Statistics
in Medicine

Research Article

Received 13 May 2014, Accepted 24 September 2015 Published online 12 October 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6760

Designing multi-arm multi-stage clinical trials using a risk–benefit criterion for treatment selection

Thomas Jaki*† and Lisa V. Hampson

Multi-arm clinical trials that compare several active treatments to a common control have been proposed as an efficient means of making an informed decision about which of several treatments should be evaluated further in a confirmatory study. Additional efficiency is gained by incorporating interim analyses and, in particular, seamless Phase II/III designs have been the focus of recent research. Common to much of this work is the constraint that selection and formal testing should be based on a single efficacy endpoint, despite the fact that in practice, safety considerations will often play a central role in determining selection decisions. Here, we develop a multi-arm multi-stage design for a trial with an efficacy and safety endpoint. The safety endpoint is explicitly considered in the formulation of the problem, selection of experimental arm and hypothesis testing. The design extends group-sequential ideas and considers the scenario where a minimal safety requirement is to be fulfilled and the treatment yielding the best combined safety and efficacy trade-off satisfying this constraint is selected for further testing. The treatment with the best trade-off is selected at the first interim analysis, while the whole trial is allowed to compose of J analyses. We show that the design controls the familywise error rate in the strong sense and illustrate the method through an example and simulation. We find that the design is robust to misspecification of the correlation between the endpoints and requires similar numbers of subjects to a trial based on efficacy alone for moderately correlated endpoints. © 2015 The Authors. *Statistics in Medicine* Published by John Wiley & Sons Ltd.

Keywords: familywise error rate; multi-arm multi-stage (MAMS); multiple endpoints; safety; treatment selection

1. Introduction

Prior to undertaking a confirmatory Phase III clinical trial, there is often uncertainty about which treatment should be selected for evaluation from a number of candidates. Here, treatments could be different doses of the same drug or different combinations of multiple drugs. Uncertainty about which treatment to select often stems from the fact that early phase trials typically evaluate medicines in different populations, using different endpoints, to those that will be the focus of confirmatory studies. The current high failure rate of Phase III trials of around 50% [1] combined with their substantial cost [2] make selecting an appropriate treatment for evaluation in Phase III of paramount importance.

As an efficient solution to this problem, designs for seamless Phase II/III multi-arm clinical trials have been proposed, which compare several active treatments with a common control group. Phase II of the study is used to learn about all treatments. At the end of this first stage, one or more of the active treatments is selected and taken forward with control for evaluation in Phase III. Data accumulated across both stages of the trial are used to test whether the selected treatment(s) is(are) superior to control at the end of the study. The simultaneous comparison of several treatments means that expected sample sizes and durations of multi-arm trials are markedly smaller than the alternative of evaluating each treatment separately. For added efficiency, solutions that incorporate a series of interim analyses to allow

Medical and Pharmaceutical Statistics Research Unit, Department of Mathematics and Statistics, Lancaster University, Lancaster, U.K.
**Correspondence to: Thomas Jaki, Medical and Pharmaceutical Statistics Research Unit, Department of Mathematics and Statistics, Lancaster University, Lancaster, U.K.*
† E-mail: jaki.thomas@gmail.com

early stopping either for efficacy or to drop ineffective treatments have recently received attention [3–7]. The approaches discussed in the literature to date can be characterized by two main differences. The first is the underlying statistical framework that either generalizes group sequential designs [8, 9] to accommodate multiple treatment arms [3] or makes use of p-value combination rules within closed testing procedures [10]. The second difference is the way in which treatments are selected. In [3], for example, only the best performing treatment is selected at the first interim analysis and subsequently compared with control over multiple stages, while in [6], all treatments surpassing a threshold at each stage are continued. Meanwhile, Kelly *et al.* [11] advocate a rule that selects all treatments close to the best performing treatment at the first interim analysis.

A further commonality of several of the approaches discussed in the literature is the assumption of normally distributed data and the fact that a single endpoint is considered. However, there are exceptions. For results for non-normal endpoints, see, for example, [12–14]. More generally, adaptive procedures using p-value combination rules within closed testing procedures make no assumptions about the distribution of patient responses nor place any constraints on the form of the treatment selection rule: the only constraint is that p-values for testing elementary and intersection null hypotheses must follow a Uniform(0,1) (or stochastically larger) distribution under the null [5]. For procedures that consider more than endpoint, see [15], which describes a seamless Phase II/III trial using a composite rule based on two hierarchically ordered efficacy endpoints to guide treatment selection decisions, as well as relevant safety data; to adjust for multiple testing, pairwise comparisons of selected treatments against placebo are adjusted using a Bonferroni correction. Early phase oncology trials also often assess efficacy and monitor toxicity, see [16] for a Phase I/II trial design combining time-to-response and time-to-toxicity endpoints into a single statistic used for interim decision making, weighting pairs of outcomes according to utilities elicited from experts. In other areas, such as mental health, co-primary efficacy endpoints are measured, and no single measure is accepted as definitive. Although it is sometimes sensible to combine different endpoints into a single test statistic, substantial gains in efficiency can be achieved if they are evaluated jointly, especially when endpoints capture the effects of a treatment on different aspects of the disease. Furthermore, combining information obtained on efficacy and safety endpoints into a single test statistic will be inappropriate because good efficacy will not compensate for poor safety in practice.

Methods for two-arm group-sequential trials with multivariate normal endpoints [17, 18], two binary endpoints [19] and a mixture of time-to-event and nonfailure endpoints [20] have been developed. In this article, we develop a multi-arm multi-stage (MAMS) design for a trial with an efficacy endpoint and a safety endpoint. The novelty of the proposed design is that it is based on a joint model for the efficacy and safety outcomes, while information on both endpoints is incorporated into treatment selection decisions. We consider the situation where a minimal safety requirement is to be fulfilled and the treatment with the best combined safety and efficacy trade-off satisfying this constraint is selected for further testing. Selection is made at the first interim analysis, while the whole trial is allowed to compose of J analyses. Final decisions about a selected treatment are based on tests of efficacy and safety relative to control. In Section 2, we show that the design controls the familywise error rate (FWER) in the strong sense and discuss methods for sample size calculations. In Section 3, we illustrate the method through an example and simulations based on the Telmisartan and Insulin Resistance in HIV (TAILoR) study, a multi-arm trial of treatments to reduce insulin resistance in human immunodeficiency virus-positive patients. We conclude in Section 4 with a discussion of our findings and avenues for future research.

2. Statistical framework

We propose MAMS designs that begin in Stage 1 by comparing K active treatments with a common control group. The overall objective of the trial is to select the 'best' of the K treatments and then make comparisons with the control. Rather than be based solely on efficacy, treatment selection decisions will often reflect a compromise between the potential benefits and side effects of a new therapy. For example, a new treatment may need to demonstrate non-inferior safety and superior efficacy to represent a clinically meaningful advantage over a well-understood control. We propose designs that explicitly account for the impact of safety considerations on decision-making. Throughout, we restrict attention to the case where a single treatment is selected at the first analysis. We begin by focusing attention on a single-stage design and discuss the natural extension to multiple stages in Section 2.4.

2.1. Treatment selection rules

Suppose the trial proceeds in Stage 1 by measuring a bivariate endpoint on each patient. Labelling control as treatment 0, let Y_{Eik} and Y_{Sik} represent the efficacy and safety responses, respectively, of subject i on treatment k, which can be modelled as

$$\begin{pmatrix} Y_{Eik} \\ Y_{Sik} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{Ek} \\ \mu_{Sk} \end{pmatrix}, \begin{pmatrix} \sigma_E^2 & \rho\sigma_E\sigma_S \\ \rho\sigma_E\sigma_S & \sigma_S^2 \end{pmatrix}\right), \quad i = 1, \ldots, n; k = 0, 1, \ldots, K,$$

where ρ is the within-subject correlation, and we assume that the variance–covariance matrix of responses is known. Let $\theta_{Ek} = \mu_{Ek} - \mu_{E0}$ and $\theta_{Sk} = \mu_{Sk} - \mu_{S0}$ measure the advantage of treatment k over control for efficacy and safety, respectively, where we will assume that increases in response are desirable for both endpoints. Thus, $\boldsymbol{\theta} = (\boldsymbol{\theta}_E, \boldsymbol{\theta}_S)$ is a vector of length $2K$ containing the efficacy and safety effects of the K treatments. For each treatment $k = 1, \ldots, K$, we define two hypotheses $H_{Ek} : \theta_{Ek} \leqslant 0$ and $H_{Sk} : \theta_{Sk} \leqslant 0$. The null hypothesis we wish to test is $H_{0k} : H_{Ek} \cup H_{Sk}$ stating that treatment k is either ineffective or unsafe in comparison with control; rejecting H_{0k} implies that treatment k is both effective and safe. The global null hypothesis $H_0 : \bigcap_{k=1}^{K} H_{0k}$ represents the case that all K treatments are either unsafe or ineffective. For ease of presentation, we consider tests of superiority, although Jennison and Turnbull [18] observe that it is straightforward to accommodate tests of non-inferiority in this framework by subtracting the non-inferiority margin (for a difference in means) from patient responses on the control treatment.

For presentational purposes, we assume a common 1:1 allocation of patients to each of the K active treatments and control and denote the number of patient responses available on each arm by n. Thus, at the end of Stage 1, for each $k = 1, \ldots, K$, Fisher's information for θ_{Tk} takes a common value denoted by $\mathcal{I}_T = n/(2\sigma_T^2)$, for $T \in E, S$. In Appendix A.1 of the Supporting Information, we outline how the procedure could be extended to accommodate a common $r : 1$ allocation of patients to active treatments and control. Define $\hat{\mu}_{Tk}$ as the maximum likelihood estimator of μ_{Tk}. Accumulated data on each treatment are summarized by the bivariate score statistic

$$\begin{pmatrix} Z_{Ek} \\ Z_{Sk} \end{pmatrix} = \begin{pmatrix} \mathcal{I}_E\left(\hat{\mu}_{Ek} - \hat{\mu}_{E0}\right) \\ \mathcal{I}_S\left(\hat{\mu}_{Sk} - \hat{\mu}_{S0}\right) \end{pmatrix} \sim N\left(\begin{pmatrix} \mathcal{I}_E\theta_{Ek} \\ \mathcal{I}_S\theta_{Sk} \end{pmatrix}, \begin{pmatrix} \mathcal{I}_E & \rho\sqrt{(\mathcal{I}_E\mathcal{I}_S)} \\ \rho\sqrt{(\mathcal{I}_E\mathcal{I}_S)} & \mathcal{I}_S \end{pmatrix}\right). \tag{1}$$

Only treatments meeting a pre-specified minimum safety requirement may be considered for selection. Let N_S denote the number of treatments eligible for selection, which are indexed by the selection set $\mathcal{SS} = \{k : Z_{Sk} > c\}$. If $N_S = 0$, the test is stopped for futility without rejecting H_0. Otherwise, we select from $\mathcal{SS}$ the treatment maximizing the objective function

$$O_k = \frac{w_E Z_{Ek}}{\sqrt{\mathcal{I}_E}} + \frac{w_S Z_{Sk}}{\sqrt{\mathcal{I}_S}}, \tag{2}$$

where w_E and w_S are pre-specified non-negative weights satisfying $w_E^2 + w_S^2 = 1$. Unplanned deviations from the pre-specified treatment selection rule could lead to inflation of the FWER above the nominal level. One of the motivations of this design is, however, to formally include safety in the decision-making so that such deviations become less frequent. Should unexpected modification be necessary, however, conditional error principle [21] can be used to maintain FWER control. It is worth pointing out that we incorporate the safety threshold because the objective function allows good efficacy to compensate for poor safety. In practice, this would only be acceptable up to a certain point, which is defined by the safety threshold. A natural choice for this threshold in our opinion is $c = 0$, that is, we only select from treatments with comparable or better safety than control in stage 1, although in principle other values could be used instead.

Let $i^\star$ index the treatment selected from Stage 1 on the basis of the objective function $O_{i^\star} = \max_{k \in \mathcal{SS}}\{O_k\}$. Because the selected treatment will only be declared preferable to control if we can reject the null hypothesis $H_{0i^\star} : H_{Ei^\star} \cup H_{Si^\star}$, a natural choice of weights is $w_E = w_S = \sqrt{0.5}$ as this ensures consistency between selection decisions and the final analysis of the trial.

We propose single-stage tests of $H_{01}, \ldots, H_{0K}$ with stopping rules of the form:

$$\begin{array}{ll} \text{If } Z_{S1}, \ldots, Z_{SK} \leqslant c & \text{Stop and accept } H_0 \\ \text{Otherwise} & \text{Select from } SS \text{ treatment } i^\star \text{ maximizing objective function } O \\ & \text{and conduct the final analysis.} \\ \text{At the final analysis:} & \\ \text{If } Z_{Ei^\star} \geqslant u_E \text{ and } Z_{Si^\star} \geqslant u_S & \text{Stop and reject } H_0 \text{ in favour of } H_{1i^\star} : \{\theta_{Ei^\star} > 0\} \cap \{\theta_{Si^\star} > 0\}, \\ \text{Otherwise} & \text{Stop and accept } H_0. \end{array} \tag{3}$$

At the final analysis of the proposed test, superiority can only be claimed for the selected treatment $i^\star$. Consequently, we define the FWER of the procedure as $\mathbb{P}\{\text{Reject } H_0 \text{ in favour of a false } H_{1i^\star}; \boldsymbol{\theta}\}$. This probability depends on both the minimum safety requirement, c, and the stopping boundaries (u_E, u_S). Our approach is to fix $c = 0$ and find the pair of critical values maintaining strong control of the FWER at level α. This criterion stipulates that $\mathbb{P}\{\text{Reject } H_0 \text{ in favour of a false } H_{1i^\star}; \boldsymbol{\theta}\} \leqslant \alpha$ for all configurations of $\boldsymbol{\theta}$ with at least one $\theta_{Tk} \leqslant 0$, for $T \in \{E, S\}$ and $k \in \{1, \ldots, K\}$. If $H_{01}, \ldots, H_{0K}$ are all true, a familywise error is made if the test terminates with rejection of $H_{0i^\star}$ whatever treatment is selected, and the FWER is given by $\mathbb{P}\{Z_{Ei^\star} \geqslant u_E, Z_{Si^\star} \geqslant u_S, N_S \geqslant 1; \boldsymbol{\theta}\}$. In the remainder of this section, we discuss how to find (u_S, u_E) maintaining strong control of the FWER.

2.2. *Specification of test boundaries*

We propose choosing the boundaries of test (3) to ensure the FWER is controlled at level α as we approach the following two 'worst-case' limiting configurations of θ: (1) $\boldsymbol{\theta}_E = (\infty, \ldots, \infty)$, $\boldsymbol{\theta}_S = (0, \ldots, 0)$ and (2) $\boldsymbol{\theta}_E = (0, \ldots, 0)$, $\boldsymbol{\theta}_S = (\infty, \ldots, \infty)$. We claim that specifying test boundaries according to this criterion ensures strong control of the FWER and prove this claim using a combination of analytical arguments and simulation. This result agrees with the findings of [18] for the case that $K = 1$. Let Γ represent a set indexing treatments with positive efficacy and safety effects. We begin considering a subset of the null parameter space comprising configurations of $\boldsymbol{\theta}$ such that

(a) For all treatments k with $k \notin \Gamma$, $\theta_{Ek} \leqslant 0$ and $\theta_{Sk} \leqslant 0$; or
(b) For all treatments k with $k \notin \Gamma$, $\theta_{Ek} \geqslant 0$ and $\theta_{Sk} \leqslant 0$; or
(c) For all treatments k with $k \notin \Gamma$, $\theta_{Ek} \leqslant 0$ and $\theta_{Sk} \geqslant 0$.

Under the global null hypothesis, $\Gamma = \emptyset$, and the constraints on $\boldsymbol{\theta}$ configurations defined previously correspond to assuming that effects of different treatments on the same endpoint have the same sign. We claim that the FWER of procedure (3) under configurations of $\boldsymbol{\theta}$ in this restricted global null parameter space is maximized under constellations with $\boldsymbol{\theta}_E = (\gamma_E, \ldots, \gamma_E)$ and $\boldsymbol{\theta}_S = (\gamma_S, \ldots, \gamma_S)$, and furthermore that local maxima of the FWER are attained in the limit as $\gamma_S \to \infty$ and $\gamma_E = 0$, and in the limit as $\gamma_E \to \infty$ and $\gamma_S = 0$.

To prove these claims, we begin by assuming that all treatments are always eligible for selection and consider the configuration of $\boldsymbol{\theta}$ with $\boldsymbol{\theta}_E = (\gamma_E, \ldots, \gamma_E)$ and $\boldsymbol{\theta}_S = (\gamma_S, \ldots, \gamma_S)$. Then, letting some elements of $\boldsymbol{\theta}_S$ fall below γ_S decrease stochastically the distribution of $(Z_{Ei^\star}, Z_{Si^\star})$ as both statistics tend to take lower values on average. To explain this, note that since all treatments remain competitive for efficacy, treatments must perform well for Z_S if they are to rank highly for the objective function O. Thus, selection decisions are, in effect, driven primarily by safety data so that a treatment may beat its competitors on the basis of O with lower values of Z_E. Letting some of the θ_{Sk}s drop below γ_S also decreases stochastically the distribution of $Z_{Si^\star}$ too: the treatment associated with the largest element of $\boldsymbol{\theta}_S$ is now 'safest' by some margin, meaning that on average, lower values of Z_S will be sufficient for it to beat the weaker competition to ensure selection. Similar arguments imply keeping $\boldsymbol{\theta}_S$ fixed at $\boldsymbol{\gamma}_S$ and letting some elements of $\boldsymbol{\theta}_E$ fall below γ_E decreases stochastically the distributions of $Z_{Ei^\star}$ and $Z_{Si^\star}$. On the other hand, simultaneously forcing elements of $\boldsymbol{\theta}_E$ below γ_E and elements of $\boldsymbol{\theta}_S$ below γ_S decreases stochastically the distribution of $(Z_{Ei^\star}, Z_{Si^\star})$: systematic differences between treatments imply that it is possible for a good safety profile to compensate for poor efficacy (and vice versa) resulting in lower average values of Z_E and Z_S for the selected treatment.

Letting $\boldsymbol{\theta}_E = (\gamma_E, \ldots, \gamma_E)$ and $\boldsymbol{\theta}_S = (\gamma_S, \ldots, \gamma_S)$, increasing γ_E or γ_S increases the probability of rejecting H_0. Thus, looking across the restricted global null parameter space, the probability of making a familywise error is maximized at the boundaries of the space, that is, in the limit as $\gamma_S \to \infty$ and $\gamma_E = 0$,

and in the limit as $\gamma_E \to \infty$ and $\gamma_S = 0$. If for some treatment k, θ_{Ek} and θ_{Sk} are both positive so that $\Gamma \neq \emptyset$, this treatment will be more likely to be selected, in which case we cannot commit a family-wise error and the FWER decreases. Therefore, controlling the FWER for all configurations of $\boldsymbol{\theta}$ in the restricted global null parameter space ensures the FWER is controlled over the wider null parameter space defined previously.

So far, we have consider the case that all treatments are always eligible for selection, in effect setting $c = -\infty$. However, we claim that for general values of c, local maxima of the FWER are attained in the limit as we approach the worst case configurations of $\boldsymbol{\theta}$ identified previously. This is because under this requirement, the expected size of SS is determined by γ_S. Therefore, increasing γ_S increases stochastically the distribution of $(Z_{Ei^\star}, Z_{Si^\star})$ as the average number of treatments from which we can select increases. In particular, as γ_S approaches ∞, SS includes all K treatments almost surely and rejects H_0 if $Z_{Ei^\star} > u_E$. The probability of falsely rejecting H_0 is then maximized for $\gamma_E = 0$. Similarly, setting $\gamma_S = 0$, the FWER reaches a second local maximum as $\gamma_E \to \infty$. To see this note that for $\gamma_S = 0$, inclusion of treatments in the selection set is random so that the probability of rejection is maximized for maximal effect on efficacy.

To complete our justification for designing test (3) to control the FWER under the 'worst-case' limiting configurations of $\boldsymbol{\theta}$, we go beyond the arguments stated previously to claim that this approach ensures strong control of the FWER. In particular, tests defined in this way will control the FWER for any configuration of $\boldsymbol{\theta}$ with $\boldsymbol{\theta}_E = (\theta_{E1}, \ldots, \theta_{EK})$ and $\boldsymbol{\theta}_S = (\theta_{S1}, \ldots, \theta_{SK})$, where one $\theta_{\cdot i}$ is zero and the other ∞. While we cannot prove these claims analytically, we evaluate them via simulation in Section 3.2. Assuming for now that these claims do hold, it is appropriate to choose boundaries (u_E, u_S) to ensure that

$$\lim_{\gamma_S \to \infty} \mathbb{P}\{Z_{Ei^\star} \geqslant u_E, Z_{Si^\star} \geqslant u_S \mid N_S \geqslant 1; \boldsymbol{\theta}_E = (0, \ldots, 0), \boldsymbol{\theta}_S = (\gamma_S, \ldots, \gamma_S)\} = \alpha, \tag{4}$$

$$\lim_{\gamma_E \to \infty} \mathbb{P}\{Z_{Ei^\star} \geqslant u_E, Z_{Si^\star} \geqslant u_S \mid N_S \geqslant 1; \boldsymbol{\theta}_E = (\gamma_E, \ldots, \gamma_E), \boldsymbol{\theta}_S = \mathbf{0}\} = \alpha / \mathbb{P}\{N_S \geqslant 1; \boldsymbol{\theta}_S = \mathbf{0}\}, \tag{5}$$

because $\lim_{\gamma_S \to \infty} \mathbb{P}\{N_S \geqslant 1; \boldsymbol{\theta}_S = (\gamma_S, \ldots, \gamma_S)\} = 1$ and the probability that at least one treatment meets the minimum safety criterion does not depend on $\boldsymbol{\theta}_E$. As γ_S and γ_E approach ∞, the bivariate probabilities on the left hand sides (LHSs) of (4) and (5) converge to univariate probabilities. We find (u_E, u_S) so that the limits of these marginal rejection probabilities are equal to the values required to ensure FWER control. Limits of rejection probabilities are found by integrating the limits of the marginal conditional densities of $Z_{Ei^\star}$ and $Z_{Si^\star}$ derived in Appendix A.2 of the Supporting Information. It is important to note that these marginal densities depend on the correlation coefficient ρ. So far, we have assumed that this parameter is known. In Section 3.3, we explore the robustness of attained FWERs to misspecification of ρ. Marginal densities of test statistics depend on variances only through the information levels $\mathcal{I}_E$ and $\mathcal{I}_S$. The effect of assuming a known variance has previously been investigated in similar settings [22], and the quantile substitution approach described in [8] has been shown to work well.

2.3. *Sample size calculations*

We wish to calculate the sample size needed for test (3) to attain a disjunctive power, that is, probability of rejecting at least one false null hypothesis [5, 23], of $1 - \beta$ under the configuration of $\boldsymbol{\theta}$ with $\boldsymbol{\theta}_E = (\delta_0, \ldots, \delta_0, \delta)$ and $\boldsymbol{\theta}_S = (\gamma_S, \ldots, \gamma_S)$ with $\gamma_S > 0$. We may approximate further by letting $\gamma_S \to \infty$, which can be justified by the belief that a potentially unsafe treatment is unlikely to be included in the trial. In this case, a test's power simplifies to

$$\begin{aligned} &\lim_{\gamma_S \to \infty} \mathbb{P}\{Z_{Ei^\star} \geqslant u_E, Z_{Si^\star} \geqslant u_S \mid N_S \geqslant 1; \boldsymbol{\theta}_E = (\delta_0, \ldots, \delta_0, \delta), \boldsymbol{\theta}_S = (\gamma_S, \ldots, \gamma_S)\} \\ &\quad = \lim_{\gamma_S \to \infty} \mathbb{P}\{Z_{Ei^\star} \geqslant u_E \mid N_S \geqslant 1; \boldsymbol{\theta}_E = (\delta_0, \ldots, \delta_0, \delta), \boldsymbol{\theta}_S = (\gamma_S, \ldots, \gamma_S)\}, \end{aligned} \tag{6}$$

and limiting probabilities are found by integrating the limits of the marginal conditional density of $Z_{Ei^\star}$. Using the results of Appendix A.1 and following the workings of Appendix A.2.1 of the Supporting Information, we can show that limiting rejection probability (6) is given by

$$
\begin{aligned}
(K-1)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} & f_{X_{E0}|X_{S0}}\left(x_1-\delta_0\mathcal{I}_E \mid y\right) f_{X_{S0}}(y) f_{X_{S0}}(y-m)\mathbb{P}\left\{U\leqslant \ell_4\right\}^{K-2}\mathbb{P}\left\{U\leqslant \ell_5\right\} \\
& \times\Phi\left(\frac{x_1-\sqrt{\mathcal{I}_E/\mathcal{I}_S}\rho(y-m)-u_E}{\sqrt{(\mathcal{I}_E/2)(1-\rho^2)}}\right)\mathrm{d}x_1\,\mathrm{d}y\,\mathrm{d}m \\
+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} & f_{X_{E0}|X_{S0}}\left(x_1-\delta\mathcal{I}_E \mid y\right) f_{X_{S0}}(y) f_{X_{S0}}(y-m)\mathbb{P}\left\{U\leqslant \ell_4\right\}^{K-1} \\
& \times\Phi\left(\frac{x_1-\sqrt{\mathcal{I}_E/\mathcal{I}_S}\rho(y-m)-u_E}{\sqrt{\left(\mathcal{I}_E/2\right)\left(1-\rho^2\right)}}\right)\mathrm{d}x_1\,\mathrm{d}y\,\mathrm{d}m,
\end{aligned}
\tag{7}
$$

where $U\sim N(0,\, w_E w_S\rho+0.5)$, $\sqrt{(\mathcal{I}_E\mathcal{I}_S)}\ell_4=\sqrt{\mathcal{I}_E}w_S y+\sqrt{\mathcal{I}_S}w_E(x_1-\mathcal{I}_E\delta_0)$ and $\sqrt{(\mathcal{I}_E\mathcal{I}_S)}\ell_5=\sqrt{\mathcal{I}_E}w_S y+\sqrt{\mathcal{I}_S}w_E(x_1-\mathcal{I}_E\delta)$. For computational convenience, we proceed assuming that $\mathcal{I}_E=\mathcal{I}_S=\mathcal{I}_1$ and conduct a one-dimensional search to find the common information level $\mathcal{I}_1^\star$ for which rejection probability (7) equals $1-\beta$; at each iteration of this search, boundaries for monitoring score statistics $Z_{Ei^\star}$ and $Z_{Si^\star}$ are updated to ensure strong control of the FWER at level α under the proposed information level. Because information level $\mathcal{I}_1^\star$ typically corresponds to requiring fractions of subjects, in practice, we propose rounding up the total sample size to $n^\star=2\max\{\sigma_E^2\mathcal{I}_1^\star,\,\sigma_S^2\mathcal{I}_1^\star\}$ patients per treatment arm. The test is then applied with critical values calculated for information levels $\mathcal{I}_E=n^\star/(2\sigma_E^2)$ and $\mathcal{I}_S=n^\star/(2\sigma_S^2)$. If a procedure's power is monotone increasing in $\mathcal{I}_E$ and $\mathcal{I}_S$, this sample size criterion will be conservative in the sense that attained power will exceed $1-\beta$. In Section 3.2, we use simulation to evaluate properties of tests designed according to the proposed sample size criterion.

2.4. *Beyond single-stage designs*

It is straightforward to extend our approach to find designs maintaining control of the FWER when multiple interim analyses are planned. Let $Z_{Tk,j}$ denote the score statistic at interim analysis j for endpoint T on treatment k. A multi-stage test of $H_{01},\ldots,H_{0K}$ has a stopping rule of the form:

At the end of stage 1:	
If $Z_{S1,1},\ldots,Z_{SK,1}\leqslant c$	Stop and accept H_0
Otherwise	Select from $\mathcal{SS}$ treatment $i^\star$ maximizing objective function O and conduct interim analysis 1.
At interim analysis $j=1,\ldots,J$:	
If $Z_{Ei^\star,j}\geqslant u_{Ej}$ and $Z_{Si^\star}\geqslant u_{Sj}$	Stop and reject H_0 in favour of $H_{1i^\star}:\{\theta_{Ei^\star}>0\}\cap\{\theta_{Si^\star}>0\}$,
If $Z_{Ei^\star,j}\leqslant l_{Ej}$ or $Z_{Si^\star,j}\leqslant l_{Sj}$	Stop and accept H_0,
Otherwise	Continue to interim analysis $j+1$.

(8)

Multi-stage tests are defined with binding futility rules so that if either $Z_{Ei^\star,j}\leqslant l_{Ej}$ or $Z_{Si^\star,j}\leqslant l_{Sj}$, the procedure must stop immediately at interim analysis j without declaring treatment $i^\star$ safe and effective. Criteria (4) and (5) imply that we can uncouple the searches needed to find critical values for monitoring efficacy and safety score statistics. Furthermore, for $T\in\{E,S\}$, increments $Z_{Ti^\star,2}-Z_{Ti^\star,1},\ldots,Z_{Ti^\star,J}-Z_{Ti^\star,J-1}$ are independent and follow the same distribution as increments in score statistics generated by a univariate group sequential test (GST) without selection [3]. Thus, we can find $(l_{E1},u_{E1}),\ldots,(l_{EJ},u_{EJ})$ as the boundaries defining a one-sided univariate GST monitoring $\{Z_{Ei^\star,1},\ldots,Z_{Ei^\star,J}\}$ with limiting conditional type I error rate α given $N_S\geqslant 1$ under $\gamma_E=0$ and letting $gamma_S\to\infty$. Following [3], we propose that an alpha-spending approach [24] be used to find the upper and lower boundaries at each stage $j=1,\ldots,J$ satisfying

$$
\begin{aligned}
\lim_{\gamma_S\to\infty}\mathbb{P}\Big\{ & Z_{Ei^\star,1}\in\left(l_{E1},\,u_{E1}\right),\ldots,Z_{Ei^\star,j-1}\in\left(l_{E(j-1)},u_{E(j-1)}\right),Z_{Ei^\star,j}\geqslant u_{Ej}\mid N_S\geqslant 1;\boldsymbol{\theta}_E=\mathbf{0},\boldsymbol{\theta}_S \\
& =\left(\gamma_S,\ldots,\gamma_S\right)\Big\}=f_U\left(t_j\right)-f_U\left(t_{j-1}\right) \\
\lim_{\gamma_S\to\infty}\mathbb{P}\Big\{ & Z_{Ei^\star,1}\in\left(l_{E1},\,u_{E1}\right),\ldots,Z_{Ei^\star,j-1}\in\left(l_{E(j-1)},u_{E(j-1)}\right),Z_{Ei^\star,j}\leqslant l_{Ej}\mid N_S\geqslant 1;\boldsymbol{\theta}_E \\
& =\mathbf{0},\boldsymbol{\theta}_S=\left(\gamma_S,\ldots,\gamma_S\right)\Big\}=f_L\left(t_j\right)-f_L\left(t_{j-1}\right),
\end{aligned}
$$

where t_j is the fraction of $\mathcal{I}_{EJ}$, the maximum information level for the efficacy treatment effect, accumulated by stage j, and f_U and f_L are monotone increasing functions satisfying $f_U(0) = f_L(0) = 0$ and, for $t \geqslant 1$, $f_U(t) = \alpha$ and $f_L(t) = 1 - \alpha$. A similar process can be used to find the boundaries $(l_{S1}, u_{S1}), \ldots, (l_{SJ}, u_{SJ})$ for monitoring $\{Z_{Si^\star,1}, \ldots, Z_{Si^\star,J}\}$. Safety boundaries are determined using f_L and f_U to spend error probabilities as a function of the observed information for $\theta_{E,i^\star}$; this ensures that $u_{EJ} = l_{EJ}$ and $u_{SJ} = l_{SJ}$, so that procedure (8) terminates properly at analysis J with a final hypothesis decision for any choice of $\mathcal{I}_{EJ}$ even when the variances of the efficacy and safety endpoints differ. To find the required sample size, we follow Section 2.3 and search for the maximum information level $\mathcal{I}_{EJ}$ for which the test has power $1 - \beta$ according to criterion (6) under an anticipated information sequence $\mathcal{I}_{E1}, \ldots, \mathcal{I}_{EJ}$, setting each $\mathcal{I}_{Sj} = \sigma_E^2 \mathcal{I}_{Ej}/\sigma_S^2$ to account for differences between the rates at which information on safety and efficacy effects accumulate. The test will recruit up to $n^\star = 2\sigma_E^2 \mathcal{I}_{EJ}$ patients on the selected treatment and control in the absence of early stopping.

3. Example

In this section, we will examine the operating characteristics of the proposed designs through a series of examples motivated by the TAILoR study, a MAMS trial comparing several doses of telmisartan with control for the reduction of insulin resistance in human immunodeficiency virus-positive patients receiving combination antiretroviral therapy [6]. The study, which is currently ongoing, uses the change in the Homeostatic model assessment - Insulin resistance (HOMA-IR) index between baseline and 24 weeks as the efficacy endpoint.

In this section, we imagine how the TAILoR study might have been designed as a single-stage procedure of the form shown in (3), using the methodology described in this paper to incorporate a safety endpoint in addition to the efficacy endpoint used in the ongoing study. A plausible safety endpoint is change in systolic blood pressure from baseline because telmisartan is licensed for the treatment of hypertension. An excessive drop in blood pressure for patients without hypertension would be considered an undesirable safety risk. With the exception of this modification, we will assume the design parameters of the original TAILoR study. We therefore stipulate an FWER of 0.05 and seek designs randomizing patients equally across treatment arms with power 0.9 to correctly reject one false null hypothesis. When the TAILoR study was first designed, four doses of telmisartan were planned. For consistency with previous publications [22, 25], we consider the scenario that $K = 4$ active treatments are to be compared with control, despite the ongoing study using three doses because of last minute changes to the study. The standardized desirable effect for efficacy, δ, used for sample size calculations is set as 0.545, and the minimum clinically important difference is defined as 0.178. Under the assumption that all treatments are truly safe, we do not require specification of an effect on safety when using (7) for sample size calculations. However, if such an assumption is undesirable, Equation (6) in Appendix A.1 of the Supporting Information can be used with the anticipated safety effect. Boundary calculations and sample size determinations require us to numerically evaluate multi-dimensional integrals. For this purpose, we used the R package `cubature` [26] and verified solutions for the obtained boundaries using 100 000-fold simulations.

3.1. Design options

Figure 1 shows how the required information per arm and safety/efficacy stopping boundaries vary as the weight w_E changes. The within-subject correlation of efficacy and safety responses is assumed to be 0.4. The information required is largest when selection of the treatment is based only on the safety endpoint ($w_E = 0$), while it decreases as the weight on efficacy increases. Similarly, both the efficacy and safety boundaries decrease as the required information decreases, as expected. There is, however, an apparent difference between the efficacy and safety boundary, depending on the weight given to each endpoint. For small weights on efficacy, the efficacy boundary is smaller than the safety boundary, while this pattern reverses once more weight is attributed to efficacy for selection. For equal weights, the boundaries for efficacy and safety are identical.

3.2. Error rates

In this section, we illustrate properties of tests of the form (3) designed and conducted with equal weights $w_E = w_S$ and correlation coefficient $\rho = 0.4$. Under this setting, the information required per arm is $\mathcal{I}_1^\star = 47.148$ ($n = 94.296$), and the stopping boundaries are $u_E = u_S = 14.466$. Empirical error rates based on 10 000 simulation runs for each point on a grid of parameters are shown in Figure 2 for

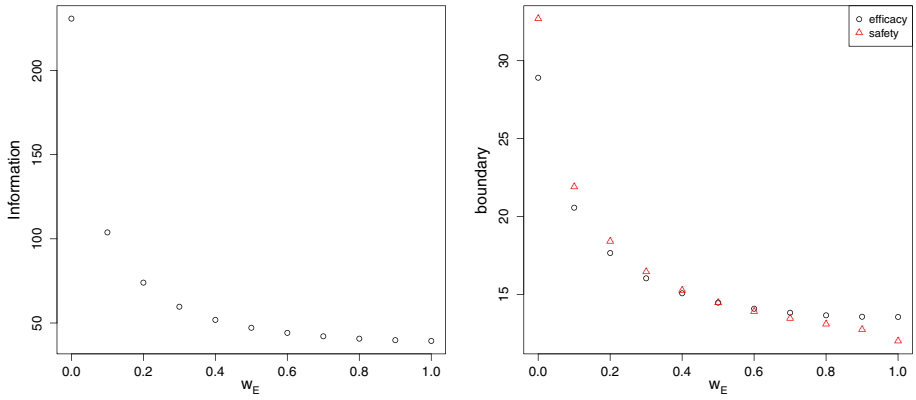

Figure 1. Information per arm, $\mathcal{I}_1^\star$ (left), and safety/efficacy stopping boundary (right) needed for tests of the form (3) to maintain strong control of the FWER at level 0.05 and attain limiting power 0.9 when $\boldsymbol{\theta}_E = (0.178, 0.178, 0.178, 0.545)$ and all treatments are safe. Designs are found for tests making treatment selection decisions according to objective function (2) assuming $\sigma_E = \sigma_S = 1$ and $\rho = 0.4$.

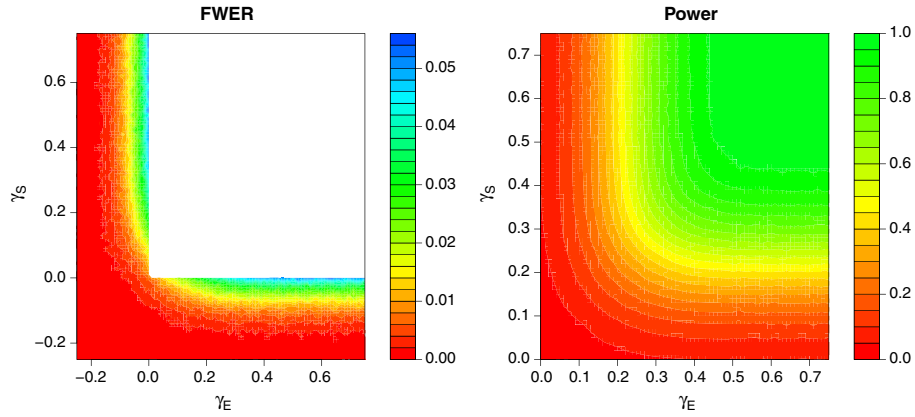

Figure 2. Empirical FWERs over the null space (left) and power to correctly reject at least one false null hypothesis (right) of tests of the form (3) designed to maintain strong control of the FWER at level 0.05 and to attain power 0.9 under $\boldsymbol{\theta}_E = (0.178, 0.178, 0.178, 0.545)$ (when all treatments are safe). Tests are designed and conducted with $K = 4$, $w_E^2 = w_S^2 = 0.5$, $\rho = 0.4$ and $\sigma_E = \sigma_S = 1$. FWER and power are evaluated under configurations of $\boldsymbol{\theta}$ with $\boldsymbol{\theta}_S = (\gamma_S, \ldots, \gamma_S)$ and $\boldsymbol{\theta}_E = (\gamma_E, \ldots, \gamma_E)$. Results are based on 10 000 simulations for each parameter configuration.

cases where all treatments have the same pair of effects (γ_E, γ_S) versus control. The left-hand panel clearly shows that the FWER of the design is maximized if one of the effects is at the boundary of the null space and the other is large. As expected, the power of the design increases as at least one of the effects increases.

Figure 3 provides empirical FWERs for parameter configurations of the from $\boldsymbol{\theta_E} = (\theta_{E1}, \ldots, \theta_{EK})$, $\boldsymbol{\theta_S} = (\theta_{S1}, \ldots, \theta_{SK})$, where one parameter of each pair $(\theta_{Ei}, \theta_{Si})$ is large and the other zero to evaluate the conjecture made in Section 2.2, which designs will control these at the nominal level α. For the purpose of this evaluation, the large effect was set to 1 million, and 100 000-fold simulations are used. From the graph, it can be seen that the FWER is well controlled for any parameter configuration, as conjectured.

Figure 4 shows how the power of the procedure changes as the safety of the experimental treatments changes. Results are presented for one to four treatments exhibiting the desired effect for efficacy of 0.545, while the remaining have the minimum clinically important effect of 0.178. Power increases as the safety of the treatments increases and reaches the desired level of 0.9 for a safety effect of around 0.5. Power also increases as the number of treatments with the desired efficacy increases, although this increase diminishes somewhat with the number of efficacious treatments.

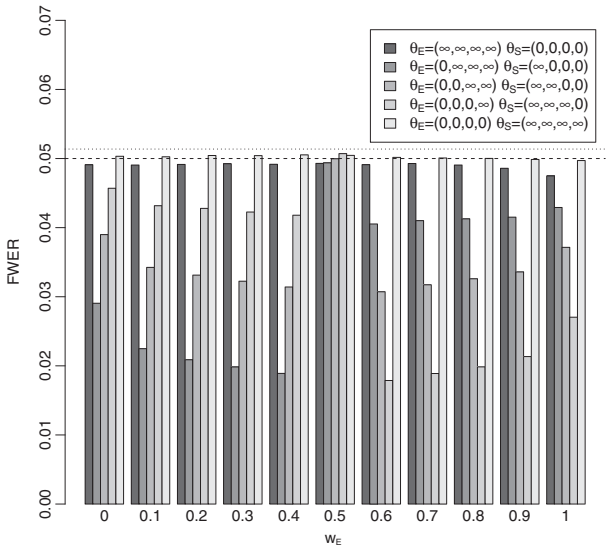

Figure 3. Empirical FWER for different configurations of the true effects for tests of the form (3) designed to maintain strong control of the FWER at level 0.05. Tests are designed and conducted with $K = 4$, $\rho = 0.4$, $w_E^2 = w_S^2 = 0.5$ and $\sigma_E = \sigma_S = 1$. Results are based on 100 000 simulations for each parameter configuration. The dashed horizontal line corresponds to the nominal FWER and the dotted line to the upper bound for simulation error.

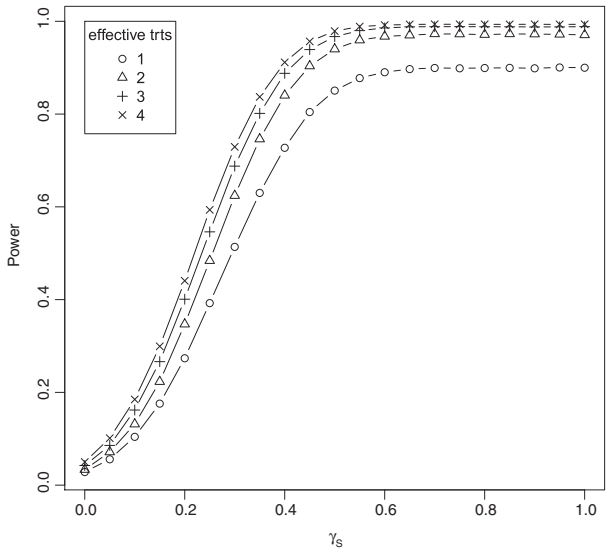

Figure 4. Empirical power to correctly reject at least one false null hypothesis of tests of the form (3) designed to maintain strong control of the FWER at level 0.05 and attain power 0.9 under $\boldsymbol{\theta}_E = (0.178, 0.178, 0.178, 0.545)$ (when all treatments are safe). Tests are designed and conducted with $K = 4$, $w_E^2 = w_S^2 = 0.5$, $\rho = 0.4$ and $\sigma_E = \sigma_S = 1$. Power is evaluated under configurations of $\boldsymbol{\theta}$ with $\boldsymbol{\theta}_S = (\gamma_S, \ldots, \gamma_S)$ and $\boldsymbol{\theta}_E = (\delta_0, \ldots, \delta_0, \delta, \ldots, \delta)$ such that j treatments have the desired effect for efficacy ($\delta = 0.545$) and the remaining have the minimum clinically important effect ($\delta_0 = 0.178$). Results are based on 100 000 simulations for each parameter configuration.

3.3. Misspecification of ρ

When specifying our model, we have so far assumed that response variances and their correlation are known. In this section, we will investigate the robustness of our design to the assumption of known correlation. Figure 5 shows the simulated FWER based on 100 000 simulations of tests as the correlation between endpoints varies. Six different true parameter constellations are considered, namely, the global null hypothesis and five 'worst-case' configurations (once again using 1 million instead of infinity for simulation purposes). For all six settings, the FWER is controlled at the design value $\rho = 0.4$, and the

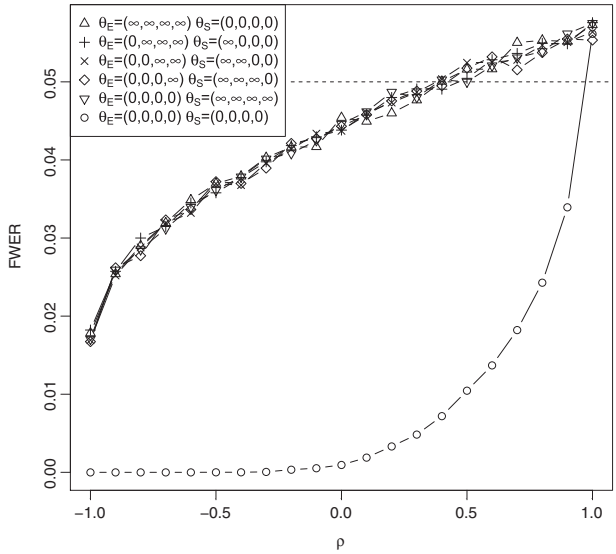

Figure 5. Empirical FWERs of tests of the form (3) designed to maintain strong control of the FWER at level 0.05 and attain power 0.9 under $\boldsymbol{\theta}_E = (0.178, 0.178, 0.178, 0.545)$ (when all treatments are safe). Tests are designed and conducted with $K = 4$, $w_E^2 = w_S^2 = 0.5$ and $\sigma_E = \sigma_S = 1$. Tests are also designed assuming $\rho = 0.4$ but conducted for a range of correlations. Empirical error probabilities are evaluated under configurations of $\boldsymbol{\theta}$ with $\boldsymbol{\theta}_S = (\gamma_S, \ldots, \gamma_S)$ and $\boldsymbol{\theta}_E = (\gamma_E, \ldots, \gamma_E)$. Results are based on 100 000 simulations for each scenario.

procedure is conservative for all correlations below this. Under the global null hypothesis, only perfect correlation results in an inflation of the FWER, while it is inflated once the true correlation is above the design value for the worst-case configurations. The maximum inflation, achieved under perfect positive correlation, is, however, small at 10% of the nominal value of the FWER.

Tamhane *et al.* [27] observe that typically either the correlation is assumed to be known (as performed here) or a correlation of one is treated as the worst-case scenario. Given the relative conservatism of the proposed procedure at reasonable values of the parameters, we believe the former is sufficient, although the latter would certainly also be possible. A more elegant solution given in [27] overcomes this problem by estimating the correlation mid-study and uses an approach due to Berger & Boos [28] to obtain an upper bound for the FWER accounting for the sampling error of the sample correlation coefficient.

4. Discussion

In this paper, we have presented an approach for designing MAMS studies based on a joint model for efficacy and safety data, which considers both endpoints when selecting the most promising treatment for further investigation and tests the efficacy and safety of the selected treatment relative to control. The main challenge with obtaining the relevant distributions of the test statistics arose from the requirement to select from treatments satisfying a minimum safety requirement. We have shown that the FWER is strongly controlled under the assumption that effects of different treatments for the same endpoint have the same sign. Our simulation results show, however, that strong control of the FWER also appears to hold when this assumption is not made.

In the presentation and derivations, we have made a number of assumptions that may not be appropriate for specific settings. For example, single-stage designs are formulated assuming patient responses follow a bivariate normal distribution with a common correlation between efficacy and safety responses across the active and control treatments. In addition, calculations assume that at the end of Stage 1, there is a common information level for $\theta_{E,1}, \ldots, \theta_{E,K}$ and a common information level for $\theta_{S,1}, \ldots, \theta_{S,K}$. This joint distribution will not in general apply if data do not follow a normal distribution because information levels and correlation coefficients may depend on unknown parameters, such as response rates in the case of binary data (see section 9 of [27]). One potential solution would be to approximate and derive test boundaries setting the correlation coefficient and information levels to the values that would apply under $\boldsymbol{\theta}_E = \boldsymbol{\theta}_S = \mathbf{0}$. However, further simulations would be needed to verify whether this approach would maintain strong control of the FWER at a level close to the nominal value.

Another simplifying assumption we have made is to propose designs setting the safety threshold to be zero, so that only treatments with better safety than control can be selected. A simple shift of the safety test statistic can be used to allow for different thresholds to be used. Similarly, it may be desirable to select treatments only based on efficacy provided that the treatment is safe enough. Simply setting the weight on safety within the objective function to zero can accomodate this situation. Finally, as outlined before, it may not be appropriate to test for superiority in terms of safety over control. Shifting the respective test statistics for safety will allow non-inferiority hypothesis to be used instead. It is also easy to envisage application of this design in other settings, such as mental health trials, where there are co-primary efficacy endpoints. In these cases, no minimal threshold would be applied to either efficacy endpoint – the ideas of this work apply to this, somewhat simpler, situation setting the threshold $c = -\infty$.

One great benefit of multi-stage clinical trials is their reduced expected sample size compared with single-stage designs. Such gains can, however, only be realized, if the primary endpoint is observed quickly relative to the recruitment time [29]. When this is not the case, it would be of interest to investigate whether methods such as the one described in [30] can be extended to make selection decisions based on intermediate endpoints in the setting discussed in this paper.

Another area for further work regards how to calculate confidence intervals on termination of tests of the form (3). The procedure based around hypothesis testing described here allows almost formulaic decisions about the superiority of experimental treatments over control. It is essential, however, that confidence intervals should also be available to inform decision makers about the probable sizes of any efficacy and safety benefits, in order to give a complete description of the evidence supporting a selected treatment. A future work will be necessary to evaluate if related work [14, 31] can be utilized to obtain interval estimates as well.

Acknowledgements

This work is an independent research arising from Dr Jaki's Career Development Fellowship (NIHR-CDF-2010-03-32) supported by the National Institute for Health Research. Funding for this work was also provided by the Medical Research Council (MR/J004979/1 and MR/J014079/1). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health. Both authors have made equal contributions to this manuscript.

References

1. Arrowsmith J. Trial watch: phase III and submission failures: 2007-2010. *Nature Reviews Drug Discovery* 2011; **10**:87.
2. EFPIA. *The pharmaceutical industry in figures*, European Federation of Pharmaceutical Industries and Associations, 2012.
3. Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine* 2003; **22**:689–703.
4. Stallard N, Friede T. A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine* 2008; **27**:6209–6227.
5. Bretz F, Koenig F, Brannath W, Glimm E, Posch M. Adaptive designs for confirmatory clinical trials. *Statistics in Medicine* 2009; **28**(8):1181–1217.
6. Magirr D, Jaki T, Whitehead J. A generalised Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika* 2012; **99**(2):494–501.
7. Magirr D, Stallard N, Jaki T. Flexible sequential designs for multi-arm clinical trials. *Statistics in Medicine* 2014; **33**(19):3269–3279.
8. Jennison C, Turnbull BW. *Group sequential methods with applications to clinical trials*. Chapman and Hall: Boca Raton, 2000.
9. Whitehead J. *The design and analysis of sequential clinical trials*. Wiley: Chichester, 1997.
10. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 1999; **18**:1833–1848.
11. Kelly PJ, Stallard N, Todd S. An adaptive group sequential design for phase II/III clinical trials that select a single treatment from several. *Journal of Biopharmaceutical Statistics* 2005; **15**:641–658.
12. Thall PF, Simon R, Ellenberg SS. Two-stage selection and testing designs for comparative clinical trials. *Biometrika* 1988; **75**:303–310.
13. Whitehead J, Jaki T. One- and two-stage design proposals for a phase II trial comparing three active treatments with a control using an ordered categorical endpoint. *Statistics in Medicine* 2009; **28**:828–847.
14. Jaki T, Magirr D. Considerations on covariates and endpoints in multi-arm multi-stage clinical trials selecting all promising treatments. *Statistics in Medicine* 2013; **32**(7):1150–1163.
15. Lawrence D, Bretz F, Pocock S. INHANCE: an adaptive confirmatory study with dose selection at interim. In *Indacaterol, Milestones in Drug Therapy*, Trifilieff A (ed.) Springer: Basel, 2014; 77–92.

16. Thall PF, Nguyen HQ, Braun TM, Qazilbash MH. Using joint utilities of the times to response and toxicity to adaptively optimize schedule-dose regimes. *Biometrics* 2013; **69**:673–682.
17. Tang DI, Gnecco C, Geller NL. Design of group sequential clinical trials with multiple endpoints. *Journal of the American Statistical Association* 1989; **84**:776–779.
18. Jennison C, Turnbull BW. Group sequential tests for bivariate response: interim analyses of clinical trials with both efficacy and safety endpoints. *Biometrics* 1993; **49**:741–752.
19. Conaway MR, Petroni GR. Bivariate sequential designs for phase II trials. *Biometrics* 1995; **51**:656–664.
20. Kosorok MR, Shi Y, DeMets DL. Design and analysis of group sequential clinical trials with multiple primary endpoints. *Biometrics* 2004; **60**:134–145.
21. Müller HH, Schäfer H. A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine* 2004; **23**:2497–2508.
22. Wason JMS, Magirr D, Law M, Jaki T. Some recommendations for multi-arm multi-stage trials. *Statistical Methods in Medical Research* 2013. Published online; DOI: 10.1177/0962280212465498.
23. Senn S, Bretz F. Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics* 2007; **6**(3): 161–170.
24. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**(3):659–663.
25. Wason JMS, Jaki T. Optimal design of multi-arm multi-stage trials. *Statistics in Medicine* 2012; **31**(30):4269–4279.
26. Johnson SG, Narasimhan B. *cubature: Adaptive Multivariate Integration Over Hypercubes*, 2013. R package version 1.1-2.
27. Tamhane AC, Wu Y, Mehta CR. Adaptive extensions of a two-stage group sequential procedure for testing primary and secondary endpoints (I): unknown correlation between the endpoints. *Statistics in Medicine* 2012; **31**(19):2027–2040.
28. Berger RL, Boos DD. P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* 1994; **89**:1012–1016.
29. Hampson LV, Jennison C. Group sequential tests for delayed responses (with discussion). *Journal of the Royal Statistical Society B* 2013; **75**:3–54.
30. Stallard N. A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Statistics in Medicine* 2010; **29**(9):959–971.
31. Magirr D, Jaki T, Posch M, Klinglmueller F. Simultaneous confidence intervals that are compatible with closed testing in adaptive designs. *Biometrika* 2013; **100**:985–996.

Supporting information

Additional supporting information may be found in the online version of this article at the publishers web site.