A comparison of methods for treatment selection in seamless phase II/III clinical trials incorporating information on short-term endpoints

Journal of Biopharmaceutical Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lbps20

A Comparison of Methods for Treatment Selection in Seamless Phase II/III Clinical Trials Incorporating Information on Short-Term Endpoints

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

[a] Warwick Medical School, University of Warwick, Coventry, United Kingdom

[b] Department of Medical Statistics, University Medical Center, Göttingen, Germany

[c] Department of Mathematics and Statistics, University of Reading, Berkshire, United Kingdom

Accepted author version posted online: 03 Apr 2014.Published online: 20 Jan 2015.

To cite this article: Cornelia Ursula Kunz, Tim Friede, Nicholas Parsons, Susan Todd & Nigel Stallard (2015) A Comparison of Methods for Treatment Selection in Seamless Phase II/III Clinical Trials Incorporating Information on Short-Term Endpoints, Journal of Biopharmaceutical Statistics, 25:1, 170-189, DOI: 10.1080/10543406.2013.840646

To link to this article: http://dx.doi.org/10.1080/10543406.2013.840646

Journal of Biopharmaceutical Statistics, 25: 170–189, 2015
Published with license by Taylor & Francis Group, LLC
ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543406.2013.840646

A COMPARISON OF METHODS FOR TREATMENT SELECTION IN SEAMLESS PHASE II/III CLINICAL TRIALS INCORPORATING INFORMATION ON SHORT-TERM ENDPOINTS

Cornelia Ursula Kunz[1], Tim Friede[2], Nicholas Parsons[1], Susan Todd[3], and Nigel Stallard[1]

[1]*Warwick Medical School, University of Warwick, Coventry, United Kingdom*
[2]*Department of Medical Statistics, University Medical Center, Göttingen, Germany*
[3]*Department of Mathematics and Statistics, University of Reading, Berkshire, United Kingdom*

In an adaptive seamless phase II/III clinical trial interim analysis, data are used for treatment selection, enabling resources to be focused on comparison of more effective treatment(s) with a control. In this paper, we compare two methods recently proposed to enable use of short-term endpoint data for decision-making at the interim analysis. The comparison focuses on the power and the probability of correctly identifying the most promising treatment. We show that the choice of method depends on how well short-term data predict the best treatment, which may be measured by the correlation between treatment effects on short- and long-term endpoints.

Key Words: Adaptive seamless design; Multi-arm multi-stage trial; Surrogate endpoints.

1. INTRODUCTION

In recent years, adaptive designs in the various phases of drug development have gained popularity. Such designs use information from accumulating data in an ongoing trial to make decisions about the conduct of the rest of the study (Gallo et al., 2006). One particular form of adaptive design is the combined phase II/III adaptive seamless design. A trial of this type is conducted in two stages. During the first stage, the exploratory stage, patients are recruited to several experimental treatments and a control treatment. One or more interim analyses are then performed, at which treatments that appear ineffective are dropped. The main objective of this first stage is to identify the most promising treatments, so that recruitment of further patients can be restricted to only those treatments and the control. At the end of the second stage, the confirmatory stage, the selected treatment(s) is (are) compared to the control within a formal testing framework, again possibly involving

Received September 18, 2012; Accepted August 17, 2013

Address correspondence to Nigel Stallard, Warwick Medical School, University of Warwick, Coventry CV4 7AL, UK; E-mail: n.stallard@warwick.ac.uk

a sequence of interim analyses, based on all data from the selected treatment(s) and the control. Several authors have developed methodology for conducting phase II/III studies that protects the overall type I error rate of the trial (see, e.g., Bauer and Kieser, 1999; Stallard and Todd, 2003; Kelly et al., 2005; Posch et al., 2005; Bretz et al., 2006; Koenig et al., 2008). Reviews of the different approaches are given by Chow et al. (2005), Friede and Stallard (2008), Bretz et al. (2009), and Stallard and Todd (2011).

In the pharmaceutical setting, adaptive designs continue to gain acceptance. Regulatory authorities have recently produced guidance documents on the topic (Food and Drug Administration (FDA), 2010; European Medicines Agency (EMEA) - Committee for Medicinal Products for Human Use (CHMP), 2007), giving further evidence that they anticipate more clinical trials will be designed using this framework. Indeed, there are a number of therapeutic areas where phase II/III seamless adaptive designs have already been implemented. Schmoll et al. (2010) describe a pharmaceutical trial in oncology that was designed using the methodology of Stallard and Todd (2003) and Todd and Stallard (2005). Barnes et al. (2010) discuss the use of a phase II/III design in chronic obstructive pulmonary disease. In other therapeutic areas adaptive designs have been proposed and promoted. Dragalin (2011) discusses the potential for the use of adaptive designs in all phases of development, including discussion of phase II/III trials, in central nervous system studies. Chataway et al. (2011) and Friede et al. (2011) propose a phase II/III seamless adaptive design for use in secondary progressive multiple sclerosis trials.

A recent area of research in the development of further methodology for phase II/III designs concerns the question of how to incorporate early endpoint data into the treatment selection part of such a trial. The desire to do this arises when the primary endpoint of interest for each patient is only available after a number of months or even years and yet there are more immediately measured endpoints available, building on earlier work on incorporation of early endpoints in sequential clinical trials comparing a single experimental treatment with a control (Cook and Farewell, 1996; Marschner and Becker, 2001; Galbraith and Marschner, 2003; Sooriyarachchi et al., 2006; Whitehead et al., 2008). An example can be found in secondary progressive multiple sclerosis, where long-term changes in disability scales are the main goal, but early evidence of treatment effect may be observed as changes to lesions in the brain detected using magnetic resonance imaging scanning technology.

Two alternative methods for incorporating early endpoint data in phase II/III clinical trials have been proposed by Stallard (2010) and Friede et al. (2011). The methods differ in the way in which the treatment to continue to the second stage is chosen. Treatment selection under the method described by Stallard (2010) makes use of short-term endpoint data combined with any available long-term data. In contrast, Friede et al. (2011) propose a method of treatment selection that uses only short-term endpoint data. Both approaches base the final inference on the long-term endpoint data only, though they differ in the way in which data from the two stages of the trial are combined. The aim of this paper is to compare the methods proposed in these two manuscripts. Since both methods have been shown to control the type I error rate, we will focus on comparison of the power of the two approaches in a range of realistic scenarios. This will inform researchers aiming to design a seamless phase II/III trial in which short-term endpoint data can be used for decision-making at an interim analysis.

The two methods under consideration are reviewed in detail in Section 2, where a common notation is also established. Sections 3 and 4 describe comparisons of the two approaches in the settings of fixed and random treatment effect models, respectively. The paper concludes with a discussion in Section 5.

2. NOTATION AND REVIEW OF METHODS

2.1. Setting and Notation

Consider a clinical trial conducted in two stages. In the first stage, patients are randomized to the control treatment T_0 or to one of k experimental treatments, T_i, $i = 1, \ldots, k$. Suppose that data on the primary, long-term, endpoint are available for n_1 patients in each treatment group, and that in addition, short-term endpoint data are observed for N_1 patients in each treatment group, with $N_1 \geq n_1$. In stage one, we therefore have $N_1 - n_1$ patients with short-term endpoint data only and n_1 patients with both short- and long-term endpoint data in each treatment group. Following an interim analysis, one experimental treatment, denoted by T_I, is chosen to continue to the second stage along with the control treatment with a further $n_2 - N_1$ patients recruited to each of these treatment groups, giving a total of n_2 patients per group in all. Two possible ways of making the treatment selection are described below.

Suppose that following the second stage, patients are followed up so that primary long-term endpoint data are available for the total of n_2 patients receiving each of treatments T_I and T_0.

Denote by $X_{i,j}$ and $Y_{i,j}$, respectively, the short-term and long-term endpoint data from patient j in group i. When both endpoints are observed, that is for $j = 1, \ldots, n_2$ for $i = 0, I$ and $j = 1, \ldots, n_1$ for other $i = 1, \ldots, k$, the two endpoints for each patient are assumed to follow a bivariate normal distribution. When only the short-term endpoint is observed, that is for $j = n_1 + 1, \ldots, N_1, i = 1, \ldots, k, i \neq I, X_{i,j}$ is assumed to follow a normal distribution so that we have

$$\begin{aligned}
&\begin{pmatrix} X_{ij} \\ Y_{ij} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{b_i} \\ \mu_{B_i} \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \rho_w\sigma_0\sigma \\ \rho_w\sigma_0\sigma & \sigma^2 \end{pmatrix}\right), \quad j = 1, \ldots, n_2, i = 0, I \\
&\begin{pmatrix} X_{ij} \\ Y_{ij} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{b_i} \\ \mu_{B_i} \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \rho_w\sigma_0\sigma \\ \rho_w\sigma_0\sigma & \sigma^2 \end{pmatrix}\right), \quad j = 1, \ldots, n_1, i = 1, \ldots, k, i \neq I \qquad (1) \\
&X_{ij} \sim N(\mu_{b_i}, \sigma_0^2), \qquad\qquad j = n_1 + 1, \ldots, N_1, i = 1, \ldots, k, i \neq I,
\end{aligned}$$

where μ_{b_i} and μ_{B_i} denote the true means on the short- and long-term endpoints, respectively, in group i; σ_0^2 and σ^2 denote the true variances for the short- and long-term endpoints, respectively; and ρ_w denotes the true correlation between the endpoints within each group.

The variances σ_0^2 and σ^2 and the correlation ρ_w will be assumed known and equal for all patients. In the calculation of selection probabilities below, the true variances and correlation will be used. In the simulations, estimates obtained from the data will be used in place of the true values, as suggested by Stallard (2010) and Friede et al. (2011).

Given the mean values, individual patients are assumed to be independent so that $cov\left(X_{i,j}, X_{i'j'}\right) = 0$, $cov\left(Y_{i,j}, Y_{i'j'}\right) = 0$, and $cov\left(X_{i,j}, Y_{i'j'}\right) = 0$ for $i \neq i'$ or $j \neq j'$.

A summary of the parameters in the fixed and random effects models are given in Table 1. The parameters of interest are the treatment effects relative to the control treatment on the long-term endpoint, that is $\mu_{B_1} - \mu_{B_0}, \ldots, \mu_{B_k} - \mu_{B_0}$, and we wish to test the null hypotheses denoted H_i: $\mu_{B_i} - \mu_{B_0} = 0$ against the one-sided alternative hypotheses denoted by H_i': $\mu_{B_i} - \mu_{B_0} > 0$ for treatment group $i = 1, \ldots, k$.

Two methods for use of short-term endpoint data for treatment selection in a two-stage trial have been proposed (Friede et al., 2011; Stallard, 2010). These methods are described briefly below.

Table 1 Summary of model parameters

Sample sizes	
N_1	Total number of patients per group with short-term data at interim analysis
n_1	Number of patients per group with short-term and long-term data at interim analysis
$N_1 - n_1$	Number of patients per group with short-term data only at interim analysis
n_2	Number of patients per group with short-term and long-term data at final analysis
Fixed or random effects model parameters	
μ_{B_i}	Long-term endpoint treatment mean for group i
σ^2	Long-term endpoint variance
μ_{b_i}	Short-term endpoint treatment mean for group i
σ_0^2	Short-term endpoint variance
ρ_w	Correlation between long-term and short-term endpoints within each treatment group
Random effects model parameters	
θ_{B_i}	Mean long-term treatment mean for group i
σ_B^2	Variance of long-term treatment mean
θ_{b_i}	Mean short-term treatment mean for group i
σ_b^2	Variance of short-term treatment mean
ρ_b	Correlation between long-term and short-term treatment means

The aim of this paper is to compare these methods. This comparison will be based on model (1). We will consider two cases. In the first case, the fixed effects model, it is assumed that the true means μ_{b_i} and μ_{B_i} can be specified so that these can be taken to be constant. And in the second, the random effects model, the means will be taken to be random and to follow a bivariate normal distribution with

$$\begin{pmatrix} \mu_{b_i} \\ \mu_{B_i} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_{b_i} \\ \theta_{B_i} \end{pmatrix}, \begin{pmatrix} \sigma_{b_i}^2 & \rho_{b_i}\sigma_{b_i}\sigma_{B_i} \\ \rho_{b_i}\sigma_{b_i}\sigma_{B_i} & \sigma_{B_i}^2 \end{pmatrix} \right), \tag{2}$$

where θ_{b_i} and θ_{B_i} denote the true means, $\sigma_{b_i}^2$ and $\sigma_{B_i}^2$ denote the true variances, and ρ_{b_i} denotes the true correlation between the means for the two endpoints for any given treatment. We assume that the random treatment means have the same variances and correlations, so we may drop the subscript and denote these by σ_b^2, σ_B^2, and ρ_b, and are independent for different treatments, that is $cov\left(\mu_{b_i}, \mu_{b_j}\right) = 0$, $cov\left(\mu_{B_i}, \mu_{B_j}\right) = 0$, and $cov\left(\mu_{b_i}, \mu_{B_j}\right) = 0$ for $i \neq j$. The random effects model will allow us to model a situation in which we envisage that the treatments being evaluated are drawn at random from the distribution given by (2). In this case, the treatment means are considered to be unknown but correlated for the two endpoints with specified correlation and variance.

2.2. Method of Friede et al. (2011)

Friede et al. (2011) propose a method for selection of the treatment that will continue to the next stage based on the short-term endpoint only, selecting the experimental treatment with the largest observed sample mean at the interim analysis.

Let

$$Z_i^* = \frac{\sum_{j=1}^{N_1}\left(X_{i,j} - X_{0,j}\right)}{\sigma_0\sqrt{2N_1}} \tag{3}$$

denote the standardized test statistic for comparison of treatment i, $i = 1, \ldots, k$ to the control in terms of the short-term endpoint only on the basis of data available at the interim analysis. The experimental treatment group with the highest value of Z_i^* is then chosen to continue to the second stage along with the control while all other treatments are dropped.

At the end of the trial, long-term endpoint data are available from all n_2 patients randomized to the selected treatment and the control. Thus, as the parameters of interest are the long-term endpoint means μ_{B_i}, only the long-term endpoint data will be used in the final analysis. In this method, then, the short-term data are thus only used for treatment selection, and the long-term data are used only for the final comparison of the selected treatment with the control. In order to control the type I error rate, the final analysis must allow for the treatment selection. Friede et al propose using a combination test approach to combine all data from those patients with any data observed at the interim analysis with the data from new patients observed at the end of the second stage, with a Dunnett correction applied to the first stage test statistics.

In detail, let

$$Z_{i,1} = \frac{\sum_{j=1}^{N_1} \left(Y_{i,j} - Y_{0,j}\right)}{\sigma\sqrt{2N_1}}$$

denote the standardized test statistic for comparison of group i to the control group based on the long-term endpoint data from the N_1 patients per group who have short-term endpoint data available at the interim analysis. Let $p_{i,1} = 1 - \Phi(Z_{i,1})$ denote a p-value based on $Z_{i,1}$.

Similarly, let

$$Z_{I,2} = \frac{\sum_{j=N_1+1}^{n_2} \left(Y_{I,j} - Y_{0,j}\right)}{\sigma\sqrt{2(n_2 - N_1)}}$$

denote the standardized test statistic for comparison of group I and the control group based on the additional long-term endpoint data observed at the end of the trial and let $p_{I,2} = 1 - \Phi(Z_{I,2})$.

Note that the $p_{1,1}, \ldots, p_{k,1}$ are based on some data not observed at the time of the interim analysis and that $p_{I,2}$ is independent of all $p_{i,1}$ and of any data available at the interim analysis.

To allow for the treatment selection at the first stage, in order to test a null hypothesis $H_{\mathcal{I}}$, where $\mathcal{I}$ is some nonempty subset of $\{1, \ldots, k\}$ and $H_{\mathcal{I}}$ denotes the intersection hypothesis $\cap_{i\in\mathcal{I}} H_i$, the stage one p-value is obtained from a Dunnett test (Dunnett, 1955) using the test statistic $Z_1^{\max} = \max_{i\in\mathcal{I}} Z_{i,1}$ in, for instance, equation (1) of Friede and Stallard (2008). This gives a stage one p-value for the test of $H_{\mathcal{I}}$, $p_{\mathcal{I},1}$ corrected for the multiple comparisons. If the selected treatment, I, is in $\mathcal{I}$, a stage two p-value for testing $H_{\mathcal{I}}$, $p_{\mathcal{I},2}$ is just that for testing the selected treatment, $p_{I,2}$. If $I \notin \mathcal{I}$, $p_{\mathcal{I},2}$ is set to one to give a conservative test (Posch et al., 2005). The stage one and stage two p-values may then be combined, for example, using the *weighted inverse normal* combination function (Lehmacher and Wassmer, 1999)

$$C\left(p_{\mathcal{I},1}, p_{\mathcal{I},2}\right) = 1 - \Phi\left(w_1\Phi^{-1}(1 - p_{\mathcal{I},1}) + w_2\Phi^{-1}\left(1 - p_{\mathcal{I},2}\right)\right) \tag{4}$$

for predefined weights w_1 and w_2 with $w_1^2 = w_2^2 = 1$, which may be used to test $H_{\mathcal{I}}$.

The construction of the p-values ensures that the stage two p-values are independent of any data available at the interim analysis, and hence of the treatment selection. The p-values obtained thus satisfy the weaker *p-clud* condition (Brannath et al., 2002), so that no further correction for the treatment selection is necessary and the combination test provides a test of $H_{\mathcal{I}}$ that controls the type I error rate at the nominal level for any treatment selection method.

If the null hypothesis H_i is rejected and if all $H_{\mathcal{I}}$ with $i \in \mathcal{I}$ are rejected, the type I error rate for the family of hypotheses H_i, $i = 1, \ldots, k$, is controlled in the strong sense (Marcus et al., 1976).

2.3. Method of Stallard (2010)

Stallard (2010) proposes basing treatment selection on the maximum likelihood estimate of the long-term treatment effects, $\mu_{B_i} - \mu_{B_0}$, $i = 1, \ldots, k$ calculated at the interim analysis.

Let $S_{i,1}$ denote the standardized score statistic for $\mu_{B_i} - \mu_{B_0}$ obtained from all data available at the interim analysis. In the case that $N_1 > n_1$, this depends on the short-term data in addition to the long-term data. If ρ, σ, and σ_0 are unknown, $S_{i,1}$ may be estimated using the double regression method proposed by Engel and Walstra (1991) (see, Stallard, 2010), in which results of regression of X on group membership for $j = 1, \ldots, N_1$ and of Y on X and group membership for $j = 1, \ldots, n_1$ are combined to give $S_{i,1}$. For known ρ, σ, and σ_0, $S_{i,1}$ is shown by Hampson and Jennison (2013) to be given by

$$S_{i,1} = \frac{\sum_{j=1}^{n_1} \left(Y_{i,j} - Y_{0,j} - \rho_w \frac{\sigma}{\sigma_0} \left(X_{i,j} - X_{0,j} - \bar{X}_i + \bar{X}_0 \right) \right) / n_1}{\sigma \sqrt{2/N_1^*}} \tag{5}$$

where

$$N_1^* = \frac{n_1 N_1}{N_1 - \rho_w^2 (N_1 - n_1)} = \frac{n_1}{1 - \rho_w^2 (1 - n_1/N_1)} \tag{6}$$

and $\bar{X}_i$ denotes the sample mean of the N_1 short-term endpoint observations from group i observed at the interim analysis.

The quantity N_1^* given by expression (6) can be viewed as an effective sample size per group, corresponding to the number of long-term observations per group that would give the same amount of information on $\mu_{B_i} - \mu_{B_0}$ as that available from the n_1 long-term and N_1 short-term responses allowing for the correlation ρ_w. If $\rho_w = 0$ so that long-term and short-term responses for any given patient are independent, and the short-term observations give no information on $\mu_{B_i} - \mu_{B_0}$, $N_1^* = n_1$. If $\rho_w = \pm 1$, so that short- and long-term responses are perfectly correlated, $N_1^* = N_1$, so that the amount of information on $\mu_{B_i} - \mu_{B_0}$ is the same as if long-term data had been observed for all patients.

In the method described by Stallard (2010), treatment selection is based on statistics $S_{i,1}$ with the treatment group with the highest value for $S_{i,1}$ being selected to continue to the second stage together with the control group. Note that, unlike the Friede et al. method, this method requires that at least some long-term endpoint data are available at the time of the interim analysis.

At the end of the trial, long-term endpoint data are available from all n_2 patients randomized to the selected treatment and the control, so that as with the Friede et al. method,

only the long-term endpoint data will be used in the final analysis. However, the final analysis using the Stallard method combines the evidence from the two stages in a different way to that suggested by Friede et al.

Suppose that treatment T_I is selected to continue with the control to the second stage. Let $S_{I,2}$ denote the standardized score test statistic for $\mu_{B_1} - \mu_{B_0}$ based on all data available at the end of the trial, that is,

$$S_{I,2} = \frac{\sum_{j=1}^{n_2} \left(Y_{I,j} - Y_{0,j}\right)/n_2}{\sigma_0\sqrt{2/n_2}}.$$

Stallard derives the joint distribution of $\left(S_{1,1}, \ldots, S_{k,1}, S_{I,2}\right)$, showing that this is similar to that for test statistics in a seamless phase II/III trial with the primary endpoint alone used at an interim analysis with N_1^* patients per group. The joint distribution of $\max_{i=1,\ldots,k} S_{i,1}$ and $S_{I,2}$ can thus be obtained, allowing a critical value c to be constructed so as to control the type I error rate if H_0 is rejected whenever $S_{I,2} \geq c$.

3. COMPARISON OF METHODS: FIXED EFFECTS MODEL

We are interested in comparing the methods proposed by Stallard (2010) and Friede et al. (2011). We first consider the fixed effects model setting and explore the properties of the two methods for fixed treatment effects on the short- and long-term endpoints.

The methods will be compared in terms of the probability of selecting an effective treatment in Section 3.1 and of the resulting power of the final analysis in Section 3.2.

3.1. Selection Probability

Although we wish to focus on the probability of selecting the correct treatment, we can define this in two different ways. For given treatment means for the long-term endpoints, $\mu_{B_0}, \ldots, \mu_{B_k}$, we could consider either the probability of selecting any effective treatment, that is choosing I to be any i with $\mu_{B_i} - \mu_{B_0} > 0$, or the probability of selecting the most effective treatment, that is choosing I to be the i that maximizes $\mu_{B_i} - \mu_{B_0}$. Throughout this paper, we will focus on the latter. Furthermore, we will, without loss of generality, generally consider scenarios in which T_1 has the best effect, and report the probability of selecting treatment T_1.

The probability of selecting treatment T_1 with the Friede et al. method based on equation (3) is equal to

$$Pr\left(Z_1^* > Z_2^*, \ldots, Z_1^* > Z_k^*\right) = Pr\left(Z_1^* - Z_2^* > 0, \ldots, Z_1^* - Z_k^* > 0\right), \tag{7}$$

while the probability of selecting treatment T_1 with the Stallard selection method based on equation (5) is equal to

$$Pr\left(S_{1,1} > S_{2,1}, \ldots, S_{1,1} > S_{k,1}\right) = Pr\left(S_{1,1} - S_{2,1} > 0, \ldots, S_{1,1} - S_{k,1} > 0\right). \tag{8}$$

These probabilities could be estimated via simulation. Alternatively, for σ and σ_0 assumed known, they can be calculated exactly from the joint distributions of $Z_1^* - Z_2^*, \ldots, Z_1^* - Z_k^*$ and $S_1 - S_2, \ldots, S_1 - S_k$, respectively. These distributions are given in

the Appendix. Selection probabilities can thus be found using standard numerical routines for calculation of multivariate normal tail areas, for example using pmvnorm in R (Genz et al. 2012). Computer code to perform these calculations and the simulations described below in R can be obtained from the corresponding author.

For the Friede et al. method, the selection probability depends only on N_1 and the standardized short-term endpoint treatment effects, $\mu_{b_1}/\sigma_0, \ldots, \mu_{b_k}/\sigma_0$, while for the Stallard method, the selection probability depends on n_1, N_1, and the correlation between the endpoints, ρ_w, through N_1^*, the standardized long-term endpoint treatment effects, $\mu_{B_1}/\sigma, \ldots, \mu_{B_k}/\sigma$, but not on $\mu_{b_1}, \ldots, \mu_{b_k}$.

The upper panels (panels A1 and B1) of Fig. 1 show the probability of selecting treatment T_1 using the Friede et al and Stallard selection methods when three experimental treatments are included in the first stage and $\mu_{b_i} = \mu_{B_i} = 0$, $i > 1$. Panel A1 gives the selection probability with $\mu_{b_1}/\sigma_0 = \mu_{B_1}/\sigma = 0.5$ for different stage one sample sizes for a a range of ρ_w values. Panel B1 gives the selection probability with $n_1 = 5$, $N_1 = 100$, and $\sigma = \sigma_0 = 1$ for a range of $\mu_{b_1} = \mu_{B_1}$ values (with $\sigma = \sigma_0 = 1$ so that these are the standardized values), again for a range of ρ_w values.

As indicated above, the probability of selection with the Friede et al. method does not depend on ρ_w. With $\mu_{b_1}/\sigma_0 = \mu_{B_1}/\sigma$, the probability of selection with the Stallard method is equal to that with the Friede et al. method when $\rho_w = \pm 1$, when the most information is obtained from the $N_1 - n_1$ observations per group for whom only short-term endpoint data are available and $N_1^* = N_1$. For $\rho_w \neq \pm 1$, the probability of selecting treatment T_1 is lower for the Stallard method than that for the Friede method when μ_{b_1} and μ_{B_1} exceed 0, so that treatment T_1 is actually the most effective, with the difference between the two methods larger for larger $N_1 - n_1$. The selection probability for the Stallard method is smallest,

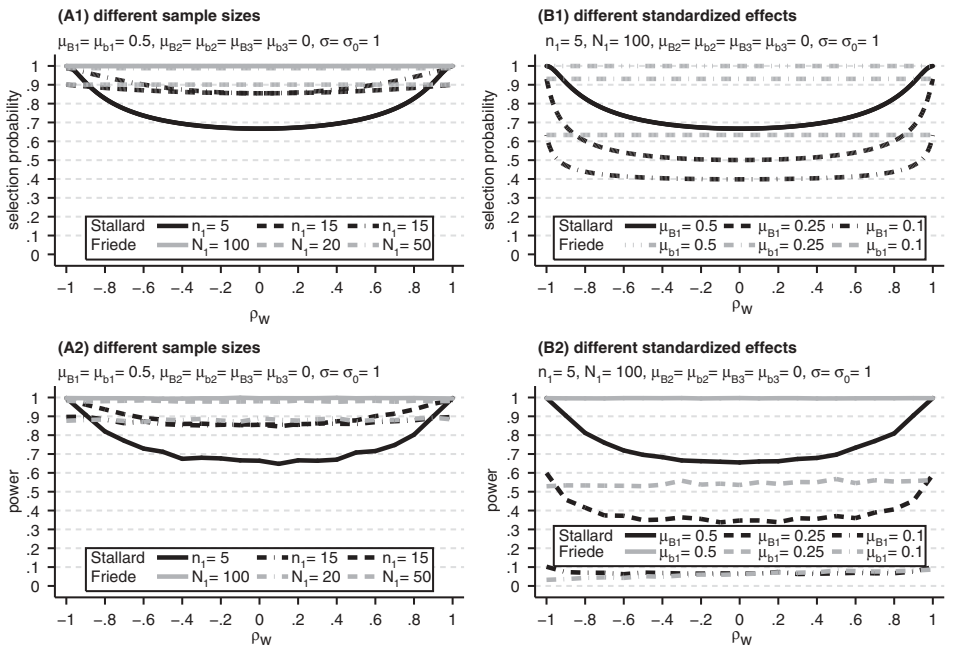

Figure 1 Probability to select treatment 1 (panels A1 and B1) and power (panels A2 and B2) for the Stallard (2010) and Friede et al. (2011) methods for different parameter settings under the fixed effects model.

and most different from that for the Friede et al. method, when $\rho_w = 0$ and $N_1^* = n_1$. The selection probability in this case is equal to that for a method that selects the best treatment solely on the basis of the long-term endpoint data from n_1 patients per group available at the interim analysis and so, unsurprisingly, decreases with decreasing n_1.

Since the selection probability for the Stallard method depends on μ_{B_1} and not on μ_{b_1}, whilst that for the Friede et al. method depends on μ_{b_1} and not on μ_{B_1}, panel B1 of Fig. 1 enables comparison of selection probabilities in settings with $\mu_{b_1}/\sigma_0 \neq \mu_{B_1}/\sigma$. Although when $\mu_{b_1}/\sigma_0 = \mu_{B_1}/\sigma$ the probability of selecting treatment T_1 with the Friede et al. is always as great at that for the Stallard method, it can be seen that this probability may be lower for the Friede et al. method when $\mu_{b_1}/\sigma_0 < \mu_{B_1}/\sigma$.

3.2. Power

As was the case with the probability of correct selection, we can define the power in different ways. In order to be consistent with the definition of selection probability, we define the power as the probability of rejecting the false null hypothesis corresponding to the most effective treatment, that is to rejecting H_I when I is the i that maximimes $\mu_{B_i} - \mu_{B_0}$. This definition is closely related to the "individual power" defined as the probability of rejecting a *particular* false null hypothesis (Westfall et al., 2011). The difference is that in the case of the individual power the null hypothesis we are interested in is specified in advance. Note that other definitions for the power are possible, such as, for example, defining the power as the probability to reject any false null hypothesis. For a discussion of different power concepts in the context of multiple testing see Westfall et al. (2011).

Assuming as above, without loss of generality, that the treatment effect on the long-term endpoint, $\mu_{B_i} - \mu_{B_0}$, is largest for $i = 1$, the power for the Friede et al. method is equal to

$$Pr(Z_2^* > Z_2^*, \ldots, Z_1^* > Z_k^*, C(p_{I,1}, p_{I,1}) \leq \alpha \text{ for all } \mathcal{I} \ni 1), \tag{9}$$

where $C(p_{\mathcal{I},1}, p_{\mathcal{I},2})$ is the combination function defined by (4) so that $C(p_{\mathcal{I},1}, p_{\mathcal{I},2}) \leq \alpha$ for all $\mathcal{I} \ni 1$ corresponds to rejection of H_1 in the Friede et al. method using the combination test and closed testing procedure as described above.

For the Stallard method, the power is equal to

$$Pr\,(S_{1,1} > S_{2,1}, \ldots, S_{1,1} > S_{k,1}, S_{1,2} \geq c), \tag{10}$$

where c is the critical value obtained to control the type I error rate using the method of Stallard (2010).

For the Stallard method, the power depends on N_1^*, n_2, and the standardized long-term endpoint treatment effects, $\mu_{B_1}/\sigma, \ldots, \mu_{Bk}/\sigma$, but not on the short-term endpoint effects $\mu_{b_1}, \ldots, \mu_{b_k}$. For the Friede et al. method, as the selection is based on the short-term endpoint data and the final test of the long-term endpoint data, the power depends on $\mu_{B_1}/\sigma, \ldots, \mu_{Bk}/\sigma$ in addition to $N_1, n_2, \mu_{b_1}/\sigma_0, \ldots, \mu_{b_k}/\sigma_0$ and ρ_w.

As (9) and (10) involve data from both stage one and stage two, analytic calculation of the power is less straightforward than that for the selection probabilities. The power values can most easily be estimated through simulation of data from the fixed effects model (1). This also allows the assumption of known σ and σ_0 to be relaxed.

Simulated power values for the two methods are shown in the lower panels (panels A2 and B2) of Fig. 1 in the same settings as the selection probabilities shown in the upper panels and discussed above with $n_2 = 200$. Estimated power values plotted are based on 10,000 simulations for each of the scenarios considered. For the larger effect sizes as shown in the panel A2 and the upper curves in panel B2, the power is very similar to the selection probability shown in the upper two panels. In this case if treatment T_1 is selected in stage one, the combination of the larger stage two sample size and the large effect size mean that it is very likely to be shown to be superior to the control. For smaller effect sizes, there is a larger chance of failing to demonstrate superiorioty even if the treatment T_1 is correctly selected, so that power values are smaller than the selection probabilities. In this case for extreme values of ρ_w or for very small treatment effects the Friede et al. method may be less powerful than the Stallard method. It is also interesting to note that while the power for the Stallard method is the same for positive and negative values of ρ_w of the same magnitude, for the Friede et al. method the power appears to be slightly lower for negative ρ_w than for positive ρ_w.

Figure 1 shows power values for $\mu_{B_1}/\sigma = \mu_{b_1}/\sigma_0$. As the power cannot exceed the selection probability, we may note, as above, that the Stallard method will be more powerful than the Friede et al. method if μ_{B_1}/σ is sufficiently large compared to μ_{b_1}/σ_0.

4. COMPARISON OF METHODS: RANDOM EFFECTS MODEL

For the fixed effects model, the distributional forms and calculated values given above show that the probability of selecting treatment T_1 and the power to reject the null hypothesis for this treatment, H_1, is higher for the Friede et al. selection method than that for the Stallard method when $\mu_{b_1}/\sigma_0 = \mu_{B_1}/\sigma \geq \mu_{b_2}/\sigma_0 = \cdots = \mu_{b_k}/\sigma_0 = \mu_{B_2}/\sigma = \cdots = \mu_{B_k}/\sigma$, but can be lower when $\mu_{b_1}/\sigma_0 < \mu_{B_1}/\sigma$. Unsurprisingly, given that the Friede et al. selection method relies solely on short-term endpoint observations, the performance of the method is good when the effects on the short-term endpoint are similar (or larger) to those on the long-term endpoint, but may be poor when they are smaller or reversed. In order to capture the relationship between the treatment effects μ_{b_i} and μ_{B_i}, it is therefore interesting to consider the random effects model introduced above, in which the correlation between the treatment means is explicitly included in the statistical model.

4.1. Selection Probability

As with the fixed effects model, we will consider the probability of selecting treatment T_1. Since the mean effect for this treatment, μ_{B_1}, is now considered to be a random variable, however, treatment T_1 might not always be the most effective even if $\theta_{B_1} > \theta_{B_i}$ for $i = 2, \ldots, k$. We will therefore focus on the probability of selecting treatment T_1 given that it is the most effective treatment, that is given that $\mu_{B_1} > \mu_{B_i}$ for all $i = 2, \ldots, k$. This is given by

$$Pr(\text{select } T_1) = Pr(Z_1^* - Z_i^* > 0, \text{ for } i = 2, \ldots, k | \mu_{B_1} - \mu_{B_i} > 0 \text{ for } i = 2, \ldots, k) \tag{11}$$

in the case in which selection is made using the Friede et al. method, and by

$$Pr(\text{select } T_1) = Pr(S_1 - S_i > 0 \text{ for } i = 2, \ldots, k | \mu_{B_1} - \mu_{B_i} > 0 \text{ for } i = 2, \ldots, k) \tag{12}$$

in the case when selection is made using the Stallard method. These probabilities may be evaluated using the joint distributions of $Z_1^* - Z_2^*, \ldots, Z_1^* - Z_k^*$ and $\mu_{B_1} - \mu_{B_2}, \ldots \mu_{B_1} - \mu_{B_2}$ or of $S_1 - S_2, \ldots, S_1 - S_k$ and $\mu_{B_1} - \mu_{B_2}, \ldots \mu_{B_1} - \mu_{B_2}$ given in the Appendix.

Figure 2 shows the probability of selecting treatment T_1 given that this is actually the most effective treatment when selection uses either the Friede et al. or the Stallard method. Selection probabilities are shown for a range of ρ_w values in the setting in which $\sigma_{B_1} = \cdots = \sigma_{B_k} = \sigma_{b_1} = \cdots = \sigma_{b_k} = 1$. In panels A1 and A2, $\theta_{B_1} = \theta_{b_1} = 0.5$ and $\theta_{B_i} = \theta_{b_i} = 0$ for $i = 2, \ldots, k$, so that on average the first treatment is effective on both endpoints and all others are not, and $\sigma = \sigma_0 = 1$. The separate lines give the treatment selection for different sample sizes. In panels B1 and B2, $\sigma = \sigma_0 = 1$, $n_1 = 5$

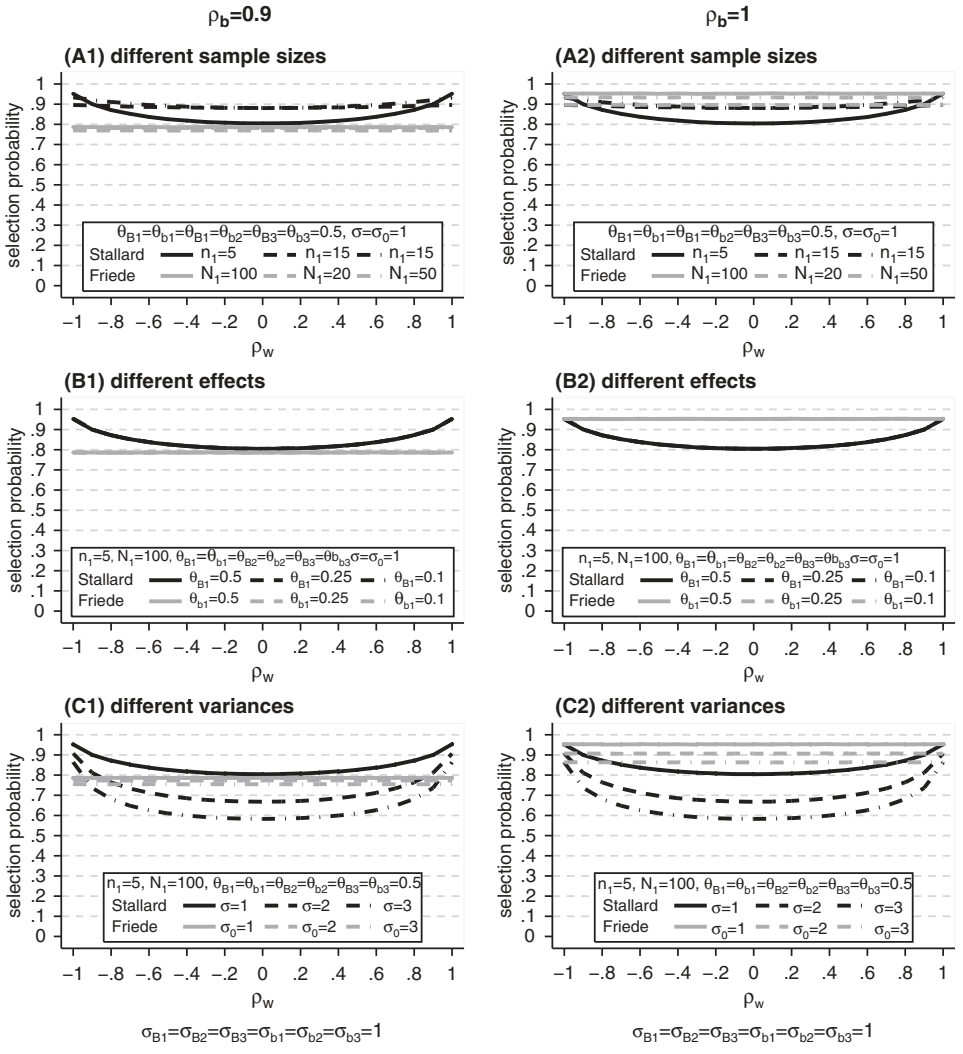

Figure 2 Probability to select treatment 1 based on the methods by Stallard (2010) and by Friede et al. (2011) for different parameter settings under the random effects model (given that treatment 1 is the most effective).

and $N_1 = 100$, $\theta_{B_i} = \theta_{b_i} = 0$ for $i = 2, \ldots, k$ and $\theta_{B_1} = \theta_{b_1}$ with different lines on the plot corresponding to different treatment effects. In panels C1 and C2, $\theta_{B_1} = \theta_{b_1} = 0.5$ and $\theta_{B_i} = \theta_{b_i} = 0$ for $i = 2, \ldots, k$, $n_1 = 20$ and $N_1 = 100$ and the different lines correspond to different values of σ and σ_0. The right-hand column in the figure shows selection probabilities for $\rho_b = 1$, that is when there is perfect correlation between the means on the two endpoints, and the left-hand column those for $\rho_b = 0.9$.

The selection probabilities for $\rho_b = 1$ shown in the right-hand column are generally similar to those for the fixed effects model with $\mu_{B_1} = \mu_{b_1}$ given in Fig. 1. The best treatment is more likely to be selected using the Friede et al. method than the Stallard method, with the two methods coinciding when $\rho_w = \pm 1$. The main difference between the selection probabilities under the random effects model and the fixed effects model in this case is that under the random effects model there is very little effect of the average treatment effect, $\theta_{B_1} = \theta_{b_1}$, in contrast to the results for the fixed effects model considered above. This is reasonable given that the figure shows the probability of selecting T_1 given that the actual treatment effect is largest for that treatment, that is given $\mu_{B_1} \geq \mu_{B_i}, i = 2, \ldots, k$. An increase in σ for the Stallard method or in σ_0 for the Friede et al. method does, however, reduce the probability of selecting treatment T_1, as the standardized average difference between the treatments on the long- or short-term endpoint, respectively, is reduced.

As the Friede et al. method uses only the short-term endpoint data for the selection, it is not surprising that it performs well when the means on the two endpoints are perfectly correlated, since the selection is based on a larger number of observations and a treatment performing well on the short-term endpoint is more likely to have a large long-term endpoint mean. The left-hand column shows selection probabilities for $\rho_b = 0.9$. The selection probabilities for the Stallard method do not depend on ρ_b, so that these are exactly the same as those in the panels in the right-hand column. The Friede et al. method selects the correct treatment with lower probability than when the short-term and long-term treatment means are perfectly correlated; in this case the short-term endpoint means are less predictive of the treatment with the largest long-term responses. In this case, the Stallard method can lead to a higher probability of correctly selecting treatment T_1, particularly when ρ_w is high. Smaller values of the correlation ρ_b will result in worse performance of the Friede et al. method.

The latter point is illustrated more clearly in Fig. 3. This shows the probability under the random effects model of correctly selecting treatment T_1 given that this is the most effective for $n_1 = 5$, $N_1 = 100$, $\sigma = \sigma_0 = 1$, $k = 3$, $\sigma_{B_1} = \cdots = \sigma_{B_k} = \sigma_{b_1} = \cdots = \sigma_{b_k} = 1, \theta_{B_1} = \cdots = \theta_{B_k} = \theta_{b_1} = \cdots = \theta_{b_k} = 0.5$ for the Stallard method for a range of ρ_w values and for the Friede et al. method for a range of ρ_b values. Since the selection probabilities for the Stallard method do not depend on ρ_b and for the Friede et al. method do not depend on ρ_w, the two lines are shown on the same graph. Comparing the two lines, we see that the Stallard method always has a higher selection probability than the Friede et al. method if $\rho_w = \rho_b$ except when $\rho_w = \rho_b = 1$, when both probabilities are the same. The three horizontal lines represent selection probabilities for the Friede et al. method where ρ_b is fixed to either 1 (short dash), 0.95 (dash dot), or 0.9 (long dash). Comparing these lines with those for the Stallard method, we observe that if $\rho_b = 1$ the Friede et al. method is always better than the Stallard method regardless of ρ_w (with the exception of $\rho_w = \pm 1$, when the selection probabilities for the two methods are again equal). If $\rho_b = 0.95$, the Friede et al. method is only better than the Stallard method if ρ_w is small. While if $\rho_b = 0.9$, the Stallard method is always better than the Friede et al. method regardless of ρ_w.

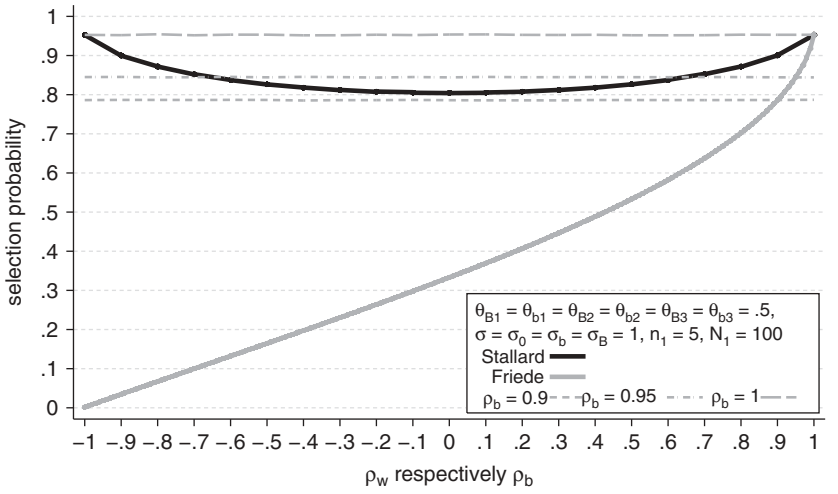

Figure 3 Probability to select treatment 1 based on the methods by Stallard (2010) for a range of ρ_w values and by Friede et al. (2011) for a range of ρ_b values under the random effects model (given that treatment 1 is the most effective).

When $\rho_b < 0$, the probability of selecting treatment T_1 using the Friede et al. method can be low. The probability approaches zero as ρ_b approaches -1 and the treatment effects on the long- and short-term endpoints consistently go in opposite directions.

4.2. Power

In a similar approach to that used for evaluation of the treatment selection probabilities, we consider the power defined to be the probability that treatment T_1 is selected at the interim analysis and found to be significantly superior to the control at the final analysis conditional on it actually being the best treatment, that is on $\mu_{B_1} \geq \mu_{B_i}, i = 2, \ldots, k$. As in the fixed effects model, the power will again be estimated via simulation. In this case, data are simulated from the random effects model given by (1) and (2). In detail, for each simulation, treatment means $\mu_{b_0}, \ldots, \mu_{b_k}, \mu_{B_0}, \ldots, \mu_{B_k}$ are first simulated from (2) then, given these treatment mean values, data are simulated from (1).

Simulated power values are shown in Fig. 4 under the same scenarios as Fig. 2. As in the fixed effects setting, for reasonably large standardized effect sizes, the power is similar to the selection probability, but is slightly lower when the standardized effect is smaller, either because of a reduction in the effect size or an increase in the within-treatment variance.

5. DISCUSSION

There has been much recent interest in adaptive seamless phase II/III clinical trials in which randomization is initially between a number of experimental treatments and a control, with less effective treatments dropped from the study on the basis of results from an interim analysis. Building on methods using short-term information to supplement

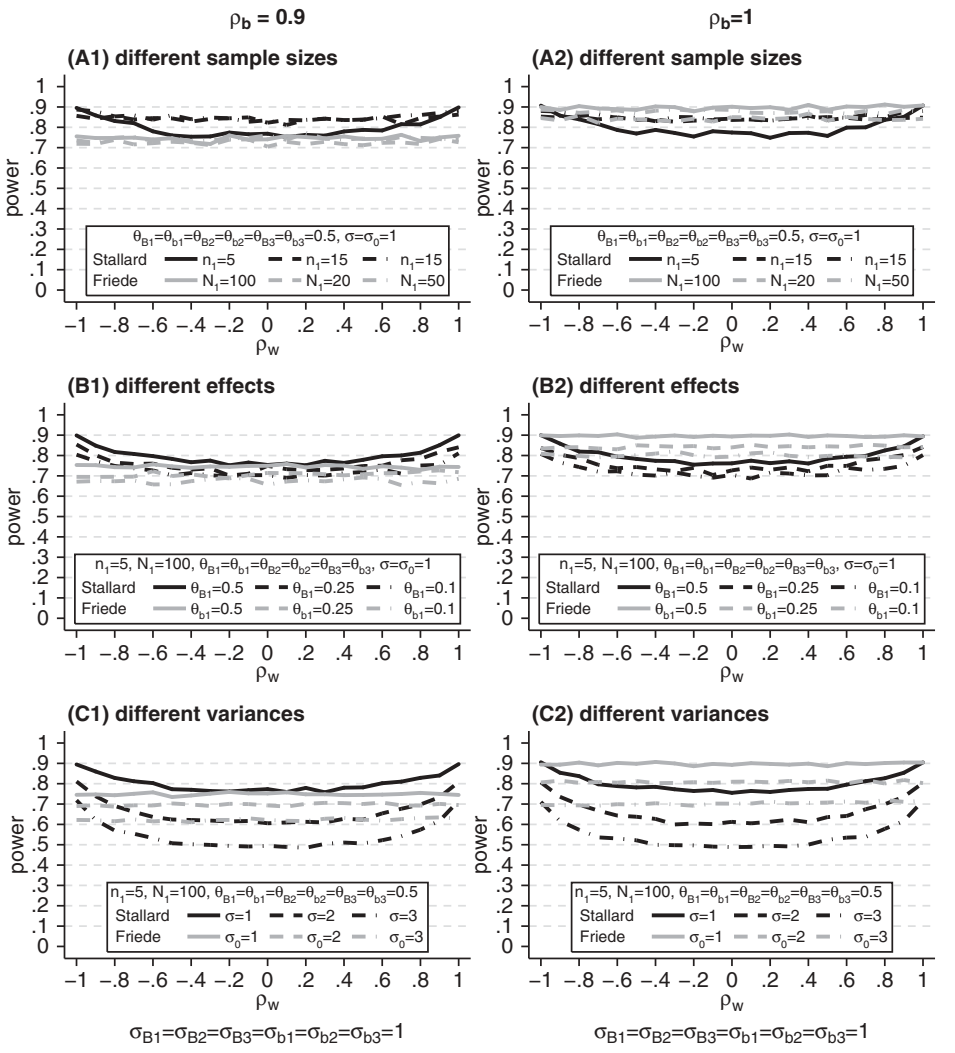

Figure 4 Power for the Stallard (2010) and Friede et al. (2011) methods for different parameter settings under the random effects model (given that treatment 1 is the most effective).

long-term information originally developed in the context of interim analyses for early stopping, two methods have been proposed for using short-term endpoint data in the treatment selection (Stallard, 2010; Friede et al., 2011). In this paper, we have compared these two methods. Our aim has been to provide a comparison that will enable choice of the most appropriate method when designing an adaptive seamless phase II/III design.

In the Friede et al. method, only the short-term endpoint data are used for the treatment selection. In contrast, the Stallard method uses a combination of short- and long-term endpoint data. The latter method can thus only be used when some long-term responses are available for inclusion in the interim analysis. In both methods, the final analysis is based on the long-term endpoint data alone from the selected treatment and control. This

Table 2 Summary impact of model parameters on selection probabilities

Sample sizes	
n_1	Larger values reduce impact of short-term endpoint data.
N_1	Larger numbers increase impact of short-term endpoint data.
n_2	Larger values increase power but do not influence treatment selection.
Fixed or random effects model parameters	
μ_{B_i}	More disperse values increase differences between treatments making treatment selection easier.
σ^2	Larger values increase variability making treatment selection harder.
μ_{b_i}	More disperse values increase differences between treatments making treatment selection easier with Friede et al. method. No impact on Stallard method.
σ_0^2	Larger values increase variability making treatment selection harder.
ρ_w	Larger values (of ρ_w^2) increase influence of short-term endpoints in Stallard method. No impact in Friede et al. method.
Random effects model parameters	
θ_{B_i}	More disperse values increase differences between treatments making treatment selection easier.
σ_B^2	Larger values make treatment means more disperse making treatment selection easier.
θ_{b_i}	More disperse values increase differences between treatments making treatment selection easier with Friede et al. method. No impact on Stallard method.
σ_b^2	Larger values make treatment means more disperse making treatment selection easier with Friede et al. method. No impact on Stallard method.
ρ_b	Larger values make treatment effects on two endpoints more closely related and improve treatment selection with Friede et al. method. No impact on Stallard method.

is in contrast to other group-sequential methods in which it is desired to draw inference on both endpoints, for example requiring both to be sufficiently promising (see, e.g., Jennison and Turnbull, 1993; Kimani et al., 2009) or with early and late observations of the same endpoint treated as a longitudinal data (see, e.g., Spiessens et al., 2000; Lee et al., 1996).

Our comparison has considered scenarios in which the treatment means are taken to be fixed, with one treatment more effective than all others and the control, which are equally effective, and scenarios in which the treatment means are taken to be random but are correlated. A summary of the effects of the different model parameters on the selection probability based on the simulations reported above is given in Table 2. Our results indicate that under the fixed effects model, if the treatment effect on the short-term endpoint is as large or larger than that on the long-term endpoint for the effective treatment, the Friede et al. method is more likely to lead to selection of the most effective treatment, and is correspondingly more powerful. If the effect on the short-term endpoint is less than that on the long-term endpoint, the Stallard method may be more likely to select the correct treatment and more powerful, particularly when the within-group correlation between the endpoints is high.

Under the random effects model, the effect of correlation between the treatment means on the two endpoints can be considered. This parameter gives an indication of the extent to which treatment effects on the long- and short-term endpoints go in the same direction. In this case, our results indicate that the Friede et al. method leads to a higher probability of selecting the best treatment and to higher power only when the correlation between the treatment means is sufficiently high. The threshold depends on the sample sizes and variances, but we have shown that even when the number of patients for whom long-term endpoint data are available at the interim analysis is small, under the scenarios we

have considered, the Friede et al. method is less powerful unless the correlation between the means is relatively high; for the scenario we considered above 0.9 when the within-group variance and between-group variance are both equal to 1.

In order to be able to choose between the different methods, some estimates of the model parameters, including the variances and correlations in (1) and (2) are required. In some cases, data from other trials will be available, particularly to give information on the parameters in (1). The correlation can vary considerably depending on the setting and endpoints chosen. Julious and Mullee (2008), for example, report a ρ_w of 0.67 between the same endpoint measured at baseline and at the end of the trial, so that the correlation between an early and final measurement of this endpoint would presumably be higher than this, whereas Chataway et al. (2011) report a ρ_w of 0.13 between two different endpoints, though it was still proposed to use the early endpoint for treatment selection. The parameters in (2) are harder to estimate since their estimation requires data from a number of different trials or treatments.

If detailed information on parameter values is unavailable, it may still be possible to make some guess of possible ranges for parameters, or to use the methods described above to conduct sensitivity analyses. We are currently working on approaches that use the data from the first stage of the trial to estimate the parameters of (1) and (2) and to decide between the different treatment selection strategies on the basis of these estimates.

Our comparison of the procedures has used a combination of analytic calculations based on multivariate normal distributions to calculate selection probabilities and simulations to estimate the power. The simulations can be time-consuming when an extensive search for an appropriate sample size is required, or when it is desirable to explore the tradeoff between patients in stages one and two of the trial. The power is bounded above by the selection probability and in many of the settings considered above, the two probabilities are quite similar. This is likely to be particularly true when the assumed effect size is relatively large and the sample size for the second stage is substantially larger than that for the first stage. For example, in the settings described above with three treatments compared to a control treatment on the basis of long-term data on 5 or 15 patients per group and short-term data on 100, 20, or 50 patients per group at the interim analysis with a final sample size of 200 per group, when the standardized effect size on both endpoints for the sole effective treatment of 0.5, we found that the estimated power was at least 97.5% of the selection probability. In such cases, an approximate sample size calculation could be based on the selection probability using the analytic calculations described. If necessary, this could be followed by a much more restricted set of simulations to confirm the power of the final design chosen.

A. APPENDIX: DISTRIBUTIONS REQUIRED FOR CALCULATION OF TREATMENT SELECTION PROBABILITIES

A.1. Fixed Effects Model

Calculation of the probability of selecting treatment T_1 using the Friede et al. and Stallard methods under the fixed effects model require the joint distribution of $\left(Z_1^* - Z_2^*, \ldots, Z_1^* - Z_k^*\right)$ and $\left(S_{1,1} - S_{2,1}, \ldots, S_{1,1} - S_{k,1}\right)$, respectively. Detailed derivations of these are given in the online Supplementary Material, leading to

$$\begin{pmatrix} Z_1^* - Z_2^* \\ \vdots \\ Z_1^* - Z_k^* \end{pmatrix} \sim N\left(\begin{pmatrix} \frac{\mu_{b_1}-\mu_{b_2}}{\sigma_0}\sqrt{\frac{N_1}{2}} \\ \vdots \\ \frac{\mu_{b_1}-\mu_{b_3}}{\sigma_0}\sqrt{\frac{N_1}{2}} \end{pmatrix}, \begin{pmatrix} 1 & 1/2 & \cdots & 1/2 \\ 1/2 & 1 & & 1/2 \\ \vdots & & \ddots & \vdots \\ 1/2 & \cdots & 1/2 & 1 \end{pmatrix}\right) \tag{A.1}$$

and

$$\begin{pmatrix} S_{1,1} - S_{2,1} \\ \vdots \\ S_{1,1} - S_{k,1} \end{pmatrix} \sim N\left(\begin{pmatrix} \frac{\mu_{B_1}-\mu_{B_2}}{\sigma}\sqrt{\frac{N_1^*}{2}} \\ \vdots \\ \frac{\mu_{B_1}-\mu_{B_3}}{\sigma}\sqrt{\frac{N_1^*}{2}} \end{pmatrix}, \begin{pmatrix} 1 & 1/2 & \cdots & 1/2 \\ 1/2 & 1 & & 1/2 \\ \vdots & & \ddots & \vdots \\ 1/2 & \cdots & 1/2 & 1 \end{pmatrix}\right). \tag{A.2}$$

A.2. Random Effects Model

Treatment selection probabilities using the Friede et al. and Stallard methods under the fixed effects model may be evaluated using the joint distributions of $Z_1^* - Z_2^*, \ldots, Z_1^* - Z_k^*$ and $\mu_{B_1} - \mu_{B_2}, \ldots \mu_{B_1} - \mu_{B_2}$ or of $S_1 - S_2, \ldots, S_1 - S_k$ and $\mu_{B_1} - \mu_{B_2}, \ldots \mu_{B_1} - \mu_{B_2}$, respectively.

The joint distribution of $Z_1^* - Z_2^*, \ldots, Z_1^* - Z_k^*$ and $\mu_{B_1} - \mu_{B_2}, \ldots \mu_{B_1} - \mu_{B_2}$ is given by

$$\begin{pmatrix} Z_1^* - Z_2^* \\ \vdots \\ Z_1^* - Z_k^* \\ \mu_{B_1} - \mu_{B_2} \\ \vdots \\ \mu_{B_1} - \mu_{B_k} \end{pmatrix} \sim N\left(\begin{pmatrix} \frac{\theta_{b_1}-\theta_{b_2}}{\sigma_0}\sqrt{\frac{N_1}{2}} \\ \vdots \\ \frac{\theta_{b_1}-\theta_{b_k}}{\sigma_0}\sqrt{\frac{N_1}{2}} \\ \theta_{B_1} - \theta_{B_2} \\ \vdots \\ \theta_{B_1} - \theta_{B_k} \end{pmatrix}, \Sigma^{(F)}\right) \tag{A.3}$$

where

$$\Sigma^{(F)} = \begin{pmatrix} \Sigma_{1,1}^{(F)} & \Sigma_{1,2}^{(F)} \\ \Sigma_{1,2}^{(F)} & \Sigma_{2,2}^{(F)} \end{pmatrix}$$

with

$$\Sigma_{1,1}^{(F)} = \begin{pmatrix} \frac{\left(\sigma_{b_1}^2+\sigma_{b_2}^2\right)}{2\sigma_0^2} + 1 & \frac{\sigma_{b_1}^2 N_1}{2\sigma_0^2} + \frac{1}{2} & \cdots & \frac{\sigma_{b_1}^2 N_1}{2\sigma_0^2} + \frac{1}{2} \\ \frac{\sigma_{b_1}^2 N_1}{2\sigma_0^2} + \frac{1}{2} & \frac{\left(\sigma_{b_1}^2+\sigma_{b_3}^2\right)N_1}{2\sigma_0^2} + 1 & & \frac{\sigma_{b_1}^2 N_1}{2\sigma_0^2} + \frac{1}{2} \\ \vdots & & \ddots & \vdots \\ \frac{\sigma_{b_1}^2 N_1}{2\sigma_0^2} + \frac{1}{2} & \cdots & \frac{\sigma_{b_1}^2 N_1}{2\sigma_0^2} + \frac{1}{2} & \frac{\left(\sigma_{b_1}^2+\sigma_{b_k}^2\right)N_1}{2\sigma_0^2} + 1 \end{pmatrix},$$

$$\Sigma_{12}^{(F)} = \begin{pmatrix} \rho_b \frac{\sigma_{b_1}\sigma_{B_1}+\sigma_{b_2}\sigma_{B_2}}{\sigma_0}\sqrt{\frac{N_1}{2}} & \rho_b \frac{\sigma_{b_1}\sigma_{B_1}}{\sigma_0}\sqrt{\frac{N_1}{2}} & \cdots & \rho_b \frac{\sigma_{b_1}\sigma_{B_1}}{\sigma_0}\sqrt{\frac{N_1}{2}} \\ \rho_b \frac{\sigma_{b_1}\sigma_{B_1}}{\sigma_0}\sqrt{\frac{N_1}{2}} & \rho_b \frac{\sigma_{b_1}\sigma_{B_1}+\sigma_{b_3}\sigma_{B_3}}{\sigma_0}\sqrt{\frac{N_1}{2}} & & \rho_b \frac{\sigma_{b_1}\sigma_{B_1}}{\sigma_0}\sqrt{\frac{N_1}{2}} \\ \vdots & & \ddots & \vdots \\ \rho_b \frac{\sigma_{b_1}\sigma_{B_1}}{\sigma_0}\sqrt{\frac{N_1}{2}} & \cdots & \rho_b \frac{\sigma_{b_1}\sigma_{B_1}}{\sigma_0}\sqrt{\frac{N_1}{2}} & \rho_b \frac{\sigma_{b_1}\sigma_{B_1}+\sigma_{b_k}\sigma_{B_k}}{\sigma_0}\sqrt{\frac{N_1}{2}} \end{pmatrix}$$

and

$$\Sigma_{22}^{(F)} = \begin{pmatrix} \sigma_{B_1}^2 + \sigma_{B_2}^2 & \sigma_{B_1}^2 & \cdots & \sigma_{B_1}^2 \\ \sigma_{B_1}^2 & \sigma_{B_1}^2 + \sigma_{B_3}^2 & & \sigma_{B_1}^2 \\ \vdots & & \ddots & \vdots \\ \sigma_{B_1}^2 & \cdots & \sigma_{B_1}^2 & \sigma_{B_1}^2 + \sigma_{B_k}^2 \end{pmatrix}.$$

The joint distribution of $S_1 - S_2, \ldots, S_1 - S_k$ and $\mu_{B_1} - \mu_{B_2}, \ldots \mu_{B_1} - \mu_{B_2}$ is given by

$$\begin{pmatrix} S_1 - S_2 \\ \vdots \\ S_1 - S_k \\ \mu_{B_1} - \mu_{B_2} \\ \vdots \\ \mu_{B_1} - \mu_{B_k} \end{pmatrix} \sim N\left(\begin{pmatrix} \frac{\theta_{B_1} - \theta_{B_2}}{\sigma}\sqrt{\frac{N_1^*}{2}} \\ \vdots \\ \frac{\theta_{B_1} - \theta_{B_k}}{\sigma}\sqrt{\frac{N_1^*}{2}} \\ \theta_{B_1} - \theta_{B_2} \\ \vdots \\ \theta_{B_1} - \theta_{B_k} \end{pmatrix}, \Sigma^{(S)} \right) \tag{A.4}$$

where

$$\Sigma^{(S)} = \begin{pmatrix} \Sigma_{1,1}^{(S)} & \Sigma_{1,2}^{(S)} \\ \Sigma_{1,2}^{(S)} & \Sigma_{2,2}^{(S)} \end{pmatrix}$$

with

$$\Sigma_{1,1}^{(S)} = \begin{pmatrix} \frac{\left(\sigma_{B_1}^2 + \sigma_{B_2}^2\right)N_1^*}{2\sigma^2} + 1 & \frac{\sigma_{B_1}^2 N_1^*}{2\sigma^2} + \frac{1}{2} & \cdots & \frac{\sigma_{B_1}^2 N_1^*}{2\sigma^2} + \frac{1}{2} \\ \frac{\sigma_{B_1}^2 N_1^*}{2\sigma^2} + \frac{1}{2} & \frac{\left(\sigma_{B_1}^2 + \sigma_{B_3}^2\right)N_1^*}{2\sigma^2} + 1 & & \frac{\sigma_{B_1}^2 N_1^*}{2\sigma^2} + \frac{1}{2} \\ \vdots & & \ddots & \vdots \\ \frac{\sigma_{B_1}^2 N_1^*}{2\sigma^2} + \frac{1}{2} & \cdots & \frac{\sigma_{B_1}^2 N_1^*}{2\sigma^2} + \frac{1}{2} & \frac{\left(\sigma_{B_1}^2 + \sigma_{B_k}^2\right)N_1^*}{2\sigma^2} + 1 \end{pmatrix},$$

$$\Sigma_{12}^{(S)} = \begin{pmatrix} \frac{\sigma_{B_1}^2 + \sigma_{B_2}^2}{\sigma}\sqrt{\frac{N_1^*}{2}} & \frac{\sigma_{B_1}^2}{\sigma}\sqrt{\frac{N_1^*}{2}} & \cdots & \frac{\sigma_{B_1}^2}{\sigma}\sqrt{\frac{N_1^*}{2}} \\ \frac{\sigma_{B_1}^2}{\sigma}\sqrt{\frac{N_1^*}{2}} & \frac{\sigma_{B_1}^2 + \sigma_{B_3}^2}{\sigma}\sqrt{\frac{N_1^*}{2}} & & \frac{\sigma_{B_1}^2}{\sigma}\sqrt{\frac{N_1^*}{2}} \\ \vdots & & \ddots & \vdots \\ \frac{\sigma_{B_1}^2}{\sigma}\sqrt{\frac{N_1^*}{2}} & \cdots & \frac{\sigma_{B_1}^2}{\sigma}\sqrt{\frac{N_1^*}{2}} & \frac{\sigma_{B_1}^2 + \sigma_{B_k}^2}{\sigma}\sqrt{\frac{N_1^*}{2}} \end{pmatrix},$$

$\Sigma_{2,2}^{(S)} = \Sigma_{2,2}^{(F)}$ and N_1^* given by (6).

Detailed derivations are again given in the online supplemental material.

ACKNOWLEDGMENTS

We are grateful to the Editor and two anonymous reviewers for their helpful comments on this paper.

FUNDING

The work was funded by UK Medical Research Council grant number G1001344.

SUPPLEMENTAL MATERIAL

Supplemental data for this article can be accessed on the publisher's website.

REFERENCES

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

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

Brannath, W., Posch, M., Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association* 97:236–244.

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

Bretz, F., Schmidli, H., König, F., Racine, A., Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal* 48:623–634.

Chataway, J., Nicholas, R., Todd, S., Miller, D., Parsons, N., Valdés-Márquez, E., Stallard, N., Friede, T. (2011). A novel adaptive design strategy increases the efficiency of clinical trials in secondary progressive multiple sclerosis. *Multiple Sclerosis* 17:81–88.

Chow, S.-C., Chang, M., Pong, A. (2005). Statistical consideration of adaptive methods in clinical development. *Journal of Biopharmaceutical Statistics* 15:575–591.

Cook, R., Farewell, V. (1996). Incorporating surrogate endpoints into group sequential trials. *Biometrical Journal* 38:119–130.

Dragalin, V. (2011). An introduction to adaptive designs and adaptation in CNS trials. *European Neuropsychopharmacology* 21(2):153–158.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* 50:1096–1121.

Engel, B., Walstra, P. (1991). Increasing precision or reducing expense in regression experiments by using information from a concomitant variable. *Biometrics* 47(1):13–20.

European Medicines Agency (EMEA) - Committee for Medicinal Products for Human Use (CHMP) (2007). CHMP reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. http://www.emea.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003616.pdf (accessed July 13, 2012).

Food and Drug Administration (FDA) (2010). Guidance for industry - adaptive design clinical trials for drugs and biologics. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm201790.pdf (accessed July 13, 2012).

Friede, T., Parsons, N., Stallard, N., Todd, S., Valdés-Márquez, E., Chataway, J., Nicholas, R. (2011). Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: an application in multiple sclerosis. *Statistics in Medicine* 30:1528–1540.

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

Galbraith, S., Marschner, I. (2003). Interim analysis of continuous long-term endpoints in clinical trials with longitudinal outcomes. *Statistics in Medicine* 22:1787–1805.

Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M., Pinheiro, J. (2006). Adaptive designs in clinical drug development - an executive summary of the pharma working group. *Journal of Biopharmaceutical Statistics* 16:275–283.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Bornkamp, B., Hothorn, T. (2012). Package 'mvtnorm'. URL http://CRAN.R-project.org, R package version 0.9-9992.

Hampson, L. V., Jennison, C. (2013). Group sequential tests for delayed responses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75:1–37.

Jennison, C., Turnbull, B. (1993). Group sequential tests for bivariate response: Interim analyses of clinical trials with both efficacy and safety endpoints. *Biometrics* 49:741–752.

Julious, S. A., Mullee, M. A. (2008). Issues with using baseline in last observation carried forward analysis. *Pharmaceutical Statistics* 7:142–146.

Kelly, P. J., Stallard, N., Todd, S. (2005). An adaptive group sequential design for phase II/III clinical trials that select a single treatment from several. *Journal of Biopharmaceutical Statistics* 15:641–658.

Kimani, P. K., Stallard, N., Hutton, J. L. (2009). Dose selection in seamless phase ii/iii clinical trials based on efficacy and safety. *Statistics in Medicine* 28:917–936.

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

Lee, S. J., Kim, K., Tsiatis, A. A. (1996). Repeated significance testing in longitudinal clinical trials. *Biometrika* 83:779–789.

Lehmacher, W., Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* 55(4):1286–1290.

Marcus, R., Peritz, E., Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 63(3):655–660.

Marschner, I., Becker, S. (2001). Interim monitoring of clinical trials based on long-term binary endpoints. *Statistics in Medicine* 20:177–192.

Posch, M., König, F., Branson, M., Brannath, W., Dunger-Baldauf, C., Bauer, P. (2005). Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine* 24:3697–3714.

Schmoll, H., Cunnighmam, D., A., S., Krapetis, C., Rougier, P., Koski, S., P., B., Mookerjee, B., Robertson, J., van Cutsem, E. (2010). mFOLFOX6 + cediranib vs mFOLFOX6 + bevacizumab in previously untreated metastatic colorectal cancer (mcrc): A randomised, double-blind, phase II/III study (HORIZON III). *Ann. Oncol*. 21(Supplement 8): vii189–vii224.

Sooriyarachchi, M., Whitehead, J., Whitehead, A., Bolland, K. (2006). The sequential analysis of repeated binary responses: A score test for the case of three time points. *Statistics in Medicine* 25:2196–2214.

Spiessens, B., Lesaffre, E., Verbeke, G., Kim, K., DeMets, D. L. (2000). An overview of group sequential methods in longitudinal clinical trials. *Statistical Methods in Medical Research* 19:497–515.

Stallard, N. (2010). A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Statistics in Medicine* 29:959–971.

Stallard, N., Todd, S. (2003). Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine* 22:689–703.

Stallard, N., Todd, S. (2011). Seamless phase II/III designs. *Statistical Methods in Medical Research* 20:623–634.

Todd, S., Stallard, N. (2005). A new clinical trial design combining phases II and III: Sequential designs with treatment selection and a change of endpoint. *Drug Information Journal* 39:109–118.

Westfall, P. H., Tobias, R. D., Wolfinger, R. D. (2011). *Multiple Comparisons and Multiple Tests Using SAS*. Cary, NC: SAS Institute Inc.

Whitehead, A., Sooriyarachchi, M., Whitehead, J., Bolland, K. (2008). Incorporating intermediate binary responses into interim analyses of clinical trials: A comparison of four methods. *Statistics in Medicine* 27:1646–1666.