Flexible selection of a single treatment incorporating short-term endpoint information in a phase II/III clinical trial

Research Article

Statistics in Medicine

Received 19 December 2014, Accepted 1 June 2015 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6567

Flexible selection of a single treatment incorporating short-term endpoint information in a phase II/III clinical trial

Nigel Stallard,[a*†] Cornelia Ursula Kunz,[a] Susan Todd,[b] Nicholas Parsons[a] and Tim Friede[c]

Seamless phase II/III clinical trials in which an experimental treatment is selected at an interim analysis have been the focus of much recent research interest. Many of the methods proposed are based on the group sequential approach. This paper considers designs of this type in which the treatment selection can be based on short-term endpoint information for more patients than have primary endpoint data available. We show that in such a case, the familywise type I error rate may be inflated if previously proposed group sequential methods are used and the treatment selection rule is not specified in advance. A method is proposed to avoid this inflation by considering the treatment selection that maximises the conditional error given the data available at the interim analysis. A simulation study is reported that illustrates the type I error rate inflation and compares the power of the new approach with two other methods: a combination testing approach and a group sequential method that does not use the short-term endpoint data, both of which also strongly control the type I error rate. The new method is also illustrated through application to a study in Alzheimer's disease.

Keywords: adaptive design; conditional error; error rate control; multiple testing; sequential clinical trial

1. Introduction

Adaptive and sequential methods are often used in clinical trials to allow changes to be made to a trial design at one or more interim analyses during the course of the trial on the basis of the data observed. Such methods are appealing because they allow data observed early in the trial to be used to ensure that the trial is as efficient as possible [1].

Several authors have developed methods for such trials that test a single hypothesis to compare an experimental treatment with a control. These include the group sequential method [1, 2], the combination test method [3–5] and the conditional error function method [6, 7]. Building on this work, there has been much recent interest in trials involving selection of the most promising of a number of treatments, in what is sometimes called a *multi-arm, multi-stage* (MAMS) design, an *adaptive seamless design*, or a *seamless phase II/III trial* [8, 9], or of a subgroup of the population in which a therapy is particularly effective [10, 11].

A desire in the analysis of trials that use interim analysis data for treatment selection is usually the control of the overall type I error rate, and this has been the focus of most of the development of statistical methodology in the area. In trials in which multiple hypotheses are tested, it is usually required that the familywise error rate is controlled in the strong sense. Several authors have developed methods based on the group sequential approach. Compared with other approaches, the group sequential method benefits

[a] *Statistics and Epidemiology, Division of Health Sciences, Warwick Medical School, University of Warwick, Coventry, U.K.*
[b] *Department of Mathematics and Statistics, University of Reading, Reading, U.K.*
[c] *Department of Medical Statistics, University Medical Center, Göttingen, Germany*
[*] *Correspondence to: Nigel Stallard, Warwick Medical School, The University of Warwick, Coventry CV4 7AL, U.K.*
[†] *E-mail: n.stallard@warwick.ac.uk*

from the fact that inference is based on sufficient statistics for the parameters of interest [12–14] and may be preferable from a regulatory perspective [15]. The group sequential approach lacks flexibility, however, in that the rules for adaptation, in this case treatment selection, must generally be specified in advance in order to ensure type I error rate control. This has been termed *pre-specified adaptivity*.

A number of authors [8, 16–22] have proposed group sequential methods for selection of one or more treatments using a pre-specified rule. The selection rules specify either the conditions under which treatments should be dropped, as proposed by Follman [17], Hellmich [18] and Magirr *et al.* [22], or the number of treatments to continue at each stage, as proposed by Thall *et al.* [16], Stallard and Todd [8], Bischoff and Miller [19] and Stallard and Friede [20]. Here, we take the latter approach and consider two-stage designs in which the data available at the interim analysis are used for selection of a single experimental treatment, which continues along with the control to the second stage.

Thall *et al.* [16] and Stallard and Todd [8] show how the type I error rate can be controlled when selection of a single experimental treatment is based on the primary endpoint data alone and the most promising treatment is selected. Jennison and Turnbull [23] and Graf *et al.* [24] point out that the method proposed by Thall *et al.* [16] and Stallard and Todd [8] also controls the type I error rate when the treatment selected is not the most promising.

In some cases, in addition to primary endpoint data, data on some short-term endpoint may also be used for treatment selection. As such data may be observed more quickly than the primary endpoint, short-term endpoint data may be available at the interim analysis for patients for whom the primary endpoint data are not yet available. Stallard [21] showed how to control the type I error rate in this case, again assuming that the most promising experimental treatment was selected to continue beyond the interim analysis. The aim of this paper is to explore the properties of this approach when some treatment other than the most promising is selected, in particular when the selection rule is not specified in advance. We show that in this case, unlike the setting in which primary endpoint data only are used at the interim analysis considered by Jennison and Turnbull [23] and Graf *et al.* [24], such selection may lead to inflation of the type I error rate. Based on consideration of a trial in which the selection is made so as to maximise the conditional type I error given the data available at the interim analysis, we show how to construct critical values for the final hypothesis test to ensure that the familywise type I error rate is controlled strongly for any selection rule.

2. Motivating case study

There is a growing number of examples of multi-arm clinical trials that incorporate treatment selection at an interim analysis, illustrating the range of areas of application and variety of details of implementation of treatment selection and interim analyses, with four examples given in the recent paper by Cuffe *et al.* [25].

Wilkinson and Murray [26] describe a multi-stage phase II trial in Alzheimer's disease in which patients are initially randomised between a placebo control and 18, 24 and 36 mg/day doses of the exploratory drug, galantamine. In this study, the primary endpoint was a 3-month change in Alzheimer's Disease Assessment Scale cognitive subsection (ADAS-Cog) score, but both 3-month and 6-week change data were used at an interim analysis for dose selection. Wilkinson and Murray found the 24 mg/day dose to be the most promising, although further research [27] indicated that lower doses may be as efficacious.

In the setting of a phase II/III trial in Alzheimer's disease, the primary endpoint might be change in ADAS-Cog score over 6 months, but, as in the trial described by Wilkinson and Murray, data on the change over a shorter period, for example, over 3 months, might be recorded and be available more rapidly than the final endpoint data. A decision at an interim analysis of which dose should continue along with the control to the second stage of the trial could thus be based on a combination of 6- and 3-month data. Trials such as this motivate the work described in this paper. The aim is to provide a method that strongly controls the familywise type I error rate while allowing for use of all interim analysis data for treatment selection.

3. Error rate control for flexible selection of a single treatment in a two-stage multi-arm clinical trial

Consider a multi-arm clinical trial comparing k ($k \geqslant 2$) experimental treatments, treatments $1, \ldots, k$, with a control treatment, treatment 0. Let θ_i be a measure of the efficacy of treatment i relative to treatment 0 for $i = 1, \ldots, k$, in terms of the primary endpoint and suppose that we wish to test a family of null

hypotheses $H_i, i \in \{1, \ldots, k\}$, with $H_i : \theta_i \leqslant \theta_0$ for specified θ_0, which, without loss of generality, we may take to be zero. We wish to control the familywise error rate in the strong sense for testing this family of null hypotheses.

Suppose that the trial is conducted in two stages, with, in the first stage, primary endpoint responses available for n_1 patients randomised to each of treatments $0, \ldots, k$. Let X_1 denote the full data observed in the first stage, which includes the primary responses from these n_1 patients in each treatment group but, as in the example described earlier, may also include data from additional patients from whom primary endpoint data are not yet available.

On the basis of the data X_1, treatment $T(X_1)$, for some $T(X_1) \in \{1, \ldots, k\}$, will be chosen to continue along with the control to the second stage with the hypothesis, $H_{T(X_1)}$ tested at the end of the trial. Primary responses are then observed for a further $n_2 - n_1$ patients randomised to each of treatments $T(X_1)$ and 0 in the second stage of the trial. Let S_i denote a test statistic for H_i based on data observed in both stages of the trial. We assume that the distribution of S_i depends on θ_i, but not on $\theta_{i'}$ with $i' \neq i$, and that a larger θ_i value leads to larger values of S_i, as formalised in the Appendix. The hypothesis $H_{T(X_1)}$ will be rejected at the end of the trial if and only if $S_{T(X_1)} \geqslant c$ for some critical value c, which will be chosen so as to provide strong familywise error rate control.

As indicated earlier, in general, the data X_1 may include data other than the primary responses summarised by $S_1, \ldots, S_k$ and so may depend on further parameters in addition to $\theta_1, \ldots, \theta_k$. When this is the case, we will denote these additional parameters by $\theta_{k+1}, \ldots, \theta_{k^*}$ with $k^* > k$. Otherwise, we will define $k^* = k$. We will write $\boldsymbol{\theta}$ for the vector of all parameters, that is, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{k^*})'$.

We require to strongly control the familywise error rate at level α, that is, to ensure that the probability of rejecting any true H_i $(i = 1, \ldots, k)$ is at most α for any $\boldsymbol{\theta}$. This is required for any data-dependent choice of $T(X_1)$. We therefore wish to find c such that

$$\mathrm{pr}_{\boldsymbol{\theta}}(S_{T(X_1)} \geqslant c, \theta_{T(X_1)} \leqslant 0) \leqslant \alpha \tag{1}$$

for all $\boldsymbol{\theta}$, and for any T in $\mathcal{T}$, where $\mathcal{T}$ denotes the set of functions from the stage 1 sample space to $\{1, \ldots, k\}$.

Given $\boldsymbol{\theta}$ with $\theta_i \leqslant 0$, that is H_i true, for some $i = 1, \ldots, k$, let

$$T^*_{\boldsymbol{\theta}}(X_1) = \arg\max_{i \in \{1,\ldots,k\} : \theta_i \leqslant 0} \left\{ \mathrm{pr}_{\boldsymbol{\theta}}(S_i \geqslant c \mid X_1) \right\}.$$

Thus, $T^*_{\boldsymbol{\theta}}$ denotes the rule that selects the treatment, treatment i, for which the conditional probability given X_1 of rejecting H_i is highest amongst those i for which H_i is true given $\boldsymbol{\theta}$. That is, $T^*_{\boldsymbol{\theta}}(X_1)$ is chosen to maximise the conditional error given X_1 under $\boldsymbol{\theta}$.

It can be shown (Appendix) that the left-hand side of (1) is maximised over $T \in \mathcal{T}$ by taking $T(x_1) = T^*_{\boldsymbol{\theta}}(x_1)$ for all x_1 and maximised over $\theta_1, \ldots, \theta_k$ by taking $\theta_1 = \cdots = \theta_k = 0$.

To satisfy (1) and control the familywise error rate in the strong sense, it is therefore sufficient to have

$$\mathrm{pr}\left(S_{T^*_{\boldsymbol{\theta}}(X_1)} \geqslant c; \theta_1 = \cdots = \theta_k = 0\right) \leqslant \alpha \tag{2}$$

for all $\theta_{k+1}, \ldots, \theta_{k^*}$.

In the case that $k = k^*$, when no short-term endpoint data are available, (2) states that the familywise error rate is controlled at level α in the weak sense, that is, under the global null hypothesis $\cap_{i=1,\ldots,k} H_i$. The result in the Appendix shows that strong control of the familywise type I error rate is also achieved. For $k = k^*$, $\boldsymbol{\theta}$ and hence $T^*_{\boldsymbol{\theta}}$ are entirely defined. In order to find the value of the critical value c to satisfy (2) and hence (1), it is thus necessary to obtain the distribution of $S_{T^*_{\boldsymbol{\theta}}(X_1)}$. Although in general this may not be a straightforward problem, the distribution is obtained for normally distributed responses by Thall *et al.* [16] and Stallard and Todd [8].

In general for $k^* > k$, the probability in (1) must be controlled for all $\theta_{k+1}, \ldots, \theta_{k^*}$. In this case, it is therefore necessary to find the values of $\theta_{k+1}, \ldots, \theta_{k^*}$ to maximise the probability in (2) to obtain the critical value c. Depending on the setting, this may be possible. For example, it may be that the values of $\theta_{k+1}, \ldots, \theta_{k^*}$ to maximise the error rate can be found directly, or that the probability in (2) does not depend on $\theta_{k+1}, \ldots, \theta_{k^*}$. The latter situation arises when data from different patients are independent and treatment selection is based either only on data from patients for whom primary endpoint data are available or on a combination of primary endpoint data and short-term endpoint data available for patients for whom the primary endpoint is not yet available, as discussed in more detail in the next section.

4. Application with normally distributed responses and short-term endpoint data

In this section, the method described earlier is illustrated through application in the setting of a clinical trial with a normally distributed primary endpoint. In many cases, similar settings with non-normal data may be handled using asymptotically normally distributed test statistics as described in the group sequential setting by Jennison and Turnbull [28].

Suppose that in stage 1, primary endpoint data, Y_{ij}, are observed for patients $j = 1, \dots, n_1$ receiving treatment i and that in stage 2, data Y_{ij} are additionally observed for patients $j = n_1 + 1, \dots, n_2$ for the control, $i = 0$ and $i = T(X_1)$. Assume $Y_{ij} \sim N(\mu_i, \sigma^2)$, with data from different patients independent, that is, $cov(Y_{ij}, Y_{i'j'}) = 0$ unless $i = i'$ and $j = j'$, and let $\theta_i = \mu_i - \mu_0$ $(i = 1, \dots, k)$.

We assume that at the first stage, we additionally observe short-term endpoint data, W_{ij}, for patients $j = 1, \dots, N_1$ in treatment group i, $i = 0, \dots, k$, where $N_1 > n_1$, and that at the end of the second stage, data are available for $n_2 \geqslant N_1$ patients per group including the N_1 with short-term endpoint data available at the first stage. We assume

$$\begin{pmatrix} Y_{ij} \\ W_{ij} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_i \\ \mu_{k+1+i} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma\sigma_0 \\ \rho\sigma\sigma_0 & \sigma_0^2 \end{pmatrix} \right), \tag{3}$$

with $cov(W_{ij}, W_{i'j'}) = 0$, $cov(Y_{ij}, Y_{i'j'}) = 0$, and $cov(W_{ij}, Y_{i'j'}) = 0$ $(i \neq i' \text{ or } j \neq j')$. Noting that μ_{k+1} is the mean for the short-term endpoint for the control treatment, the short-term treatment effect for treatment i is given by $\mu_{k+1+i} - \mu_{k+1}$, which will be denoted θ_{k+i}, $(i = 1, \dots, k)$. The parameters of interest are θ_i, $i = 1, \dots, k$, and it is desired to test the null hypotheses $H_i : \theta_i \leqslant 0$ $(i = 1, \dots, k)$. The short-term endpoint treatment effects, θ_{k+i} $(i = 1, \dots, k)$, are not of interest, so that we are interested in testing k null hypotheses, with the distribution of the data depending on $k^* = 2k$ parameters.

Assume that at the end of the trial a test statistic $S_i = \sum_{j=1}^{n_2} \left(Y_{ij} - Y_{0j}\right)$ will be used for testing hypothesis H_i. The results stated earlier and proved in the Appendix show that in order to find c to satisfy (1), we need to consider only the case $\theta_1 = \cdots = \theta_k = 0$ and $T(x_1) = T^*_{\boldsymbol{\theta}}(x_1) = \arg\max_{i \in \{1,\dots,k\} : \theta_i \leqslant 0} \left\{ \text{pr}_\theta(S_i \geqslant c \mid X_1 = x_1) \right\}$. The resulting form of the rule, $T^*_{\boldsymbol{\theta}}(x_1)$, that maximises the conditional error over $\theta_{k+1}, \dots, \theta_{k^*}$ is obtained later.

We consider first the case with σ, σ_0 and ρ, assumed known. It can be shown that, given $X_1 = (W_1, \dots, W_{N_1}, Y_1, \dots, Y_{n_1})' = (w_1, \dots, w_{N_1}, y_1, \dots, y_{n_1})'$, the conditional distribution of S_i is given by

$$S_i \sim N\left((n_2 - n_1)\theta_i + n_1\tilde{\theta}_i, 2\left((n_2 - n_1) - (N_1 - n_1)\rho^2\right)\sigma^2 \right),$$

where

$$\tilde{\theta}_i = \breve{\theta}_i + \rho\frac{\sigma}{\sigma_0}\sum_{j=n_1+1}^{N_1}(w_{ij} - w_{0j} - \theta_{k+i}))/n_1$$

with $\breve{\theta}_i = \sum_{j=1}^{n_1}(y_{ij} - y_{0j})/n_1$ denoting the estimate of θ_i based on the interim primary endpoint data alone.

Thus, if $\theta_1 = \cdots = \theta_k = 0$, $T^*_{\boldsymbol{\theta}}(x_1) = \arg\max_{i \in \{1,\dots,k\}} \tilde{\theta}_i$. The selection rule $T^*_{\boldsymbol{\theta}}(X_1)$ would maximise the error rate if $\theta_{k+1}, \dots, \theta_{k^*}$ were known. Even if $\theta_{k+1}, \dots, \theta_{k^*}$ are unknown, however, full flexibility over the choice of T means that this rule might be chosen. The critical value, c, must thus be calculated based on this rule to ensure error rate control.

The joint distribution of $\tilde{\theta}_1, \dots, \tilde{\theta}_k$ and $S_1, \dots, S_k$ is given by

$$(\tilde{\theta}_1, \dots, \tilde{\theta}_k, S_1, \dots, S_k)' \sim N\left((\theta_1, \dots, \theta_k, \theta_1, \dots, \theta_k)', \begin{pmatrix} V_1 & \rho_e\sqrt{V_1V_2} \\ \rho_e\sqrt{V_1V_2} & V_2 \end{pmatrix} \otimes \boldsymbol{\Sigma} \right)$$

with $V_1 = 2\sigma^2(n_1 + \rho^2(N_1 - n_1))/n_1^2$, $V_2 = 2n_2\sigma^2$ and $\rho_e = \sqrt{(n_1 + \rho^2(N_1 - n_1))/n_2}$. As this distribution does not depend on $\theta_{k+1}, \dots, \theta_{k^*}$, the distribution of $S_{T^*_{\boldsymbol{\theta}}(X_1)}$ does not depend on $\theta_{k+1}, \dots, \theta_{k^*}$ and is entirely specified by taking $\theta_1 = \cdots = \theta_k = 0$.

Noting that the conditional variance of S_i given $\tilde{\theta}_i$ is $V_2\left(1 - \rho_e^2\right)$, $S_{T^*_{\boldsymbol{\theta}}(X_1)}$ has density

$$\sum_{i=1}^{k}\int_{-\infty}^{\infty}\frac{1}{\sqrt{V_2\left(1 - \rho_e^2\right)}}\phi\left(\frac{s - (n_2 - n_1)\theta_i - n_1x}{\sqrt{V_2\left(1 - \rho_e^2\right)}}\right)f_i(x, V_1)dx$$

and distribution function

$$\sum_{i=1}^{k} \int_{-\infty}^{\infty} \Phi\left(\frac{s-(n_2-n_1)\theta_i - n_1 x}{\sqrt{V_2\left(1-\rho_e^2\right)}} \right) f_i(x, V_1) dx,$$

where

$$f_i(x, V) = \int_{-\infty}^{\infty} \frac{1}{V} \phi\left(\frac{x-y}{\sqrt{V}} \right) \phi\left(\frac{y-\theta_i}{\sqrt{V}} \right) \prod_{i' \neq i} \Phi\left(\frac{y-\theta_i}{\sqrt{V}} \right) dy$$

and ϕ and Φ denote the density and distribution functions of the standard normal distribution [8]. Numerical integration and a simple numerical search, for example, using the R function `uniroot`, can thus be used to find the critical value c to control the error rate as required.

When σ, σ_0 and ρ are not known, estimates obtained at the interim analysis can be used in this expression to find c.

5. Example

In this section, the method described earlier is illustrated through application to a simulated dataset based on the setting of the phase II/III trial described in Section 2. We assume that three doses, 16, 24 and 32 mg/day of galantamine, are compared with a placebo control. In the first stage of the trial, $N_1 = 100$ patients per group are randomised between the four treatment arms, with an interim analysis being conducted when w_{ij}, the 3-month endpoint data, are available for all of these patients. It is assumed that at this time y_{ij}, the primary, 6 month, endpoint data, are available for $n_1 = 40$ patients per arm. Data were simulated from a model based on the Cochrane review of clinical trials of galantamine by Loy and Schneider [27].

The simulated interim analysis data are shown in Figure 1 giving the change from baseline to 3- and 6-month ADAS-Cog scores with the sign of the change chosen so that positive changes correspond to an improvement. Plotted points indicate values of the long-term and short-term endpoint data for the n_1 patients per group for whom these are both available. The tick marks at the bottom of each plot give the values of the short-term endpoint data for the remaining $N_1 - n_1$ patients per group for whom long-term endpoint data are not available at this time. The mean long-term response for the n_1 patients in each group are given in Table I along with the mean short-term response for these n_1 patients per group, for the additional $N_1 - n_1$ patients per group for whom only short-term endpoint data are available and for all N_1 patients per group included in the interim analysis.

Using the approach of Stallard and Todd [8], treatment selection would be based on the long-term endpoint data alone. The estimate of the treatment effect for treatment i relative to the control, θ_i, is then given by $\breve{\theta}_i = \sum_{j=1}^{n_1}(y_{ij} - y_{0j})/n_1$, which is given in Table I. In this case, this would lead to selection of the 32 mg/kg dose. Assuming treatment selection is always made using this approach, the critical value for the final test statistic needs to be adjusted as described by Stallard and Todd [8] in order to control the type I error rate. In this case, the critical value depends on n_1 and N. The critical value, $c/\sqrt{V_2}$, with which a standardised test statistic should be compared with give a one-sided type I error rate of $\alpha = 0.025$ is 2.19.

Use of the additional 3-month endpoint data would improve the treatment selection. The method of Stallard [21] would base treatment selection on the estimate of θ_i given by

$$\hat{\theta}_i = \breve{\theta}_i - \rho \frac{\sigma}{\sigma_0} \sum_{j=1}^{n_1} \left(w_{ij} - w_{0j} - \hat{\theta}_{k+i} \right) / n_1,$$

with $\hat{\theta}_{k+i} = \sum_{j=1}^{N_1}(w_{ij} - w_{0j})/N_1$. These estimates are also given in Table I. Figure 1 shows that for $i = 3$ values of w_{ij} for $j = n_1, \ldots, N_1$ are typically smaller than those for $j = 1, \ldots, n_1$, as is reflected in the mean values given in Table I. This leads to $\hat{\theta}_3$ being considerably smaller than $\breve{\theta}_3$, so that $\hat{\theta}_3 < \hat{\theta}_2$. Incorporation of the additional short data would thus lead to selection of the 24 mg/day dose rather than 32 mg/day. If this approach is always used for treatment selection, the critical value needs to be adjusted to allow for the use of the additional 3-month endpoint information as described by Stallard [21]. In this case, the

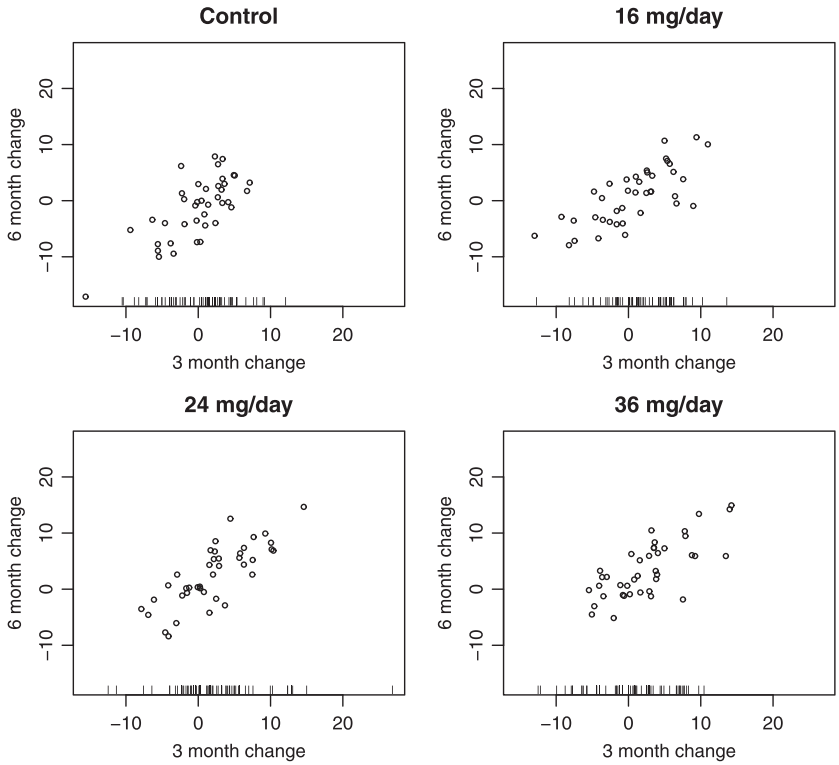

Figure 1. Interim analysis data for simulated example (see main text for details).

Table I. Summary of interim analysis results for simulated example.

Dose	$\frac{\sum_{j=1}^{n_1} y_{ij}}{n_1}$	$\frac{\sum_{j=1}^{n_1} w_{ij}}{n_1}$	$\frac{\sum_{j=n_1+1}^{N_1} w_{ij}}{N_1-n_1}$	$\frac{\sum_{j=1}^{N_1} w_{ij}}{N_1}$	$\breve{\theta}_i$	$\hat{\theta}_i$	$\tilde{\theta}_k$
Control ($i=0$)	−1.24	−0.05	0.06	0.02			
16 mg/day ($i=1$)	0.91	0.56	1.48	1.11	2.15	2.53	2.06
24 mg/day ($i=2$)	2.64	2.15	2.43	2.32	3.88	3.96	3.74
32 mg/day ($i=3$)	3.62	2.59	0.36	1.25	4.86	3.77	2.30

critical value depends on n_1, N_1, N and the correlation between the endpoints, ρ. Using an estimate of ρ from the interim analysis data, in this case, 0.77, leads to a critical value with which a standardised test statistic should be compared of 2.23.

As described earlier, the conditional type I error rate is maximised by selecting the treatment based on the estimate $\tilde{\theta}_i$. This requires specification of the true short-term endpoint treatment effects $\theta_{k+i}, i = 1, \ldots, 3$. Based on [27], for example, a researcher might guess that the true treatment effects relative to the control might be close to 1.5, 2.5 and 2.5 for $i = 1, 2$ and 3, that is, for the 16, 24 and 32 mg/day doses. The values of $\tilde{\theta}_k$ obtained using these values for $k + i$ are given in Table I. In this case, this would again lead to selection of the 24 mg/day dose. If a larger value was assumed for θ_{k+i}, $\tilde{\theta}_i$ would be reduced, reflecting the fact that the observed values of w_{ij} are smaller relative to their expected value. Thus, for example, if it was assumed that the true treatment effects for the three doses were 1.5, 4.0 and 4.0, $\tilde{\theta}_2$ and $\tilde{\theta}_3$ would be smaller than $\tilde{\theta}_1$, so that the 16 mg/day dose would be selected. Note that using the estimated values for θ_{k+i}, $i = 1, \ldots, 3$ based on the interim analysis data, denoted by $\hat{\theta}_{k+i}$ earlier, leads to $\tilde{\theta}_i$ equal to $\hat{\theta}_i$ given earlier, reflecting the fact that no additional information is available in this case.

As this selection maximises the type I error given the observed data, in the case of fully flexible treatment selection, selection based on $\tilde{\theta}_i$ could lead to inflation of the type I error rate for appropriate choice of θ_{k+i}, $i = 1, \ldots, 3$ unless the critical value was adjusted to allow for this. In this case, the critical value that controls the type I error rate again depends on n_1, N_1, N and ρ. Using the estimated value of ρ as mentioned earlier, the critical value for a standardised test statistic is 2.25.

Assuming selection of the 24 mg/day dose to continue to the second stage, simulated means at the end of the trial for the 6-month change for control and 24 mg/day arms were −1.39 and 2.07, respectively,

with an estimated standard error for the difference between them, $\sqrt{V_2}$, of 0.525, leading to standardised test statistic of 6.60. A significant treatment effect is thus indicated for this dose irrespective of the way in which it was selected.

6. Simulation study

In order to illustrate the properties of the method described and to compare them with those of alternative approaches, it is of interest to consider outcomes from a large number of simulated trials under specified scenarios.

One alternative to the phase II/III trial considered earlier would be to conduct separate phase II and phase III trials, the first used for selection of the most promising dose and the second comparing this dose with a placebo. As previous simulations (for example, [29]) have shown that such an approach can require considerably more patients than a two-stage phase II/III design, this option was not included in our simulation study.

We considered two-stage trials with, as mentioned earlier, $\alpha = 0.025, k = 3, n_1 = 40, N_1 = 100$ and $n_2 = 200$. Critical values with which standardised test statistics, $S_i/\sqrt{V_2}$, would be compared in this trial using the earlier approach to control the type I error rate for any treatment selection rule are given in Table II for a range of ρ values calculated assuming σ, σ_0 and ρ are known. The critical values increase with ρ to control the type I error rate. Results from simulations of this procedure are shown in columns 3 to 5 of the table. Estimated error rates were based on 100 000 simulations so that if the true type I error rate was 0.025, we would expect the estimated error rate to be below 0.026 with probability 0.975. Columns 3 and 4 show the type I error rates simulated under $\theta_1 = \cdots = \theta_k = 0$ with selection of $\arg\max\{\tilde{\theta}_i\}$ assuming $\theta_{k+1}, \ldots, \theta_{k^*}$ known, when the type I error rate is maximised, both using the known values of σ, σ_0 and ρ and using estimates of these obtained from the stage 1, respectively, confirming that the type I error rate is controlled at the nominal 0.025 level in both cases.

In practice, as $\theta_{k+1}, \ldots, \theta_{k^*}$ are unknown, selection based on $\tilde{\theta}_i$ is impossible. An alternative would be to select the treatment with the largest estimated effect $\hat{\theta}_i$ as described in the earlier example, with $\hat{\theta}_{k+i} = \sum_{j=1}^{N_1}(w_{ij} - w_{0j})/N_1$ as in Stallard [21]. Column 5 shows the type I error rate in this case, when the test is conservative.

The last column in Table II shows the power to select treatment 1 and reject H_1 when $\theta_1 = 1/3, \theta_2 = \theta_3 = 0$ again selecting $\arg\max_{i=1}^{k}\{\hat{\theta}_i\}$ with $\theta_{k+1}, \ldots, \theta_{k^*}$ unknown. The power shows how the gain from using the short-term endpoint increases with ρ relative to using the long-term endpoint alone, equivalent to $\rho = 0$.

For comparison with Table II, Table III gives critical values, $c/\sqrt{V_2}$, from the method proposed by Stallard [21], assuming the selection of $\arg\max_{i=1}^{k}\{\hat{\theta}_i\}$, and Stallard and Todd [8], assuming the selection of $\arg\max_{i=1}^{k}\{\breve{\theta}_i\}$, and the resulting simulated type I error rate under $\theta_1 = \cdots = \theta_k = 0$ with the treatment selected so as to maximise the error rate, that is, selecting $\arg\max\{\tilde{\theta}_i\}$, if $\theta_{k+1}, \ldots, \theta_{k^*}$ were known, in this case using the known values of σ, σ_0 and ρ. The error rate inflation from use of the Stallard and Todd [8] method when selecting $\arg\max_{i=1}^{k}\{\hat{\theta}_i\}$ was demonstrated by Stallard [21]. The new results show that the error rate may be inflated further by using some other selection rule and illustrate the maximal error rate inflation from the use of any selection rule based on X_1 for these two designs.

Table II. Properties of the flexible design with $\alpha = 0.025, k = 3, n_1 = 40, N_1 = 100, n_2 = 200$ and $\sigma = \sigma_0 = 1$ for a range of ρ values (error rates are based on 100 000 simulations).

ρ	$c/\sqrt{V_2}$	Type I error selecting $\arg\max\{\tilde{\theta}_i\}$	Type I error selecting $\arg\max\{\tilde{\theta}_i\}$	Type I error selecting $\arg\max\{\hat{\theta}_i\}$	Power selecting $\arg\max\{\hat{\theta}_i\}$
		Variances known	Variances unknown	Variances known	Variances unknown
0.0	2.19	0.0242	0.0246	0.0243	0.7827
0.5	2.22	0.0241	0.0244	0.0234	0.7974
0.6	2.23	0.0245	0.0248	0.0233	0.8071
0.7	2.24	0.0249	0.0252	0.0239	0.8119
0.8	2.25	0.0253	0.0255	0.0240	0.8251
0.9	2.27	0.0245	0.0248	0.0236	0.8358

Table III. Properties of the Stallard [21] and Stallard and Todd [8] designs and the combination test with parameters as in Table I (error rates are based on 100 000 simulations).

ρ	Stallard test		Stallard and Todd test		Combination test	
	$c/\sqrt{V_2}$	Type I error selecting $\arg\max\{\tilde{\theta}_i\}$	$c/\sqrt{V_2}$	Type I error selecting $\arg\max\{\tilde{\theta}_i\}$	Type I error selecting $\arg\max\{\tilde{\theta}_i\}$	Power selecting $\arg\max\{\hat{\theta}_i\}$
0.0	2.19	0.0242	2.19	0.0242	0.0197	0.7617
0.5	2.20	0.0252	2.19	0.0259	0.0200	0.7807
0.6	2.21	0.0256	2.19	0.0270	0.0200	0.7912
0.7	2.22	0.0262	2.19	0.0284	0.0216	0.8010
0.8	2.23	0.0266	2.19	0.0296	0.0218	0.8134
0.9	2.25	0.0252	2.19	0.0298	0.0231	0.8265

Additional simulations were conducted with $n_1 = 20, N_1 = 50$ and $n_2 = 100$ and $n_1 = 10, N_1 = 25$ and $n_2 = 50$ to explore the impact of smaller sample sizes when σ, σ_0 and ρ were considered unknown. There was some indication of type I error rate inflation in these cases with maximum simulated error rates for the new method of 0.02601 and 0.02645, respectively. This suggests that the new method may not be suitable in trials of very rare diseases or orphan drugs when sample sizes as small as this may be used. In other settings, it is unlikely that such small sample sizes would be used for a confirmatory clinical trial.

An alternative method to that proposed here would be to use a combination testing method [3, 9, 30], as this is known to strongly control the type I error rate for any treatment selection. To apply such an approach in this case requires some care to ensure that the p-values that are combined satisfy the 'p-clud' condition [31]. Friede *et al.* [32] describe one way in which this can be performed, considering the 'stage 1' p-value to be that obtained from the analysis of the primary endpoint of all those N_1 patients per group for whom some data were available at the interim analysis, and the 'stage 2' p-value to be from the analysis of the primary endpoint for the $n_2 - N_1$ patients per group recruited following the treatment selection. The power for this method to select treatment 1 and reject H_1 when $\theta_1 = 1/3, \theta_2 = \theta_3 = 0$ again selecting $\arg\max_{i=1}^{k}\{\hat{\theta}_i\}$ and assuming σ, σ_0 and ρ are known is also shown in Table III for comparison with that of the new procedure shown in Table II. It can be seen that the combination test has slightly lower power in this case. This is consistent with the findings of Friede and Stallard [33]. Although the power gain for the new procedure over the combination test is modest, in settings in which it is known that a single experimental treatment will be selected to continue along with the control to the second stage, the new method is to be preferred. If additional flexibility, for example, over the number of treatments to continue, is required, the combination test could be used with only a small loss in power.

7. Discussion

This paper has considered trials in which treatment selection at an interim analysis may be made using data for patients for whom the primary endpoint has not yet been observed. Stallard [21] showed how the error rate can be controlled in this case if the most promising treatment is selected. We have shown that this does not provide error rate control for an arbitrary selection rule and illustrated how this can be achieved.

It is interesting to consider possible extensions of the method proposed. One extension is to allow the possibility of early stopping at the interim analysis with additional testing of H_i $(i = 1, \ldots, k)$ on the basis of X_1. In this case, H_i could be tested using some test statistic $S_i^{(1)}(X_1)$, rejecting H_i if and only if $S_i^{(1)}(X_1)$ is at least as large as some specified value c_1. If any H_i is rejected, that hypothesis will not be tested again, so that treatment i could be dropped from the trial, and if not all hypotheses are rejected, $H_{T(x_1)}$ is selected and tested at the end of the second stage, being rejected if and only if $S_{T(x_1)}(X_2) \geqslant c$. In this case, as there is an opportunity to reject hypotheses at both stages, requirement (1) is insufficient to give both c_1 and c uniquely. A stronger requirement is that the probability to reject any true H_i $(i = 1, \ldots, k)$ at or before analysis j is controlled to be at most some specified α_j $(j = 1, 2)$ with $\alpha_1 \leqslant \alpha_2 = \alpha$. The values α_1 and α_2 may be considered as a simple α-spending function as proposed by Slud and Wei [34].

A further extension would be to allow additional interim analyses, so that the two-stage design becomes a multi-stage design. With a single hypothesis selected at the first interim analysis, as no further selection is possible, the extension to allow additional stages with the opportunity for early stopping, for example,

using an α-spending function approach, is relatively straightforward. The extension of the method proposed to select more than one treatment to continue, and hence more than one hypothesis to be tested at the end of the trial, or in a multi-stage trial to allow selection of treatments over several stages, while maintaining strong control of the familywise error rate, for example, to add flexibility to the methods of Follman *et al.* [17], Hellmich [18] or Magirr *et al.* [22], is more difficult. If c is fixed as above, with full flexibility over the choice of hypotheses from the family H_i $(i = 1, \ldots, k)$, the type I error rate is maximised by selecting all k hypotheses, so that c must be chosen to control the type I error in this case. This can be achieved using a test similar to the Dunnett test. An alternative is to have c dependent on the number of hypotheses selected, say $k_1 \leqslant k$. One such approach, using the conditional error method, has been proposed by Magirr *et al.* [35]. If k_1 is specified in advance, the method described can be extended for $k_1 > 1$. In this case, a result analogous to Theorem 1 in the Appendix, stating that the type I error rate is maximised by selecting the k_1 treatments corresponding to values of i with the largest values of $\mathrm{pr}_\theta(S_i \geqslant c \mid X_1)$, can be shown to hold.

In this paper, we have assumed that the short-term endpoint information available is based on the same length of follow-up for all patients. An anonymous referee has suggested that short-term information based on different follow-up duration for different patients might be available. This would also be an interesting area for future research.

In this paper, we have considered flexible treatment selection. Other adaptations could, in principle, be handled in a similar way. Graf and Bauer [36] and Graf *et al.* [24], for example, considered sample size reestimation based on interim data in two-stage studies both with and without treatment selection.

The method introduced in this paper enables strong control of the familywise error rate when the treatment selection rule is unspecified by constructing the selection rule that maximises the conditional error rate and then obtaining a critical value such that strong error rate control is achieved if this rule is used.

When treatment selection is made on the basis of the primary endpoint data available at the interim analysis, that is, in the setting considered by Thall *et al.* [16] and Stallard and Todd [8], this maximum occurs for selection of the most promising experimental treatment as noted by Jennison and Turnbull [23] and Graf *et al.* [24], so that the method of Thall *et al.* [16] and Stallard and Todd [8] strongly controls the familywise error rate if any other treatment is selected.

In addition, if the data available at the interim analysis, X_1, include additional data available from the same patients from whom primary endpoint data are available, given the final endpoint data, S_i is conditionally independent of these additional data. The distribution of $S_i \mid X_1$ thus depends on θ_i alone, and the probability in (2) does not depend on $\theta_{k+1}, \ldots, \theta_{k^*}$, so that strong control of the familywise error rate is again obtained. Similarly, the distribution of S_i given X_1 can also reasonably be assumed to be independent of any data obtained from sources external to the trial, so that this can also be used for decision-making without inflating the type I error rate. The latter point allows, for example, treatment selection using Bayesian methods where prior distributions may be informed, either formally or informally, by results from other trials of the same or similar treatments.

Although we have considered group sequential approaches, with c fixed, as indicated earlier, alternative methods with c depending on X_1 can also be used to allow flexibility with strong control of the familywise error rate. Both the combination testing approach and the conditional error approach have been proposed for use in the treatment selection setting. Combination tests generally allow greater flexibility than the group sequential approach, in this case allowing fully flexible selection of any number of hypotheses for testing at the second stage, with some modification [32] even in the case of correlated data. As demonstrated in the earlier example, however, this can be at the loss of efficiency. Koenig *et al.* [37], Posch *et al.* [38] and Magirr *et al.* [35] have proposed methods based on the conditional error approach of Müller and Schäfer [7]. Friede *et al.* [33] showed that the method of Koenig *et al.* [37] performed well in terms of power. Extending such methods to more general settings, such as multi-stage designs or correlated data at different stages, could be difficult, however.

Appendix

In a study in two stages, let X_1 denote the first stage data with the distribution of X_1 depending on $\theta = (\theta_1, \ldots, \theta_{k^*})'$. We wish to test a family of null hypotheses $H_i, i \in \{1, \ldots, k\}$, for some $k \leqslant k^*$ with $H_i : \theta_i \leqslant 0$, controlling the familywise error rate in the strong sense.

The study will be conducted so as to test a hypothesis, $H_{T(x_1)}$ $(T(x_1) \in \{1, \ldots, k\})$, chosen on the basis of X_1. Further data will then be observed, leading to a final dataset X_2, and $H_{T(x_1)}$ will be rejected if and

only if $S_{T(x_1)}(X_2) \geqslant c$ for some specified c and test statistic $S_{T(x_1)}(X_2)$. Writing S_i for $S_i(X_2)$, we assume that the distribution of S_i depends on θ_i, but on no other elements of θ and that for any possible stage 1 dataset, x_1, S_i and $S_i \mid X_1 = x_1$ is stochastically non-decreasing in θ_i.

Let $T^*_\theta(X_1) = \arg\max_{i\in\{1,\ldots,k\}:\theta_i\leqslant 0}\left\{\mathrm{pr}_\theta(S_i \geqslant c \mid X_1)\right\}$.

We prove the following result.

Theorem 1

(i) For given θ, the left-hand side of (1) is maximised over $T \in \mathcal{T}$ by taking $T(x_1) = T^*_\theta(x_1)$ for all x_1.

(ii) The left-hand side of (1) is maximised over $\theta_1, \ldots, \theta_k$ by taking $\theta_i = 0$ $(i = 1, \ldots, k)$.

The proof of Theorem 1 is based on the following two lemmas.

Lemma 1

Let $\mathcal{T}(t)$ be the set of functions from the sample space of X_1 to $\{1, \ldots, t\}$, for $t \leqslant k$, and then for fixed θ, the left-hand side of (1) is maximised over $T \in \mathcal{T}(t)$ by $T = T^{*(t)}_\theta$, where

$$T^{*(t)}_\theta(x_1) = \arg\max_{i\in\{1,\ldots,t\}:\theta_i\leqslant 0}\left\{\mathrm{pr}_\theta(S_i \geqslant c \mid X_1 = x_1)\right\}.$$

Proof

The probability $\mathrm{pr}_\theta(S_{T(X_1)} \geqslant c, \theta_{T(X_1)} \leqslant 0)$ is equal to $E_\theta(\mathrm{pr}_\theta(S_{T(X_1)} \geqslant c, \theta_{T(X_1)} \leqslant 0 \mid X_1))$.

This is maximised by taking $T \in \mathcal{T}(t)$ to maximise $\mathrm{pr}_\theta(S_{T(X_1)} \geqslant c, \theta_{T(X_1)} \leqslant 0 \mid X_1)$, that is by taking $T = T^{*(t)}_\theta$ as stated. □

Lemma 2

For fixed $\theta_{k+1}, \ldots, \theta_{k^*}$, the left-hand side of (1) is non-decreasing in $\theta_i, i = 1, \ldots, k$.

Proof

As T^*_θ takes values in $\{i : \theta_i \leqslant 0\}$, we have $\mathrm{pr}_\theta\left(S_{T^*_\theta(X_1)} \geqslant c, \theta_{T^*_\theta(X_1)} \leqslant 0\right) = \mathrm{pr}_\theta\left(S_{T^*_\theta(X_1)} \geqslant c\right)$. This is equal to $E_\theta\left(\mathrm{pr}_\theta\left(S_{T^*_\theta(X_1)} \geqslant c \mid X_1\right)\right)$, which, by Lemma 1, is equal to $E_\theta\left(\max_{i:\theta_i\leqslant 0}\left\{\mathrm{pr}_\theta(S_i \geqslant c \mid X_1)\right\}\right)$.

It has been assumed that for any x_1 and i, $S_i \mid X_1 = x_1$ is stochastically non-decreasing in θ_i. Therefore, $\max_{i:\theta_i\leqslant 0}\left\{\mathrm{pr}_\theta\left(S_i \geqslant c \mid X_1 = x_1\right)\right\}$ is non-decreasing in θ_i $(i = 1, \ldots, k)$. As this holds for all x_1, the expected value must also be non-decreasing in θ_i as required. □

Proof of Theorem 1

Setting $t = k$, the first part of Theorem 1 follows from Lemma 1.

For the second part, writing $\theta^{(t)} = \left(0, \ldots, 0, \theta_{t+1}, \ldots, \theta_{k^*}\right)'$, where the first t elements are equal to 0, we require

$$\mathrm{pr}_\theta\left(S_{T^*_\theta(X_1)} \geqslant c, \theta_{T^*_\theta(X_1)} \leqslant 0\right) \leqslant \mathrm{pr}_{\theta^{(k)}}\left(S_{T^*_{\theta^{(k)}}(X_1)} \geqslant c, \theta_{T^*_{\theta^{(k)}}(X_1)} \leqslant 0\right). \tag{1}$$

If $\theta_i > 0 (i = 1, \ldots, k)$, the left-hand side of (1) is equal to zero, so that the inequality is trivially satisfied. Suppose, therefore, that $\theta_i < 0$ for some $i \leqslant k$ and, without loss of generality, reorder such that $\theta_i \leqslant 0$ for $i = 1, \ldots, t$ and $\theta_i > 0$ for $i = t + 1, \ldots, k$ for some t.

As, by Lemma 2, $\mathrm{pr}_\theta(S_{T^*_\theta(X_1)} \geqslant c, \theta_{T^*_\theta(X_1)} \leqslant 0)$ is non-decreasing in θ_i, the probability on the left-hand side of (1) must be non-decreased by taking $\theta_1 = 0 (i = 1, \ldots, t)$. Thus,

$$\mathrm{pr}_\theta\left(S_{T^*_\theta(X_1)} \geqslant c, \theta_{T^*_\theta(X_1)} \leqslant 0\right) \leqslant \mathrm{pr}_{\theta^{(t)}}\left(S_{T^*_{\theta^{(t)}}(X_1)} \geqslant c, \theta_{T^*_{\theta^{(t)}}(X_1)} \leqslant 0\right). \tag{2}$$

As $T^*_{\theta^{(t)}}$ maximises the left-hand side of (1) under $\theta^{(t)}$, we must have $T^*_{\theta^{(t)}}(X_1) \in \{1, \ldots, t\}$ because if, for any x_1, it took a value outside this range, the probability would be increased by moving mass into this set. Thus, $T^*_{\theta^{(t)}} = T^{*(t)}_{\theta^{(t)}}$ and the right-hand side of (2) is

$$\mathrm{pr}_{\theta^{(t)}}\left(S_{T^{*(t)}_{\theta^{(t)}}(X_1)} \geqslant c, \theta_{T^{*(t)}_{\theta^{(t)}}(X_1)} \leqslant 0\right).$$

The function $T^{*(t)}_{\theta^{(t)}}$, and hence $T^*_{\theta^{(t)}}$, as defined by Lemma 1, depends only on $S_1, \ldots, S_t$, and by Assumption 1 therefore not on $\theta_{t+1}, \ldots, \theta_k$. Thus, the right-hand side of (2) is unchanged by setting $\theta_i = 0 (i = t + 1, \ldots, k)$ and is, as required, equal to

$$\mathrm{pr}_{\theta^{(k)}}\left(S_{T^*_{\theta^{(k)}}(X_1)} \geqslant c, \theta_{T^*_{\theta^{(k)}}(X_1)} \leqslant 0\right).$$

□

Acknowledgements

The authors are grateful to the UK Medical Research Council for financial support (grant number G1001344) and to the anonymous referees of an earlier draft for their helpful comments.

References

1. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall: Boca Raton, FL, 2000.
2. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663.
3. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994; **50**(4):1029–1041.
4. Cui L, Hung HMJ, Wang S-J. Modification of sample size in group sequential clinical trials. *Biometrics* 1999; **55**:853–857.
5. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics* 1999; **55**(4): 1286–1290.
6. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* 1995; **51**:1315–1324.
7. Müller H-H, Schäfer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* 2001; **57**:686–891.
8. Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine* 2003; **22**:689–703.
9. Bretz F, Schmidli H, König F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. *Biometrical Journal* 2006; **48**:623–634.
10. Brannath W, Zuber E, Branson M, Bretz F, Gallo P, Posch M, Racine-Poon A. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine* 2009; **28**:1445–1463.
11. Friede T, Parsons N, Stallard N. A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in Medicine* 2012; **31**:4309–4320.
12. Tsiatis AA, Mehta CR. On the inefficiency of adaptive design for monitoring clinical trials. *Biometrika* 2003; **90**:367–378.
13. Jennison C, Turnbull BW. Adaptive and nonadaptive group sequential tests. *Biometrika* 2006; **93**:1–21.
14. Burman C-F, Sonesson C. Are flexible designs sound? *Biometrics* 2006; **62**:664–683.
15. Food and Drug Administration (FDA). *Guidance for industry—adaptive design clinical trials for drugs and biologics*, 2010. Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm201790.pdf [accessed 13 Jul 2012].
16. Thall PF, Simon R, Ellenberg SS. Two-stage selection and testing designs for comparative clinical trials. *Biometrika* 1988; **75**:303–310.
17. Follman DA, Proschan MA, Geller NL. Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics* 1994; **50**:325–336.
18. Hellmich M. Monitoring clinical trials with multiple arms. *Biometrics* 2001; **57**:892–898.
19. Bischoff W, Miller F. Adaptive two-stage test procedures to find the best treatment in clinical trials. *Biometrika* 2005; **92**:197–212.
20. Stallard N, Friede T. A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine* 2008; **27**:6209–6227.
21. Stallard N. A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Statistics in Medicine* 2010; **29**:959–971.
22. Magirr D, Jaki T, Whitehead J. A generalised Dunnett test for multi-arm, multi-stage clinical studies with treatment selection. *Biometrika* 2012; **99**:494–501.
23. Jennison C, Turnbull BW. Confirmatory seamless phase II/III trials with hypothesis selection at interim: opportunities and limitations. *Biometrical Journal* 2006; **48**:650–655.
24. Graf AC, Bauer P, Glimm E, Koenig F. Maximum type 1 error rate inflation in multiarmed clinical trials with interim sample size modifications. *Biometrical Journal* 2014; **56**:614–630.
25. Cuffe RL, Lawrence D, Stone A, Vandemeulebroecke M. When is a seamless study desirable? Case studies from different pharmaceutical sponsors. *Pharmaceutical Statistics* 2014; **13**:229–237.
26. Wilkinson D, Murray J. Galantamine: a randomized, double-blind, dose comparison in patients with Alzheimer's disease. *International Journal of Geriatric Psychiatry* 2001; **16**:852–857.
27. Loy C, Schneider L. Galantamine for Alzheimer's disease and mild cognitive impairment. *Cochran Database of International Reviews* 2006; (1). Art. No.: CD001747.
28. Jennison C, Turnbull BW. Group-sequential analysis incorporating covariate information. *Journal of the American Statistical Association* 1997; **92**(440):1330–1341.
29. Todd S, Stallard N. A new clinical trial design combining phases II and III: sequential designs with treatment selection and a change of endpoint. *Drug Information Journal* 2005; **39**:109–118.
30. Posch M, König 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.

31. Brannath W, Posch M, Bauer P. Recursive combination tests. *Journal of the American Statistical Association* 2002; **97**: 236–244.
32. Friede T, Parsons N, Stallard N, Todd S, Valdés-Márquez E, Chataway J, Nicholas R. Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: an application in multiple sclerosis. *Statistics in Medicine* 2011; **30**:1528–1540.
33. Friede T, Stallard N. A comparison of methods for adaptive treatment selection. *Biometrical Journal* 2008; **50**:767–781.
34. Slud EV, Wei LJ. Two-sample repeated significance tests based on the modified Wilcoxon statistics. *Journal of the American Statistical Association* 1982; **77**:862–868.
35. Magirr D, Stallard N, Jaki T. Flexible sequential designs for multi-arm clinical trials. *Statistics in Medicine* 2014; **33**: 3269–3279.
36. Graf AC, Bauer P. Maximum inflation of the type 1 error rate when sample size and allocation rate are adapted in a pre-planned interim look. *Statistics in Medicine* 2011; **30**:1637–1647.
37. Koenig F, Brannath W, Bretz F, Posch M. Adaptive Dunnett tests for treatment selection. *Statistics in Medicine* 2008; **27**:1612–1625.
38. Posch M, Maurer W, Bretz F. Type I error rate control in adaptive designs for confirmatory clinical trials with treatment selection at interim. *Pharmaceutical Statistics* 2010; **10**:96–104.