's work to allow design modifications which include interim changes to the primary endpoint as well as addition of experimental treatments. All these adaptive designs (see also [8]) propose tests that control the familywise type I error rate in the strong sense; that is, the probability that any treatment is erroneously declared significantly superior to the control is maintained below a prespecified significance level under all possible configurations of effective and ineffective treatments.

In the next 2 sections we discuss hypothesis testing and estimation in two-stage adaptive designs with interim treatment selection. These sections serve as a summary of the topics developed in the two articles.

1 Hypothesis testing

1.1 Two-stage combination tests

Let H_0 be a one-sided null hypothesis to be tested in two sequential stages. Let α be the overall significance level of the test. Let p and q be the p-values for testing H_0 based on, respectively, stage 1 and stage 2 data. A two-stage combination test is defined by a combination function $\mathcal{C}(\cdot,\cdot)$, early stopping boundaries α_1 and α_0 and a critical value c for the final analysis. We assume that $0 \leq \alpha_1 < \alpha < \alpha_0 \leq 1$ and that $\mathcal{C}(\cdot,\cdot)$ is monotonically increasing in both arguments. The combination test is defined as follows: if $p < \alpha_1$ or $p \geq \alpha_0$, we stop the trial at the end of stage 1 with, respectively, a rejection of H_0 or a failure to reject H_0. If $\alpha_1 \leq p < \alpha_0$, the study continues into the second stage and we reject H_0 at the end of the study if and only if $\mathcal{C}(p,q) \leq c$. The constant c, which depends on α, α_1, α_0 and $\mathcal{C}$, is determined so that the following level condition is satisfied:

$$P_{H_0}(p \leq \alpha_1) + P_{H_0}\left(\mathcal{C}(p,q) \leq c, \alpha_1 < p \leq \alpha_0\right) \leq \alpha. \tag{1}$$

We further assume that, under H_0, the distribution of p and the conditional distribution of q given p are stochastically larger than or equal to the uniform distribution on $[0,1]$; which means that p and q satisfy

$$P_{H_0}(p \leq x) \leq x \quad \text{and} \quad P_{H_0}(q \leq x \,|\, p) \leq x \quad \text{for all} \quad 0 \leq x \leq 1. \tag{2}$$

(Brannath et al. [9] called this property "p clud"). This is the case, for instance, when disjoint sample units are recruited at the different stages and conservative tests are used at each stage. The level condition (1) can now be replaced by the equation

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 \mathcal{I}_{\{\mathcal{C}(x,y) \leq c\}} dy dx = \alpha, \tag{3}$$

where the indicator function $\mathcal{I}_{\{\cdot\}}$ in equation (3) equals 1 when $\mathcal{C}(x, y) \leq c$ and 0 otherwise. The combination test defined above is a level α test even when the design of stage 2 (e.g. sample size or test statistic) depends on stage 1 data. The decision function of the combination test is defined as

$$\varphi_{\mathcal{C}}(p, q) = \begin{cases} 1 & \text{if } p < \alpha_1 \text{ or both } \alpha_1 \leq p < \alpha_0 \text{ and } \mathcal{C}(p, q) \leq c \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The cases $\varphi_{\mathcal{C}} = 1$ and $\varphi_{\mathcal{C}} = 0$ correspond, respectively, to the rejection and non-rejection of H_0. Bauer [10] and Bauer & Köhne [11] suggested using the combination function given by the product of p-values $\mathcal{C}(p, q) = pq$ [12]. The procedure of Proschan and Hunsberger [13] can be described by the combination function $\mathcal{C}(p, q) = 1 - \Phi\left([\Phi^{-1}(1 - p)]^2 + [\max\{0, \Phi^{-1}(1 - q)\}]^2\right)$, where $\alpha_1 < \alpha_0 \leq 0.5$, $0 < p \leq 0.5$ and $0 \leq q \leq 1$. In our work we use the weighted inverse normal combination function [14]

$$\mathcal{C}(p, q) = 1 - \Phi\left[w_1 \Phi^{-1}(1 - p) + w_2 \Phi^{-1}(1 - q)\right], \tag{5}$$

where w_1 and w_2 are pre-defined non-negative weights that satisfy $w_1^2 + w_2^2 = 1$ and Φ is the cumulative distribution function of the standard normal distribution.

1.2 Closed testing principle

Suppose that we wish to compare k experimental treatments to a control and denote H_i; $i = 1, ..., k$, the corresponding one-sided null hypotheses. In order to account for the multiplicity of the testing procedure, we use the closed testing principle [15], which works as follows. We first must define level α tests for all intersection hypotheses $H_\Psi = \cap_{i \in \Psi} H_i$, where $\Psi \subseteq \mathcal{F} = \{1, ..., k\}$. Then, an individual null hypothesis H_j, $j \in \mathcal{F}$, can be rejected at multiple level α if and only if, for all subsets $\Psi \subseteq \mathcal{F}$ containing j, we have that the hypothesis H_Ψ has been rejected at level α. The testing procedure described above protects the family-wise error rate at level α in the strong sense; that is, the probability of rejecting at least one true null hypothesis is bounded above by α, under any configuration of true and false null hypotheses.

Let p_i be the p-value associated with the testing of individual null hypothesis H_i; $i = 1, ..., k$. Let $H_\Psi = \cap_{i \in \Psi} H_i$ be an intersection hypothesis, where $\Psi \subseteq \mathcal{F}$. Let $r = |\Psi|$ be the cardinality of Ψ. A p-value for testing H_Ψ can be defined in several ways:

$$p_\Psi = \min\left[1, r \min_{i \in \Psi}(p_i)\right] \qquad \text{(Bonferroni test)}$$

$$p_\Psi = 1 - \left[1 - \min_{i \in \Psi}(p_i)\right]^r \qquad \text{(Sidak test)}$$

$$p_\Psi = \min_{i \in \Psi} \frac{r}{i} p_{(i)} \qquad \text{(Simes test)}$$

$$p_\Psi = \min_{i \in \Psi}(r + 1 - i) p_{(i)} \qquad \text{(Hochberg test)}$$

where $p_{(i)}$ denotes the i-th smallest p-value, for $i \in \Psi$. For the case of (approximately) normally distributed test statistics, the Dunnett test [16] can be used to test the intersection hypotheses. The Dunnett test accounts for the correlation between the test statistics due to the fact that the different treatments are compared to the same control.

As an example, consider the case of two treatments. Let H_1 and H_2 be the individual null hypotheses and let p_1 and p_2 be the corresponding p-values. We must construct a test for the intersection hypothesis $H_1 \cap H_2$. By Hochberg's procedure, the p-value $p_{\text{int}} = \min\{2p_{(1)}, p_{(2)}\} = \min\{2\min\{p_1, p_2\}, \max\{p_1, p_2\}\}$ provides a conservative test for $H_1 \cap H_2$. By the closed testing principle, a multiplicity-adjusted p-value for testing H_j is given by $p_{j,\text{adj}} = \max\{p_j, p_{\text{int}}\}$; $j = 1, 2$.

1.3 Combination tests applied after interim treatment selection

Let us now consider the comparison of k experimental treatments to a control in two stages. Let H_i; $i \in \mathcal{F} = \{1, ..., k\}$ denote the one-sided null hypotheses. Let p_Ψ be the stage 1 p-value for testing the intersection hypothesis H_Ψ, $\Psi \subseteq \mathcal{F}$, computed by one of the methods described in 1.2. Let $\mathcal{G} \subseteq \mathcal{F}$ be the indices of the treatments selected at the interim analysis to continue into the second stage. For all hypotheses H_Ψ, $\Psi \subseteq \mathcal{G}$, let q_Ψ be the corresponding stage 2 p-value. For all other hypotheses H_Ψ, $\Psi \subseteq \mathcal{F}$, we define

$$q_\Psi = q_{\Psi \cap \mathcal{G}}$$

where $q_\emptyset = 1$. These p-values are conservative for testing H_Ψ because of the "p clud" condition

$$P_{H_\Psi}\left(q_\Psi \leq x \,|\, p_\Psi\right) \leq P_{H_{\Psi\cap\mathcal{G}}}\left(q_{\Psi\cap\mathcal{G}} \leq x \,|\, p_{\Psi\cap\mathcal{G}}\right) \leq x.$$

Now, for each $\Psi \subseteq \mathcal{F}$, the intersection hypothesis H_Ψ is rejected if and only if $\varphi_\mathcal{C}(p_\Psi, q_\Psi) = 1$, where $\varphi_\mathcal{C}(\cdot,\cdot)$ is defined in (4) for a pre-specified combination function $\mathcal{C}(\cdot,\cdot)$, early stopping boundaries α_0 and α_1 and critical value c, which satisfy (3) for a given overall significance level α. Finally, by the closed testing principle, an individual hypothesis H_j, $j \in \mathcal{F}$ is rejected at multiple level α if and only if all intersection hypotheses H_Ψ, with $\Psi \subseteq \mathcal{F}$ and $j \in \Psi$ are rejected at level α with their combination tests; that is $\varphi_\mathcal{C}(p_\Psi, q_\Psi) = 1$.

As an example, let us consider again the case of two treatments. Suppose that, after stage 1, treatment 1 is selected to continue into stage 2 and treatment 2 is stopped. At the end of the study we have p_1 and p_2, the stage 1 p-values for testing H_1 and H_2, and q_1, the stage 2 p-value for testing H_1. We set $q_2 = 1$, because there is no stage 2 data comparing treatment 2 with control. We compute the stage 1 p-value for testing the intersection hypothesis $H_1 \cap H_2$ using Hochberg's method by $p_{\text{int}} = \min\{2\min\{p_1, p_2\}, \max\{p_1, p_2\}\}$. The stage 2 p-value for the intersection hypothesis is set to $q_{\text{int}} = q_1$. Finally, H_1 is rejected at the end of the study if both $\varphi_\mathcal{C}(p_1, q_1) = 1$ and $\varphi_\mathcal{C}(p_{\text{int}}, q_{\text{int}}) = 1$. In the case that both treatments are selected to continue into stage 2, q_2 would be the stage 2 p-value for testing H_2 and $q_{\text{int}} = \min\{2\min\{q_1, q_2\}, \max\{q_1, q_2\}\}$ would be the stage 2 p-value for testing $H_1 \cap H_2$. Hypothesis H_2 could be rejected at the end of the study if and only if both $\varphi_\mathcal{C}(p_2, q_2) = 1$ and $\varphi_\mathcal{C}(p_{\text{int}}, q_{\text{int}}) = 1$.

1.4 Time-to-event setting

In this section, we follow ideas described in [17]. Suppose that we are interested in comparing k experimental treatments to a control in 2 stages with regards to a time-to-event endpoint. Let H_i be the null hypothesis of no difference in the survival distribution of the time-to-event endpoint between treatment i and control; $i = 1, ..., k$. Let us assume one-sided alternative hypotheses that the treatments prolong the time to an event compared to control. A common test applied in this situation is the log-rank test, which we now describe. Let d_{ij} be the cumulative number of events observed at the end of stage j among patients recruited into treatment i and control; $i = 1, ..., k$ and $j = 1, 2$. Let N_{ijt} and N_{0jt} be the number of patients at risk, respectively, in treatment i and control at stage j when the t-th event occurred in treatment i or control. Under the assumption of no ties, the log-rank statistic at stage j for the comparison of treatment i with control is given by

$$Z_{ij} = \frac{\sum_{t=1}^{d_{ij}} I_{ijt} - \frac{N_{ijt}}{N_{ijt}+N_{0jt}}}{\sqrt{\sum_{t=1}^{d_{ij}} \frac{N_{ijt}N_{0jt}}{(N_{ijt}+N_{0jt})^2}}}; \qquad i = 1, ..., k \qquad j = 1, 2,$$

where $I_{ijt} = 1$ if the event ocurred in the treatment i and 0 otherwise. Note that 'large' negative values of Z_{ij} are evidence against H_i. For fixed d_{ij}, Z_{ij} has an asymptotic normal distribution with variance 1 and mean $\sqrt{d_{ij}}\frac{\sqrt{r_i}}{r_i+1}\ln\theta$, where θ is the true hazard ratio between treatment i and control and r_i is the ratio between the number of patients recruited into treatment i and control groups respectively [18]. If the numbers at risk in each treatment group remain nearly equal over time, as it is to be expected under the null hypothesis, then $N_{ijt} \approx N_{0jt}$ and we can use the approximation

$$\sum_{t=1}^{d_{ij}} \frac{N_{ijt}N_{0jt}}{(N_{ijt}+N_{0jt})^2} \approx \frac{d_{ij}}{4}.$$

The sequence of test statistics Z_{i1}, Z_{i2} approximately has an independent and normally distributed increments structure (see, e.g. [19], [20]). Thus,

$$Z_{i2}^* = \frac{\sqrt{d_{i2}}Z_{i2} - \sqrt{d_{i1}}Z_{i1}}{\sqrt{d_{i2}-d_{i1}}}$$

is independent of Z_{i1} and we have that

$$\sqrt{d_{i2}}Z_{i2} = \sqrt{d_{i1}}Z_{i1} + \sqrt{d_{i2}-d_{i1}}Z_{i2}^*;$$

for $i = 1, ..., k$. The stagewise p-values for the comparison of treatment i with control are calculated by

$$p_i = \Phi(Z_{i1}), \qquad \text{and} \qquad q_i = \Phi(Z_{i2}^*).$$

Stagewise p-values for the intersection hypotheses can be computed as described in section 1.3 using one of the methods described in section 1.2. We suggest applying the inverse normal combination function $\mathcal{C}$ given in 5, where $w_1 = \sqrt{\frac{\xi_1}{\xi_2}}$, $w_2 = \sqrt{\frac{\xi_2-\xi_1}{\xi_2}}$ and ξ_1 and ξ_2 are the planned number of events among all treatment arms after stage 1 and stage 2 respectively. The decision function is given in (4), where $\mathcal{C}$, α_0, α_1 and c satisfy the level equation (3). If the observed numbers of events d_{i1} and d_{i2} are equal to, respectively, the planned numbers of events ξ_{i1} and ξ_{i2}, then the inverse normal test statistic at the end of the study equals the log-rank test statistic. In order to preserve the type I error rate, the weights w_1 and w_2 must be pre-fixed and remain unchanged throughout the study.

Schäfer & Müller [21] extended the adaptive design methodology to time-to-event data by using the conditional error approach [13] and applying the independent increments structure of the log-rank statistics described above. But Bauer & Posch [22] pointed out that caution must be exercised when applying these flexible designs: in the particular setting of a two-stage design with treatment selection after the first stage, it means that stage 2 p-values for the comparison of the selected treatments with the control at the end of the study cannot be influenced by surrogate information of stage 1 patients who are at risk at the interim treatment selection analysis. Therefore, the standard testing procedure of Schäfer & Müller [21] that combines stagewise p-values computed from increments of the log-rank statistic (see also [17]) may not protect the type I error rate unless treatment selection is based only on the score statistics of the primary endpoint, which is unrealistic in practice.

In the context of enrichment designs, Jenkins et al. [23] proposed a testing procedure that preserves type I error rate while allowing all information collected until the interim analysis to be used for treatment selection. König et al. [24] suggested a testing procedure based on a modification of the classical Dunnett [16] test which guarantees strict control of type I error rate without imposing any restrictions on the information that can be used at the interim analysis for treatment selection purposes. The method by König et al. [24] has, however, not been developed for time-to-event data so far. Di Scala and Glimm [25] extended the classical Dunnett [16] test to time-to-event data for adaptive seamless designs with interim treatment selection.

In the second article of this thesis, the planning of an oncology clinical study with an adaptive seamless phase II/III design is discussed. Two regimens of an experimental treatment are compared to a control at an interim analysis and the most-promising regimen is selected to continue, together with control, until the end of the study. Since the study's primary endpoint will be immature at the regimen selection analysis, it is of interest to investigate whether the incorporation of surrogate information can help improving the regimen selection process and thus the study's probability of success. To this end, designs are considered which include the primary as well as surrogate endpoints (e.g. exposure to treatment) in the regimen selection analysis. At the end of the study, testing of efficacy is carried out to compare the selected regimen to the control with respect to the primary endpoint, utilizing all relevant data collected both before and after the interim analysis. Since the operating characteristics of these designs depend on the specific regimen selection rules considered, on the correlation between primary and surrogate endpoints and on the true standardized mean difference of the surrogate endpoint(s), benchmark scenarios are proposed in which a perfect surrogate and no surrogate is used at the regimen selection analysis. The operating characteristics of these benchmark scenarios provide a range where those of the actual study design are expected to lie.

The above three approaches ([21],[23],[24]) were assessed with regard to power and type I error rate for testing the primary null hypothesis comparing the selected regimen to the control at the end of the study. The standard testing procedure that combines stage-wise p-values based on the independent increments property of the log-rank statistic (Schäffer & Müller [21], Wassmer [17]) did not protect the type I error when treatment selection was based on exposure and correlation between exposure and survival was high. One possibility would be to adjust the significance level to achieve type I error control to the specified level α for the perfect surrogate approach, which provides an upper bound for the type I error. This adjustment will of course affect the power of the study. The procedure proposed by Jenkins et al. [23] protects the type I error when testing the primary endpoint as long as the follow-up of stage 1 patients remains unchanged after the regimen selection analysis, which may be difficult to achieve in practice. If, for instance, it is decided to discontinue treatment in the dropped arm (e.g. for ethical reasons), this may have an impact on the type I error rate: (i) because the p-value for the comparison of the dropped regimen with the control may no longer be uniformly distributed under the null hypothesis (if e.g. these patients drop out and get an alternative treatment) and (ii) because this adaptation may affect the time point of the final analysis. The conservative Dunnett test procedure suggested in [24] protects the type I error without placing any strong restrictions to the study design. It only requires that the recruitment of the second stage patients is independent from the interim data. It can also be extended to a procedure for testing primary and secondary endpoints which protects the multiple type I error rate. Moreover, its conservatism provides a safeguard against unintended type I error inflations, for instance, due to unintended changes in the recruitment rate. Changes to treatment and/or follow-up of patients in the dropped arm do not affect the type I error as these patients are not used in the final analysis. The conservative Dunnett test also allows construction of simultaneous confidence intervals. A disadvantage of the Dunnett procedure is that it does not permit data driven changes of the preplanned overall event number. All three procedures have similar performance with regard to the power to reject the primary hypothesis for the comparison of the selected treatment to the control at the end of the study.

We have not found a clear advantage with regard to power in using adaptive designs compared to a 3-arm phase III design. However, even though the probability of correct treatment selection is maximized with the 3-arm phase III design, an adaptive design allows us to select a treatment early, avoiding an unnecessary large recruitment of patients to the non-selected arm, which is important from an ethical and practical point of view. If efficacy differences in survival are reflected by substantial differences in exposure, then a selection rule based on exposure can improve power compared to a rule based on survival only. However, power could be dramatically reduced when the surrogate endpoint is not able to select the correct treatment with high probability, independently of how highly correlated the primary and surrogate endpoints may be. A solution to this problem could be selection rules that incorporate both, exposure and the immature primary endpoint. How to combine these two endpoints in a sensible selection rule is an interesting open research question. Treatment selection rules should always include the primary endpoint independently how immature the primary endpoint may be. In practice, selection rules will also incorporate safety parameters such as overall death rates, adverse events, serious adverse events, treatment withdrawals and exposure to treatment, which may be important drivers in the dosing decision. One design feature we have not considered in our investigation is the possibility to continue with both experimental arms to the end of the study if interim results do not show a clear advantage for one or the other. In this situation, the conservative Dunnett procedure could inflate the type I error because the overall number of events could not be pre-fixed.

2 Estimation

2.1 One-stage design

Consider the comparison of $k \geq 2$ treatments in a single-stage design. At the end of the study, we select the treatment that has the largest observed effect and we wish to estimate the effect of this treatment. We assume that we have n independent observations $X_{ij} \sim N(\theta_i, \sigma^2), j = 1, ..., n$ in each treatment group $i, i = 1, ..., k$ and that σ^2 is the common and known variance. For simplicity, we do not consider a control treatment but the methods presented here can be extended to designs with a control arm. We let $X_i = \sum_{j=1}^{n} X_{ij}/n$ be the sample mean of the observations in group i. We denote with $S \in \{1, ..., k\}$ the index of the selected treatment so that $X_S = \max\{X_1, ..., X_k\}$. The objective is to estimate θ_S, which is a random variable that depends on $X_1, ..., X_n$. The performance of a generic estimator Q_S of θ_S can be assessed by the selection bias $b_{\underline{\theta}}(Q_S) = E_{\underline{\theta}}[Q_S - \theta_S]$ and the selection mean squared error $\text{MSE}_{\underline{\theta}}(Q_S) = E_{\underline{\theta}}\left[(Q_S - \theta_S)^2\right]$, as it was originally introduced by Putter & Rubinstein [26]. Both quantities depend on the unknown mean vector $\underline{\theta} = (\theta_1, ..., \theta_k)$.

The maximum-likelihood estimator (MLE) for the effect of the selected treatment, θ_S, is the sample mean of the selected treatment, X_S, which ignores the fact that a selection has been performed. This estimator has well known undesirable properties. It has positive selection bias and it is highly missleading when all true treatment effects are equal and the number of treatments being compared is large, as $X_S \to \infty$ in probability when $k \to \infty$ for $\theta_1 = ... = \theta_k$. On the one hand, Putter & Rubinstein [26] showed that there is no unbiased estimator for θ_S. On the other hand, Brown [27] proved that X_S is an admissible estimator of θ_S; that is, no other estimator of θ_S can have smaller MSE than X_S for all values of the parameters $\theta_1, ..., \theta_k$. Therefore, the aim should be to find estimators of θ_S that have smaller bias than the MLE and that at the same time improve upon the MLE with respect to the MSE in realistic scenarios.

In order to better assess the magnitude of the problem, it is useful to know for which values of the parameters the selection bias and MSE of the MLE will be largest. Cohen & Sackrowitz [28] showed that the MSE is largest when all true effects are equal. In the first paper of this thesis, we prove that selection bias is also maximal in this case.

Hwang [29] focused on single-stage multi-armed trials with $k \geq 4$ experimental treatments and proposed Lindley's estimator [30] for estimating the mean θ_S of the treatment S with largest average. This gives the estimator

$$Q_S^{\text{L}} = \hat{C}_+ X_S + (1 - \hat{C}_+)\bar{X} \quad \text{where} \quad \hat{C}_+ = \max\left(\hat{C}, 0\right) \quad \text{with} \quad \hat{C} = 1 - \frac{(k-3)\sigma^2}{n\sum_{j=1}^{k}(X_j - \bar{X})^2}\,. \tag{6}$$

Lindley's estimator shrinks the conventional MLE X_S towards the overall sample mean $\bar{X} = \sum_{i=1}^{k} X_i/k$.

One can see by heuristic arguments that this estimator is reasonable. For $\theta_1 = ... = \theta_k$, the overall mean $\bar{X}$ is the most-efficient estimator for θ_S and $n\sum_{j=1}^{k}(X_j - \bar{X})^2/\sigma^2 \sim \chi^2(k-1)$ has its mode at $k-3$. Therefore $\hat{C}_+$ is likely to be close to 0 in which case shrinkage to $\bar{X}$ is strong. If $\theta_1, ..., \theta_k$ are far from each other, i.e. $\max_{i \neq j} |\theta_i - \theta_j|$ is large, then $\sum_{j=1}^{k}(X_j - \bar{X})^2$ is likely to be large and $\hat{C}_+$ is likely to be close to 1. As a consequence Q_S^{L} is close to X_S which is reasonable because the bias of X_S is small for large $\max_{i \neq j} |\theta_i - \theta_j|$.

Hwang considered the problem from an empirical Bayesian point of view and showed that Q_S^{L} uniformly improves X_S within this frame work. To this end, he assumed conjugate normal priors $\theta_i \sim N(\mu, \tau^2)$ and that $X_i \,|\, \theta_i \sim N(\theta_i, \sigma^2/n)$, $i = 1, ..., k$, for a known variance σ^2. These assumptions correspond to a random effects model for the replicates X_{ij} with treatment as a random factor and a known residual variance σ^2. The Bayes estimator for θ_S is

$$Q_S^{\text{B}} = E\left[\theta_S \,|\, X_1, ..., X_k\right] = CX_S + (1-C)\mu\,, \quad \text{where} \quad C = n\tau^2/(\sigma^2 + n\tau^2) = 1 - \sigma^2/(\sigma^2 + n\tau^2)\,. \tag{7}$$

The estimator Q_S^{B} coincides with the best linear unbiased predictor (BLUP) for the random effects model and is known to have minimal Bayes risk

$$R_{\mu,\tau}(Q_S^{\text{B}}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} E_\theta\left(Q_S^{\text{B}} - \theta_S\right) \phi\left(\frac{\theta_1 - \mu}{\tau^2}\right) \cdots \phi\left(\frac{\theta_k - \mu}{\tau^2}\right) d\theta_1 \cdots d\theta_k\,.$$

The Bayes risk is the integral of $\text{MSE}_\theta(Q_S^{\text{B}})$ with regard to the normal prior distribution for $\theta_1, \ldots, \theta_k$. Replacing in (7) the unknown μ with its mean unbiased estimator $\bar{X}$ and the unkown C with the positive part

of $\hat{C} = 1 - (k-3)\sigma^2/[n\sum_{j=1}^{k}(X_j - \bar{X})^2]$ ($\hat{C}$ can be shown to be mean unbiased for C) leads to Lindley's estimator (6). Hwang verified the following domination result.

Theorem 2.1 *If $k \geq 4$, then Q_S^L has uniformly smaller Bayes risk than X_S, i.e. $R_{\mu,\tau}(Q_S^L) \leq R_{\mu,\tau}(X_S)$ for all $N(\mu, \tau^2)$-distributed priors.*

2.2 Two-stage design

Consider now two-stage designs with treatment selection at the end of stage 1. Let $X_1, ..., X_k$ be the stage 1 sample means, based on a sample size of n_1 subjects per group (equal for all treatments). Assume that S is the index of the best-performing treatment at the interim analysis, i.e. $X_S = \max(X_1, ..., X_k)$. Suppose further that n_2 subjects are recruited in the selected treatment arm in the second stage. Let Y_S be the stage 2 sample mean for the selected treatment. The overall MLE at the end of the study can be written as $Q_S^{\text{MLE}} = tX_S + (1-t)Y_S$ where $t = n_1/(n_1 + n_2)$. Note that $b_{\underline{\theta}}(Q_S^{\text{MLE}}) = tb_{\underline{\theta}}(X_S)$ because Y_S is an unbiased estimator of θ_S. Therefore, the larger the first stage sample size the larger the bias. Note that the larger the first stage sample size the better the treatment selection at the end of stage 1 because of the larger amount of information to make the treatment selection. This dichotomy between bias and selection was pointed out by Bauer et al [31].

Cohen & Sacrowitz [32] proposed a two-stage estimator of the mean of the best-performing treatment that is conditionally unbiased given the order statistics of the first-stage sample means and has uniformly minimum variance among all such conditionally unbiased estimators. Their estimator is of course also unconditionally unbiased. Bowden & Glimm [33] extended Cohen & Sackrowitz's work to unequal group variances and selecting not only the most-promising but also j-th most-promising treatment for $j \leq k$.

Stallard & Todd [34] considered the mean bias of the MLEs conditional on the selection of treatment S, i.e. $b_\theta(i; S) = E_\theta(Q_i^{\text{MLE}}|S)$ for $i = 1, \ldots, k$, and derived numerical expressions for these conditional selection biases. They suggested to subtract the selection bias $b_\theta(i; S)$ from each MLE Q_i^{MLE}. Since bias depends the unknown $\theta = (\theta_1, \ldots, \theta_k)$, only an estimator of the conditional biases can be subtracted. Stallard & Todd proposed an iterative approach similar to the iterative scheme in Whitehead [35].

In the first paper of this thesis, we extend Hwang's domination result [29] to two-stage designs with treatment selection whereby we focus on the case of selecting a single experimental treatment at the interim analysis. We define for $k \geq 4$ the following two-stage version of Lindley's estimator

$$Q_S^{\text{L}} = tQ_S^{\text{L},1} + (1-t)Y_S \tag{8}$$

where $Q_S^{\text{L},1}$ is Lindley's estimator (6) from the first stage data. Since the second stage continues with only one experimental treatment there is no need to adjust the second stage MLE. We show that Hwang's domination result also applies to the two-stage version (8) of Lindley's estimator.

Theorem 2.2 *If $k \geq 4$, then Lindley's two-stage estimator (8) has uniformly smaller Bayes risk than Q_S^{MLE}, i.e. $R_{\mu,\tau}(Q_S^L) \leq R_{\mu,\tau}(Q_S^{\text{MLE}})$ for all $N(\mu, \tau^2)$-distributed priors.*

Lindley's estimator is defined only for $k \geq 4$ whereas the Bayes estimator (7) is defined and optimal with regard to the Bayes risk for all $k \geq 2$. The restriction of Lindley's estimator to $k \geq 4$ is caused by the use of the mean unbiased estimator $\hat{C}$ for C which is defined only for $k \geq 4$. It appears natural to consider the Bayes estimator with another estimator for C when $k = 2, 3$. To this end, we suggest to adopt the random effects model point of view and use for the first stage estimator the standard estimator of the BLUP which is defined for all $k \geq 2$ and equals Lindley's estimator (6) with $k - 3$ replaced by $k - 1$ (see e.g. Searle, Casella and McCulloch [36]). More precisely, we consider replacing Lindley's estimator $Q_S^{\text{L},1}$ in (8) by the estimated BLUP from the random effects model for the first stage data. No domination result is known for this type of estimator.

We have investigated by simulation studies the performance of the two-stage version of Lindley's estimator and the estimated BLUP and have found favorable properties of these estimators in comparison to the MLE and the bias-adjusted estimators of Cohen & Sackrowitz [32] and Stallard & Todd [34]. Whereas Cohen & Sackrowitz's estimator perfectly removes the selection bias, it has increased variance in a way that the mean square error is generally larger than for the MLE. Stallard & Todd's estimator was found to have the tendency for an overcorrection of bias, at least in the two-stage setting, and it is also in general inferior to the MLE and shrinkage estimators in terms of mean square error. The two-stage shrinkage estimators did, on the one hand, reduce selection bias of the MLE substantially (although not removing it completely) and,

on the other hand, improved or was equivalent to the MLE in terms of mean square error. Since we believe that mean square error is more appropriate as measure of precision than the bias itself, we suggest using shrinkage estimators to deal with the problem of selection bias.

References

[1] Dragalin, V. *Adaptive designs: terminology and classification.* Drug Information Journal 2006; **40** : 425-435.

[2] *Reflection paper on methodological issues in confirmatory clinical trials with flexible design and analysis plan.* Committe for Medicinal Products for Human Use (CHMP), European Medicines Agency (EMEA), 2006. Doc. Ref. CHMP/EWP/2459/02.

[3] *Guidance for industry: adaptive design clinical trials for drugs and biologics.* U.S. Department of Health and Human Services, Food and Drug Administration (FDA), Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), 2010.

[4] Bauer P. and Brannath W. *The advantages and disadvantages of adaptive designs for clinical trials.* Drug Discovery Today 2004; **9** : 351-357.

[5] Bretz F., König F., Brannath W., Glimm E. and Posch M. *Adaptive designs for confirmatory clinical trials.* Statistics in Medicine 2009; **28** : 1181-1217.

[6] Bauer P. and Kieser M. *Combining different phases in the development of medical treatments within a single trial.* Statistics in Medicine 1999; **18** : 1833-1848.

[7] Hommel G. *Adaptive modifications of hypotheses after an interim analysis.* Biometrical Journal 2001; **43** : 581-589.

[8] Posch M., König F., Branson M., Brannath W., Dunger-Baldauf C. and Bauer P. *Testing and estimation in fexible group sequential designs with adaptive treatment selection.* Statistics in Medicine 2005; **24** : 3697-3714.

[9] Brannath, W., Posch, M. and Bauer, P. *Recursive combination tests*. Journal of the American Statistical Association 2002; **97** : 236-244.

[10] Bauer, P. (1989). *Multistage testing with adaptive designs (with discussion).* Biometrie und Informatik in Medizin und Biologie 1989; **20** : 130-148.

[11] Bauer, P. and Köhne, K. *Evaluation of experiments with adaptive interim analyses.* Biometrics 1994; **50** : 1029-1041.

[12] Fisher R. A. *Statistical methods for research workers.* London: Oliver & Boyd.

[13] Proschan, M.A. and Hunsberger, S.A. *Designed extension of studies based on conditional power.* Biometrics 1995; **51** : 1315-1324.

[14] Lehmacher, W., and Wassmer, G. *Adaptive sample size calculations in group sequential trials*. Biometrics 1999; **55** : 1286-1290.

[15] Marcus R., Peritz E., Gabriel K.R. *On closed testing procedures with special reference to ordered analysis of variance*. Biometrika 1976; **63** : 655-660.

[16] Dunnett D.W. *A multiple comparison procedure for comparing several treatments with a control*. Journal of the American Statistical Association 1955; **50** : 1096-1121.

[17] Wassmer G. *Planning and analyzing adaptive group sequential survival trials*. Biometrical Journal 2006; **48** : 714-729.

[18] Schoenfeld D. A. *The asymptotic properties of nonparametric tests for comparing survival distributions.* Biometrika 1981; **68** : 316-319.

[19] Tsiatis, A. A. (1981). *The asymptotic joint distribution of the efficient scores test for the proprtional hazards model calculated over time.* Biometrika 1981; **68** : 311-315.

[20] Tsiatis, A. A. (1982). *Repeated significance testing for a general class of statistics used in censored survival analysis.* Journal of the American Statistical Association 1982; **77** : 855-861.

[21] Schäfer H. and Müller H. *Modication of the sample size and the schedule of interim analyses in survival trials based on data inspections*. Statistics in Medicine 2001; **20** : 3741-3751.

[22] Bauer P. and Posch M. *Letter to the Editor: Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections.* Statistics in Medicine 2004; **23** : 1333-1335.

[23] Jenkins M., Stone A. and Jennison, C. *An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints*. Pharmaceutical Statistics 2011; **10** : 347-356.

[24] König F., Brannath W., Bretz F. and Posch M. *Adaptive Dunnett tests for treatment selection*. Statistics in Medicine 2008; **27** : 1612-1625.

[25] Di Scala L. and Glimm E. *Time-to-event analysis with treatment arm selection at interim*. Statistics in Medicine 2011; **30** : 3067-3081.

[26] Putter J, Rubinstein D. *On estimating the mean of a selected population.* Technical Report No 165, University of Wisconsin, Department of Statistics, 1968.

[27] Brown L. D. *Personal communication to J. T. Hwang.* 1987.

[28] Cohen A., Sackrowitz H.B. *Estimating the mean of the selected population.* Statistical Decision Theory and Related Topics III 1982; **1** : 243-270.

[29] Hwang J.T. *Empirical Bayes estimation for the means of the selected populations.* The Indian Journal of Statistics 1993; **55** : 285-311.

[30] Lindley D. V. *Discussion of Professor Stein's paper "Confidence sets for the mean of a multivariate normal distribution"*. Journal of the Royal Statistical Society, Serie B 1962; **24** : 265-296.

[31] Bauer P, Koenig F, Brannath W, Posch M. *Selection and bias - two hostile brothers*. Statistics in Medicine 2010; **29** : 1-13.

[32] Cohen A., Sackrowitz H.B. *Two stage conditionally unbiased estimators of the selected mean.* Statistics & Probability Letters 1989; **8** : 273-278.

[33] Bowden J., Glimm E. *Unbiased estimation of selected treatment means in two-stage trials.* Biometrical Journal 2008; **4** : 515-527.

[34] Stallard N., Todd S. *Point estimators and confidence regions for sequential trials involving selection.* Journal of Statistical Planning and Inference 2005; **135** : 402-419.

[35] Whitehead J. *On the bias of maximum-likelihood estimation following a sequential test.* Biometrika 1986; **73** : 573-81.

[36] Searle, S.R., Casella G. and McCulloch C. *Variance Components.* Wiley Series in Probability and Statistics 1992.

Special Issue Paper

Statistics in Medicine

Received 15 May 2011, Accepted 1 April 2012 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5463

Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection‡

Máximo Carreras[a] and Werner Brannath[b]*†

We consider the problem of estimation in adaptive two-stage designs with selection of a single treatment arm at an interim analysis. It is well known that the standard maximum-likelihood estimator of the selected treatment is biased. We prove that selection bias of the maximum-likelihood estimator is maximal when all treatment effects are equal and the most-promising treatment is selected. Furthermore, we consider shrinkage estimation as a solution for the selection bias problem. We thereby extend previous work of Hwang on Lindley's estimator for single-stage multi-armed trials with four or more treatments and post-trial treatment selection. Following Hwang's ideas, we show that a simple two-stage version of Lindley's estimator has uniformly smaller Bayes risk than the maximum-likelihood estimator when assuming an empirical Bayesian framework with independent normal priors for the group means. For designs that start with two or three treatment groups, we suggest using a two-stage version of the common estimator of the best linear unbiased predicator of the corresponding random effects model. We show by an extensive simulation study that the shrinkage estimators perform well compared with maximum-likelihood and previously suggested bias-adjusted estimators in terms of selection bias and mean squared error.

Keywords: best linear unbiased predictor (BLUP); Bayes risk; clinical trial; empirical Bayes; Lindley's estimator; random effects model; seamless phase II/III design; selection bias; selection mean squared error

1. Introduction

There has been increasing interest over the last years in two-stage clinical trial designs in which $k \geq 2$ experimental treatments are compared with a control in an interim analysis after a first stage and the most-promising treatments are selected at the interim analysis to continue, together with the control, into a second stage (for reviews, see, e.g., [1, 2]). Hypothesis testing methods have been developed that allow the treatment selection rule to be defined at the interim analysis integrating all available information, including information external to the study, without affecting the overall multiple type I error rate [3–6]. Whereas testing can be viewed as a resolved issue, the equally important task of parameter estimation still lacks a satisfactory solution (e.g., [5, 7]). In this article, we provide a solution to the important problem of point estimation of the effect of the selected treatments at the end of such trials by considering shrinkage methods.

It is well known that maximum-likelihood estimators (MLEs) exhibit statistical bias, and therefore, attempts have been undertaken to remove or reduce bias of treatment effect estimators in designs with midtrial treatment selection [8–11]. Although these methods are successful in removing bias [8, 11] or at least reducing it substantially [9, 10], they do not provide estimators with generally improved mean squared error (MSE) compared with the MLE. We will illustrate by simulation results that they can even

[a] *F. Hoffmann-La Roche AG, Malzgasse 30, 4052 Basel, Switzerland*
[b] *Competence Center for Clinical Trials Bremen, Faculty 3, University of Bremen, Bremen, Germany*
**Correspondence to: Werner Brannath, Competence Center for Clinical Trials Bremen, Faculty 3, University of Bremen, Bremen, Germany.*
† *E-mail: brannath@uni-bremen.de*
‡ *Supporting information may be found in the online version of this article.*

lead to an inflation of the MSE. Because the MSE is a combination of bias and variance and more reasonably describes the precision of an estimator than the bias by itself, this implies that bias reduction with the current methods may come for a too high price of a substantial increase in the variance and these methods are not recommendable for practical use. The focus of our work is on utilizing estimation methods that reduce the bias and, equally important, control the MSE of estimation.

There are two sources of statistical estimation bias in designs with interim treatment selection. One type of bias is produced by the data-dependent sample size induced by the treatment selection process. Because recruitment to a specific treatment arm is stopped if the observed interim effect is small (namely, smaller than the maximum treatment effect), there is a tendency to underestimate the treatment effect of that arm. This is similar to the bias that is produced by a stopping for futility rule. This negative bias applies to each of the k treatment estimators and becomes apparent when we report effect estimates for all treatments simultaneously, irrespective of which treatments have been selected, with the use of the interim estimate for the dropped treatments. We call this negative bias the *always-reporting bias* (see [7] for simulation results).

Another type of bias is caused by the selection process itself: when selecting the apparently most-efficient treatment at the interim analysis, the interim estimator of the selected treatment equals the maximum of all interim estimators and hence is positively biased. This positive bias is carried forward to the second stage when using data from both stages in the final estimator. This type of bias is called *selection bias*. It usually remains relevant even though it is diluted by the independent second-stage sample for the selected treatment (e.g., [7]). In contrast to the negative always-reporting bias, the positive selection bias indicates that treatment effects are overestimated.

The contradictory sign of always-reporting and selection bias may appear confusing. However, it has a simple and important implication. It implies that selection bias is solely caused by our focus on the selected treatment and our neglect of the other treatment effect estimators. We may therefore ask whether we can improve the MLE of the selected treatment with regard to bias and MSE by using an estimator that is computed from all available information including the data from the dropped treatment groups. If the different treatments are different doses of a medicine, then a way to utilize all data is to fit a dose–response model to the complete data and predict the mean efficacy of the selected treatment from this model. In this paper, we follow an alternative and more general approach, namely, we consider using a shrinkage estimator.

Our focus on shrinkage methods has been motivated by two references. The first reference is Chapter 4.5 of Harrell's modeling strategy book [12, p. 62–63] in which the selection bias problem is used as a motivating example for shrinkage techniques in regression analysis. He considered k group means that satisfy the analysis of variance assumptions and draws a plot of the ordered group means against their ranks. This always gives an increasing curve whose slope is strongly up biased because of selection bias. Shrinkage in the regression modeling context means to shrink down the slope towards zero to reduce the selection bias. This points to an interesting connection between selection bias and the problem of overfitting. The other reference is a paper of Hwang [13] who considered Lindley's shrinkage estimator [14] in the same context and showed strong preferable properties with regard to the MSE in comparison with the MLE. We will review Hwang's result in Section 2. Both references apply to the setup of multi-arm clinical trials where treatments are selected at the end of the trial, that is, designs with a single stage. In this article, we propose an extension of shrinkage estimation to two-stage designs with midtrial treatment selection.

We organize the paper as follows. In Section 2, we review the definition of selection bias and selection MSE in the context of single-stage multi-armed trials where treatment selection is performed at the end of the trial. We also prove that selection bias of the MLE is maximal when all treatment effects are equal and the best-performing treatment is selected. This fact has frequently been observed; however, no proof has been given yet. We then review Hwang's result on Lindley's estimator. In Section 3, we introduce a two-stage version of the shrinkage estimator and show that Hwang's result essentially remains valid for two-stage designs with midtrial selection of a single treatment. In Section 4, we present simulation results and conclude with a discussion in Section 5. We provide proofs in Appendix A.

2. Selection bias and mean squared error in single-stage designs

To introduce the problem and to review Hwang's result, we start by considering a randomized single-stage design with k parallel treatment arms where the apparently most-efficacious treatment is selected at the end of the trial. We assume n independent observations $X_{ij} \sim N(\theta_i, \sigma^2)$, $j = 1, \ldots, n$, in

each treatment group i, where σ^2 is the common and known variance. For simplicity, we do not consider a control treatment, but the methods proposed in this article can easily be extended to designs with a control arm. We will discuss this in Section 5. We let $X_i = \sum_{j=1}^{n} X_{ij}/n$ be the sample mean of the observations in group i. Denoting by $S \in \{1,\ldots,k\}$ the index of the selected treatment, we have $X_S = \max(X_1,\ldots,X_k)$. The objective is to estimate or rather predict θ_S, which is a random variable that depends on $X_1,\ldots,X_k$. The performance of a generic estimator Q_S of θ_S will be assessed by the selection bias $b_\theta(Q_S) = E_\theta[Q_S - \theta_S]$ and the selection mean squared error $\mathrm{MSE}_\theta(Q_S) = E_\theta[(Q_S-\theta_S)^2] = \mathrm{Var}_\theta(Q_S) + b_\theta^2(Q_S)$, as originally introduced in [15]. Both quantities depend on the unknown mean vector $\theta = (\theta_1,\ldots,\theta_k)$.

2.1. *Maximum-likelihood estimator of θ_S*

The MLE for the effect of the selected treatment, θ_S, is the sample mean of the selected treatment, X_S, which ignores the fact that a previous selection has been performed. This estimator has well-known undesirable properties. It has positive selection bias and is highly misleading when all true treatment effects are equal and the number of treatments being compared is large, as $X_S \to \infty$ in probability when $k \to \infty$ for $\theta_1 = \cdots = \theta_k$. On the other hand, Brown (personal communication to Hwang, [13]) proved that X_S is an admissible estimator of θ_S; that is, no other estimator of θ_S can have smaller MSE than X_S for all values of the parameters $\theta_1,\ldots,\theta_k$. Therefore, the aim should be to find estimators of θ_S that improve upon the MLE with respect to the MSE in realistic scenarios.

To better assess the magnitude of the problem, it is useful to know for which values of the parameters the selection bias and MSE of the MLE will be largest. Cohen and Sackrowitz [16] proved that the MSE is largest when all true effects are equal. We present here a theorem that states that the selection bias is also maximal in this case. Although intuitively clear and observed in numerous simulation studies (e.g., [7]), no proof of this fact has been given before. We provide a simple proof of this theorem in Appendix A.

Theorem 2.1
The bias $b_\theta(X_S) = E[X_S - \theta_S]$ of X_S is maximal when $\theta_1 = \cdots = \theta_k$.

The same theorem applies to two-stage designs, where at an interim analysis, a single treatment is selected and recruitment to this group (and the control) is continued up to a prespecified sample size. This follows from the fact that, in such two-stage designs, the second-stage MLE (computed from the second-stage data only) is unbiased, and hence the selection bias of the overall MLE, which is a weighted mean of the first-stage and second-stage MLE, is solely driven by the bias of the first-stage MLE.

2.2. *Shrinkage estimator of θ_S for $k \geq 4$ treatments*

Hwang [13] focused on multi-armed trials with $k \geq 4$ experimental treatments and proposed Lindley's estimator [14] for estimating the mean θ_S of the treatment S with largest average. This gives the estimator

$$Q_S^{\mathrm{L}} = \hat{C}_+ X_S + (1-\hat{C}_+)\bar{X}, \quad \text{where} \quad \hat{C}_+ = \max(\hat{C},0) \quad \text{with} \quad \hat{C} = 1 - \frac{(k-3)\sigma^2}{n\sum_{j=1}^{k}(X_j-\bar{X})^2}. \quad (1)$$

Lindley's estimator shrinks the conventional MLE X_S towards the overall sample mean $\bar{X} = \sum_{i=1}^{k} X_i/k$.

One can see by heuristic arguments that this estimator is reasonable. For $\theta_1 = \cdots = \theta_k$, the overall mean $\bar{X}$ is the most efficient estimator for θ_S and $n\sum_{j=1}^{k}(X_j-\bar{X})^2/\sigma^2 \sim \chi^2(k-1)$ has its mode at $k-3$. Therefore, $\hat{C}_+$ is likely to be close to 0 in which case shrinkage to $\bar{X}$ is strong. If $\theta_1,\ldots,\theta_k$ are far from each other, that is, $\max_{i\neq j}|\theta_i-\theta_j|$ is large, then $\sum_{j=1}^{k}(X_j-\bar{X})^2$ is likely to be large and $\hat{C}_+$ is likely to be close to 1. As a consequence, Q_S^{L} is close to X_S, which is reasonable because the bias of X_S is small for large $\max_{i\neq j}|\theta_i-\theta_j|$.

Hwang considered the problem from an empirical Bayesian point of view and showed that Q_S^{L} uniformly improves X_S within this framework. To this end, he assumed conjugate independent and identically distributed (i.i.d.) normal priors $\theta_i \sim N(\mu,\tau^2)$ and that $X_i \mid \theta_i \sim N(\theta_i,\sigma^2/n)$, $i = 1,\ldots,k$,

for a known variance σ^2. These assumptions correspond to a random effects model for the replicates X_{ij} with treatment as a random factor and a known residual variance σ^2. The Bayes estimator for θ_S is

$$Q_S^{\mathrm{B}} = E\left[\theta_S \,|X_1, \ldots, X_k\right] = CX_S + (1-C)\mu\,, \quad \text{where} \quad C = n\tau^2/(\sigma^2+n\tau^2) = 1-\sigma^2/(\sigma^2+n\tau^2)\,. \tag{2}$$

The estimator Q_S^{B} is an unbiased predictor for θ_S in the sense that

$$\begin{aligned} E(Q_S^{\mathrm{B}}) &= E\{E(\theta_S|X_1, \ldots, X_k)\} = E(\theta_S) \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\sum_{i=1}^{k} \theta_i \; P_\theta(S=i)\right) \phi\left(\frac{\theta_1-\mu}{\tau^2}\right) \cdots \phi\left(\frac{\theta_k-\mu}{\tau^2}\right) \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_k, \end{aligned}$$

where the expectations are with respect to the data and i.i.d. normal priors as indicated for $E(\theta_S)$. It coincides with the best linear unbiased predictor (BLUP) for the random effects model and is known to have minimal Bayes risk

$$R_{\mu,\tau}(Q_S^{\mathrm{B}}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} E_\theta\left\{\left(Q_S^{\mathrm{B}} - \theta_S\right)^2\right\} \phi\left(\frac{\theta_1-\mu}{\tau^2}\right) \cdots \phi\left(\frac{\theta_k-\mu}{\tau^2}\right) \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_k\,.$$

The Bayes risk is the integral of $MSE_\theta(Q_S^{\mathrm{B}})$ with regard to the normal prior distribution for $\theta_1, \ldots, \theta_k$. Replacing in (2) the unknown μ with its mean unbiased estimator $\bar{X}$ and the unknown C with the positive part of $\hat{C} = 1-(k-3)\sigma^2/\left[n\sum_{j=1}^{k}(X_j-\bar{X})^2\right]$ ($\hat{C}$ can be shown to be mean unbiased for C) leads to Lindley's estimator (1). Hwang verified the following domination result.

Theorem 2.2
If $k \geqslant 4$, then Q_S^{L} has uniformly smaller Bayes risk than X_S, that is, $R_{\mu,\tau}(Q_S^{\mathrm{L}}) \leqslant R_{\mu,\tau}(X_S)$ for all i.i.d. $N(\mu, \tau^2)$-distributed priors.

We finally comment on the relationship between this and Brown's [13] admissibility result for the MLE. Brown follows the frequentist's point of view and considers the MSE for each parameter configuration asking for an estimator that dominates the MLE uniformly with regard to all mean configurations. Hwang takes the Bayesian's point of view and considers the Bayes risk where the MSE is averaged over the parameter space with regard to i.i.d. normal priors. Such averaging allows him to find an estimator (namely, Lindley's estimator) that outperforms the MLE with regard to the Bayes risk for all i.i.d. normal priors, even though according to Brown, no estimator can outperform the MLE with respect to the MSE for all parameter configurations.

2.3. *Extensions*

Hwang considered several extensions of the aforementioned domination result. One extension is for the case of unknown variance σ^2. He showed that the domination theorem 2.2 remains true if we replace in Q_S^{L} the unknown σ^2 by the estimator $\hat{\sigma}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-X_i)^2/\{k(n-1)+2\}$. The latter estimator is biased but more efficient (in terms of MSE) than the unbiased estimator $s^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij}-X_i)^2/\{k(n-1)\}$.

Hwang also considered estimation of all ordered group means $\theta_{(1)}, \ldots, \theta_{(k)}$ where the ordering is according to the MLEs, that is, $X_{(1)} < X_{(2)} < \cdots < X_{(k)}$. He showed the same domination result for Lindley's estimator $Q_{(i)}^{\mathrm{L}}$ of $\theta_{(i)}$ for all $i \leqslant k$, where $Q_{(i)}^{\mathrm{L}}$ is defined as in (1) but with S equal to the index of the group (i) from the ith smallest MLE. This leads to the question for which selection rules Theorem 2.2 remains true. We investigated Hwang's proof and found that it applies to any selection rule $S = S\left((X_1-\bar{X})/\sqrt{V}, \ldots, (X_k-\bar{X})/\sqrt{V}, W\right)$, where $V = \sum_{j=1}^{k}(X_j-\bar{X})^2$ and W is a random vector that is stochastically independent from the groups means and does not influence the choice of the prior distribution. Examples for W are external data and/or an estimator of the common unknown variance σ^2. We provide the proof of this result in Appendix A.

An interesting application of the latter extension are cases where the selection depends on a multivariate normal outcome with one component representing the primary efficacy variable X and the other components the secondary or safety variables. Because the multivariate normal outcome vector can be decomposed into the vector of the primary variable and the projection of the secondary and safety variables into the orthogonal complement of the primary outcome vector, which is independent from the

primary outcome variable, our extension covers selection rules that depend on the standardized means $(X_i - \bar{X})/\sqrt{V}$ and the secondary and safety endpoints. One particularly interesting case is where the selection depends on the ordering of the means and the secondary and safety variables. This covers the case of sometimes selecting the second (or even third) best-performing treatment (instead of the best-performing one) on the basis of the secondary and safety endpoints.

3. Point estimation in two-stage designs with treatment selection

We consider now two-stage designs with treatment selection at the interim analysis. To this end, we let $X_1, \ldots, X_k$ be the first-stage sample means, based on a sample size of n_1 subjects per group (equal for all treatments). We assume that S is the index of the best-performing treatment at the interim analysis, that is, $X_S = \max(X_1, \ldots, X_k)$, or any other selection rule $S = S\left((X_1 - \bar{X})/\sqrt{V}, \ldots, (X_k - \bar{X})/\sqrt{V}, W\right)$ where W is independent from $X_1, \ldots, X_k$. Suppose further that n_2 subjects are recruited in the selected treatment arm in the second stage. Let Y_S be the second-stage sample mean for the selected treatment. The overall MLE at the end of the two-stage design can be written as $Q_S^{\mathrm{MLE}} = tX_S + (1-t)Y_S$, where $t = n_1/(n_1 + n_2)$.

3.1. Extension of Lindley's estimator

Our goal is to extend Hwang's domination result [13] to two-stage designs with treatment selection whereby we focus on the case of selecting a single experimental treatment at the interim analysis. We define for $k \geqslant 4$ the following two-stage version of Lindley's estimator

$$Q_S^{\mathrm{L}} = tQ_S^{\mathrm{L},1} + (1-t)Y_S, \tag{3}$$

where $Q_S^{\mathrm{L},1}$ is Lindley's estimator (1) from the first-stage data. Because the second stage continues with only one experimental treatment, there is no need to adjust the second-stage MLE. We show in Appendix A that Hwang's domination result also applies to the two-stage version (3) of Lindley's estimator.

Theorem 3.1
If $k \geqslant 4$, then Lindley's two-stage estimator (3) has uniformly smaller Bayes risk than Q_S^{MLE}, that is, $R_{\mu,\tau}(Q_S^{\mathrm{L}}) \leqslant R_{\mu,\tau}(Q_S^{\mathrm{MLE}})$ for all $N(\mu, \tau^2)$-distributed priors.

It can be seen from the proof in Appendix A that the theorem applies whenever the first-stage estimator $Q_S^{\mathrm{L},1}$ has uniformly smaller Bayes risk than the first-stage mean X_S. Hence, the result covers all extensions mentioned in Section 2.3.

3.2. The case $k = 2, 3$

Lindley's estimator is defined only for $k \geqslant 4$, whereas the Bayes estimator (2) is defined and optimal with regard to the Bayes risk for all $k \geqslant 2$. The restriction of Lindley's estimator to $k \geqslant 4$ is caused by the use of the mean unbiased estimator $\hat{C}$ for C, which is defined only for $k \geqslant 4$. It appears natural to consider the Bayes estimator with another estimator for C when $k = 2, 3$. To this end, we suggest to adopt the random effects model point of view and use for the first-stage estimator the standard estimator of the BLUP, which is defined for all $k \geqslant 2$ and equals Lindley's estimator (1) with $k - 3$ replaced by $k - 1$ (e.g., [17]). More precisely, we consider replacing Lindley's estimator $Q_S^{\mathrm{L},1}$ in (3) by the estimated BLUP from the random effects model for the first-stage data. No domination result is known for this type of estimator; however, simulation results shown in the next section indicate that this estimator performs well compared with the MLE and previously proposed bias-adjusted estimators.

4. Simulation results

We have performed an extensive simulation study to investigate the performance of our two-stage version (3) of Lindley's estimator and of the estimated BLUP for $k = 2, 3$ in comparison with the MLE and the two different mean bias-adjusted estimators of Cohen and Sackrowitz [8] and Stallard and Todd [10]. For completeness, we start with a brief description of the two mean bias-adjusted estimators. We then present the simulation results.

4.1. Cohen and Sackrowitz's two-stage estimator of θ_S

Cohen and Sacrowitz [8] proposed a two-stage estimator of the mean of the best-performing treatment that is conditionally unbiased given the order statistics of the first-stage sample means and has uniformly minimum variance among all such conditionally unbiased estimators. Of course, a conditionally unbiased estimator is also unconditionally unbiased. One interesting feature of this estimator is that it depends on the first-stage data only through the first two order statistics of the first-stage sample means. Therefore, for $k \geqslant 3$, the method does not utilize all first-stage data. Hence, there is a potential for an improvement (in terms of MSE) by estimators that utilize all interim data. Our two-stage version of Lindley's estimator and the BLUP utilize all interim data.

Bowden and Glimm [11] extended Cohen and Sackrowitz's work to unequal group variances and selecting not only the most-promising but also jth most-promising treatment for $j \leqslant k$.

4.2. Stallard and Todd's estimator of θ_S

Stallard and Todd [10] considered the mean bias of the MLEs conditional on the selection of treatment S, that is, $b_\theta(i;S) = E_\theta(Q_i^{\mathrm{MLE}}|S)$ for $i = 1,\ldots,k$, and derived numerical expressions for these conditional selection biases. They suggested to subtract the selection bias $b_\theta(i;S)$ from each MLE Q_i^{MLE}. Because bias depends on the unknown $\theta = (\theta_1,\ldots,\theta_k)$, only an estimator of the conditional biases can be subtracted. Stallard and Todd proposed an iterative approach similar to the iterative scheme in Whitehead [18]. Their method is computationally intensive as it requires numerical integration within each iteration. It can also happen that the iterative approach diverges in which case no estimator is provided. We have observed such divergence problems in several realistic scenarios.

4.3. Simulation input

We have considered two-stage designs in which the total sample size for the selected treatment is fixed to $N = 100$ subjects and the standard deviation to $\sigma = 3.155$ such that a difference in means of 1.25 could be detected in a two-treatment comparison with a power of 80%. We have varied the information fraction at the interim analysis by considering a range of first-stage sample sizes from 10 to 100 in steps of 10 subjects. For each scenario, we have run 10^6 iterations for all methods except Stallard and Todd's for which we have run 10^4 iterations because of its extensive numerical calculations. We have considered the procedures of selecting the best-performing and the second best-performing treatments at interim analysis for all methods except Stallard and Todd's, which only works when selecting the best-performing treatment. The results of the simulation study are given in units of the total standard error, that is, in terms of $b(Q_S)/\sqrt{\frac{\sigma^2}{N}} = E[Q-\theta_S]/\sqrt{\frac{\sigma^2}{N}}$ and $\sqrt{MSE(Q_S)}/\sqrt{\frac{\sigma^2}{N}} = \sqrt{E[(Q_S-\theta_S)^2]}/\sqrt{\frac{\sigma^2}{N}}$.

4.4. Selection of the best-performing treatment

On the top row of Figure 1, we present a scenario with $k = 6$ equally effective treatments ($\theta_1 = \cdots = \theta_k = 0$). The graphs show the information fraction on the x-axis. The y-axis gives the standardized selection bias on the left graph and the square root of the standardized MSE on the right graph. We see that Lindley's estimator considerably reduces the selection bias as well as the MSE compared with the MLE. Cohen and Sackrowitz's estimator perfectly corrects for the bias but at the price of a substantial increase in the MSE. The curves end at $t = 0.9$, because the method requires an interim analysis. Stallard and Todd's method overcorrects the bias, which leads to an inflation of the MSE compared with the MLE.

On the middle row of Figure 1, we present a scenario in which there is a linear relationship between the effects of the treatments; that is, $\theta_i = 0.25\cdot(i-1)$, for $i = 1,\ldots,6$. We observe that the selection bias of the MLE is less problematic than in the previous scenario, but it is still present. Lindley's estimator still has a reduced bias and a slightly reduced MSE compared with the MLE. The performance of Cohen and Sackrowitz's method in this scenario is similar to that of the previous one; the method perfectly corrects for the bias and has an inflated MSE although this inflation is less pronounced. The performance of Stallard and Todd's method in this scenario is also similar to that of the previous one.

On the bottom row of Figure 1, we present a scenario in which only one treatment is effective; $\theta_1 = \cdots = \theta_5 = 0$ and $\theta_6 = 1.25$. Lindley's estimator overcorrects the bias in this case but still performs reasonably well with respect to the MSE compared with the MLE. Lindley's estimator is at least conservative with regard to bias.

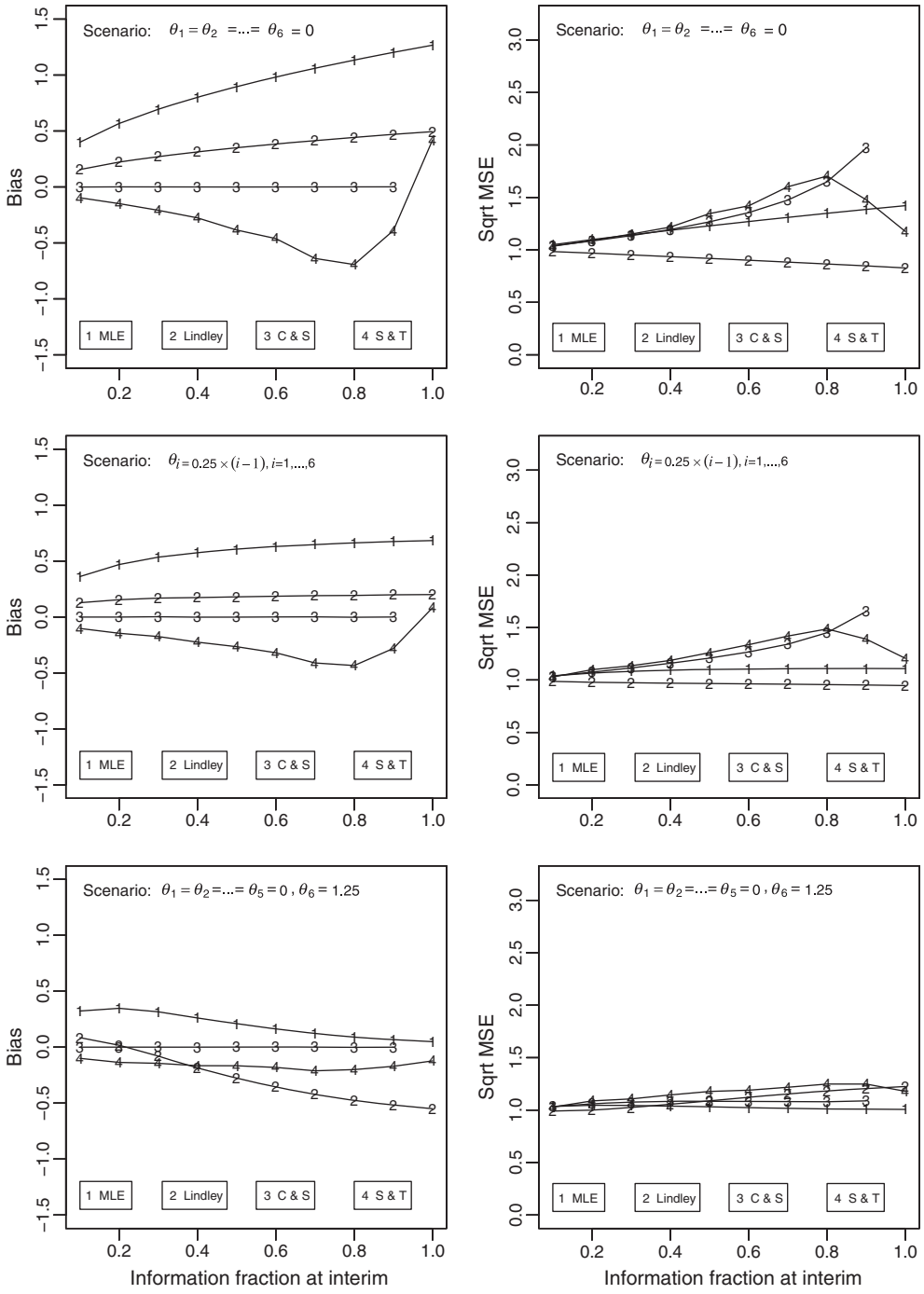

Figure 1. Selecting the best-performing treatment among $k = 6$ treatments at the interim analysis. MLE, maximum-likelihood estimator; MSE, mean squared error; C & S, Cohen and Sackrowitz; S & T, Stallard and Todd.

Stallard and Todd's method had severe divergence problems when the interim analysis is performed late. The divergence rates were up to 5% when the interim analysis is performed at an information fraction of $t = 0.8$, and it was between 20% and 30% when the interim analysis is performed at $t = 0.9$. Divergence problems have also been observed with low rates for earlier interim analyses. The simulation results shown for this method are conditional on convergence.

We have performed similar simulations assuming σ is unknown and have estimated it in $Q_S^{\mathrm{L},1}$ by the estimator $\hat{\sigma}$ mentioned in Section 2.3. We found essentially no difference between the known and unknown variance case.

We finally performed a similar simulation study for $k = 2, 3$ experimental treatments, replacing Lindley's estimator with the two-stage version of the estimated BLUP from the corresponding random effects model. Figure 2 shows the results for the case $k = 3$. One can see from the figure that the estimated BLUP performs similarly well as Lindley's estimator, improving the MLE and the two bias-adjusted estimators in terms of MSE, although no domination result is available for this estimator. We obtain similar results for $k = 2$.

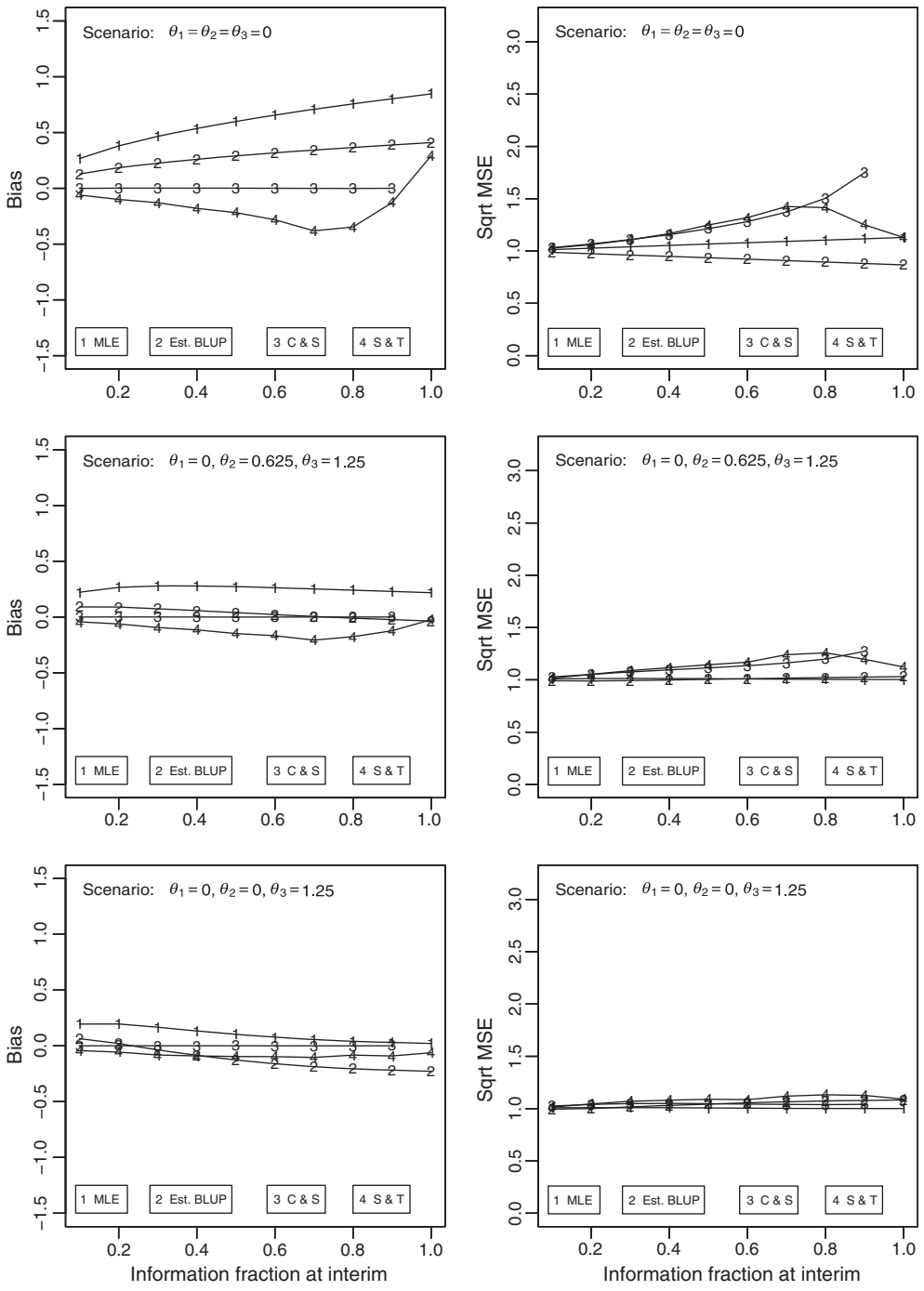

Figure 2. Selecting the best-performing treatment among $k = 3$ treatments at the interim analysis. MLE, maximum-likelihood estimator; MSE, mean squared error; BLUP, best linear unbiased predictor; C & S, Cohen and Sackrowitz; S & T, Stallard and Todd.

4.5. *Selection of the second best-performing treatment*

In this section, we present the same three scenarios as in the previous section with $k = 6$ treatments but we select the second best-performing treatment instead. (We performed similar simulations for $k = 3$ and found essentially the same results.)

Figure 3 presents standardized bias and square root of standardized MSE for the maximum-likelihood, Lindley's and Bowden and Glimm's estimators. We see that Lindley's estimator reduces the selection bias as well as the MSE compared with the MLE. Bowden and Glimm's estimator perfectly corrects for the bias and also at the price of a dramatic increase in the MSE.

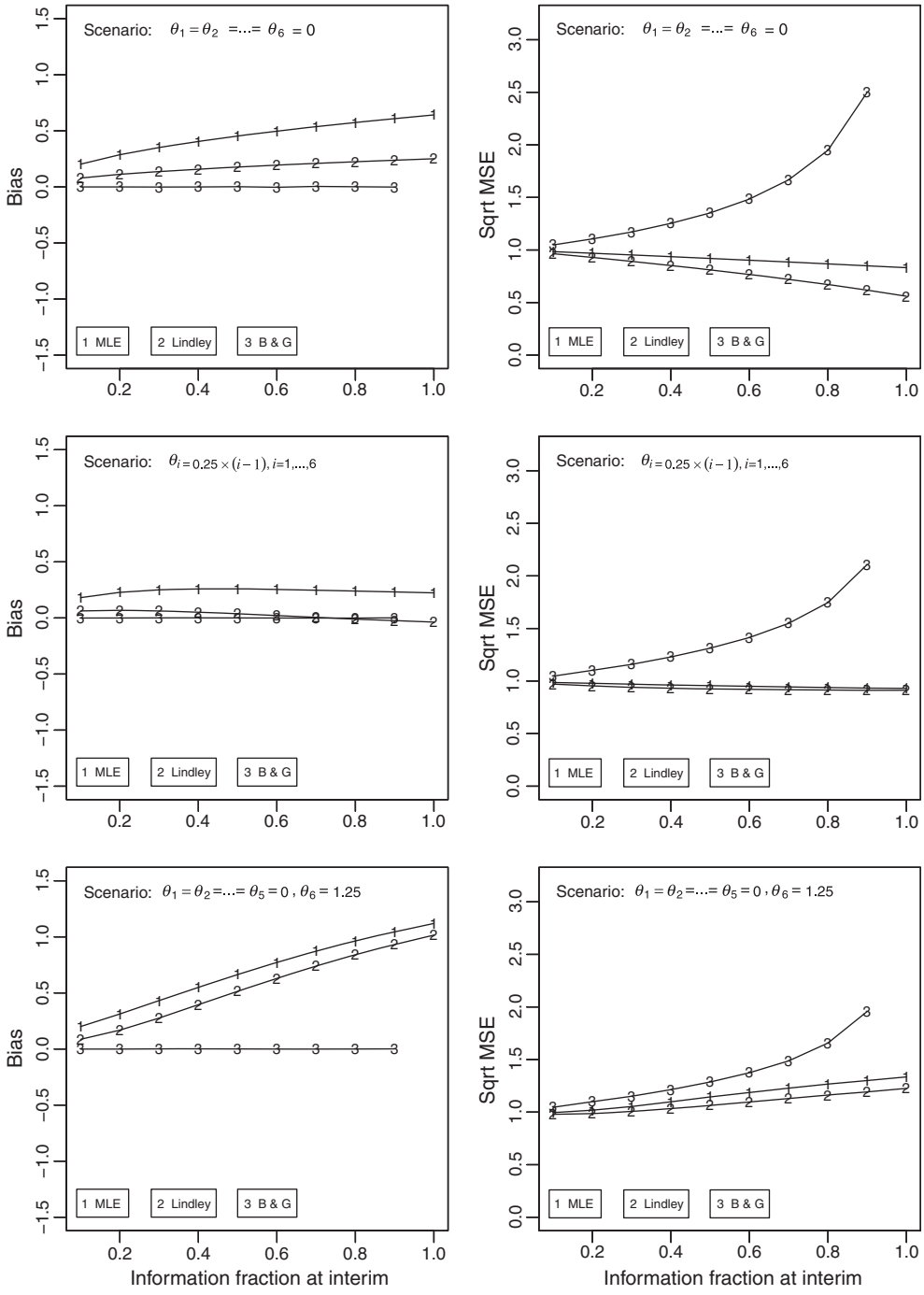

Figure 3. Selecting the second best-performing treatment among $k = 6$ treatments at the interim analysis. MLE, maximum-likelihood estimator; MSE, mean squared error; B & G, Bowden and Glimm.

For the scenario with only one effective treatment, at the bottom line of Figure 3, we see that the MLE exhibits a large bias. This is because most of the time the best-performing treatment will be the effective one and the second best-performing treatment will be the one that shows the largest effect among the five noneffective treatments. Lindley's estimator reduces the bias as well as the MSE compared with the MLE, and both methods clearly outperform Bowden and Glimm's method with respect to the MSE.

The relatively large bias of Lindley's estimator in the last scenario has motivated us to consider a modification of this estimator for the case when the second best treatment is selected. In this case, we could adopt a conservative strategy in which we ignore the mean of the best-performing treatment. This means to compute Lindley's first-stage estimator only with the second largest and all smaller group means. Of course, we suggest to apply such modification only when selecting the second-best treatment. More generally, when selecting the jth best treatment at interim, the shrinkage estimator could be constructed from this and smaller interim means. This will always lead to an estimator that is more conservative than Lindley's estimator and the MLE.

We have run simulations for this modification when always selecting the second-best treatment. Figure 4 shows the result for the scenario where only one treatment is effective. In this scenario, the unmodified Lindley's estimator performed not so well. One can see that the modified estimator performs much better with regard to bias and MSE. We observed only small differences in bias and MSE between the modified and the unmodified Lindley's estimator for the other mean patterns.

4.6. Additional simulation results

We considered more scenarios than the ones presented in the previous subsections. Our simulation study included, for instance, the same mean patterns as presented previously but with an increased variance such that the power of the two-group comparison between the most effective and control groups was only 60%. We also considered step functions with two and three plateaus, several scenarios where one treatment was clearly more effective than the others and patterns with an exponential increase of the treatment effect. We furthermore performed simulations with only two experimental treatments. We obtained the same results in all scenarios: the shrinkage estimator outperformed all other methods in terms of the MSE, except in the case where one treatment was clearly more effective than the others, in which case, the MSE of the shrinkage estimator was comparable with that of the other methods. The shrinkage estimator also showed a substantial reduction of bias compared with the MLE. In the case where one treatment was clearly more effective than the others, Lindley's estimator overcorrected the bias. Given that the results are invariant with respect to permutations of the treatment effects, we have covered a broad range of realistic scenarios.

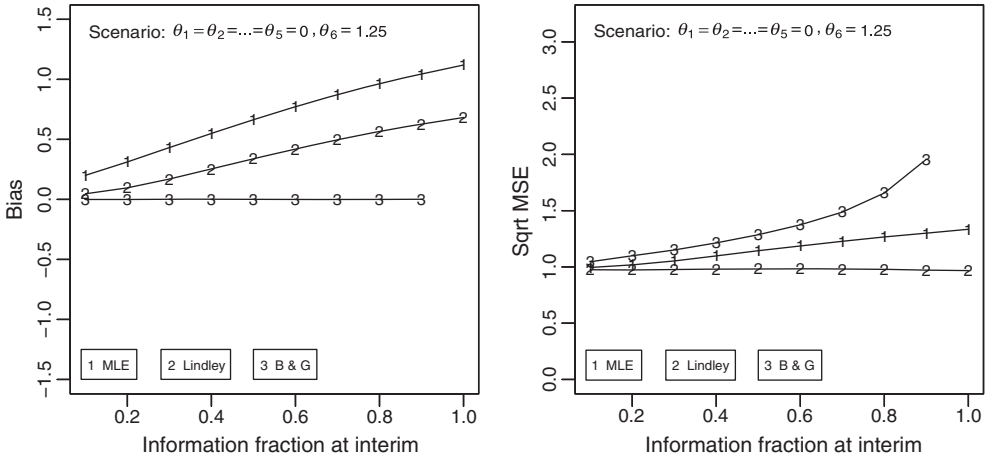

Figure 4. Selecting the second best-performing treatment among $k = 6$ treatments at the interim analysis when the shrinkage estimate is computed only from the second best and all worse treatments. MLE, maximum-likelihood estimator; MSE, mean squared error; B & G, Bowden and Glimm.

5. Discussion

On the basis of previous work on shrinkage estimation in single-stage multi-armed trials, we have considered shrinkage estimation in adaptive two-stage designs with selection of a single experimental treatment arm at the interim analysis. We have particularly extended the work of Hwang [13] who showed that Lindley's estimator [14] dominates the MLE in terms of Bayes risk within an empirical Bayesian framework with independent normal priors. We have shown that the same domination result applies to a natural extension of Lindley's estimator to the two-stage design. The two-stage version of Lindley's estimator is the weighted mean of the first-stage Lindley's estimator and the single second-stage MLE of the selected treatment. Lindley's estimator applies to designs that start with at least four experimental treatments. For designs with less treatments, we considered using the standard estimator of the BLUP of the corresponding random effects model for the first-stage estimator.

We have investigated by simulation studies the performance of the two-stage version of Lindley's estimator and the estimated BLUP and have found favorable properties of these estimators in comparison with the MLE and the bias-adjusted estimators of Cohen and Sackrowitz [8] and Stallard and Todd [10]. Although Cohen and Sackrowitz's estimator perfectly removes the selection bias, it has increased variance in a way that the MSE is generally larger than for the MLE. Stallard and Todd's estimator was found to have the tendency for an overcorrection of bias, at least in the two-stage setting, and it is also in general inferior to the MLE and shrinkage estimators in terms of MSE. The two-stage shrinkage estimators did, on the one hand, reduce selection bias of the MLE substantially (although not removing it completely) and, on the other hand, improved or was equivalent to the MLE in terms of MSE. Because we believe that MSE is more appropriate as measure of precision than the bias itself, we suggest using shrinkage estimators to deal with the problem of selection bias.

In practice, we not only report the estimate but also a measure of precision of the estimate such as its standard error. This is possible only if the treatment selection process follows a known prespecified rule because, otherwise, the distribution of the estimate remains unknown. This applies to any treatment effect estimate. When the rule is known, then the variance of the shrinkage estimate can be determined, for instance, by parametric or nonparametric bootstrapping. We suggest to report the MSE instead of the variance, because the MSE also accounts for the bias and hence is a more conservative measure for the precision of the estimate. Reporting the variance would lead to an overestimation of the precision. Another related topic is the construction of simultaneous confidence intervals. Simultaneous confidence intervals should be consistent with the test decision of the multiple hypothesis test for the superiority or noninferiority hypotheses of primary interest. Consistency means that there should be no contradiction between the parameters excluded by the confidence intervals and those excluded by the hypothesis tests. Authors have suggested efficient hypothesis testing methods for designs with treatments selection (e.g., [3–6]), and any simultaneous confidence interval should be discussed in comparison with such test procedures (e.g., [5]).

We did not directly address other important extensions of our work. One important extension is to designs with a control group where the treatment effect is measured in terms of the mean difference between the selected experimental treatment and control group. If the control group is continued with the selected treatment to the second stage, then we can subtract the MLE of the control group from the shrinkage estimator to obtain an estimator that performs similarly well in terms of bias and MSE. The reason why we can expect the same favorable properties is that the MLE of the control group is mean unbiased (and hence, all bias is driven by the estimator for the treatment group) and the MSE of the difference between treatment and control group is just the sum of the MSEs of the estimators for the treatment and control group. Hence, the domination result applies, and simulation results will be similar for designs with control group (see [7] for similar arguments).

Another important extension is to designs where, in addition to treatment selection, the second-stage sample size is assessed at the interim analysis. One important case is covered by our exposition, namely, the case where the sample sizes of the dropped treatment arms are reshuffled to the selected treatment and control arm in equal (or some other prespecified) parts. In this case, the second-stage sample sizes will always be the same and hence are as if they where fixed in advance. More general sample size adaptations rules, for example, based on conditional power, are not covered by our work and require additional research. A particularly difficult but interesting problem is how far the domination result can be extended to designs including data-driven sample size adaptations.

Another important extension that is covered only partly by our research are designs where more than a single treatment is selected at the interim analysis. If we are concerned only with the bias caused by the

interim selection process, then we can use the two-stage shrinkage estimators considered here for each of the treatments selected at the interim analysis. However, if we wish to account also for the bias caused by the second selection process at the end of the trial (where we select one of the multiple treatments considered until the end of the trial), then the selection bias problem becomes much more difficult. In this case, a natural approach would be to use the weighted mean of first-stage and second-stage shrinkage estimators. However, we have not been able to prove a domination theorem for such an estimator. This is another interesting open question for future research.

Appendix A

Proof of Theorem 2.2

Because the joint distribution of $\{X_1-\theta_1,\ldots,X_k-\theta_k\}$ is independent of the true mean vector $\theta=(\theta_1,\ldots,\theta_k)$, also $C=E\left[\max(X_1-\theta_1,\ldots,X_k-\theta_k)\right]$ is independent of θ. Because $\max(X_1-\theta_1,\ldots,X_k-\theta_k)\geqslant X_S-\theta_S$, we obtain $C\geqslant E_\theta\left[X_S-\theta_S\right]$ for any θ and any selection rule S. Therefore, C is an absolute upper bound for the selection bias that is independent of the true θ as well as of the selection rule. Finally, if $\theta_1=\theta_2=\cdots=\theta_k=\theta$ and $X_S=\max(X_1,\ldots,X_k)$, we obtain that $\max(X_1-\theta,\ldots,X_k-\theta)=\max(X_1,\ldots,X_k)-\theta=X_S-\theta_S$, which implies that $C=E[X_S-\theta_S]$. Hence, $b_\theta(X_S)$ attains the absolute upper bound C if $\theta_1=\theta_2=\cdots=\theta_k=\theta$ and $X_S=\max(X_1,\ldots,X_k)$. □

Proof of Theorem 3.1

Observe that

$$\begin{aligned}R_{\mu,\tau}(Q_S^{\mathrm{L}})&=E_{\mu,\tau}\left[\left(Q_S^{\mathrm{L}}-\theta_S\right)^2\right]=E_{\mu,\tau}\left\{\left[t\left(Q_S^{\mathrm{L},1}-\theta_S\right)+(1-t)\left(Y_S-\theta_S\right)\right]^2\right\}\\&=t^2R_{\mu,\tau}\left(Q_S^{\mathrm{L},1}\right)+(1-t)^2E_{\mu,\tau}\left\{(Y_S-\theta_S)^2\right\}+2t(1-t)E_{\mu,\tau}\left\{\left(Q_S^{\mathrm{L},1}-\theta_S\right)\left(Y_S-\theta_S\right)\right\}.\end{aligned}$$

Now,

$$\begin{aligned}E_{\mu,\tau}\left\{\left(Q_S^{\mathrm{L},1}-\theta_S\right)\left(Y_S-\theta_S\right)\right\}&=E_{\mu,\tau}\left\{E_{\mu,\tau}\left[\left(Q_S^{\mathrm{L},1}-\theta_S\right)\left(Y_S-\theta_S\right)\mid X_1,\ldots,X_k,W\right]\right\}\\&=E_{\mu,\tau}\left\{\left(Q_S^{\mathrm{L},1}-\theta_S\right)E_{\mu,\tau}\left[Y_S-\theta_S\mid X_1,\ldots,X_k,W\right]\right\}=0\end{aligned}$$

because $E_{\mu,\tau}[Y_S-\theta_S\mid X_1,\ldots,X_k,W]=0$. Therefore,

$$E\left[\left(Q^{\mathrm{L}}-\theta_S\right)^2\right]=t^2R_{\mu,\tau}\left(Q_S^{\mathrm{L},1}\right)+(1-t)^2E_{\mu,\tau}\left\{(Y_S-\theta_S)^2\right\}\ .$$

Similarly, $E\left[(Q^{\mathrm{N}}-\theta_S)^2\right]=t^2R_{\mu,\tau}(X_S)+(1-t)^2E_{\mu,\tau}\left\{(Y_S-\theta_S)^2\right\}$. The domination result in the two-stage setting is a direct consequence of Theorem 2.2, which implies that for $k\geqslant 4$, $R_{\mu,\tau}(Q_S^{\mathrm{L},1})\leqslant R_{\mu,\tau}(X_S)$ for all μ and τ. □

Hwang's domination result with more general selection rules

Let $S\in\{1,\ldots,k\}$, $k\geqslant 4$, be the index of the selected treatment and suppose that $S=h(Z_1,\ldots,Z_k,W)$, where $Z_i=(X_i-\bar{X})/\sqrt{V}$, for $i=1,\ldots,k$, $\bar{X}=\sum_{j=1}^k X_j/k$, $V=\sum_{j=1}^k(X_j-\bar{X})^2$ and W is independent from the group mean vector $X=(X_1,\ldots,X_k)$ for any given mean vector θ. Because W is independent from X and the a priori distribution does not depend on W, we can assume that W is a fixed (nonrandom) vector and consider S as function of the Z_i only. We present the proof of Theorem 2.2 for such selection rules. The proof follows the lines of the proof in Hwang [13].

Before we prove the theorem, we note that, for any estimator $\delta=\delta(X)$ of θ_S, the posterior risk, namely, the risk conditional on X, is $E\left\{\{\theta_S-\delta(X)\}^2\mid X\right\}=\{\delta(X)-Q_S^B\}^2+\tau^2\sigma^2/(\sigma^2+n\tau^2)$.

Hence, the Bayes risk of δ is $R(\delta) = E\left[\left\{\delta(X) - Q_S^B\right\}^2\right] + \tau^2\sigma^2/(\sigma^2 + n\tau^2)$. We now consider estimators of the form $\delta(X) = g(V)(X_S - \bar{X}) + \bar{X}$. From simple algebra, we obtain

$$\delta(X) - Q_S^B = [g(V) - C]\,(X_S - \bar{X}) + (1 - C)(\bar{X} - \mu)\,. \tag{A1}$$

Because $X_i - \bar{X}$, $1 \leqslant i \leqslant k$ and $\bar{X}$ are stochastically independent and S is a function of $X_i - \bar{X}$, $1 \leqslant i \leqslant k$, it follows that also $X_S - \bar{X}$ and $\bar{X}$ are stochastically independent. Therefore, the first term on the right-hand side of Equation (A1) is independent of the second term, and hence

$$E\left[\left\{\delta(X) - Q_S^B\right\}^2\right] = E\left[\{g(V) - C\}^2\left(X_S - \bar{X}\right)^2\right] + (1 - C)\,\sigma^2/(nk)\,. \tag{A2}$$

Now, $E\left\{(X_S - \bar{X})^2 \mid V\right\} = E\left\{Z_S^2 \mid V\right\} V$. Because the distribution of $Z_S^2 = Z^2_{h(Z_1,\ldots,Z_k)}$ is independent from σ^2, and V is a complete sufficient statistic for σ^2 (with regard to the group means $X_1, \ldots, X_k$), we obtain from Basu's theorem that Z_S^2 and V are independent. Hence,

$$E\left\{(X_S - \bar{X})^2 \mid V\right\} = c_S V \tag{A3}$$

for some constant $c_S > 0$ that does not depend on μ, σ^2, or τ^2. Using conditional arguments, Equations (A2) and (A3) and the identity $\tau^2\sigma^2/(\sigma^2 + n\tau^2) = (1 - C)\tau^2$, we obtain

$$R(\delta) = c_S E\left[\{g(V) - C\}^2 V\right] + (1 - C)\,\left\{\sigma^2/(nk) + \tau^2\right\}\,. \tag{A4}$$

We consider now the estimator $\delta(a) = \{1 - a\sigma^2/(nV)\}\left(X_S - \bar{X}\right) + \bar{X}$ for arbitrary $a \geqslant 0$. Note that $\delta(0) = X_S$, and $\delta(k-3)$ is close to Lindley's estimator but with $\hat{C}$ instead of $\hat{C}_+$. To complete the proof, we show that (i) Q_S^L has smaller Bayes risk than $\delta(k-3)$ and (ii) that $\delta(k-3)$ has smaller Bayes risk than $\delta(0)$. To show (i), note that by Equation (A4), we have that $R(Q_S^L) < R\{\delta(k-3)\}$ if and only if

$$E\left\{\left(\left[1 - \frac{(k-3)\sigma^2}{nV}\right]_+ - C\right)^2 V\right\} < E\left\{\left(\left[1 - \frac{(k-3)\sigma^2}{nV}\right] - C\right)^2 V\right\}$$

and that the last inequality is obvious. For (ii), note that by (A4), we obtain $R\{\delta(a)\} = c_S E[\{1 - C - a\sigma^2/(nV)\}^2 V] + (1 - C)\,\{\sigma^2/(nk) + \tau^2\}$. From the identities $E(V) = (\sigma^2/n + \tau^2)(k-1)$ and $E(1/V) = \{(\sigma^2/n + \tau^2)(k-3)\}^{-1}$ and direct evaluations, we get that $E[\{1 - C - a\sigma^2/(nV)\}^2 V] = \sigma^2(1 - C)\left[k - 1 - 2a + a^2/(k-3)\right]$, which is uniquely minimized for $a = k - 3$. This establishes (ii), completing the proof.

References

1. Bauer P, Brannath W. The advantages and disadvantages of adaptive designs for clinical trials. *Drug Discovery Today* 2004; **9**:351–357.
2. Bretz F, Koenig F, Brannath W, Glimm E, Posch M. Adaptive designs for confirmatory clinical trials. *Statistics in Medicine* 2009; **28**:1181–1217.
3. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 1999; **18**:1833–1848.
4. Hommel G. Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal* 2001; **43**:581–589.
5. 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.
6. Stallard N, Friede T. A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine* 2008; **27**:6209–6227.
7. Bauer P, Koenig F, Brannath W, Posch M. Selection and bias—two hostile brothers. *Statistics in Medicine* 2010; **29**:1–13.
8. Cohen A, Sackrowitz HB. Two stage conditionally unbiased estimators of the selected mean. *Statistics & Probability Letters* 1989; **8**:273–278.
9. Shen L. An improved method of evaluating drug effect in a multiple dose clinical trial. *Statistics in Medicine* 2001; **20**:1913–1929.
10. Stallard N, Todd S. Point estimators and confidence regions for sequential trials involving selection. *Journal of Statistical Planning and Inference* 2005; **135**:402–419.
11. Bowden J, Glimm E. Unbiased estimation of selected treatment means in two-stage trials. *Biometrical Journal* 2008; **4**:515–527.
12. Harrell Jr., FE. *Regression Modelling Strategies: With Applications to Linear Models, Logistic Regression and Survival Analysis*, Series in Statistics. Springer: New York, 2001.

13. Hwang JT. Empirical Bayes estimation for the means of the selected populations. *The Indian Journal of Statistics* 1993; **55**:285–311.
14. Lindley DV. Discussion of Professor Stein's paper "Confidence sets for the mean of a multivariate normal distribution". *Journal of the Royal Statistical Society, Series B* 1962; **24**:265–296.
15. Putter J, Rubinstein D. On estimating the mean of a selected population. *Technical Report No 165*, University of Wisconsin, Department of Statistics, 1968.
16. Cohen A, Sackrowitz HB. Estimating the mean of the selected population. *Statistical Decision Theory and Related Topics III* 1982; **1**:243–270.
17. Searle, S R, Casella G, McCulloch C. *Variance Components*, Series in Probability and Statistics. Wiley: New York, 1992.
18. Whitehead J. On the bias of maximum-likelihood estimation following a sequential test. *Biometrika* 1986; **73**:573–581.

Research Article

Statistics in Medicine

Received 17 March 2014, Accepted 10 December 2014 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6407

Adaptive seamless designs with interim treatment selection: a case study in oncology

Máximo Carreras,[a*†] Georg Gutjahr[b] and Werner Brannath[b]

The planning of an oncology clinical trial with a seamless phase II/III adaptive design is discussed. Two regimens of an experimental treatment are compared to a control at an interim analysis, and the most-promising regimen is selected to continue, together with control, until the end of the study. Because the primary endpoint is expected to be immature at the interim regimen selection analysis, designs that incorporate primary as well as surrogate endpoints in the regimen selection process are considered. The final testing of efficacy at the end of the study comparing the selected regimen to the control with respect to the primary endpoint uses all relevant data collected both before and after the regimen selection analysis. Several approaches for testing the primary hypothesis are assessed with regard to power and type I error rate. Because the operating characteristics of these designs depend on the specific regimen selection rules considered, benchmark scenarios are proposed in which a perfect surrogate and no surrogate is used at the regimen selection analysis. The operating characteristics of these benchmark scenarios provide a range where those of the actual study design are expected to lie. A discussion on family-wise error rate control for testing primary and key secondary endpoints as well as an assessment of bias in the final treatment effect estimate for the selected regimen are also presented.

Keywords: adaptive seamless designs; treatment selection; surrogate endpoints; family wise error rate; confidence intervals; selection bias

1. Introduction

Adaptive seamless designs with interim treatment selection have been considered extensively in the literature; see Bauer & Brannath [1] and Bretz *et al.* [2] for comprehensive reviews. Hypothesis testing methods have been developed that allow the treatment to be selected at the interim analysis, incorporating all available information, including information external to the study, while controlling the overall multiple type I error rate at a prespecified level α [3–6]. Schäfer and Müller [7] extended the adaptive design methodology to survival data by using the conditional error approach [8] and applying the independent increments structure of the log-rank statistics. But Bauer and Posch [9] pointed out that caution must be exercised when applying these flexible designs: in the particular setting of a two-stage design with treatment selection after the first stage, it means that stage 2 p-values for the comparison of the selected treatments with the control at the end of the study cannot be influenced by surrogate information of stage 1 patients who are at risk at the interim treatment selection analysis. Therefore, the standard testing procedure of Schäfer and Müller [7] that combines stage-wise p-values computed from increments of the log-rank statistic (see also Wassmer [10]) may not protect the type I error rate unless treatment selection is based only on the score statistics of the primary endpoint, which is unrealistic in practice. In the context of enrichment designs, Jenkins *et al.* [11] proposed a testing procedure that preserves type I error rate while allowing all information collected until the interim analysis to be used for treatment selection. König *et al.* [12] suggested a testing procedure based on a modification of the classical Dunnett [13] test, which guarantees strict control of type I error rate without imposing any restrictions on the information

[a] *F. Hoffmann-La Roche AG, Basel, Switzerland*
[b] *Competence Center for Clinical Trials Bremen, Faculty 3, University Bremen, Bremen, Germany*
[*] *Correspondence to: Máximo Carreras, F. Hoffmann-La Roche AG, Grenzacherstrasse 124, 4058 Basel, Switzerland.*
[†] *E-mail: maximo.carreras@roche.com*

that can be used at the interim analysis for treatment selection purposes. The method by König *et al.* [12] has, however, not been developed for time-to-event data so far. Di Scala and Glimm [14] extended the classical Dunnett [13] test to survival data for adaptive seamless designs with interim treatment selection.

In this article, the planning of an oncology clinical study with an adaptive seamless phase II/III design is discussed. Two regimens of an experimental treatment are compared to a control at an interim analysis, and the most-promising regimen is selected to continue, together with control, until the end of the study. Because the study's primary endpoint will be immature at the regimen selection analysis, it is of interest to investigate whether the incorporation of surrogate information can help improve the regimen selection process and thus the study's probability of success. To this end, designs are considered, which include the primary as well as surrogate endpoints in the regimen selection analysis. At the end of the study, testing of efficacy is carried out to compare the selected regimen to the control with respect to the primary endpoint, utilizing all relevant data collected both before and after the interim analysis.

The aforementioned three approaches [7, 11, 12] are assessed with regard to power and type I error rate for testing the primary null hypothesis comparing the selected regimen to the control at the end of the study. Because the operating characteristics of these designs depend on the specific regimen selection rules considered, on the correlation between primary and surrogate endpoints and on the true standardized mean difference of the surrogate endpoint(s), benchmark scenarios are proposed in which a 'perfect surrogate' and no surrogate is used at the regimen selection analysis. The operating characteristics of these benchmark scenarios provide a range where those of the actual study design are expected to lie.

This article is organized as follows. In Section 2, the case study is introduced. Section 3 discusses hypothesis testing methods for the primary endpoint. Section 4 presents the treatment selection rules that are used in our simulation study, which are a simplification of the actual selection rules used in the case study. A discussion on family-wise error rate (FWER) control for testing primary and key secondary endpoints is presented in Section 5. In Section 5, estimation of treatment effect for the selected regimen at the end of the study is considered, including an assessment of bias as well as the construction of confidence intervals. In Section 7, the assumptions and results of a thorough simulation study are presented. Finally, Section 8 presents our concluding remarks.

2. The case study

The case study is an open label study that aims to evaluate the efficacy and safety of two dose regimens of an experimental treatment as a single-agent targeted therapy compared to an active control in patients with previously treated HER2-positive locally advanced or metastatic gastric cancer (MGC). The low-dose regimen of the experimental treatment was shown to be effective in metastatic breast cancer (MBC). But results of another study investigating a closely-related compound showed that drug exposure levels in MGC could be up to 50% lower than those in MBC. Therefore, in order to mitigate the risk of insufficient drug exposure with the low-dose regimen, the sponsor decided to also incorporate a high-dose regimen in the study.

Several options were considered for the development program. A standard program with phases II and III run in separate studies was considered to have too long timelines and was regarded as a moderate business case for a fast-to-market product development. A three-arm phase III design with regimen selection at the end of the study was deemed to carry regulatory and payer challenges (e.g., the high-dose regimen could be more efficacious but could have a worse safety profile compared to the low-dose regimen, leading to similar benefit/risk profiles. If both regimens were approved by regulatory authorities, negotiations with payers about price of each regimen would be difficult). The seamless phase II/III adaptive design with interim regimen selection analysis was considered to provide the best option for the program in terms of timelines and probability of success. In this paper, we compare these three options with respect to power.

Because the study's primary endpoint, overall survival (OS), will be immature at the interim analysis (only 50 deaths are expected across the three arms) and secondary efficacy endpoints such as progression-free survival or objective response rate are not considered to be good surrogates for OS in this indication, the regimen selection criteria incorporate drug exposure as an important decision driver. The regimen selection analysis will be carried out after approximately 100 patients across all three arms have been enrolled and followed for a minimum of 12 weeks, which should provide sufficient pharmacokinetics data for decision making. An Independent Data Monitoring Committee (IDMC) will recommend one of

the two experimental regimens based on all pharmacokinetics, safety, and efficacy data available at that time. The IDMC is guided to select the high-dose regimen only when the following three conditions are satisfied:

- Mean (cycle 1) area under the plasma drug concentration versus time curve (AUC) of the high-dose regimen is at least 50% larger than that of the low-dose regimen (mean cycle 1 AUC for the experimental arms is expected to be in the range of 200–350 days$\times\mu g/mL$)
- Efficacy of the high-dose regimen is comparable or superior to that of the low-dose regimen
- High-dose regimen has an acceptable safety profile.

Otherwise, the IDMC is guided to select the (default) low-dose regimen. No early stopping for efficacy or futility is allowed at the regimen selection analysis. The IDMC has the authority to recommend stopping the study due to safety at any time during its course. Accrual will continue into all three treatment arms until the regimen selection analysis has been completed.

The sample size requirements for the study have been obtained using computer simulations. A total sample size of approximately 410 patients will be recruited into the study. A median OS of 9 months in both experimental arms and of 6 months in the control arm are expected to be observed in the study. Under the assumption of proportional hazards, this corresponds to a hazard ratio (HR) of 0.67 between each experimental arm and the control.

3. Hypothesis testing for the primary endpoint

In order to avoid conflicting directional effects at the two stages of the study to combine into a single final directional test decision, a one-sided hypothesis testing framework is adopted for OS. The null hypotheses $H_i^0 : \theta_i \geq 0; i = 1, 2$, where θ_i is the log-HR for the parameter OS between the i-th experimental treatment and the control, are considered.

In this section, we present the three testing procedures mentioned in the introduction. In order to do this, the following test statistics U_{ij}, $i = 1, 2$ and $j = a, b, c, d$ are considered, where:

(a) U_{ia} is the log-rank statistic to test H_i^0 at the regimen selection analysis, defined as the difference between the number of deaths observed in the i-th experimental arm and the number of deaths expected under H_i^0 at that time;
(b) U_{ib} is the log-rank statistic to test H_i^0 at the final analysis using stage 1 and stage 2 patients;
(c) U_{ic} is the log-rank statistic to test H_i^0 at the final analysis using only stage 1 patients;
(d) U_{id} is the log-rank statistic to test H_i^0 at the final analysis using only stage 2 patients.

The statistics U_{ia}, U_{ib}, U_{ic}, and U_{id} are computed using data from the regions depicted in Figure 1 for treatment i and control. Let V_{ij} be the variance of U_{ij} and define $Z_{ij} = U_{ij}/\sqrt{V_{ij}}$ for all i, j. Let $S = s$

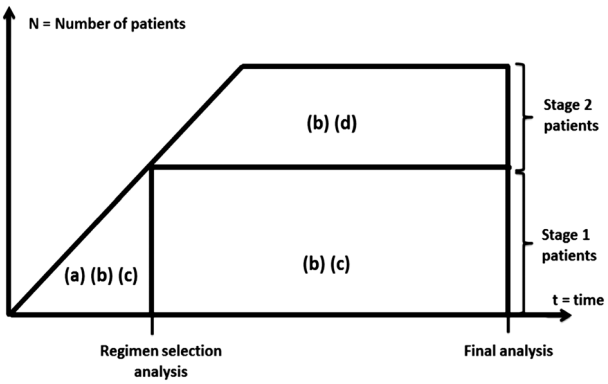

Figure 1. Graphical representation of the data used to compute the test statistics U_{ij}, $i = 1, 2$, and $j = a, b, c, d$. U_{ia} is computed using data from the bottom left triangle in the figure for the i-th experimental treatment and the control. U_{ib} is computed using data from all three regions in the figure. Similar interpretations follow for U_{ic} and U_{id}.

denote the index of the regimen selected at the regimen selection analysis and let Φ denote the standard normal CDF.

3.1. Testing procedure 1 (follow-up-wise staging)

In this testing procedure, stage 1 p-values are calculated by $p_{ia} = \Phi\left(Z_{ia}\right)$. Using the independent and normally distributed increments structure of the log-rank statistic [10], the stage 2 p-value for the selected treatment is calculated by $p_{s(b-a)} = \Phi\left(Z_{s(b-a)}\right)$, where $Z_{s(b-a)} = (U_{sb} - U_{sa})/\sqrt{V_{sb} - V_{sa}}$, incorporating all OS data from stage 2 patients as well as from continued OS follow-up of stage 1 patients in the selected treatment and the control.

In order to account for the multiplicity of the testing procedure, the closure principle [15] is used, which requires the calculation of a p-value for the intersection null hypothesis $H_1^0 \cap H_2^0$ of no difference in OS between both treatments and the control. Using Hochberg's procedure [16], the stage 1 p-value for testing $H_1^0 \cap H_2^0$ is calculated by $p_{\text{int},1} = \min\{2\min(p_{1a}, p_{2a}), \max(p_{1a}, p_{2a})\}$ (other methods such as that of Dunnett and Tamhane [17] could also be used). The stage 2 p-value for testing $H_1^0 \cap H_2^0$ is defined as $p_{\text{int},2} = p_{s(b-a)}$ (note that this is a conservative p-value).

The inverse normal combination method [10] is used to combine stage 1 and stage 2 p-values. The combined p-value for testing H_s^0 is calculated by $p_{s,\text{comb}} = 1 - \Phi(Z_{s,\text{comb}})$, where $Z_{s,\text{comb}} = w_1\Phi^{-1}(1 - p_{sa}) + w_2\Phi^{-1}(1 - p_{s(b-a)})$. Similarly, the combined p-value for testing $H_1^0 \cap H_2^0$ is calculated by $p_{\text{int,comb}} = 1 - \Phi(Z_{\text{int,comb}})$, where $Z_{\text{int,comb}} = w_1\Phi^{-1}(1 - p_{\text{int},1}) + w_2\Phi^{-1}(1 - p_{\text{int},2})$. Finally, H_s^0 is rejected at the end of the study if both $p_{s,\text{comb}}$ and $p_{\text{int,comb}}$ are smaller than α. The weights w_1 and w_2 are calculated by $w_1 = \sqrt{\frac{e_1}{e_1+e_2}}$ and $w_2 = \sqrt{\frac{e_2}{e_1+e_2}}$ where e_1 denotes the number of deaths expected to be observed from stage 1 patients across the three arms at the regimen selection analysis and e_2 denotes the number of deaths expected to be observed from stage 1 and stage 2 patients in the selected arm and control after the regimen selection analysis and until the final analysis. In this paper, e_1 and e_2 are estimated by simulations.

As originally pointed out by Bauer and Posch [9] and further discussed by Jenkins *et al.* [11] in the specific setting of seamless phase II/III adaptive designs with hypothesis selection, the method described in this section does not necessarily protect the type I error because surrogate information, such as drug exposure, of stage 1 patients who are at risk at the interim analysis influences both the treatment selection decision and the stage 2 p-value for the comparison of the selected treatment with the control at the end of the study. In the simulations, we investigate how serious the type I error inflation can be in some specific settings.

3.2. Testing procedure 2 (patient-wise staging)

In this alternative testing procedure, p-values are computed separately for patients recruited into stage 1 and those recruited into stage 2. In particular, after the regimen selection analysis, stage 1 patients continue to be followed-up for survival and stage 1 p-values are computed by $p_{ic} = \Phi\left(Z_{ic}\right)$; $i = 1, 2$. The stage 2 p-value for the comparison of the selected experimental arm with the control is calculated by $p_{sd} = \Phi\left(Z_{sd}\right)$. The stage 1 and stage 2 p-values for testing $H_1^0 \cap H_2^0$ are calculated, respectively, by $p_{\text{int},1} = \min\{2\min(p_{1c}, p_{2c}), \max(p_{1c}, p_{2c})\}$ and $p_{\text{int},2} = p_{sd}$. The combined p-value for testing H_s^0 is calculated by $p_{s,\text{comb}} = 1 - \Phi(Z_{s,\text{comb}})$, where $Z_{s,\text{comb}} = w_1\Phi^{-1}(1 - p_{sc}) + w_2\Phi^{-1}(1 - p_{sd})$. Similarly, the combined p-value for testing $H_1^0 \cap H_2^0$ is calculated by $p_{\text{int,comb}} = 1 - \Phi(Z_{\text{int,comb}})$, where $Z_{\text{int,comb}} = w_1\Phi^{-1}(1 - p_{\text{int},1}) + w_2\Phi^{-1}(1 - p_{\text{int},2})$. Finally, H_s^0 is rejected at the end of the study if both $p_{s,\text{comb}}$ and $p_{\text{int,comb}}$ are smaller than α. The weights w_1 and w_2 are calculated by $w_1 = \sqrt{\frac{e_1}{e_1+e_2}}$ and $w_2 = \sqrt{\frac{e_2}{e_1+e_2}}$ where now e_1 denotes the number of deaths expected to be observed among stage 1 patients across the three arms by the time of the final analysis and e_2 denotes the number of deaths expected to be observed among stage 2 patients by the time of final analysis. Again, e_1 and e_2 are estimated by simulations.

As discussed in Jenkins *et al.* [11], in order to ensure independence of stage-wise p-values, the total length of survival follow-up of stage 1 patients must be pre-specified. It is not necessary to specify it before the study starts (as suggested by Jenkins *et al.*), but it must be specified before the regimen selection analysis takes place. This gives the flexibility to adjust the total length of survival follow-up of stage 1 (and stage 2) patients based on observed overall recruitment patterns until the regimen selection analysis as well as on a potentially updated sample size requirements for stage 2. This restriction on the total length of survival follow-up of stage 1 patients does not allow to define the date of the final analysis in the

standard way as the date when a certain total number of deaths among stage 1 and stage 2 patients have been observed. We see four options for defining the date of final analysis (options 1 and 3 have already been mentioned by Jenkins *et al.*):

(1) A pre-fixed calendar date.
(2) The date when a pre-fixed number of deaths has been observed among stage 1 patients.
(3) The latest of the dates when pre-fixed numbers of deaths have been observed, respectively, among stage 1 patients and stage 2 patients (separate stage-wise final analyses).
(4) The date when a pre-fixed number of deaths have been observed among patients in the selected experimental arm and the control across both stages.

In this article, 'pre-fixed' ('pre-specified') means fixed (specified) before the interim analysis. Options (1) and (4) require that the recruitment rate after the regimen selection analysis is not influenced by the interim results; for example, under option (1), if interim results for the experimental treatments are good, artificially slowing down the recruitment after the interim analysis will make the evidence from the additional patients more immature and therefore less likely to influence the good interim results by the fixed calendar date of the final analysis (similar arguments can be made for option (4). Recruitment rate may be influenced by interim results in options (2) and (3) without affecting the type I error. Options (1) and (2) have the problem that the total number of deaths, and consequently the power of the study, cannot be controlled by the study team. Option (3) has the problem that, if the date when the pre-fixed number of deaths among stage 1 patients is observed occurs before the corresponding date for stage 2 patients, any death occurring among stage 1 patients in between the two dates cannot be used in the final analysis. On the other hand, if the date when the pre-fixed number of deaths among stage 2 patients is observed occurs before the corresponding date for stage 1 patients, all deaths occurring among stage 2 patients in between the two dates can be used in the final analysis without compromising the type I error. The main advantage of option (4) is that the study team has more control over power compared to options (1) and (2), because we can guarantee enough power for the comparison of the selected treatment with the control.

3.3. Testing procedure 3 (conservative Dunnett test)

The testing procedure presented in this section is a modification of the conventional Dunnett [13] test for the comparison of several treatments with a control as suggested in the introduction of König *et al.* [12]. In this testing procedure, the final p-value for the comparison of the selected treatment with the control is calculated as $p_{s,\mathrm{fn}} = 1 - \tilde{\Phi}_{\rho}\left(-Z_{sb}, -Z_{sb}\right)$, where $\tilde{\Phi}_{\rho}$ denotes the CDF of the bivariate normal distribution with mean $(0,0)'$ and variance covariance matrix $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. The correlation ρ is computed using the corrected formulas in Section 3 of Di Scala and Glimm [14] paper under the assumption of the global null hypothesis, which is the same as the estimated correlation that would be obtained under the situation of normally distributed data. In the simulations presented in this paper, we assume a 1:1:1 randomization ratio and therefore $\rho = 1/2$.

This testing procedure is basically the classical Dunnett test for the comparison of two treatments against a control where the z-statistic for the deselected treatment is set to $+\infty$. As mentioned by König *et al.* [12], this procedure protects the type I error rate because the missing test statistic is imputed in a strictly conservative way. It was also mentioned in that paper that the only possible adaptation with this method is treatment selection.

The date of the final analysis can be defined as a pre-fixed calendar date or as a pre-fixed total number of deaths to be observed in the selected experimental arm and the control. Both options again require the recruitment rate after the regimen selection analysis not to be driven by interim results.

4. Treatment selection

When treatment selection is based on OS, we consider an obvious selection rule; select the treatment with the largest OS effect observed at the interim analysis. When the treatment selection is based on exposure, we consider the following selection rule: if the mean exposure in the high-dose regimen observed at the interim analysis is at least 50% larger than that in the low-dose regimen, select the high-dose regimen. Otherwise, select the low-dose regimen.

The latter treatment selection rule is a simplified version of the actual rule used in the study. As the low-dose regimen was the standard regimen in MBC, the team considered that there was no justification in using the high-dose regimen in MGC unless exposure levels achieved with the high-dose regimen were considerably higher than those achieved with the low-dose regimen.

4.1. Perfect surrogate approach

We consider a hypothetical situation in which drug exposure is a 'perfect surrogate'; that is, a surrogate endpoint that can perfectly predict survival time at the end of the study [9]. This can be achieved in the simulations by considering the treatment selection rule that selects at the interim analysis the treatment with the largest OS effect (smallest HR) observed at the final analysis, assuming that both treatment groups would be continued until the end of the study. Note that the selection rule based on this 'perfect surrogate' is not the same as a rule based on a surrogate endpoint, which is perfectly correlated to the main endpoint: given the simulation model presented in Section 7.1, the latter rule selects at interim analysis the treatment with the largest mean of a specific transformation of the survival times. As we show in Section 7.3, the 'perfect surrogate' scenario provides an upper bound for type I error but not necessarily for the power. In the simulations, we also consider scenarios where the primary and surrogate endpoints have different degrees of correlation.

5. Testing primary and main secondary endpoints

Even though the primary endpoint of the study is OS, which is the only clinically relevant endpoint in this indication, secondary efficacy endpoints such as progression-free survival or objective response rate can help characterize the efficacy profile of the experimental treatment and results of analyses of these secondary endpoints may be included in the product label.

Let H_{11} and H_{12} be the primary null hypotheses for regimens 1 and 2, and let H_{21} and H_{22} be the corresponding secondary null hypotheses. We want to select one of the two regimens by a data-dependent selection rule, test the primary hypothesis for the selected regimen, and then also test the secondary hypothesis for the selected regimen only when the primary hypothesis has been rejected. The FWER for the family $\mathcal{F} = \{H_{11}, H_{12}, H_{21}, H_{22}\}$ is the probability of rejecting at least one true hypothesis in $\mathcal{F}$ under any configuration of true and false hypotheses in $\mathcal{F}$. Because the actual regimen selection rule of the study is unknown, the FWER must be controlled at level α under any regimen selection rule.

As testing procedure 1 described in Section 3.1 does not control the FWER for the sub-family $\mathcal{F}_1 = \{H_{11}, H_{12}\}$ of primary hypotheses, there is no point to extend this procedure to testing the family $\mathcal{F}$ of primary and secondary hypotheses. Testing procedure 2 described in Section 3.2 controls the FWER for $\mathcal{F}_1$ but it does not for family $\mathcal{F}$ as the example in Appendix A shows (assuming that the primary hypothesis is tested at full level α).

Define Dunnett's critical value c_1 for the primary hypotheses so that $P_{H_{11}\cap H_{12}}(\min\{T_{11}, T_{12}\} \leq c_1) = \alpha$, where T_{ij} denotes the test statistic to test H_{ij} and T_{ij} is assumed to be standard normally distributed under H_{ij}; $i,j = 1,2$. Set $\alpha_1 = \Phi(c_1)$ and $\alpha_2 = \alpha - \alpha_1$, with an associated critical value $c_2 = \Phi^{-1}(\alpha_2)$. In Appendix B, we demonstrate that FWER is controlled for the family $\mathcal{F}$ with the conservative Dunnett procedure presented in Section 3.3 when testing the primary hypothesis at level α and the secondary hypothesis at level α_2.

Dimitrienko *et al.* [18] propose a similar parallel gatekeeping [19] procedure based on the Dunnett test [13], which accounts for the inherent correlation among all four test statistics. The method by Dmitrienko *et al.* [18] cannot be applied directly to our situation. It requires estimation of the correlation between secondary and primary endpoints, which is difficult for time to event endpoints like overall and progression free survival. Moreover, Dmitrienko *et al.* [18] assume a multivariate t-distribution for the test statistics, which is unsuitable for the corresponding log-rank tests. We therefore prefer to control type I error across endpoints by the Bonferroni inequality.

6. Estimation of treatment effect

It is well known that the treatment selection process in this type of seamless designs can induce bias in the final treatment effect estimate of the selected treatment(s). The literature on this topic has mostly addressed the problem of estimation of the selected treatment means in designs where treatment selection

is based on the rank order of the observed treatment effects of the primary outcome measure ([20–24]). Carreras and Brannath [24] showed that the maximum selection bias is achieved when all experimental treatments are equally efficacious, and in particular, when the global null hypothesis is true.

As treatment selection in the designs proposed in this article is based not on OS (primary endpoint) but on drug exposure (surrogate endpoint), and the joint (asymptotic) distribution of primary and surrogate endpoints is not known, no relevant theoretical results are available regarding bias or bias correction. Therefore, we assess selection bias in the final estimate of the HR between the selected treatment and the control via simulations.

Let $\beta_i = \exp(\theta_i)$ be the HR between the i-th experimental treatment and the control. Let $\hat{\beta}_i$ be the estimate of β_i; $i = 1, 2$, calculated at the end of the study. The selection bias for the estimate of the HR in the comparison of the selected treatment to the control at the end of the study is given by $b(\hat{\beta}_S) = E\left[\hat{\beta}_S - \beta_S\right]$. In the simulations, we estimate $b(\hat{\beta}_S)$ by computing the average estimation error over all simulation iterations.

Simultaneous (one-sided) confidence intervals that are consistent with the conservative Dunnett testing procedure presented in Section 3.3 can be constructed. The lower confidence limits with joint confidence coefficient $1 - \alpha$ for the two treatment effects θ_i are given by $U_{ib} + c_1\sqrt{V_{ib}}$; $i = 1, 2$, where c_1 is the Dunnett's critical value introduced in Section 5. No informative confidence intervals exist that are consistent with the testing procedures introduced in Sections 3.1 and 3.2, when the hypothesis for the selected regimen is rejected ([25, 26]).

7. Simulations

This section presents a simulation study that was performed to evaluate type I error, power and selection bias in the proposed seamless phase II/III adaptive design using the three testing procedures described in Sections 3.1, 3.2, and 3.3 as well as the treatment selection rules described in Section 4. We also compare the seamless adaptive design to the standard design in which phases II and III are run in separate studies as well as to the three-arm phase III design with regard to power. In order to compare the different designs in a fair manner, we fix the total number of patients recruited in the simulated study as well as the date of the final analysis. Note that, as long as recruitment rate is not data-driven, the conservative Dunnett and the patient-wise procedures will protect type I error rate.

The simulation input presented in the next two sections represent variations of assumptions made in the study protocol with regard to design characteristics such as recruitment rate, treatment allocation ratio, sample size, median survival, mean exposure, timing of interim and final analyses, and so on. As it was mentioned in the introduction, the study team was concerned that exposure of the low-dose regimen could be too low to ensure sufficient treatment effect, and therefore, a high-dose regimen was incorporated in the study. The team also expected the safety profiles of both regimens to be appropriate, and therefore, exposure was considered to be the main decision driver at the interim treatment selection.

This simulation study attempts to answer two key questions: (i) Is there any advantage of using an adaptive seamless design compared to a standard phase II + phase III design or a 3-arm phase III design? and 2) Is there any gain in using exposure to make the treatment selection at the interim analysis compared to a treatment selection based on OS?

7.1. Data generation

The study data for the seamless phase II/III adaptive design were generated in the following way. Study enrollment was simulated using piece-wise uniform distributions assuming a total recruitment of 7 patients per month in the first 4 months, 15 patients per month in the following 4 months and 22 patients per month thereafter. In the first stage of the study, patients were recruited in a 1:1:1 ratio to either experimental arm or the control, respectively. After a certain number of patients across all three arms was recruited (Section 7.2) and followed-up for at least 3 months, an interim regimen selection analysis was performed in which one of the experimental treatments was selected to continue, together with control, into the second stage. Patients in the second stage of the study were recruited in a 1:1 ratio to the selected experimental arm and control, respectively. A total of 410 patients was recruited across both stages. The final analysis was performed 29 months after the first patient was recruited. The one-sided significance level was set to $\alpha = 0.025$. All data from both stages were used in the final testing of efficacy.

Survival time was simulated for each patient in the study. Exposure was simulated for each patient in the experimental arms. The data generation for the experimental arms can be described using copulas [27].

Let X be the survival time of a generic patient in one of the experimental arms, and let Y be his or her log exposure. A copula function C relates the joint CDF, $H(x,y) = P(X \leqslant x, Y \leqslant y)$, to the marginal CDF's $F(x)$ and $G(y)$ by $H(x,y) = C(F(x), G(y))$. For the present simulation, we chose $F(x) = 1 - \exp(-\lambda x)$ the exponential CDF with rate λ, $G(y) = \Phi(\frac{y-\mu}{\sigma})$ the normal CDF with mean μ and variance σ^2 and $C(u,v) = \Phi_\rho(\Phi^{-1}(u), \Phi^{-1}(v))$, the normal copula, where Φ_ρ denotes the bivariate normal CDF with means 0, variances 1 and correlation ρ. Note that the normally distributed log-exposure assumption was based on pharmacology experts' opinion. Survival time for each patient in the control arm was generated using an exponential distribution.

The study data for the design of the standard program running phase II separate from phase III were generated in the same way as described earlier but the final testing of efficacy used only the data collected on stage 2 patients.

The generation of study data for the three-arm phase III design was performed in a slightly different way. A total of 410 patients was recruited across the three arms in a 1/1/1 treatment allocation ratio. There was no interim regimen selection. Only the survival time for each patient was generated in this case. Assumptions on study enrollment, timing of final analysis, distribution of survival times and significance level were the same as before. The classical Dunnett test was used for the final testing of efficacy.

7.2. Scenarios

We considered the following combinations of median OS (in months) for high-dose (H), low-dose (L), and control (C) arms and of mean exposure (in days$\times\mu g/mL$) for high-dose and low-dose arms:

(1) Median OS $(H, L, C) = (6, 6, 6)$ and mean exposure $(H, L) = (300, 200)$.
(2) Median OS $(H, L, C) = (9, 7.5, 6)$ and mean exposure $(H, L) = (300, 175)$.
(3) Median OS $(H, L, C) = (9, 7.5, 6)$ and mean exposure $(H, L) = (300, 185)$.
(4) Median OS $(H, L, C) = (9, 7.5, 6)$ and mean exposure $(H, L) = (300, 200)$.
(5) Median OS $(H, L, C) = (9, 7.5, 6)$ and mean exposure $(H, L) = (300, 210)$.
(6) Median OS $(H, L, C) = (9, 9, 6)$ and mean exposure $(H, L) = (300, 200)$.
(7) Median OS $(H, L, C) = (9, 9, 6)$ and mean exposure $(H, L) = (300, 210)$.

We also considered all possible combinations of the following parameters:

(1) standard deviation for exposure (in days$\times\mu g/mL$): 120 and 170;
(2) Copula correlation parameter ρ: 0.2, 0.5, 0.8, and 1;
(3) timing of interim analysis: after 100, 150, and 200 patients were recruited and followed for 3 months;

which resulted in a total of 168 simulation scenarios. In terms of data generation by the marginal distributions F and G and copula C in Section 7.1, we have

$$\begin{aligned} \lambda &= \ln(2)/\text{median(OS)} \\ \mu &= 2\ln(m) - \frac{1}{2}\ln\left(m^2 + \tau^2\right) \\ \sigma^2 &= \ln\left(m^2 + \tau^2\right) - 2\ln(m) \end{aligned} \tag{1}$$

where m is the mean exposure and τ^2 is the variance of exposure, $\exp(Y)$.

7.3. Simulation results

For every scenario described in Section 7.2, every testing procedure describe in Section 3 and every treatment selection rule described in Section 4, we simulated 1 million trials and computed type I error, power, probability of selecting the correct regimen, and selection bias. All graphs presented in this section (except those of Figure 6) use the same line styles. Benchmark scenarios are presented in bold. The dotted bold curve represents the perfect surrogate scenario described in Section 4.1. The dashed bold curve represents the case where treatment selection is based on the primary endpoint, OS. The remaining (non-bold) curves represent the approach where treatment selection is based on the surrogate endpoint, drug exposure. In particular, the curves with plotting symbols given by two dots and one dash, two dashes and one dot, three dots and one dash, and three dashes and one dot represent, respectively, the cases of copula correlation parameter ρ equal to 0.2, 0.5, 0.8, and 1.

7.3.1. Type I error, power and probability of 'correct' treatment selection. Figure 2 presents the type I error for testing the primary endpoint for the follow-up-wise (left graph), the patient-wise (middle graph), and the conservative Dunnett (right graph) testing procedures. When treatment selection is based on exposure (non-bold curves), we observe, as expected, a pattern of increased type I error with increased copula correlation between exposure and survival. We also see that the perfect surrogate approach provides an upper bound for the type I error probability. Type I error can be inflated for the follow-up-wise procedure if treatment selection is based on exposure and the copula correlation between exposure and survival is high. The patient wise and conservative Dunnett procedures control type I error as expected. The conservative Dunnett procedure is the most conservative procedure, which provides some safeguards against (minor) potential type I error inflation such as, for instance, due to changes in recruitment rate patterns driven by the interim results.

In these simulations, power is defined as the probability to reject the selected primary null hypothesis at the end of the study, which is correct as long as the selected treatment is more effective than control. Figure 3 presents power results for the conservative Dunnett procedure. Median OS of $(H, L, C) = (9, 7.5, 6)$ months, respectively, for the high-dose, low-dose, and control arms were assumed in all graphs. Mean exposure, m in equation (1), was assumed to be 300 for the high-dose arm in all graphs. Mean exposure of 175 (top-left graph), of 185 (top-right graph), of 200 (bottom-left graph), and of 210 (bottom-right graph) were assumed for the low-dose arm. A standard deviation for exposure, τ in equation (1), was set to 120 for both experimental regimens in all graphs. We observe that there is again a pattern of higher power for higher copula correlation between exposure and survival, when treatment selection is based on exposure. The performance of the treatment selection rule based on exposure depends strongly on the probability to select the correct regimen (high-dose regimen in this case). For instance, the probability of selecting the high-dose regimen was around 0.9, 0.78, 0.5, and 0.31, respectively, for the top-left, top-right, bottom-left, and bottom-right graphs, when treatment selection was based on exposure and when performing the treatment selection after 100 patients had been recruited and followed-up for 3 months (see left graph of Figure 6). On the one hand, treatment selection based on exposure performs better with regard to power than treatment selection based on OS in the two top graphs of Figure 3, independently of the copula correlation between survival and exposure. On the other hand, power is dramatically decreased in the two bottom graphs of Figure 3 when treatment selection is based on exposure compared to when treatment selection is based on OS because the standardized difference in exposure between the treatment groups is too small for a correct treatment selection with high probability. Hence, the performance of the selection rule based exposure depends strongly on the standardized mean exposure difference between the two dose regimens. We could conclude that the difference in exposure needs to be at least one standard deviation for treatment selection based on exposure to perform better than that based on OS. Note that Figures 3 and 6 indicate that both the probability of selecting the high-dose regimen and power do not depend strongly on the interim sample size. This means that, in this situation in which treatment selection is based only on exposure or survival, treatment selection should be performed rather early during the course of the study, thus avoiding unnecessary recruitment of patients to the non-selected arm (of course, this argument may not apply to the situation in which treatment selection is also based, for instance, on safety evidence). The other two testing procedures performed similarly with regard to power.

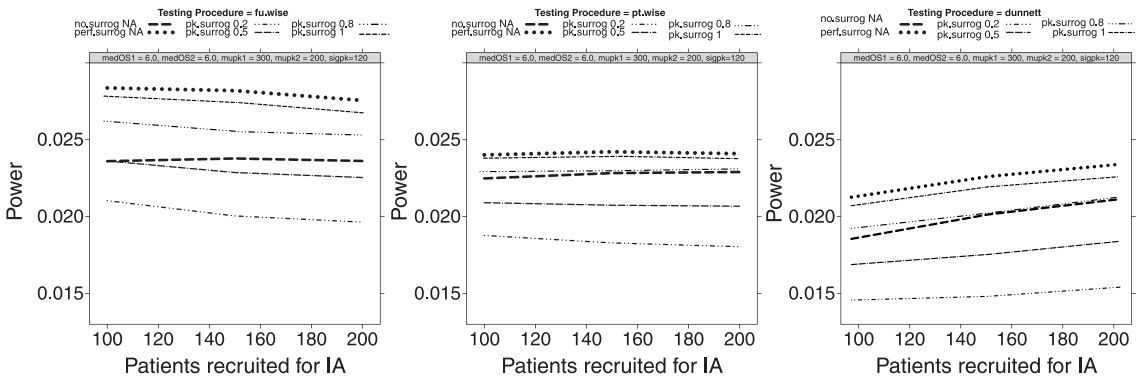

Figure 2. Type I error for follow-up-wise (left graph), patient-wise (middle graph), and conservative Dunnett (right graph) procedures. Median OS $(H, L, C) = (6, 6, 6)$, mean exposure $(H, L) = (300, 200)$, and standard deviation for exposure of 120 for all graphs.

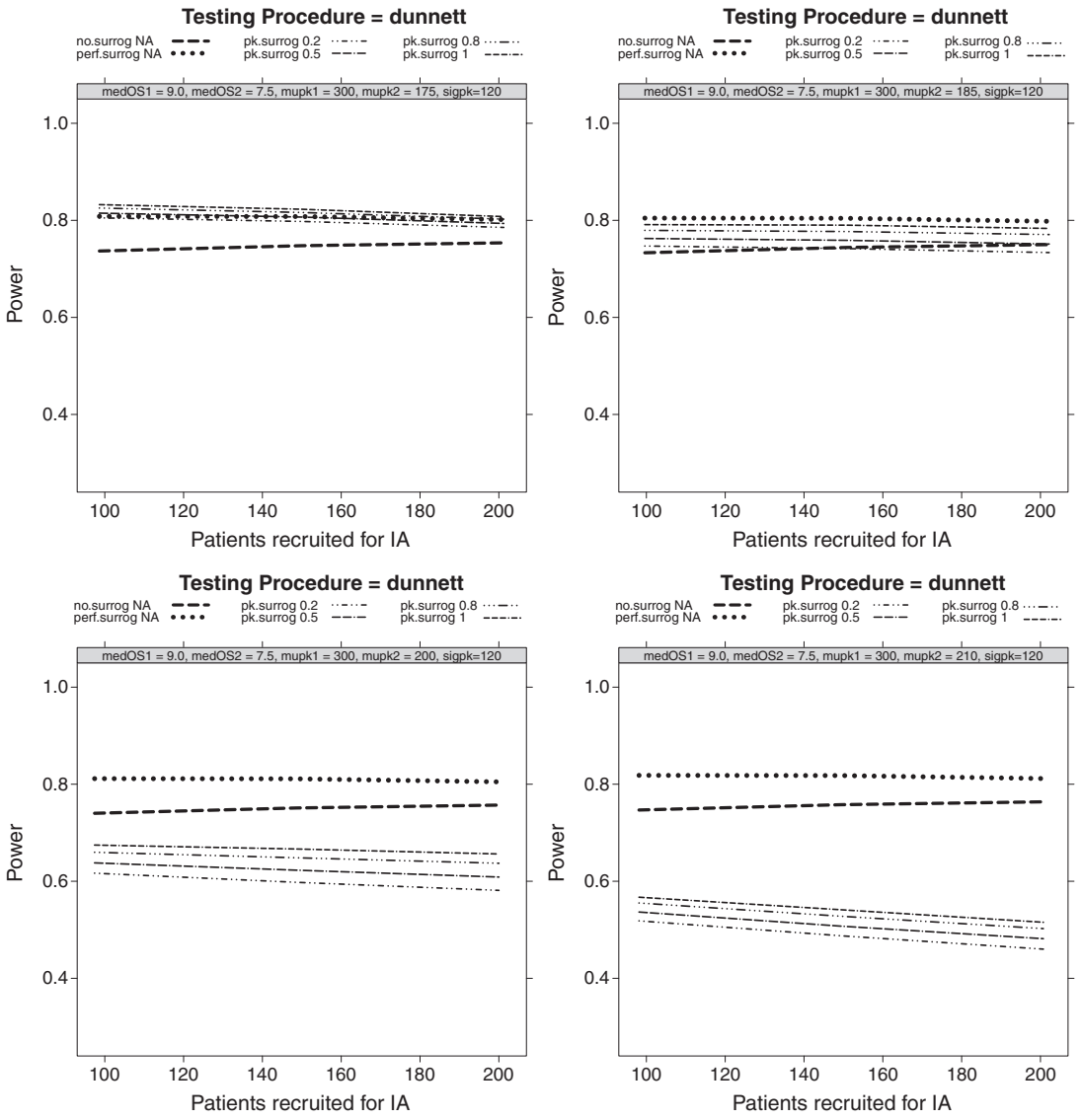

Figure 3. Power for the conservative Dunnett testing procedure. Median OS: $(H, L, C) = (9, 7.5, 6)$ in all graphs. Mean exposure of 300 for high-dose regimen. Mean exposure of 175, 185, 200, and 210 for the top-left, top-right, bottom-left, and bottom-right graphs, respectively. Standard deviation for exposure of 120 for all graphs.

Power was considerably lower for the standard phase II + phase III design than for the seamless phase II/III design, as Figure 4 shows. Because, in the standard phase II + phase III design, only stage 2 patients are used in the final testing of efficacy at the end of the study, the copula correlation between exposure and survival of stage 1 patients does not affect the power. We again observe that treatment selection based on exposure performs better with regard to power than treatment selection based on OS in the two top graphs of Figure 4 but power is dramatically decreased in the two bottom graphs of Figure 4 when treatment selection is based on exposure compared to when treatment selection is based on OS. We can conclude that, on the one hand, the seamless design is always preferred to the separate phase II + phase III design, and on the other hand, the good or bad performance of the treatment selection rule based on exposure remains the same whether a seamless design or a separate phase II + phase III design is used.

Power for the three-arm phase III design was 0.77 when median OS was set to $(H, L, C) = (9, 7.5, 6)$ months, respectively, for the high-dose, low-dose, and control arms, which is similar to the power that was obtained with the seamless phase II/III design when treatment selection was based on OS. Therefore, when the difference in exposure between the treatment groups is at least one standard deviation, the seamless design with treatment selection based on exposure is preferred, from a power perspective, compared to the 3-arm phase III design.

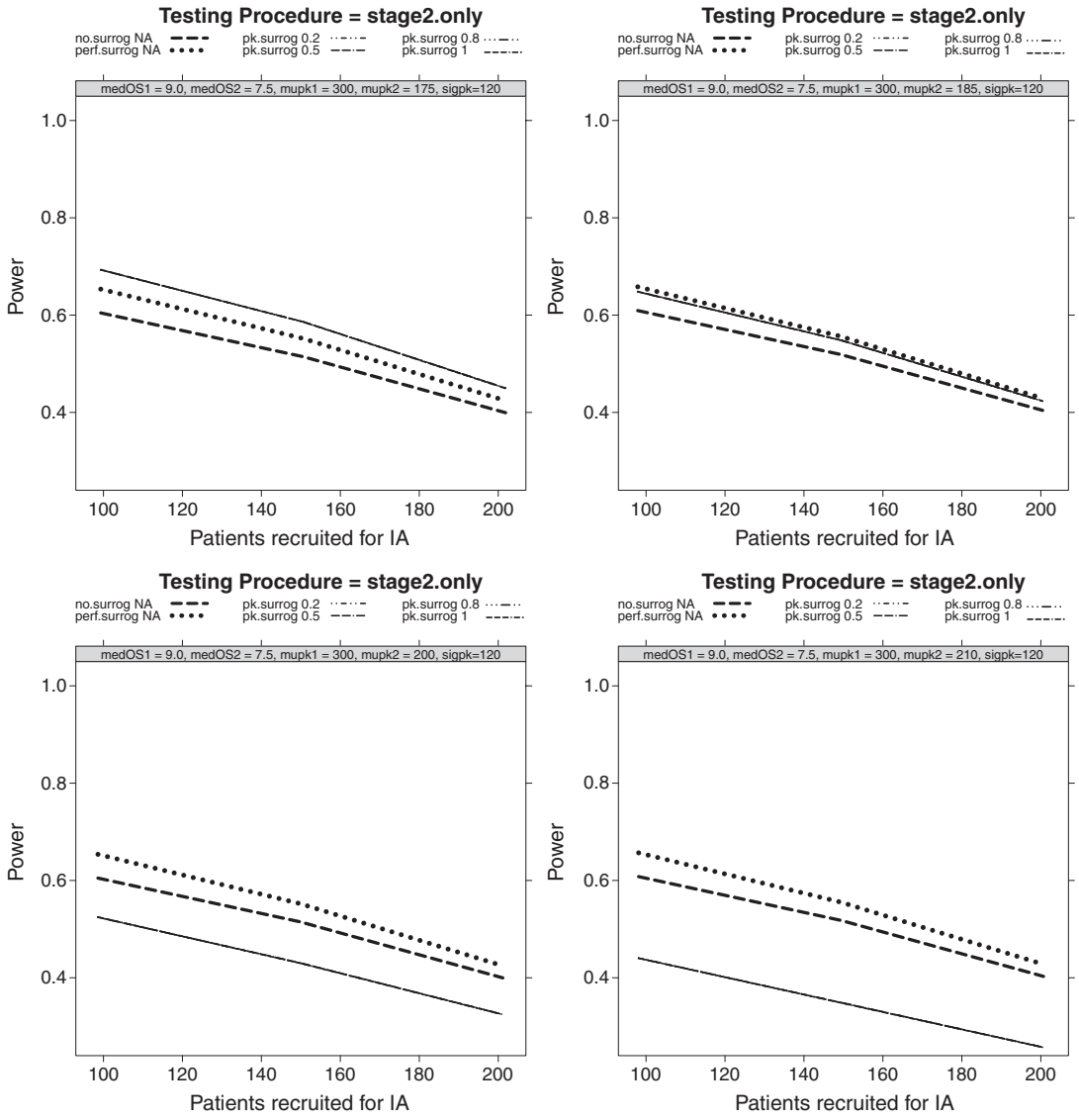

Figure 4. Power for the separate phase II + phase III design. Median OS: $(H, L, C) = (9, 7.5, 6)$ in all graphs. Mean exposure of 300 for high-dose regimen. Mean exposure of 175, 185, 200, and 210 for the top-left, top-right, bottom-left and bottom-right graphs, respectively. Standard deviation for exposure of 120 for all graphs.

Figure 5 presents power results for the conservative Dunnett procedure (left graph) and the separate phase II + phase III design (right graph) for the following scenario in both graphs: median OS was assumed to be $(H, L, C) = (9, 9, 6)$ months, respectively, for the high-dose, low-dose, and control arms, mean exposure of $(H, L) = (300, 200)$, respectively, for the high-dose and low-dose arms and a standard deviation for exposure of 120 was assumed. We observe again that the seamless design performs better than the separate phase II + phase III design. In this scenario, because both experimental regimens are equally effective, the correct regimen would be the low-dose regimen, which is expected to have a better safety profile. The probability of selecting the low-dose regimen when treatment selection is based on exposure is 0.5 in both graphs. We also considered a scenario in which mean exposure was assumed to be $(H, L) = (300, 210)$, respectively, for the high-dose and low-dose arms, and all other parameters remained the same compared to the previous scenario. Power results for these two scenarios were, as expected, very similar on the left graph and identical on the right graph, even though the probability of selecting the low-dose regimen in the latter scenario was, for instance, 0.69 when treatment selection was performed after 100 patients were recruited and followed up for at least 3 months. Once again, the other two testing procedures performed similarly with regard to power in this case.

Power for the three-arm phase III design was 0.88 when median OS was $(H, L, C) = (9, 9, 6)$, which was slightly lower than the power obtained with the seamless phase II/III adaptive design when treatment selection was based on OS. This is reasonable because, when both experimental regimens are equally effective, it is more efficient to select one of the two regimens quickly and use most of the patients for the comparison of the selected regimen with the control.

Power results for the cases in which the standard deviation for exposure was increased to 170 were similar to the corresponding results presented in this section. In Figure 6, we observe that, when treatment selection was based on exposure, the probability of selecting the correct regimen slightly decreased (increased) when mean exposure for the low-dose regimen was smaller (larger) than 200 in the case of standard deviation for exposure of 170 compared to the case of standard deviation for exposure of 120.

The aforementioned discussion about power provides only a narrow view of the main objective of the study. The objective should not really be maximizing the power at all costs but rather to find the correct

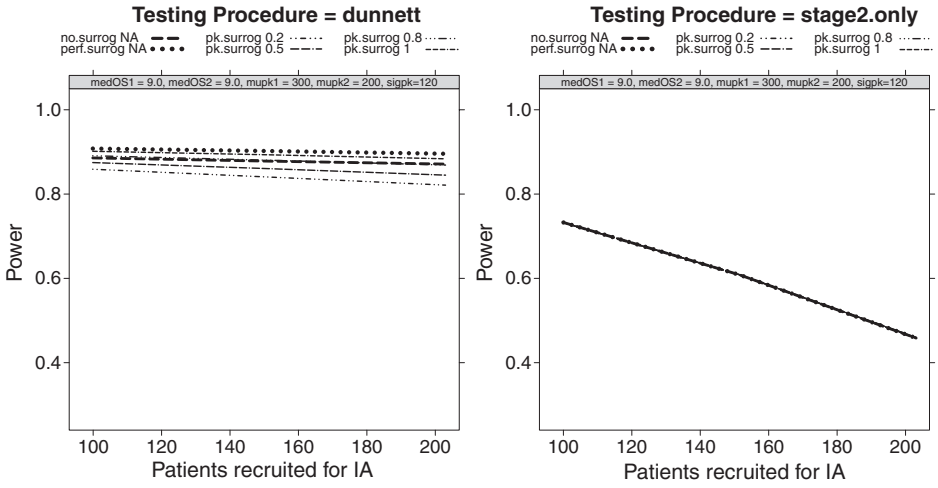

Figure 5. Power for the conservative Dunnett testing procedure (left graph) and for the separate phase II + phase III design (right graph). Median OS: $(H, L, C) = (9, 9, 6)$ in both graphs. Mean exposure of 300 and 200, respectively, for high-dose and low-dose regimens. Standard deviation for exposure of 120 for both graphs.

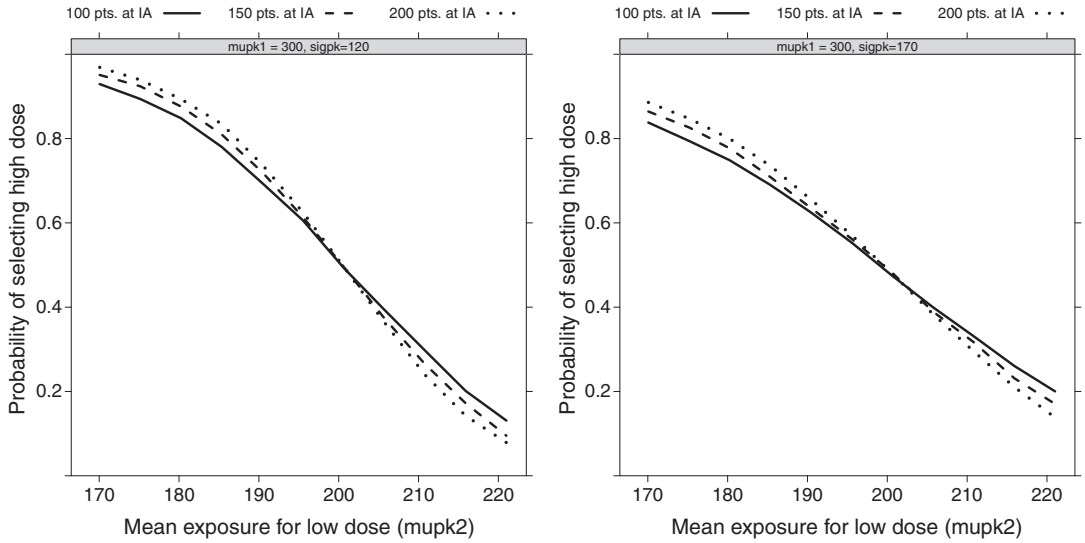

Figure 6. Probability to select the high-dose regimen when treatment selection is based on exposure. Mean exposure of 300 was assumed for the high-dose regimen in both graphs. Standard deviation for exposure was assumed to be 120 and 170, respectively, on the left and right graphs. Probabilities were obtained by simulations using a grid of mean exposures for the low-dose regimen from 170 to 220 in steps of 5.

dose regimen; that is, the regimen that provides an appropriate efficacy at the lowest feasible dose. Indeed, in the case study, the default preference is for the low-dose regimen, which is the standard regimen in MBC and for which the safety profile is well established. Having this in mind, one relevant question is: what is the probability of selecting the correct regimen at the interim analysis? This, of course, depends on the considered scenario, on the treatment selection rule and (though slightly) on the number of patients included in the interim analysis. For instance, in the scenarios in which median OS is $(9, 7.5, 6)$ months for the high dose, low dose and control respectively, mean exposure for the high-dose regimen is 300 and standard deviation for exposure is 120, we would need to have a (true) mean exposure for the low-dose of 185 or smaller in order to select the correct (high-dose) regimen with 0.8 or larger probability, when treatment selection is based on exposure and interim analysis is performed after 100 patients have been recruited and followed for 3 months (see left graph of Figure 6). Increasing the interim sample size up to 200 patients does not much improve the probability of a correct selection. Another relevant question is, how much does the probability of selecting the correct dose regimen increase if we waited until the end of the study to make the treatment selection (3-arm phase III design)? The probability of selecting the correct regimen (based on OS) at the end of the trial when median OS is assumed to be $(9, 7.5, 6)$ months respectively for the high-dose, low-dose and control treatments is 0.89, which is higher than the probability of selecting the correct regimen at interim analysis based on OS, when treatment selection is performed after 100, 150 and 200 patients are recruited and followed for 3 months, which is, respectively, 0.69, 0.73 and 0.77. Note that interim treatment selection based on exposure has the potential of improving the probability of selecting the correct regimen, and therefore, the power of the study (see two top graphs in Figure 3).

7.3.2. Selection bias. Selection bias is defined as the expected difference between the estimated and the true HR for the comparison of the selected regimen with the control at the end of the study [28]. In this definition, both the estimated and true HRs are random variables because of the data driven selection process. Selection bias comes from the fact that the selection process may tend to choose a regimen when its HR estimate is by chance smaller than expected. One example is when treatment selection is based on the HR for OS and both experimental regimens are equally effective. In this case, selecting the treatment that by chance has the smallest HR at the interim analysis and then reporting the estimated HR for this regimen will clearly overestimate the true HR. When one of the experimental regimens is more effective than the other, it will be more likely to select the most effective regimen at the interim analysis, but there will still be instances in which the least effective regimen will by chance have a smaller estimated HR than that of the most effective regimen. In this case, it is likely that the estimated HR of the (selected) least effective regimen will be an overestimation of its true HR. This will cause a negative bias. Similar arguments to those described earlier can be made when the regimen selection is based on exposure and the copula correlation between exposure and survival is high.

The most relevant selection bias results are presented in Figure 7 (results are the same for all testing procedures as bias is affected by the treatment selection rule but not by the testing procedure). Median OS was 6 months for all treatments (left graph). Median OS was 9, 7.5, and 6 months (middle graph) and 9, 9, and 6 months (right graph), respectively, for the high-dose, low-dose, and control arms. Mean exposure was assumed to be 300 and 200, respectively, for the high-dose and low-dose arms and standard deviation for exposure was assumed to be 120 in all graphs. We observe an expected pattern of larger (negative) bias for higher copula correlation between survival and exposure, when regimen selection is based on exposure.

In practice, neither the treatment effects nor the mean and standard deviation for exposure will be known. Therefore, one option would be to correct the final estimate of the treatment effect for the maximum bias observed in some plausible simulated scenarios. In our simulations, the worst bias was between -0.017 and -0.025 depending on the number of patients recruited before the interim analysis. (See left graph of Figure 7 for treatment selection rule based on exposure and copula correlation parameter ρ equal to 1. Bias was less severe in all other scenarios.) As an example, suppose that the interim analysis took place after 100 patients were recruited and supposed that the estimated HR for the comparison of the selected treatment to the control at the end of the study was, say, 0.67. Then, this estimated HR should be corrected to 0.69.

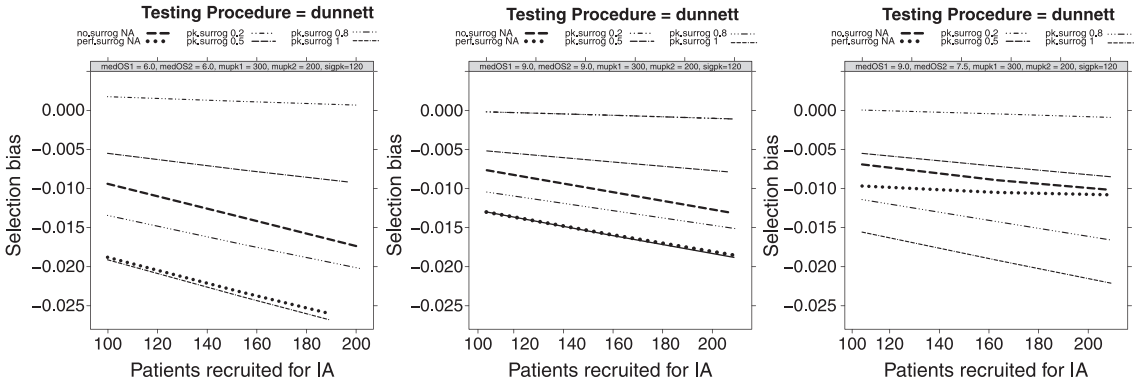

Figure 7. Selection bias. Median OS: $(H, L, C) = (6, 6, 6)$ (left graph), $(H, L, C) = (9, 9, 6)$ (middle graph), and $(H, L, C) = (9, 7.5, 6)$ (right graph). Mean exposure of $(H, L) = (300, 200)$ and standard deviation for exposure of 120 was assumed in all graphs.

8. Conclusions

In this article, we have discussed the planning of an oncology clinical trial with a seamless phase II/III adaptive design. Three testing procedures were considered. The standard follow-up-wise testing procedure that combines stage-wise p-values based on the independent increments property of the log-rank statistic (Schäffer & Müller [7], Wassmer [10]) did not protect the type I error when treatment selection was based on exposure and correlation between exposure and survival was high. One possibility would be to adjust the significance level to achieve type I error control to the specified level α for the perfect surrogate approach suggested in Section (4.1), which provides an upper bound for the type I error. This adjustment will of course affect the power of the study. The patient-wise testing procedure proposed by Jenkins *et al.* [9] protects the type I error when testing the primary endpoint as long as the follow-up of stage 1 patients remains unchanged after the regimen selection analysis, which may be difficult to achieve in practice. If, for instance, it is decided to discontinue treatment in the dropped arm (e.g., for ethical reasons), this may have an impact on the type I error rate: (i) because the p-value for the comparison of the dropped regimen with the control may no longer be uniformly distributed under the null hypothesis (if, e.g., these patients drop out and obtain an alternative treatment) and (ii) because this adaptation may affect the time point of the final analysis. The conservative Dunnett test procedure protects the type I error without placing any strong restrictions to the study design. It only requires that the recruitment of the second stage patients is independent from the interim data. It can also be extended to a procedure for testing primary and secondary endpoints, which protects the multiple type I error rate. Moreover, its conservatism provides a safeguard against unintended type I error inflations, for instance, due to unintended changes in the recruitment rate. Changes to treatment and/or follow-up of patients in the dropped arm do not affect the type I error as these patients are not used in the final analysis. The conservative Dunnett test also allows construction of simultaneous confidence intervals. A disadvantage of the Dunnett procedure is that it does not permit data driven changes of the preplanned overall event number. All three procedures have similar performance with regard to the power to reject the primary hypothesis for the comparison of the selected treatment to the control at the end of the study.

We have not found a clear advantage with regard to power in using adaptive designs compared to a three-arm phase III design. However, even though the probability of correct treatment selection is maximized with the three-arm phase III design, an adaptive design allows us to select a treatment early, avoiding an unnecessary large recruitment of patients to the non-selected arm, which is important from an ethical and practical point of view. If efficacy differences in survival are reflected by substantial differences in exposure, then a selection rule based on exposure can improve power compared to a rule based on survival only. However, power could be dramatically reduced when the surrogate endpoint is not able to select the correct treatment with high probability, independently of how highly correlated the primary and surrogate endpoints may be. A solution to this problem could be selection rules that incorporate both exposure and the immature primary endpoint. How to combine these two endpoints in a sensible selection rule is an interesting open research question. Treatment selection rules should always include the primary endpoint independently how immature the primary endpoint may be. In practice, selection rules will also incorporate safety parameters such as overall death rates, adverse events, serious adverse events,

treatment withdrawals, and exposure to treatment, which may be important drivers in the dosing decision. One design feature we have not considered in our investigation is the possibility to continue with both experimental arms to the end of the study if interim results do not show a clear advantage for one or the other. In this situation, the conservative Dunnett procedure could inflate the type I error because the overall number of events could not be pre-fixed.

Selection bias in the settings described in this paper was found not to be a serious problem. This may be related to the fact that only two experimental treatments are compared to a control and that the interim analysis is performed relatively early in the study. (Naturally, selection bias increases with an increased number of patients used at the treatment selection analysis as well as with a larger number of experimental treatments being compared.)

Finally, in order to minimize the risk of operational bias, it is important to restrict the knowledge of interim results as well as of detailed interim decision-making rules to the iDMC and potentially to a reduced number sponsor personnel that is not involved with the conduct of the study. However, as this study is open label, investigators may be subjectively influenced by the knowledge of the selected treatment regimen, and, therefore, operational bias cannot be completely ruled out. The sponsor should provide evidence that patients have been treated, managed, and evaluated according to the protocol throughout the entire study.

Appendix A. Example of FWER inflation for testing procedure 2

A.1. Example of family-wise error rate inflation

The following example shows that the closed testing procedure for testing the primary hypotheses exhausts the α level when the regimen selection rule can vary arbitrarily.

Let $I \subseteq \mathcal{F}$ be the set of true null hypotheses. Assume that $H_{11} \notin I$ and $H_{12} \in I$. Assume also that $H_{21} \in I$. If the effect of regimen 1 for the primary endpoint is very large, we can assume that H_{11} and $H_{11} \cap H_{12}$ can always be rejected. Consider the following regimen selection rule: 'select regimen 2 only when H_{12} and $H_{11} \cap H_{12}$ can be rejected at level α. Otherwise, select regimen 1'. Then,

$$\begin{aligned}\text{FWER} &\geqslant P\left(\{\text{Reject } H_{12}\} \cup \{\text{Reject } H_{21}\}\right)\\ &= P(\{\text{Reject } H_{12}\} \mid \text{select regimen 2})P(\text{select regimen 2})\\ &\quad + P(\{\text{Reject } H_{21}\} \mid \text{select regimen 1})P(\text{select regimen 1})\\ &= 1 \times \alpha + \alpha' \times (1-\alpha) > \alpha\end{aligned}$$

The aforementioned result is true for any level α' with which we test the secondary hypotheses. The aforementioned argument works when the regimen selection is performed at the end of the study. Otherwise, it is not possible to know for sure that H_{12} will be rejected at the end of the study when the regimen selection is performed at an interim analysis. But, some error inflation must be expected when the regimen selection is performed at the interim analysis. Of course, this example is based on an extreme regimen selection rule as well as on extreme assumptions about the effects of the two regimens for the different endpoints. But, as the actual regimen selection rule and treatment effects in the study are not known, we cannot rule out a potential FWER inflation.

Appendix B. Proof of FWER control for testing procedure 3

Define Dunnett's critical value c_1 for the primary hypotheses so that $P_{H_{11}\cap H_{12}}(T_{11} \vee T_{12} < c_1) = \alpha$. Set $\alpha_1 = \Phi(c_1)$ and $\alpha_2 = \alpha - \alpha_1$, with an associated critical value $c_2 = \Phi^{-1}(\alpha_2)$. Then $\alpha_1 > \alpha_2$, because

$$\alpha_2 = P_{H_{11}\cap H_{12}}\left(\{T_{11} < c_1\} \setminus \{T_{12} < c_1\}\right) \leqslant P_{H_{11}}(T_{11} < c_1) = \alpha_1,$$

and therefore

$$P_{H_{21}\cap H_{22}}(\min\{T_{21}, T_{22}\} < c_2) \leqslant 2\alpha_2 \leqslant \alpha_1 + \alpha_2 = \alpha \tag{B.1}$$

We can now verify that the FWER ξ is controlled by enumerating the possible cases for the set $I \subseteq \{H_{11}, H_{12}, H_{21}, H_{22}\}$ of true null hypotheses ($i, j \in \{1, 2\}$ with $i \neq j$):

(1) If $H_{1i} \in I$ and $H_{1j} \in I$, then $\xi = P_{H_{11} \cap H_{12}}(\min\{T_{11}, T_{12}\} < c_1) = \alpha$.
(2) If $H_{1i} \in I$ and $H_{1j} \notin I$, distinguish two cases:

   (a) If $H_{2j} \in I$, then $\xi \leqslant P_{H_{1i} \cap H_{2j}}\left(\{T_{1i} < c_1\} \cup \{T_{2j} < c_2\}\right) \leqslant \alpha_1 + \alpha_2 = \alpha$.
   (b) If $H_{2j} \notin I$, then $\xi = P_{H_{1i}}(T_{1i} < c_1) = \alpha_1 < \alpha$.

(3) If $H_{1i} \notin I$ and $H_{1j} \notin I$, distinguish three cases:

   (a) If $H_{2i} \in I$ and $H_{2j} \in I$, then $\xi \leqslant \alpha$ by Equation (B.1).
   (b) If $H_{2i} \in I$ and $H_{2j} \notin I$, then $\xi \leqslant P_{H_{2i}}(T_{2i} < c_2) = \alpha_2 < \alpha$.
   (c) If $H_{2i} \notin I$ and $H_{2j} \notin I$, then $I = \emptyset$ and $\xi = 0$.

Acknowledgements

Dr. Georg Gutjahr's research has been funded by Deutsche Forschungsgemeinschaft (DFG) Project BR 373/1-1.We would like to thank the referees for their constructive comments.

References

1. Bauer P, Brannath W. The advantages and disadvantages of adaptive designs for clinical trials. *Drug Discovery Today* 2004; **9**:351–357.
2. Bretz F, Koenig F, Brannath W, Glimm E, Posch M. Adaptive designs for confirmatory clinical trials. *Statistics in Medicine* 2009; **28**:1181–1217.
3. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 1999; **18**:1833–1848.
4. Hommel G. Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal* 2001; **43**:581–589.
5. Posch M, Koenig F, Branson M, Brannath W, Dunger-Baldauf C, Bauer P. Testing and estimation in fexible group sequential designs with adaptive treatment selection. *Statistics in Medicine* 2005; **24**:3697–3714.
6. Stallard N, Friede T. A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine* 2008; **27**:6209–6227.
7. Schäfer H, Müller H. Modication of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in Medicine* 2001; **20**:3741–3751.
8. Proschan MA, Hunsberger S. Designed extension of studies based on conditional power. *Biometrics* 1995; **51**:1315–1324.
9. Bauer P, Posch M. Letter to be editor: modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in Medicine* 2004; **23**:1333–1335.
10. Wassmer G. Planning and analyzing adaptive group sequential survival trials. *Biometrical Journal* 2006; **48**:714–729.
11. Jenkins M, Stone A, Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics* 2011; **10**:347–356.
12. König F, Brannath W, Bretz F, Posch M. Adaptive Dunnett tests for treatment selection. *Statistics in Medicine* 2008; **27**:1612–1625.
13. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical* 1995; **50**(272):1096–1121.
14. Di Scala L, Glimm E. Time-to-event analysis with treatment arm selection at interim. *Statistics in Medicine* 2011; **30**: 3067–3081.
15. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
16. Hochberg Y. A Sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988; **75**:800–802.
17. Dunnett CW, Tamhane AC. A step-up multiple test procedure. *Journal of the American Statistical Association* 1992; **87**:162–170.
18. Dmitrienko A, Offen W, Wang O, Xiao D. Gatekeeping procedures in dose-response clinical trials based on the Dunnett test. *Pharmaceutical Statistics* 2006; **5**(1):19–28.
19. Dmitrienko A, Offen W, Westfall P. Gatekeeping strategies for clinical trials that do not require all primary effects to be signifficant. *Statistics in Medicine* 2003; **22**:2387–2400.
20. Cohen A, Sackrowitz HB. Two stage conditionally unbiased estimators of the selected mean. *Statistics & Probability Letters* 1989; **8**:273–278.
21. Shen L. An improved method of evaluationg drug effect in a multiple dose clinical trial. *Statistics in Medicine* 2001; **20**:1913–1929.
22. Stallard N, Todd S. Point estimators and confidence regions for sequential trials involving selection. *Journal of Statistical Planning and Inference* 2005; **135**:402–419.
23. Bowden J, Glimm E. Unbiased estimation of selected treatment means in two-stage trials. *Biometrical Journal* 2008; **4**: 515–527.
24. Carreras M, Brannath W. Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection. *Statistics in Medicine* 2013; **32**(10):1677–1690.
25. Strassburger K, Bretz F. Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni-based closed tests. *Statistics in Medicine* 2008; **27**:4914–4927.

26. Guilbaud O. Simultaneous Confidence Regions Corresponding to Holm's Step-Down Procedure and Other Closed-Testing Procedures. *Biometrical Journal* 2008; **50**:678–692.
27. Nelsen RB. *An Introduction to Copulas*. Springer: New York, 1999. 75.4.800.
28. Putter J, Rubinstein D. On estimating the mean of a selected population. *Technical Report No. 165*, University of Wisconsin, Department of Statistics, 1968.