Optimizing the data combination rule for seamless phase II/III clinical trials

Lisa V. Hampson[a][*][†] and Christopher Jennison[b]

We consider seamless phase II/III clinical trials that compare K treatments with a common control in phase II then test the most promising treatment against control in phase III. The final hypothesis test for the selected treatment can use data from both phases, subject to controlling the familywise type I error rate. We show that the choice of method for conducting the final hypothesis test has a substantial impact on the power to demonstrate that an effective treatment is superior to control. To understand these differences in power, we derive decision rules maximizing power for particular configurations of treatment effects. A rule with such an optimal frequentist property is found as the solution to a multivariate Bayes decision problem. The optimal rules that we derive depend on the assumed configuration of treatment means. However, we are able to identify two decision rules with robust efficiency: a rule using a weighted average of the phase II and phase III data on the selected treatment and control, and a closed testing procedure using an inverse normal combination rule and a Dunnett test for intersection hypotheses. For the first of these rules, we find the optimal division of a given total sample size between phases II and III. We also assess the value of using phase II data in the final analysis and find that for many plausible scenarios, between 50% and 70% of the phase II numbers on the selected treatment and control would need to be added to the phase III sample size in order to achieve the same increase in power. © 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

Keywords: Bayes decision problem; combination test; closed testing procedure; multiple hypothesis testing; seamless phase II/III trial; treatment selection

1. Introduction

In the traditional framework for drug development, a phase II clinical trial compares several doses or formulations of a new treatment against a control. The most promising of these, in terms of efficacy, safety and possibly other considerations, is taken forward to phase III where investigators hope to confirm the benefits of the new treatment in one or two 'pivotal' clinical trials.

There has been significant recent interest in combining these two stages of the development process. If a trial follows a 'seamless' design, merging the usual phase II and phase III components, there is opportunity to gain additional value from the phase II data by using these together with phase III data in the final analysis. Regulators are liable to treat a combined phase II/III trial as a single study and require a complete protocol to be specified at the outset. This allows a monitoring committee to respond to results on all aspects of the treatments and patient responses at interim points during the trial without further input from the sponsors, who remain blinded to interim results. Seamless designs can be complex, and substantial effort may be required to plan their smooth conduct and establish the validity of the proposed analysis. Thus, it is important that the gains from using phase II data in the final analysis justify this investment.

A variety of methods is available to combine data from the two stages of a seamless design with proper protection of the type I error probability. Thall *et al.* [1] propose two-stage designs with treatment selection at the interim analysis. Sampson and Sill [2] derive most powerful procedures within a certain

[a]*Medical and Pharmaceutical Statistics Research Unit, Department of Mathematics and Statistics, Lancaster University, Lancaster, U.K.*
[b]*Department of Mathematical Sciences, University of Bath, Bath, U.K.*
[*]*Correspondence to: Lisa V. Hampson, Department of Mathematics and Statistics, Fylde College, Lancaster University, Lancaster, LA1 4YG, U.K.*
[†]*E-mail: l.v.hampson@lancaster.ac.uk*

class of tests combining data from two stages. Bretz *et al.* [3] and Schmidli *et al.* [4] present seamless phase II/III designs that use closed testing procedures [5] to control the familywise type I error rate and combination tests [6] to combine data from the two stages in the final hypothesis test.

It is not obvious how to choose between the various options for combining data across two phases of a seamless trial. Our aims are to identify efficient ways of doing this and, hence, quantify the potential benefits of using phase II data in a final combined analysis. We shall show how to derive an optimal final decision rule, maximizing the probability of selecting the best treatment and declaring it efficacious, under a particular configuration of treatment effects. In some cases, the decision rules we derive only control the familywise error rate over part of the parameter space—but they are still useful as they provide an upper bound on the attainable power and this can be enough to show that certain decision rules, which do control the familywise error rate, are very close to optimal.

In view of the high dimensionality of the parameter space, one would not expect a single data combination rule to be optimal for all parameter vectors. Nevertheless, we have found rules with robust efficiency across a wide range of scenarios. Having an efficient final decision rule is an important pre-requisite for investigating other aspects of phase II/III designs: given such a rule, one can optimize the division of resources between phases or assess the benefits of other phase II options, such as response adaptive allocation of patients to treatments.

In Sections 2 and 3, we formulate the two-stage problem and describe a selection of the final decision rules. We apply these rules to an example in Section 4 and compare their power functions. In Section 5, we derive optimal decision rules for particular configurations of treatment effects: the form of the optimal rules in Section 5.1 helps explain the rather surprising results seen in Section 4, and we solve more general optimization problems in Sections 5.2 and 5.3. In Section 6, we compare final decision rules across a range of parameter configurations and identify rules that are highly efficient across a wide variety of situations. In Section 7, we focus on one of these robustly efficient rules and show how to determine the most efficient division of resources between phases II and III. In Section 8, we assess the benefits of using phase II data in the final decision by computing the number of additional phase III observations that would be needed to produce the same improvement in power. We conclude with a discussion of the implications of our results to extensions of the seamless phase II/III design that we have considered.

2. Problem formulation

Henceforth, we shall refer to the two parts of a seamless phase II/III design as stage 1 and stage 2. We consider the format of Thall *et al.* [1] in which K experimental treatments are compared with a control in stage 1 and one of these is selected to be tested against the control in stage 2.

We suppose patient responses are normally distributed with known variance σ^2 and means μ_0 on the control arm and μ_i, $i = 1, \ldots, K$, on treatment arms, with a high mean indicating a successful treatment. We assume that the primary endpoint, the study population and the treatment definitions remain unchanged throughout the trial, so the response distribution for a given treatment in stage 2 is the same as in stage 1. The treatment effects are $\theta_i = \mu_i - \mu_0$, $i = 1, \ldots, K$, and for now, we make no assumptions about the structure of the vector $\theta = (\theta_1, \ldots, \theta_K)$. There are K one-sided null hypotheses $H_{0,1}: \theta_1 \leqslant 0$, $\ldots$, $H_{0,K}: \theta_K \leqslant 0$ that may be tested, depending on which treatment is selected at the end of stage 1.

Following Thall *et al.* [1] (hereafter TSE), we proceed as follows:

In stage 1, randomize m_1 patients to each treatment $i = 1, \ldots, K$ and the control arm and calculate maximum likelihood estimates $\hat{\theta}_{1,i} = \hat{\mu}_{1,i} - \hat{\mu}_{1,0}$, $i = 1, \ldots, K$, of the K treatment effects. Let $i^\star$ denote the treatment with maximum $\hat{\theta}_{1,i}$. If

$$\hat{\theta}_{1,i^\star} = \max_{i=1,\ldots,K} \left\{ \hat{\theta}_{1,i} \right\} < \ell, \tag{1}$$

stop the trial for futility, rejecting no null hypotheses. Otherwise, continue to stage 2 selecting treatment $i^\star$ for comparison with the control.

In stage 2, randomize m_2 patients to each of treatment $i^\star$ and the control. Denote the estimate of $\theta_{i^\star}$ based on stage 2 data only by $\hat{\theta}_{2,i^\star} = \hat{\mu}_{2,i^\star} - \hat{\mu}_{2,0}$.

In the final analysis, reject $H_{0,i^\star}: \theta_{i^\star} \leqslant 0$ in favour of $\theta_{i^\star} > 0$ if

$$T\left(\hat{\theta}_{1,1}, \ldots, \hat{\theta}_{1,K}, \hat{\theta}_{2,i^\star}\right) \geqslant C_T(K, m_1, m_2), \tag{2}$$

where the function T and critical value $C_T(K, m_1, m_2)$ are pre-specified.

The familywise error rate (FWER) under θ of such a procedure is defined as $\mathrm{pr}\{\text{reject any true } H_{0,i}; \theta\}$. We shall consider procedures that control the FWER strongly at level α; that is, they have the property

$$\mathrm{pr}\{\text{reject any true } H_{0,i}; \theta\} \leqslant \alpha \quad \text{for all parameter vectors } \theta.$$

Let $i_{\max}$ be the index of the treatment with the highest effect θ_i. Under parameter vectors θ for which $i_{\max}$ is unique and $\theta_{i_{\max}} > 0$, we define the power of a procedure to be

$$\mathrm{pr}\left\{\text{select treatment } i_{\max} \text{ and reject } H_{0,i_{\max}}; \theta\right\}. \tag{3}$$

Methods of data combination differ in the definition of the function T in (2) and the associated critical value $C_T(K, m_1, m_2)$. Our aims are to compare the power of different final decision rules and identify those with close to optimal power for a variety of vectors θ.

Searching for an optimal decision rule is a complex task because power depends on the K-dimensional θ. It may be appropriate to focus on achieving high power under certain forms of θ, particularly if 'treatments' are doses of a single compound. Rules attuned to situations where the treatment effects $\theta_1, \ldots, \theta_K$ are high or low together may be thought of as 'borrowing strength' for inference about $\theta_{i^\star}$ from other stage 1 estimates $\hat{\theta}_{1,i}$, $i \neq i^\star$. However, the correlations between $\hat{\theta}_{1,1}, \ldots, \hat{\theta}_{1,K}$, because of the common control arm in stage 1, also affect how these estimates should be weighted in the overall statistic $T\left(\hat{\theta}_{1,1}, \ldots, \hat{\theta}_{1,K}, \hat{\theta}_{2,i^\star}\right)$.

3. Methods for data combination

In this section, we outline the decision rules underlying six methods for data combination used in our numerical investigations of power. In our simulations, we have applied the futility stopping rule in (1) with $\ell = 0$. We calibrated the critical values of all six decision rules so that tests attain overall type I error rate α when $\theta = (0, \ldots, 0)$ adjusting for the possibility of early stopping, arguing for each test that this ensures strong control of the FWER at level α. Therefore, the higher power achieved by a decision rule can be attributed to an efficient use of the available data rather than a higher type I error rate.

Conventional test: In the conventional approach with separate phase II and phase III studies, only phase III data are used in making the final decision to accept or reject $H_{0,i^\star}$. Let $Z_{2,i^\star} = \hat{\theta}_{2,i^\star}/\sqrt{(2\sigma^2/m_2)}$ denote the standardized test statistic based on stage 2 data. To account for the possibility of stopping after stage 1 for futility, we reject $H_{0,i^\star}$ if

$$Z_{2,i^\star} \geqslant \Phi^{-1}\left(1 - \frac{\alpha}{\mathrm{pr}\left\{\max_i\{\hat{\theta}_{1,i}\} \geqslant 0; \theta = \mathbf{0}\right\}}\right) = \Phi^{-1}\left(1 - \frac{K+1}{K}\alpha\right),$$

where Φ denotes the standard normal cumulative distribution function and $\mathbf{0}$ denotes the parameter vector $(0, \ldots, 0)$. The overall type I error rate under $\theta = \mathbf{0}$ is exactly α, and it follows from the arguments of Jennison and Turnbull [7, Section 3] that the FWER is controlled strongly at level α.

TSE decision rule: Adapting the procedure of Thall, Simon and Ellenberg [1] to a normal response, we define

$$Z_{1,i^\star} = \frac{\hat{\theta}_{1,i^\star}}{\sqrt{(2\sigma^2/m_1)}} \quad \text{and} \quad Z_{2,i^\star} = \frac{\hat{\theta}_{2,i^\star}}{\sqrt{(2\sigma^2/m_2)}}.$$

and reject $H_{0,i^\star}$ if

$$w_1 Z_{1,i^\star} + w_2 Z_{2,i^\star} > C_{TSE}(K, m_1, m_2),$$

where $w_i = \sqrt{\{m_i/(m_1 + m_2)\}}$, $i = 1, 2$, and $C_{TSE}(K, m_1, m_2)$ is chosen to give FWER α when $\theta = \mathbf{0}$. Jennison and Turnbull [7] note that this ensures the FWER is controlled strongly at level α.

Combination tests: Bretz *et al.* [3] and Schmidli *et al.* [4] present a variety of adaptive designs for seamless phase II/III clinical trials. These proposals follow the general approach of Bauer and Kieser [8] for making mid-study data-dependent adaptations while preserving trial integrity. They use closed testing procedures [5] to control the FWER and combination tests [6] to conduct hypothesis tests using data from the two stages.

Denote the set of indices i of null hypotheses $H_{0,i}$ by $S = \{1, \ldots, K\}$. A closed testing procedure requires an α-level test of the intersection hypothesis $H_{0,I} = \bigcap_{i \in I} H_{0,i}$ for each subset I of S; this test will reject $H_{0,I}$ with probability at most α when all $H_{0,i}$ with indices $i \in I$ are true. Tests of intersection hypotheses $H_{0,I}$ combine data from the two stages of the trial. A combination test of $H_{0,I}$ is defined in terms of one-sided p-values $P_{1,I}$ and $P_{2,I}$ for $H_{0,I}$ based on stage 1 and stage 2 data, respectively. Note that 'stage 2 data' refers to new data in stage 2, not the cumulative data at the end of stage 2. Using the inverse χ^2 rule (originally proposed by Fisher [9] for combining separate experiments) to combine p-values in the test for each intersection hypothesis in the closed testing procedure, we obtain an overall decision rule that rejects $H_{0,i^\star}$ if

$$\min_{\{I \subseteq S : i^\star \in I\}} \{-\log(P_{1,I}) - \log(P_{2,I})\} > 0.5\,\chi^2_{4,1-\alpha}, \tag{4}$$

where $\chi^2_{4,1-\alpha}$ is the value exceeded with probability α by a χ^2_4 random variable. We refer to this as the 'BK inverse χ^2' decision rule. Alternatively, using the inverse normal combination rule [10, 11] to combine p-values gives the 'BK inverse normal' decision rule, which rejects $H_{0,i^\star}$ if

$$\min_{\{I \subseteq S : i^\star \in I\}} \{w_1 Z_{1,I} + w_2 Z_{2,I}\} > \Phi^{-1}(1 - \alpha), \tag{5}$$

where, as in the TSE method, $w_i = \sqrt{\{m_i/(m_1 + m_2)\}}$, $i = 1, 2$. There are various choices for defining the p-values for intersection hypotheses in the previous methods. In our simulations, we compared the efficiencies of methods when Simes [12] and Dunnett [13] p-values are used for intersection hypotheses.

Because the trial may stop for futility after stage 1, all the previous procedures have an FWER below α. Additional conservatism arises from using Simes' method to define p-values for intersection hypotheses arising from multiple comparisons with a common control [14]. So as not to disadvantage methods using Simes' test in our investigations of decision rules, we have adjusted the critical values on the right-hand sides of (4) and (5) so that the FWER is α under $\theta = \mathbf{0}$ proceeding on the assumption that this is a sufficient condition to ensure strong control of the FWER. In fact, it is difficult to give a general proof that the probability of a type I error is decreased when one treatment effect, say θ_i, is increased above zero: because $H_{0,i}$ is now false, rejecting it no longer counts as a type I error but, against this, a low p-value for treatment i may reduce the p-value for an intersection hypothesis involving a selected treatment $i^\star \neq i$. We have used simulation to check the implications of these lower critical values in our examples, and in all cases, we found the type I error rate to be controlled at level α with some conservatism: see the supporting information accompanying this manuscript for further discussion.

We now illustrate the application of the previous testing procedures in an example.

4. Illustrative example

Liu and Pledger [15] discuss a seamless phase II/III trial comparing five doses of a treatment for migraine headaches against placebo. We simplify this example by assuming that both stages of the trial measure the same clinical endpoint, the decrease in monthly headache rate over 4 months. Responses are assumed to be normally distributed with standard deviation $\sigma = 5$. A reduction of 2 in the average monthly headache rate, compared with placebo, is taken to be clinically meaningful, and high power is desired to detect a dose with such an effect.

In our notation, $K = 5$ and for each dose $i = 1, \ldots, K$, we wish to test $H_{0,i} : \theta_i \leqslant 0$ against $\theta_i > 0$. While controlling the FWER strongly at $\alpha = 0.025$, we desire high power to select and declare efficacious a dose with a treatment effect $\theta_i = 2$. Suppose the trial follows a two-stage design with $m_1 = 28$ patients randomized to each dose and placebo in stage 1 and a further $m_2 = 140$ allocated to each of dose $i^\star$ and placebo when sampling continues to stage 2. Such unequal division of resources between phases is common in practice, with larger sample sizes devoted to confirming efficacy of the selected dose in phase III.

Table I. Critical values for six decision rules testing $H_{0,i\star}: \theta_{i\star} \leqslant 0$ against $\theta_{i\star} > 0$.

Design	Test statistic	Critical value
Conventional	$Z_{2,i\star}$	1.881
TSE	$w_1 Z_{1,i\star} + w_2 Z_{2,i\star}$	2.245
BK inverse χ^2 (Simes)	$\min_{\{I \supseteq i\star\}}\{-\log(P_{1,I}P_{2,I})\}$	5.342
BK inverse χ^2 (Dunnett)	$\min_{\{I \supseteq i\star\}}\{-\log(P_{1,I}P_{2,I})\}$	5.539
BK inverse normal (Simes)	$\min_{\{I \supseteq i\star\}}\{w_1 Z_{1,I} + w_2 Z_{2,I}\}$	1.851
BK inverse normal (Dunnett)	$\min_{\{I \supseteq i\star\}}\{w_1 Z_{1,I} + w_2 Z_{2,I}\}$	1.958

These rules control familywise type I error rate at level $\alpha = 0.025$ for $K = 5$, $m_1 = 28$, $m_2 = 140$, $\ell = 0$ and $\sigma = 5$.
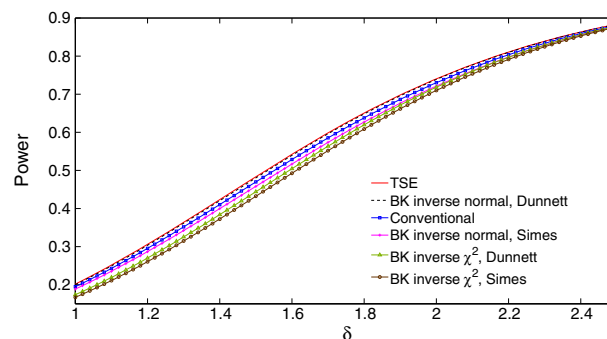
Figure 1. Power of six decision rules under $\theta = (0,0,0,0,\delta)$ when $m_1 = 28$, $m_2 = 140$, $\ell = 0$, $\sigma = 5$, and FWER is controlled strongly at $\alpha = 0.025$. All estimates are based on 1 million simulations. The legend lists rules in order of decreasing power.

Table I lists the critical values needed to implement the six decision rules described in Section 3 with a familywise type I error rate of 0.025.

Suppose that only the highest dose gives an improvement over placebo and the vector of treatment effects has the form $\theta = (0,0,0,0,\delta)$. Figure 1 shows the power of each decision rule, as a function of δ, to select dose 5 and reject $H_{0.5}$. Results are based on 1 million replicates in each scenario considered, so standard errors of estimated probabilities are at most 0.0005. The TSE procedure is most powerful at all values of δ, closely followed by the BK inverse normal combination test using Dunnett p-values for intersection hypotheses. Surprisingly, the three other combination tests have lower power than the conventional test that does not use the phase II data at all. Differences in power are as high as 0.05 in places: the values of δ at which different rules attain the same power differ by up to 5% and, supposing the sample size needed to achieve a given power to be roughly proportional to δ^{-2}, this translates into differences in sample size of up to 10%.

The results in Figure 1 parallel those of Jennison and Turnbull [16, Section 5.3] for an example with $K = 4$, $m_1 = 100$, $m_2 = 500$ and $\sigma = 5$ (although those authors did not consider methods using Dunnett tests). The failure in both examples of some decision rules to improve on the conventional test, which ignores stage 1 data, motivated our investigation of the underlying decision rules. We shall also investigate whether the same patterns of relative efficiency occur for other forms of θ and consider what is the optimal division of resources between phases II and III when the total sample size has been fixed. In order to explore these issues, we shall derive optimal decision rules for particular forms of θ and study the structure of these rules.

5. Optimal data combination rules

5.1. Optimizing power for a family of configurations of θ

In the framework of Section 2, the function $T\left(\hat{\theta}_{1,1}, \ldots, \hat{\theta}_{1,K}, \hat{\theta}_{2,i\star}\right)$ in (2) specifies a decision rule. We seek to optimize this function while protecting the FWER. We shall consider a variety of configurations of θ,

chosen to cover a wide range of plausible possibilities. We first consider vectors θ, which are permutations of $(\gamma\delta, \ldots, \gamma\delta, \delta)$, where $0 \leqslant \gamma < 1$ and the value δ is equally likely to appear in each of the K positions. We seek the decision rule with the highest probability, averaged over the K permutations of θ, of selecting the treatment with effect δ in stage 1 and rejecting the associated null hypothesis in stage 2. This is in keeping with the definition of power used by TSE that focuses on selecting the treatment with the highest effect when other treatment effects are lower by a specified margin. We shall require rules to have type I error rate at most α when $\theta = \mathbf{0}$: we have seen in Section 3 that this is a sufficient condition for some rules to provide strong control of the FWER but we shall have to check this property for the new rules that we derive.

We proceed by defining a Bayes decision problem with a prior distribution for θ and costs for each possible decision. We then search over values of these costs to find a version of this problem for which the optimal Bayes rule has type I error rate α under $\theta = \mathbf{0}$ and so solves the problem originally stated in frequentist terms. The method of re-casting a frequentist problem as a Bayes decision problem has been used to find optimal group sequential tests; see, for example, [17–21]. In our problem, power depends on a vector of treatment effects, and we handle this by dealing with a one-dimensional subset of vectors θ at a time. This provides a benchmark for each family of θ vectors, against which other decision rules can be compared. While it is desirable to have a single rule with robust efficiency in a wide variety of situations, it could be that quite different rules are needed to achieve high power for different configurations of θ, in which case, the importance of these different scenarios should guide the overall choice.

For our first problem, with θ a permutation of $(\gamma\delta, \ldots, \gamma\delta, \delta)$, let ξ_i denote the vector with $\theta_i = \delta$ and the other $K - 1$ elements equal to $\gamma\delta$. Define a prior distribution for θ with discrete mass function $\pi(\theta)$ placing probability $1/(K + 1)$ on each of the cases $\theta = \mathbf{0}$ and $\theta = \xi_i$, $i = 1, \ldots, K$. The only hypothesis that can be rejected when treatment i is selected for stage 2 is $H_{0,i}$. Thus, the set of possible actions is $\{A_0, A_1, \ldots, A_K\}$ where, for $i \geqslant 1$, A_i means that treatment i is selected after stage 1 and $H_{0,i}$ is rejected at the end of stage 2, while A_0 indicates stopping for futility at stage 1 or continuing to stage 2 but failing to reject any $H_{0,i}$. We define the loss function $L(\theta, A)$ as

$$
\begin{aligned}
L(\mathbf{0}, A_i) &= c_1 \text{ for } i = 1, \ldots, K, \\
L(\xi_i, A_i) &= -c_2 \text{ for } i = 1, \ldots, K, \\
L(\theta, A) &= 0 \text{ otherwise.}
\end{aligned}
$$

The reward for correctly rejecting $H_{0,i}$ appears as the negative cost $-c_2$, and the penalty for failing to reject $H_{0,i}$ when $\theta = \xi_i$ is the absence of this reward. Our original criteria concern power to declare efficacy of treatment i when $\theta = \xi_i$ but do not differentiate between ways of failing to reject $H_{0,i}$ when $\theta = \xi_i$; hence, we define the same loss, of zero, for actions A_0 and A_j, $j \geqslant 1$ and $j \neq i$, in this case.

The Bayes rule for the problem that we have defined minimizes the Bayes risk

$$
c_1 \, \pi(\mathbf{0}) \sum_{i=1}^{K} \text{pr}\{A_i \mid \theta = \mathbf{0}\} - c_2 \sum_{i=1}^{K} \pi(\xi_i) \, \text{pr}\{A_i \mid \theta = \xi_i\}
$$

$$
= c_1 \, \pi(\mathbf{0}) \, \text{pr}\{\text{select any treatment } i \text{ and reject } H_{0,i} \mid \theta = \mathbf{0}\} \tag{6}
$$

$$
- c_2 \sum_{i=1}^{K} \pi(\xi_i) \, \text{pr}\{\text{select treatment } i \text{ and reject } H_{0,i} \mid \theta = \xi_i\}.
$$

Suppose that treatment $i^\star$ is selected and data at the end of stage 2 are summarized as

$$
D_{i^\star} = \left(\hat{\theta}_{1,1}, \ldots, \hat{\theta}_{1,K}, \hat{\theta}_{2,i^\star}\right).
$$

Either action A_0 or action $A_i^\star$ must be taken. Let $\pi\left(\theta \mid D_{i^\star}\right)$ denote the posterior distribution of θ given data $D_{i^\star}$. If action $A_i^\star$ is chosen, so $H_{0,i^\star}$ is rejected, the posterior expected loss is

$$
c_1 \, \pi\left(\mathbf{0} \mid D_{i^\star}\right) - c_2 \, \pi\left(\xi_{i^\star} \mid D_{i^\star}\right). \tag{7}
$$

All costs associated with action A_0 are zero, so if this action is chosen and $H_{0,i^\star}$ is not rejected, the loss is exactly zero. Thus, the Bayes rule that minimizes (6) rejects $H_{0,i^\star}$ if and only if (7) is negative or, equivalently, if

$$\frac{\pi\left(\xi_{i^\star} \mid D_{i^\star}\right)}{\pi\left(\mathbf{0} \mid D_{i^\star}\right)} > \frac{c_1}{c_2}. \tag{8}$$

Because the prior probabilities of $\theta = \mathbf{0}$ and $\theta = \xi_{i^\star}$ are equal, the left-hand side of (8) is simply the likelihood ratio of the observed data under $\theta = \xi_{i^\star}$ and $\theta = \mathbf{0}$.

Given θ, the stage 1 estimates $\hat{\theta}_1 = \left(\hat{\theta}_{1,1}, \ldots, \hat{\theta}_{1,K}\right)$ are distributed as $\hat{\theta}_1 \sim N(\theta, V)$, where $V_{ii} = (2\sigma^2/m_1)$ for $i = 1, \ldots, K$ and $V_{ii'} = (\sigma^2/m_1)$ for $i \neq i'$. The inverse of V has elements

$$V_{ii}^{-1} = \frac{K}{K+1}\frac{m_1}{\sigma^2}, i = 1, \ldots, K, \text{ and } V_{ii'}^{-1} = \frac{-1}{K+1}\frac{m_1}{\sigma^2}, i \neq i'.$$

The log likelihood ratio of $\hat{\theta}_1$ under $\theta = \xi_{i^\star}$ and $\theta = \mathbf{0}$ is

$$\hat{\theta}_1^T V^{-1} \xi_{i^\star} - \frac{1}{2}\xi_{i^\star}^T V^{-1} \xi_{i^\star} = \frac{m_1 \delta}{(K+1)\sigma^2}\left[\{K - (K-1)\gamma\}\hat{\theta}_{1,i^\star} + (2\gamma - 1)\sum_{i \neq i^\star}\hat{\theta}_{1,i}\right] - g\frac{\delta^2}{\sigma^2} \tag{9}$$

for some constant g. The log likelihood ratio for the stage 2 data $\hat{\theta}_{2,i^\star}$ is

$$\frac{m_2}{2\sigma^2}\hat{\theta}_{2,i^\star}\delta - \frac{m_2}{4\sigma^2}\delta^2. \tag{10}$$

Adding (9) and (10) gives the log likelihood ratio of $D_{i^\star}$ under $\theta = \xi_{i^\star}$ and $\theta = \mathbf{0}$. It follows that the condition for the Bayes test to reject $H_{0,i^\star}$ can be written as

$$\frac{m_1}{K+1}\left[\{K - (K-1)\gamma\}\hat{\theta}_{1,i^\star} + (2\gamma - 1)\sum_{i \neq i^\star}\hat{\theta}_{1,i}\right] + \frac{m_2}{2}\hat{\theta}_{2,i^\star} \geqslant c, \tag{11}$$

where

$$c = \frac{\sigma^2}{\delta}\log\left(\frac{c_1}{c_2}\right) + \left(g + \frac{m_2}{4}\right)\delta. \tag{12}$$

The constant c is an increasing function of the ratio c_1/c_2. Also, c depends on δ, but the expression on the left-hand side of (11) does not.

Suppose c is such that the rule given by (11) has type I error rate α when $\theta = \mathbf{0}$. For any given δ, there are costs c_1 and c_2 that satisfy (12) with this c. Hence, the decision rule (11) minimizes (6) for this δ and the pair (c_1, c_2), and so, it maximizes

$$\sum_{i=1}^{K} \pi(\xi_i) \text{ pr}\{\text{select treatment } i \text{ and reject } H_{0,i} \mid \theta = \xi_i\}$$

amongst all rules with type I error rate less than or equal to α when $\theta = \mathbf{0}$. Thus, this decision rule solves the problem posed at the start of this section, and we note that, by construction, the same rule is optimal for all values of δ.

We can find this optimal rule by searching for the constant c in (11) that gives type I error rate α under $\theta = \mathbf{0}$. Because we use simulation to estimate error rates, we have applied the Robbins–Monro algorithm [22] to search for the value of c that satisfies this condition.

Setting $\gamma = 0$ gives the example of Section 4 where θ is a permutation of $(0, \ldots, 0, \delta)$. In this case, estimates $\hat{\theta}_{1,i}$ for treatments other than $i^\star$ have negative weights in (11). This is a consequence of the correlation between estimates $\hat{\theta}_{1,i}$ and $\hat{\theta}_{1,i^\star}$ for $i \neq i^\star$ caused by the common control arm: for the values $\theta = \xi_{i^\star}$ and $\theta = \mathbf{0}$ appearing in the likelihood ratio in (8), all θ_i for $i \neq i^\star$ are zero, and, because positive values of $\hat{\theta}_{1,i}$ for $i \neq i^\star$ may be due to lower than average responses on the control arm in stage 1,

this possibility detracts from the evidence a positive $\hat{\theta}_{1,i^\star}$ provides in support of $\theta_{i^\star} = \delta$. If some of the θ_i are decreased while others remain equal to zero, the negative weights for $\hat{\theta}_{1,i}$, $i \neq i^\star$, are likely to lead to higher values of the left-hand side of (11), increasing the probability of rejecting $H_{0,i^\star}$. Thus, the decision rule (11) does not control the FWER strongly at level α. The form of this decision rule is, nevertheless, instructive. The desirability of negative weights for the $\hat{\theta}_{1,i}$, $i \neq i^\star$, explains the poor performance of tests using Simes' rule, which treats good performance of other treatments as supporting evidence in favour of treatment $i^\star$. The conventional procedure ignores all first-stage data, so at least it does not give weights of the wrong sign to $\hat{\theta}_{1,i}$, $i \neq i^\star$. Dunnett p-values for intersection hypotheses are appropriate when only one treatment is efficacious as they focus on the single treatment with the highest estimated effect. If we retain the form of the test statistic in (11) but modify it so that all weights are non-negative in order to maintain strong control of the FWER, we obtain a linear combination of $\hat{\theta}_{1,i^\star}$ and $\hat{\theta}_{2,i^\star}$. The TSE decision rule has this form: it is the likelihood ratio test between $\boldsymbol{\theta} = \boldsymbol{\xi}_{i^\star}$ and $\boldsymbol{\theta} = \mathbf{0}$ based on $\hat{\theta}_{1,i^\star}$ and $\hat{\theta}_{2,i^\star}$ only, and we conjecture that the TSE procedure is very close to optimal for the case $\gamma = 0$.

In (11), the weights for estimates $\hat{\theta}_{1,i}$ with $i \neq i^\star$ are negative if $\gamma < 0.5$, zero for $\gamma = 0.5$ and positive if $\gamma > 0.5$. We show in the Appendix how these weights can be obtained by fitting a linear regression model to the stage 1 data and the signs of the weights follow from this representation. For $\gamma < 0.5$, the negative weights imply that tests of this form do not control the FWER strongly and, as for $\gamma = 0$, we conjecture that the TSE procedure is close to optimal.

If $\gamma = 0.5$, the optimal decision rule has zero weights for $\hat{\theta}_{1,i}$, $i \neq i^\star$, and is precisely the TSE rule. Thus, the TSE rule provides the ideal solution in this case where there is one high treatment and other effects are at an intermediate level. As noted in Section 3, strong control of the FWER does follow from controlling the type I error rate at $\boldsymbol{\theta} = \mathbf{0}$ in this case. When $\gamma \geqslant 0.5$, the weight of each $\hat{\theta}_{1,i}$ is positive and, by the arguments applied for Simes' rule in Section 3, we expect that controlling the type I error rate at $\boldsymbol{\theta} = \mathbf{0}$ implies strong control of the FWER for all possible vectors $\boldsymbol{\theta}$.

Figure 2 compares power curves of optimal decision rules and the six methods introduced in Section 3 in the example of Section 4, with $K = 5$ treatments and a control. Panels (a) and (b) of Figure 2 show power curves for $\gamma = 0.5$, the case in which the TSE rule is optimal.

We see that the two BK inverse normal combination tests have almost the same power as the TSE rule: for the test using Simes' rule, this is a significant improvement over the case $\gamma = 0$ seen in Figure 1.
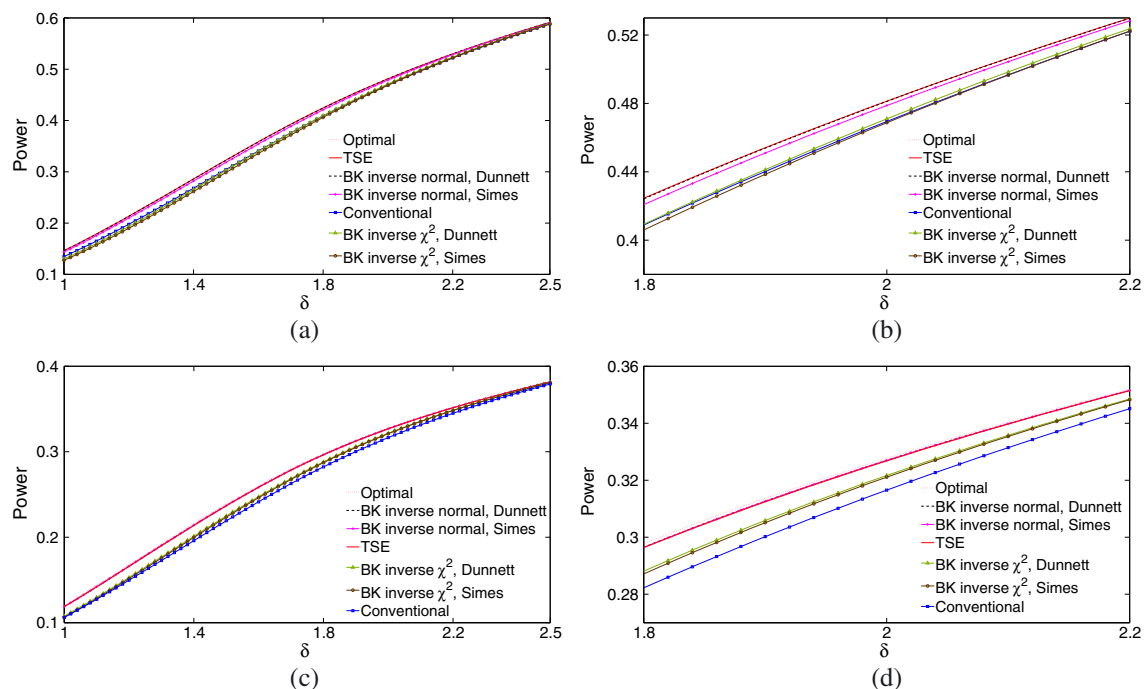
Figure 2. Power achieved by decision rules in the example of Section 4 when $\boldsymbol{\theta}$ is a random permutation of $(\gamma, \gamma, \gamma, \gamma, 1)\delta$ with (a) - (b) $\gamma = 0.5$ and (c) - (d) $\gamma = 0.75$. Decision rules are listed in order of decreasing power. Results are based on 1 million simulations for each scenario.

However, the two BK inverse χ^2 combination tests still have lower power than the conventional test using stage 2 data only. The power curves for $\gamma = 0.75$ in panels (c) and (d) show the TSE rule and the two BK inverse normal combination tests to be almost as powerful as the optimal decision rule for this case, and now the two BK inverse χ^2 combination tests have a small advantage over the conventional test.

The power curves for $\gamma = 0.75$ are noticeably lower than for $\gamma = 0.5$ because of the higher probability of a sub-optimal treatment being selected after stage 1. If a final decision in favour of a sub-optimal treatment with sufficiently high effect size is deemed acceptable, the definition of power could be modified to include this. Such a definition would certainly be reasonable in the limit as $\gamma \to 1$.

5.2 Optimizing power for general configurations of $\boldsymbol{\theta}$

The approach of Section 5.1 can be extended to obtain decision rules that maximize power averaged over the $K!$ permutations of $(\gamma_1, \ldots, \gamma_{K-1}, 1)\,\delta$, where $0 \leqslant \gamma_1 < \ldots < \gamma_{K-1} < 1$, subject to a type I error rate of at most α when $\boldsymbol{\theta} = \mathbf{0}$. As before, power is defined to be the probability of selecting the treatment $i_{\max}$ with the highest effect and then rejecting $H_{0, i_{\max}}$. The optimal decision rule can then be examined to check whether controlling the type I error at $\boldsymbol{\theta} = \mathbf{0}$ implies strong control of the FWER.

Let Q denote the set of $K!$ parameter vectors $\boldsymbol{\theta}$ obtained by permuting the elements of $(\gamma_1, \ldots, \gamma_{K-1}, 1)\,\delta$. In our Bayes decision problem, we define the prior distribution $\pi(\boldsymbol{\theta})$ on $\boldsymbol{\theta} \in Q$ to give probability $1/(K+1)$ to $\boldsymbol{\theta} = \mathbf{0}$ and $1/\{(K+1)(K-1)!\}$ to each element of Q. For $i = 1, \ldots, K$, let Q_i be the subset of Q containing vectors $\boldsymbol{\theta}$ with $\theta_i = \delta$. With actions $A_0, A_1, \ldots, A_K$ as defined in Section 5.1, we define the loss function $L(\boldsymbol{\theta}, A)$ to be

$$L(\mathbf{0}, A_i) = c_1 \text{ for } i = 1, \ldots, K,$$
$$L(\boldsymbol{\theta}, A_i) = -c_2 \text{ for all } \boldsymbol{\theta} \in Q_i, \ i = 1, \ldots, K,$$
$$L(\boldsymbol{\theta}, A) = 0 \text{ otherwise.}$$

When treatment $i^\star$ is selected in stage 1, either action A_0 or $A_{i^\star}$ must be taken after stage 2. We seek the Bayes rule that minimizes the Bayes risk

$$c_1\, \pi(\mathbf{0}) \sum_{i=1}^{K} \text{pr}\{A_i \mid \boldsymbol{\theta} = \mathbf{0}\} - c_2 \sum_{i=1}^{K} \sum_{\boldsymbol{\phi} \in Q_i} \pi(\boldsymbol{\phi})\, \text{pr}\{A_i \mid \boldsymbol{\theta} = \boldsymbol{\phi}\}$$

$$= c_1\, \pi(\mathbf{0})\, \text{pr}\{\text{select any treatment } i \text{ and reject } H_{0,i} \mid \boldsymbol{\theta} = \mathbf{0}\}$$

$$- c_2 \sum_{i=1}^{K} \sum_{\boldsymbol{\phi} \in Q_i} \pi(\boldsymbol{\phi})\, \text{pr}\{\text{select treatment } i \text{ and reject } H_{0,i} \mid \boldsymbol{\theta} = \boldsymbol{\phi}\}.$$

If c_1 and c_2 are chosen so that the Bayes optimal rule has type I error rate α when $\boldsymbol{\theta} = \mathbf{0}$, we can deduce that this rule maximizes

$$\sum_{i=1}^{K} \sum_{\boldsymbol{\phi} \in Q_i} \pi(\boldsymbol{\phi})\, \text{pr}\{\text{select treatment } i \text{ and reject } H_{0,i} \mid \boldsymbol{\theta} = \boldsymbol{\phi}\},$$

and it therefore maximizes the average power over the $K!$ permutations of $(\gamma_1, \ldots, \gamma_{K-1}, 1)\,\delta$, amongst all decision rules with type I error rate at most α at $\boldsymbol{\theta} = \mathbf{0}$.

As before, taking action A_0 after stage 2 has cost zero. The posterior expected loss for action $A_{i^\star}$ given that treatment $i^\star$ is selected and data $D_{i^\star}$ are observed is

$$c_1\, \pi\left(\mathbf{0} \mid D_{i^\star}\right) - c_2 \sum_{\boldsymbol{\phi} \in Q_{i^\star}} \pi\left(\boldsymbol{\phi} \mid D_{i^\star}\right).$$

It follows that the Bayes rule rejects $H_{0,i^\star}$ if

$$\sum_{\boldsymbol{\phi} \in Q_{i^\star}} \frac{\pi\left(\boldsymbol{\phi} \mid D_{i^\star}\right)}{\pi\left(\mathbf{0} \mid D_{i^\star}\right)} > \frac{c_1}{c_2}.$$

or, equivalently, if

$$\frac{1}{(K-1)!} \sum_{\phi \in Q_{i\star}} LR\left(D_{i\star}; \phi, 0\right) > \frac{c_1}{c_2}, \tag{13}$$

where $LR\left(D_{i\star}; \phi, 0\right)$ denotes the likelihood ratio of data $D_{i\star}$ under parameter vectors $\theta = \phi$ and $\theta = 0$.

Given θ, the stage 1 estimates $\hat{\theta}_1$ follow an $N(\theta, V)$ distribution. The inverse of V can be written as

$$V^{-1} = \{(K+1)I_K - 1_{K,K}\} \frac{m_1}{(K+1)\sigma^2},$$

where I_K is the $K \times K$ identity matrix and $1_{K,K}$ the $K \times K$ matrix with all elements equal to 1. Thus, the log likelihood ratio of the stage 1 data under $\theta = \phi$ and 0 is

$$\hat{\theta}_1^T V^{-1} \phi - \frac{1}{2} \phi^T V^{-1} \phi = \hat{\theta}_1^T \lambda(\phi) - h\frac{\delta^2}{\sigma^2}, \tag{14}$$

where

$$\lambda(\phi) = \frac{m_1}{(K+1)\sigma^2} \{(K+1)\phi - 1_{K,K}\phi\} \tag{15}$$

and the constant h is the same for all vectors $\phi \in Q$.

When $\phi \in Q_{i\star}$, and so $\phi_{i\star} = \delta$, the log likelihood ratio for the stage 2 data under $\theta = \phi$ and $\theta = 0$ is

$$\frac{m_2}{2\sigma^2} \hat{\theta}_{2,i\star} \delta - \frac{m_2}{4\sigma^2} \delta^2. \tag{16}$$

Combining (14) and (16), the condition (13) can be written as

$$\sum_{\phi \in Q_{i\star}} \exp\left\{\hat{\theta}_1^T \lambda(\phi)\right\} \exp\left\{\frac{m_2}{2\sigma^2} \hat{\theta}_{2,i\star} \delta\right\} > c, \tag{17}$$

where c depends on δ. The value of c for which the type I error rate is α under $\theta = 0$ varies with δ. Therefore, no uniformly most powerful decision rule exists for the θ configuration, and we find the appropriate critical value at each δ value of interest using the Robbins–Monro algorithm. Although the left-hand side of (17) involves a sum of $(K-1)!$ terms, this poses no real computational difficulty for typical values of K.

In order for an optimal decision rule to protect the FWER over the whole parameter space, coefficients of all elements of $\hat{\theta}_1$ must be non-negative in each term $\hat{\theta}_1^T \lambda(\phi)$. Because the smallest coefficient is $m_1/\{(K+1)\sigma^2\}$ times

$$(K+1)\gamma_1 - (\gamma_1 + \ldots + \gamma_{K-1} + 1), \tag{18}$$

we simply require that the expression (18) is greater than or equal to zero. If the γ_js are equally spaced between γ_1 and 1, this condition reduces to $\gamma_1 \geq K/(K+2)$, while a sufficient condition for any pattern of γ_js is $\gamma_1 \geq (K-1)/K$. Finally, we appeal to the argument of Section 3 to claim that a rule of the form (17) also protects the FWER when some of the elements of θ are greater than zero.

Figure 3 shows power, averaged over permutations of θ, when the methods of Section 3 are applied to the example of Section 4.

In panels (a) and (b), the γ_js are equally spaced between 0.3 and 1. As the form of the optimal rule varies with δ, each point on the power curve evaluates the rule maximizing power at that particular value of δ. As some $\hat{\theta}_{1,j}$s have negative weights in (17), these rules do not provide strong control of the FWER and we label them as 'pseudo-optimal'. The curve sets an upper bound for the power that can be attained and we deduce that the TSE rule and the two BK inverse normal combination tests have close to the maximum possible power. Indeed, the performance of these three procedures is impressive in view of the fact that they do not have the flexibility of the 'pseudo-optimal' rules to adapt to δ. In panels (c) and (d), where the γ_js are equally spaced between 0.75 and 1, $\gamma_1 > K/(K+2)$ so the rules given by (17) attach
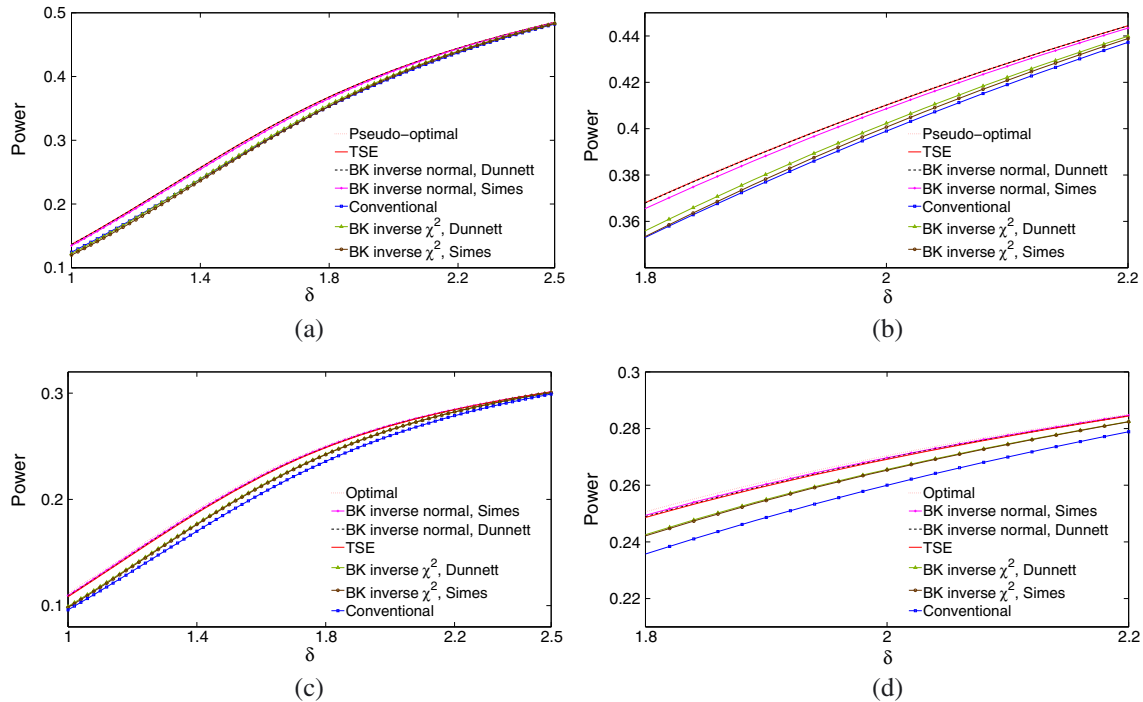
Figure 3. Power achieved by decision rules in the example of Section 4 when $\boldsymbol{\theta}$ is a random permutation (a) - (b) of $(0.3, 0.475, 0.65, 0.825, 1)\,\delta$ and (c) - (d) of $(0.75, 0.8125, 0.875, 0.9375, 1)\,\delta$. Decision rules are listed in order of decreasing power. Results are based on 1 million simulations for each scenario.

positive weights to all θ_js and we take them to be truly optimal. Again, the TSE rule and the two BK inverse normal combination tests have close to maximum power. The efficiency of the inverse normal combination test using Simes' rule in these examples indicates that the treatment effects of sub-optimal treatments are now sufficiently high that it is beneficial for the final decision rule to 'borrow strength' from $\hat{\theta}_{1,i}$, $i \neq i^\star$.

5.3. Optimizing power under dose–response assumptions

We now consider the situation where investigators suspect that a particular pattern of treatment effects may occur but these views are not held strongly enough to change the form of the study design from that described in Section 2. We shall consider the case where the treatment effect is expected to increase steadily with dose but side-effects or poorer compliance at higher doses may disrupt this relationship. It is of interest to know whether using such information about the likely pattern of treatment effects can lead to a significant increase in power.

We capture this somewhat equivocal view about possible treatment effects by formulating a Bayes decision problem with a special prior distribution. Let $0 \leqslant \gamma_1 < \ldots < \gamma_{K-1} < 1$ be specified and suppose the parameter vector $\boldsymbol{\theta}$ is either $\mathbf{0}$ or a permutation of $(\gamma_1, \ldots, \gamma_{K-1}, 1)\,\delta$. We assign prior probability $1/(K+1)$ to $\boldsymbol{\theta} = \mathbf{0}$ and allocate probability $1/(K+1)$ to each of the cases $\theta_i = \delta$, $i = 1, \ldots, K$. For $i = K$, the maximum treatment effect is $\theta_K = \delta$, and we assign all the probability $1/(K+1)$ to the case $\boldsymbol{\theta} = (\gamma_1, \ldots, \gamma_{K-1}, 1)\,\delta = \boldsymbol{\theta}_{ord}$, say, in line with the assumption that treatment effects increase with dose. For $i < K$, the maximum effect is not at the maximum dose, and, because the pattern cannot be monotone, we divide the prior probability $1/(K+1)$ evenly across the $(K-1)!$ permutations of $(\gamma_1, \ldots, \gamma_{K-1}, 1)\,\delta$ with $\theta_i = \delta$. Thus, the difference between this prior and that used in Section 5.2 is that the probabilities $1/\{(K+1)(K-1)!\}$ for vectors $\boldsymbol{\theta}$ in the subset Q_K are re-allocated to the single vector $\boldsymbol{\theta}_{ord}$, capturing the desired knowledge about the order of treatment effects in this case.

We define the same loss function as in Section 5.2 and find the Bayes optimal decision rule. The choice of prior implies that the Bayes rule rejects $H_{0,i^\star}$ if

$$LR\left(D_{i^\star}; \boldsymbol{\theta}_{ord}, \mathbf{0}\right) > \frac{c_1}{c_2} \text{ for } i^\star = K$$

and if

$$\frac{1}{(K-1)!} \sum_{\boldsymbol{\phi} \in Q_{i^\star}} LR\left(D_{i^\star}; \boldsymbol{\phi}, \mathbf{0}\right) > \frac{c_1}{c_2} \text{ for } i^\star \neq K.$$

As in Section 5.2, the likelihood ratio for data $D_{i^\star}$ can be written as

$$LR\left(D_{i^\star}; \boldsymbol{\phi}, \mathbf{0}\right) = \exp\left\{\hat{\boldsymbol{\theta}}_1^T \lambda(\boldsymbol{\phi})\right\} \exp\left\{\frac{m_2}{2\sigma^2} \hat{\theta}_{2,i^\star} \delta\right\} \exp\{-(h + m_2/4)(\delta^2/\sigma^2)\},$$

where $\lambda(\boldsymbol{\phi})$ is as defined in (15) and this is used with vectors $\boldsymbol{\phi} \in Q_{i^\star}$ for $i^\star \neq K$ and with $\boldsymbol{\phi} = \boldsymbol{\theta}_{ord}$ for $i^\star = K$.

The form of the Bayes rule depends on δ so that no uniformly most powerful test exists for treatment effect configuration $\boldsymbol{\theta}_{ord}$. At each positive value of δ, the Robbins–Monro algorithm can be used to find the appropriate choice of c_1/c_2 that gives an optimal decision rule with type I error rate α at $\boldsymbol{\theta} = \mathbf{0}$. In order for optimal decision rules to protect the FWER over the whole parameter space, coefficients of elements of $\hat{\boldsymbol{\theta}}_1$ must be non-negative in each term $\hat{\boldsymbol{\theta}}_1^T \lambda(\boldsymbol{\phi})$ for $\boldsymbol{\phi} \in Q_{i^\star}$ or $\boldsymbol{\phi} = \boldsymbol{\theta}_{ord}$, so $\gamma_1, \ldots, \gamma_K$ must satisfy the same conditions discussed in Section 5.2.

We have calculated power curves for decision rules of the previous form when $\boldsymbol{\theta} = \boldsymbol{\theta}_{ord}$ and the values of $\gamma_1, \ldots, \gamma_{K-1}$ are as in cases (a) and (c) of Figure 3. When the correct treatment, $i^\star = K$, is selected, the decision depends on $LR(D_{i^\star}; \boldsymbol{\theta}_{ord}, \mathbf{0})$ and so takes full advantage of the monotonicity assumption. We compared power under $\boldsymbol{\theta} = \boldsymbol{\theta}_{ord}$ with that of the optimum rules with no monotonicity assumption, derived in Section 5.2. In case (a), where effect sizes range from 0.3δ to δ, the maximum increase in power from use of dose–response information is 0.005; although the conditions for strong control of the FWER over the whole parameter space are not met, this is the case for both types of procedure so comparisons are fair. In case (c), effect sizes are closer, all tests control the FWER strongly, and the maximum increase in power is much smaller at 0.0005 (coupling of simulations of the different methods implies that this difference is still estimated reliably). The other six methods are unaffected by assumptions about the possible monotonicity of $\boldsymbol{\theta}$. However, because these assumptions lead to such small improvements, power curves for the new optimal rules are barely distinguishable from those shown in Figure 3 for the pseudo-optimal rules in (a) and the optimal rules in (c), and the TSE rule and both inverse normal rules remain very close to optimal.

Bretz *et al.* [23] propose methods that accommodate uncertain information about a dose–response curve by assuming that this curve belongs to a specified set of models $\mathcal{M}$. In their multiple comparison procedures with modelling techniques (MCP-Mod) approach, they define a test statistic T_m appropriate to each model $m \in \mathcal{M}$ and use $\max_m T_m$ as a global statistic to test for a positive dose–response relationship. The adjusted p-value is calculated using the joint distribution of the $T_m, m \in \mathcal{M}$, when the treatment effect is zero at all doses, that is, $\boldsymbol{\theta} = \mathbf{0}$ in our notation. Because the T_m are weighted sums of mean responses at each dose and some means can have negative coefficients, the FWER is not controlled strongly for all treatment effect vectors $\boldsymbol{\theta}$. Assuming that all treatment effects have the same sign resolves this problem: the same assumption would justify use of the 'pseudo-optimal' tests in case (a).

Our results show that robustly efficient methods such as the TSE rule achieve most of the potential gains from additional dose–response assumptions: the parallel in the setting of Bretz *et al.* [23] would be to take the maximum observed effect over all doses as the global test statistic. Bretz *et al.* found their method to have comparable power to a certain likelihood ratio test in many cases. Their method has a noticeable advantage when the effect size decreases at high doses, which is to be expected as the likelihood ratio test relies on a monotonicity assumption: the TSE rule makes no such assumption and should not be misled in such cases.

We acknowledge that our setting differs from that of Bretz *et al.* [23] in having two stages, and the gains from model information in stage 1 become diluted in the overall power. Also, Bretz *et al.* [23] made further use of their modelling framework by identifying the model producing the maximum T_m and using this model to select the minimum dose achieving a certain specified effect size for further testing. We shall return to discussion of such objectives in Section 9.

6. Relative efficiencies of data combination rules

We can express the power differences between decision rules in terms of the sample size needed to achieve a specific power. With the design of Section 2 and group sizes m_1 and m_2, we have derived optimal

Table II. Efficiencies of six decision rules when the treatment vector is a permutation of (a) $(0, \ldots, 0, 2)$, (b) $(0.5, \ldots, 0.5, 2)$, (c) $(1, \ldots, 1, 2)$, (d) $(1.5, \ldots, 1.5, 2)$, (e) $(0.6, 0.95, 1.3, 1.65, 2)$, (f) $(1.5, 1.625, 1.75, 1.875, 2)$, (g) $(0.6, 0.95, 1.3, 1.65, 2)$ and (h) $(1.5, 1.625, 1.75, 1.875, 2)$.

| | Treatment effect vector | | | | | | | |
Combination rule	a	b	c	d	e[1]	f	g[1]	h
TSE	100	100	100	99	100	98	99	98
BK inverse normal, Dunnett	100	100	100	99	100	99	99	99
BK inverse normal, Simes	95	98	99	99	99	99	98	99
BK inverse χ^2, Dunnett	95	96	96	96	96	95	95	95
BK inverse χ^2, Simes	93	95	96	96	95	95	95	95
Conventional test	97	97	96	93	94	90	94	90

In cases (g) and (h), optimal decision rules use information about the order of the elements of $\boldsymbol{\theta}$. Group sizes are $m_1 = 28$ and $m_2 = 140$. Results are based on 1 million simulations.
[1]Comparisons in (e) and (g) are with pseudo-optimal decision rules, which do not provide strong control of the familywise error rate, so entries are lower bounds on actual efficiencies.

Table III. Efficiencies of six decision rules when the treatment vector is a permutation of (a) $(0, \ldots, 0, 2)$, (b) $(0.5, \ldots, 0.5, 2)$, (c) $(1, \ldots, 1, 2)$, (d) $(1.5, \ldots, 1.5, 2)$, (e) $(0.6, 0.95, 1.3, 1.65, 2)$, (f) $(1.5, 1.625, 1.75, 1.875, 2)$, (g) $(0.6, 0.95, 1.3, 1.65, 2)$ and (h) $(1.5, 1.625, 1.75, 1.875, 2)$.

| | Treatment effect vector | | | | | | | |
Combination rule	a	b	c	d	e[1]	f	g[1]	h
TSE	100	100	100	99	99	98	98	97
BK inverse normal, Dunnett	99	99	100	99	99	98	98	98
BK inverse normal, Simes	91	95	98	99	98	99	97	98
BK inverse χ^2, Dunnett	96	97	97	97	97	96	95	96
BK inverse χ^2, Simes	93	95	97	97	96	96	95	96
Conventional test	89	89	89	85	86	82	85	82

In cases (g) and (h), optimal decision rules use information about the order of the elements of $\boldsymbol{\theta}$. Group sizes are $m_1 = 56$ and $m_2 = 112$. Results are based on 1 million simulations.
[1]Comparisons in (e) and (g) are with pseudo-optimal decision rules, which do not provide strong control of the familywise error rate, so entries are lower bounds on actual efficiencies.

decision rules under particular assumptions about the vector of treatment effects, $\boldsymbol{\theta}$. Suppose the optimal rule achieves power $1 - \beta$ for a given form of $\boldsymbol{\theta}$ with maximum treatment effect δ. If another rule requires group sizes to be increased to ρm_1 and ρm_2 in order to achieve the same power, the relative efficiency of this rule, expressed as a percentage, is $100/\rho$.

We have calculated the efficiency of decision rules applied to the example of Section 4 where $K = 5$, $\alpha = 0.025$, $m_1 = 28$ and $m_2 = 140$. Table II lists relative efficiencies of the six decision rules of Section 3 for eight configurations of $\boldsymbol{\theta}$ in which the highest treatment effect is $\delta = 2$ (fixing power at a different value of δ has only a small effect on our conclusions).

Cases (a) to (d) are for $\boldsymbol{\theta}$ of the form considered in Section 5.1. The TSE rule is optimal for case (c), and, in line with the conjecture made in Section 5.1, we also treat it as being optimal for cases (a) and (b). The form of $\boldsymbol{\theta}$ in cases (e) and (f) is that considered in Section 5.2. In case (e), we calculated efficiency relative to the 'pseudo-optimal' decision rule. This rule does not control the FWER strongly for all $\boldsymbol{\theta}$, but it provides an upper bound on the attainable power, and this is a rather tight upper bound as two rules that do protect FWER have close to 100% efficiency. Cases (g) and (h) concern the situation of Section 5.3 where there is partial information about the order of treatment effects; in (g), we are only able to derive a 'pseudo-optimal' decision rule, and we report efficiency relative to the upper bound this rule provides.

We have carried out the same efficiency assessments with group sizes $m_1 = 56$ and $m_2 = 112$, and the parallel results are presented in Table III. Here, the conventional procedure is less efficient, which is to be expected because more patients are treated in stage 1 and there is greater potential benefit in using their data in the final analysis. The inverse χ^2 tests, which give equal weight to stage 1 and stage 2 data summaries, fare better with these values of m_1 and m_2. However, in all cases where a Dunnett test is used for intersection hypotheses, the inverse χ^2 combination test still lags behind the inverse normal rule.

There is just one example (case (a) of Table III) of a Simes test, for which the inverse χ^2 combination test is superior to the inverse normal, but then the Dunnett test with an inverse normal combination rule is far superior.

Results in Tables II and III show that use of stage 1 data in the final analysis can lead to worthwhile gains in efficiency over the conventional test based on stage 2 data alone. The methods of choice are the TSE procedure and the BK inverse normal combination test using Dunnett tests for intersection hypotheses: these decision rules attain close to the maximum possible power in all scenarios, with relative efficiency of at least 97% in all cases and 99% or more in the majority of cases. The BK inverse normal method using Simes' tests for intersection hypotheses performs poorly when there is a single treatment with a high effect size, but this rule can be close to optimal in other situations. We do not recommend decision rules based on inverse χ^2 combination tests: in all but extreme cases, these are dominated by the rules using inverse normal combination tests (using Dunnett or Simes p-values for intersection hypotheses) and, in some situations, gain no advantage at all from the use of stage 1 data.

## 7. Optimal division of sample size between phases II and III

Suppose that in the previous example $K = 5$ and $\alpha = 0.025$ are fixed but the group sizes m_1 and m_2 can be chosen freely subject to an upper bound on the total sample size $(K + 1)m_1 + 2m_2$. We shall restrict attention here to the robustly efficient TSE procedure. For this decision rule, we have calculated values of m_1 and m_2 that optimize power when the total sample size is fixed at 448, as in the example of Section 4. Figure 4 plots the value of m_1 that maximizes power, as defined in (3), for a variety of treatment means θ.

Optimal values of m_1 were found by a direct search over the integers between 1 and 74; the accuracy of comparisons was enhanced by using the same sequence of pseudo-random numbers to simulate the power of each design. Thall, Simon and Ellenberg [1] report design settings that minimize the expected

(a)

(b)
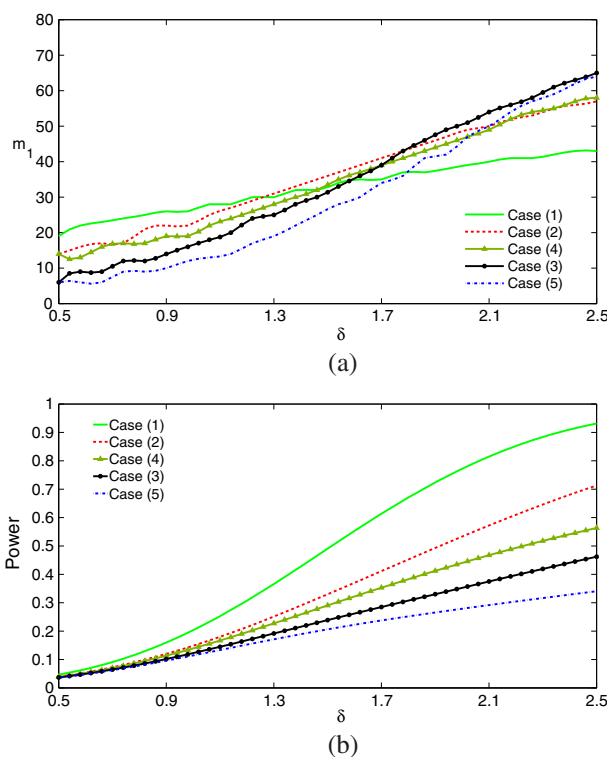
Figure 4. (a) Stage 1 group sizes maximizing the power of the TSE procedure when the total sample size is fixed at 448 and θ is a random permutation (1) of $(0, 0, 0, 0, 1)\delta$, (2) of $(0.5, 0.5, 0.5, 0.5, 1)\delta$, (3) of $(0.75, 0.75, 0.75, 0.75, 1)\delta$, (4) of $(0.3, 0.475, 0.65, 0.825, 1)\delta$ and (5) of $(0.75, 0.8125, 0.875, 0.9375, 1)\delta$. (b) Power achieved by the optimized TSE procedures. Decision rules are listed in order of decreasing power. Designs are specified with $K = 5$, $\ell = 0$, $\sigma = 5.0$ and $\alpha = 0.025$. Results are based on 1 million simulations for each scenario.
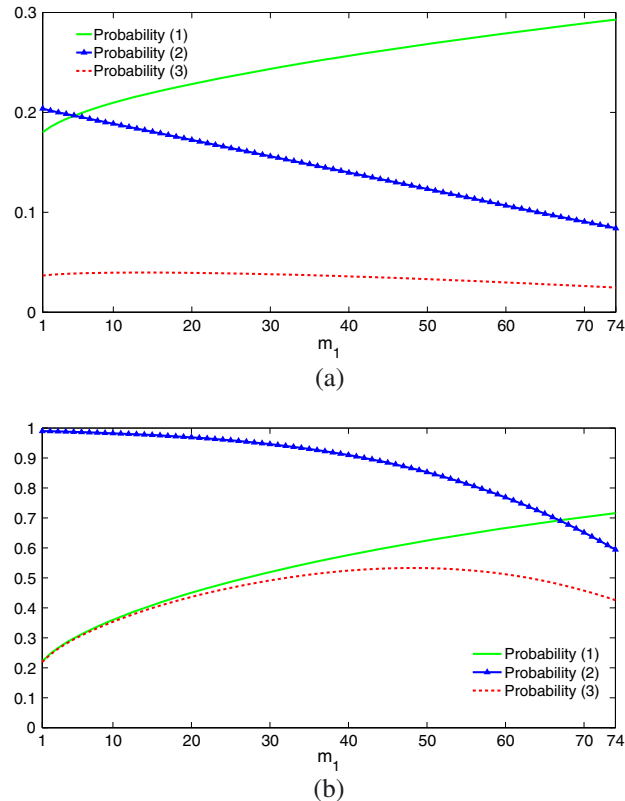
Figure 5. Operating characteristics of the TSE procedure with fixed total sample size $N = 448$ when $\theta = (0.5, \ldots, 0.5, 1) \delta$ with (a) $\delta = 0.5$ and (b) $\delta = 2.0$. Plotted probabilities are as follows: (1) $\mathrm{pr}\{\text{select treatment } K; \theta\}$, (2) $\mathrm{pr}\{\text{reject } H_{0,K} | \text{treatment } K \text{ selected}; \theta\}$ and (3) the product of these, namely $\mathrm{pr}\{\text{select treatment } K \text{ and reject } H_{0,K}; \theta\}$. Designs have $K = 5$, $\ell = 0$, $\sigma = 5.0$ and $\alpha = 0.025$. Results are based on 1 million simulations for each scenario.

sample size of the TSE procedure when θ has the form $(\gamma_1, \ldots, \gamma_1, 1)\delta$. In Figure 4, we present results for a wider variety of configurations for θ. Given the robust efficiency of the TSE procedure, we expect these values of m_1 will also be close to optimal for the optimal tests of Section 5 and for the inverse normal combination rule with Dunnett p-values.

The optimum m_1 varies with both the shape of the vector θ and the scale factor δ. When selecting m_1, we trade accuracy in selecting the best treatment, i_{max}, in stage 1 with sample size for comparing this treatment against control in stage 2. Figure 5 illustrates the consequences of this trade-off when $\theta = (0.5, \ldots, 0.5, 1) \delta$ with $\delta = 0.5$ and $\delta = 2.0$.

Initially, power increases with m_1 because of the increased selection accuracy. However, increasing m_1 also reduces the total number of observations on treatment i_{max} when this treatment is selected, and this eventually results in a loss of overall power to reject $H_{0,i_{max}}$. The same considerations help explain why the optimum m_1 increases with δ: when δ is large, modest values of m_2 still give a high conditional probability of rejecting $H_{0,i_{max}}$ when treatment i_{max} is selected, thus we can take a larger value of m_1 to improve the probability of selecting treatment i_{max} in stage 1. In the example, the optimum m_1 for $\delta = 0.5$ is 14, while that for $\delta = 2.0$ is 49.

Because optimum values of m_1 can be below 10 or above 60, we conclude that no single choice is close to ideal in all scenarios. Rather, investigators should consider the most likely scenarios, not necessarily of the form $(\gamma, \ldots, \gamma, 1) \delta$, for their trial and choose group sizes that will give the best average power across these cases. Such scenarios could be established by conducting a Bayesian prior elicitation meeting ahead of the phase II/III trial to explore experts' prior opinion on the efficacy of the K treatments relative to control. With a given decision rule, it is straightforward to run simulations to compare different choices of m_1 and choose a value that will provide good power under an anticipated set of treatment effects. Note that our definition of power gives no reward for selecting a good second-best treatment and rejecting its null hypothesis although, in practice, this might be considered a successful outcome. This is not a
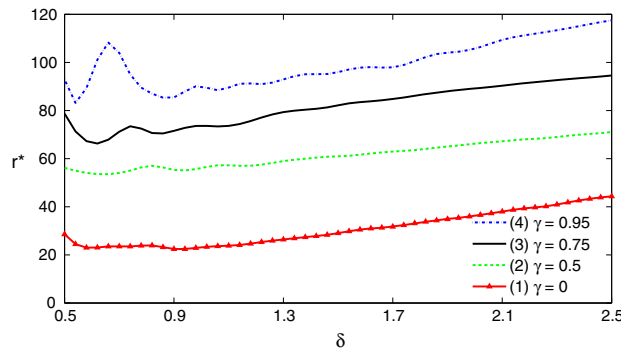
Figure 6. Percentage value, $r^\star$, of stage 1 data used in the TSE rule, relative to additional stage 2 observations for $\theta = (\gamma, \ldots, \gamma, 1)\,\delta$ with (1) $\gamma = 0$, (2) $\gamma = 0.5$, (3) $\gamma = 0.75$ and (4) $\gamma = 0.95$. Designs are as specified in Section 2 with $K = 5$, fixed total sample size $N = 448$, (m_1, m_2) chosen to maximize the power of the TSE procedure, $\ell = 0$, $\sigma = 5.0$ and $\alpha = 0.025$. Results are based on 1 million simulations for each scenario.

major issue for most of the configurations of θ described in Figure 4, where treatment i_{max} is superior to its nearest competitor by some margin. However, when considering cases where several treatments are competitive, it may be appropriate to use an alternative definition of power and, for example, choose group sizes to maximize the probability of selecting any treatment i with a treatment effect within 10% of the largest treatment effect and then rejecting $H_{0,i}$.

8. Value of phase II data in the final analysis

The relative efficiencies in Tables II and III are based on comparing trial designs where both stage 1 and stage 2 group sizes, m_1 and m_2, are multiplied by a common factor. Another way to assess the benefits of a seamless phase II/III design is to determine how many additional phase III observations would be needed to achieve the increase in power gained by using phase II data in the final analysis. We shall make this assessment when the TSE decision rule is used.

For a given vector of treatment effects θ, we can calculate the stage 2 sample size $\tilde{m}_2$ such that selecting a treatment based on m_1 stage 1 observations and then applying a conventional test with $\tilde{m}_2$ stage 2 observations on the selected treatment and control gives the same overall power as the TSE rule with group sizes, m_1 and m_2. Thus, the $2\,m_1$ stage 1 observations on treatment $i^\star$ and control in the TSE decision rule have the same benefit as an additional $2\,(\tilde{m}_2 - m_2)$ stage 2 observations for the conventional test. We express the percentage value of the stage 1 observations on treatment $i^\star$ and control relative to extra stage 2 observations in a conventional design as

$$r^\star = 100\,(\tilde{m}_2 - m_2)\,/(m_1).$$

Figure 6 shows plots of $r^\star$ against δ for the example of Section 4 when θ is of the form $(\gamma, \ldots, \gamma, 1)\,\delta$ and, for each value of δ, m_1 and m_2 are chosen to maximize the power of the TSE procedure subject to a fixed total sample size $N = 448$.

Results vary with the form of θ and values of $r^\star$ at $\delta = 1$ rise from 22 when $\gamma = 0$ to almost 100 when $\gamma = 0.95$. The critical value in the TSE decision rule, calculated under $\theta = (0, \ldots, 0)$, adjusts for multiple testing and so avoids any bias from selecting the treatment with the best stage 1 results. When $\gamma = 0$, the treatment with effect size δ is very likely to be chosen and the adjustment for multiple testing reduces power, leading to a low $r^\star$. For higher values of γ, the treatment with effect size δ must outperform its rivals in order to be selected after stage 1: it is then likely to have an above average estimate, $\hat{\theta}_{1,i}$, and this balances the effect of the multiplicity adjustment. As $\gamma \to 1$, $r^\star$ can exceed 100, indicating that information from all K treatments, not just treatment $i^\star$ and the control, contributes to the final decision.

In a trial involving multiple treatments or several doses of a single treatment, one might expect the treatment effects to be spread out between zero and the highest value. Thus, of the scenarios in Figure 6, case (2), with $\gamma = 0.5$, represents the most plausible situation. In this case, the stage 1 responses on treatment $i^\star$ have an equivalent value to around 60% of their number of stage 2 observations. Recognizing

the usual uncertainty about likely treatment effects, we suggest that stage 1 data on treatment $i^\star$ should typically be viewed as offering around 50% to 70% of their face value as stage 2 observations in many situations. Since the TSE rule is close to optimal under general configurations of θ, the results of this section should also provide an accurate reflection of the value of using stage 1 data in other efficient decision rules.

9. Discussion

We have sought optimal data combination rules for seamless two-stage designs, making the problem tractable by focusing on one configuration of θ at a time. In many situations, the optimal rules that we have derived control the FWER strongly; in other cases, our results provide upper bounds on the attainable power that serve as benchmarks for other procedures. We have identified two decision rules that we would recommend for use in practice, namely the TSE procedure and the BK inverse normal combination test using Dunnett p-values for intersection hypotheses. These rules are highly efficient in a variety of situations and tailoring the decision rule to the particular configuration of θ can give only very small additional efficiency. Furthermore, because both rules can be expressed as closed testing procedures, they can be used flexibly, still controlling the FWER when additional criteria are used to select a treatment for stage 2. We have also demonstrated how observations can be divided between the two stages to maximize power in a given scenario. Comparisons with the conventional practice of using phase III data alone in a final hypothesis test confirm that combining data across phases can improve power: for typical vectors of treatment effects, the increase in power is comparable with that achieved by adding 50% to 70% of the subjects on two stage 1 treatments to the stage 2 sample size and performing a conventional analysis. We have reached similar conclusions in simulations with different numbers of treatments and different sample sizes. In some situations, the benefits of data combination may be deemed insufficient to compensate for the planning and logistical effort involved in a seamless phase II/III trial; in others, particularly clinical trials for rare diseases, the power gained from stage 1 data may be deemed very worthwhile.

Sampson and Sill [2] derived a conditionally unbiased most powerful test for this problem. Their conditioning event is rather complex, as is the resulting test, and their procedure does not include stopping for futility after stage 1. If adapted to our problem, this method could not do better than our optimal rules for particular θ vectors. Bretz *et al.* [24] asked whether the approach of Sampson and Sill [2] could be extended to find an unconditionally unbiased most powerful test: our results show that different tests are optimal for different configurations of treatment effects, so this is not the case.

Extensions of the problem described in Section 2 have been proposed and studied. Optimizing procedures in more complex settings may not be feasible, but, to the extent that these new problems retain core elements of the basic problem, we expect our conclusions to remain relevant. As an example, Bischoff and Miller [25] consider the case of two treatments and a control with normal responses of unknown variance, and they tailor the design to minimize total expected sample size. Their test statistic combines estimates of the effect of the selected treatment from stages 1 and 2 in the same way as the TSE rule, so our results suggest that using this estimate in a t-statistic will give good power.

Stallard and Todd [26] consider testing multiple treatments against a control in a sequential design in which the most promising treatment is selected at the first analysis and subsequent interim analyses allow early stopping for a final decision. Calculations follow similar lines to those of standard group sequential tests; see, for example, [27, Ch. 19]. With just two analyses, this method reduces to the TSE procedure, and we conclude that it combines data before and after treatment selection in an efficient way. The approach can be extended in various directions: these include allowing treatments to be dropped over several analyses [28, 29] or basing the treatment selection on a short-term endpoint [30].

Magirr *et al.* [31] propose a new type of trial design for comparing multiple treatments with a control at multiple analyses. As in the TSE procedure, decision rules are defined in terms of the means of cumulative data on each treatment and the control. An innovative approach to computation makes it feasible to create designs comparing many treatments with several interim analyses. Wason and Jaki [32] use numerical search methods to optimize features of these designs, including the allocation ratio between active treatments and the control.

DiScala and Glimm [33] consider an adaptive trial design with a survival endpoint, in which treatment selection is based on a more rapidly observed event. When analysing follow-up data on subjects who have already contributed to a decision about treatment choice, there is a danger of type I error

inflation ([34]) but the methods of Jenkins *et al.* [35] and Irle and Schäfer [36] can be used to avoid this problem.

In Section 5.3, we considered the case where 'treatments' represent dose levels and a dose–response model may be used. The smaller risk of safety problems at lower dose levels motivates the decision in the MCP-mod procedure of Bretz *et al.* [23] to select the lowest dose that produces a specified treatment effect, even when safety responses are not considered directly. Some authors have considered treatment selection and testing based on both efficacy and safety data: Liu and Pledger [15] refer to safety outcomes when deciding on the treatment to take forward from the first stage of a seamless phase II/III design; König *et al.* [37] and Kimani *et al.* [38] propose further procedures for this case. If a model for efficacy, and possibly safety, is specified and the benefits of demonstrating a new treatment to be effective are quantified, the question of how best to design two (or more) phases of a drug development programme can be clearly stated. The problem is challenging, even without the combination of data across stages of a seamless design. This is an area of considerable current activity that is starting to produce important insights; see, for example, [39, 40].

Appendix A

Connection between optimal data combination rules and parameter estimates for a linear model

We consider the case of Section 5.1 where θ is a permutation of $(\gamma, \ldots, \gamma, 1)\delta$ for known $\gamma \in (0, 1)$. Let $\hat{\mu}_{1,i}$, $i = 0, 1, \ldots, K$, denote the first stage sample means on the control and K experimental treatments, so the estimated treatment effects in stage 1 are $\hat{\theta}_{1,i} = \hat{\mu}_{1,i} - \hat{\mu}_{1,0}$, $i = 1, \ldots, K$. For simplicity, suppose $\hat{\theta}_{1,K}$ is the largest stage 1 estimate and treatment $i^\star = K$ is compared against control in stage 2.

The optimal decision rule (8) in Section 5.1 is based on the likelihood ratio of the combined stage 1 and stage 2 data under $\theta = \xi_{i^\star} = (\gamma, \ldots, \gamma, 1)\delta$ and $\theta = 0$, and this can be written as the product of separate terms for stage 1 and stage 2 data. The stage 1 estimates $\hat{\theta}_{1,1}, \ldots, \hat{\theta}_{1,K}$ follow a normal linear model with a single unknown parameter δ; also, given $i^\star = K$, the stage 2 estimate $\hat{\theta}_{1,K}$ is normally distributed with mean δ. Let $\hat{\delta}_1$ and $\hat{\delta}_2$ be the maximum likelihood estimates of δ based on stages 1 and 2 data, respectively, with variances $\mathrm{var}\left(\hat{\delta}_1\right)$ and $\mathrm{var}\left(\hat{\delta}_2\right)$. Standard algebra shows that the log likelihood ratio between $\theta = \xi_{i^\star}$ and $\theta = 0$ for stage 1 data is a constant plus $\delta\,\hat{\delta}_1/\mathrm{var}\left(\hat{\delta}_1\right)$, and for stage 2, data it is a constant plus $\delta\,\hat{\delta}_2/\mathrm{var}\left(\hat{\delta}_2\right)$. Combining these terms, we find the log likelihood ratio based on the stage 1 and stage 2 data together is an increasing function of

$$\delta\left(\frac{\hat{\delta}_1}{\mathrm{var}\left(\hat{\delta}_1\right)} + \frac{\hat{\delta}_2}{\mathrm{var}\left(\hat{\delta}_2\right)}\right),$$

a multiple of the maximum likelihood estimate of δ for the pooled stages 1 and 2 data. It follows that the first-stage estimates $\hat{\theta}_{1,1} \ldots, \hat{\theta}_{1,K}$ contribute to the optimal decision rule with weights proportional to their weights in $\hat{\delta}_1$, the estimate of δ obtained by fitting a normal linear model to $\hat{\theta}_{1,1}, \ldots, \hat{\theta}_{1,K}$.

At this point, it helps to represent the stage 1 data as the sample means $\hat{\mu}_{1,i}$, $i = 0, 1, \ldots, K$, on the control and K experimental treatments. In the case we are considering, $\theta = (\gamma, \ldots, \gamma, 1)\delta$ and $\hat{\mu}_{1,0}, \ldots, \hat{\mu}_{1,K}$ follow a linear regression model with $E\left(\hat{\mu}_{1,i}\right) = \alpha + \delta x_i$, where $x_0 = 0$, $x_1 = \ldots = x_{K-1} = \gamma$ and $x_K = 1$. The estimate $\hat{\delta}_1$ is a linear combination of $\hat{\mu}_{1,0}, \ldots, \hat{\mu}_{1,K}$ with weights summing to zero, so

$$\begin{aligned}
\hat{\delta}_1 &= w_0\,\hat{\mu}_{1,0} + w_1\,\hat{\mu}_{1,1} + \ldots + w_K\,\hat{\mu}_{1,K} \\
&= w_1\,\hat{\theta}_{1,1} + \ldots + w_K\hat{\theta}_{1,K}
\end{aligned}$$

and we see $\hat{\theta}_{1,1}, \ldots, \hat{\theta}_{1,K}$ contribute to $\hat{\delta}_1$ with the same weights as $\hat{\mu}_{1,1}, \ldots, \hat{\mu}_{1,K}$. If $\gamma = 0$, so $\theta = (0, \ldots, 0, 1)\delta$, it is straightforward to show

$$\hat{\delta}_1 = \hat{\mu}_{1,K} - \frac{1}{K}\sum_{i=0}^{K-1}\hat{\mu}_{1,i} = \hat{\theta}_{1,K} - \frac{1}{K}\sum_{i=1}^{K-1}\hat{\theta}_{1,i},$$

in agreement with the contributions of $\hat{\theta}_{1,0}, \ldots, \hat{\theta}_{1,K}$ in the decision rule (11), which can be regarded as a test of $H_0: \delta = 0$. If $\gamma = 0.5$, $\hat{\mu}_{1,1}, \ldots, \hat{\mu}_{1,K-1}$ make no contribution to the estimate of the slope δ in the

linear regression $E(\hat{\mu}_{1,i}) = \alpha + \delta\, x_i$ and so have zero weight in $\hat{\delta}_1$, in keeping with their absence from the TSE decision rule, which is optimal in this case. For other values of γ, inspection of the linear regression model shows that $\hat{\mu}_{1,1}, \ldots, \hat{\mu}_{1,K-1}$ contribute to $\hat{\delta}_1$ with negative weights if $\gamma < 0.5$ and with positive weights if $\gamma > 0.5$, which agrees with the pattern of weights for unselected first-stage treatments in the data combination rule (11).

Acknowledgements

References

1. Thall PF, Simon R, Ellenberg SS. Two-stage selection and testing designs for comparative clinical trials. *Biometrika* 1988; **75**:303–310.
2. Sampson AR, Sill MW. Drop-the-losers design: normal case. *Biometrical Journal* 2005; **3**:257–268.
3. Bretz F, Schmidli H, König F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts (with discussion). *Biometrical Journal* 2006; **48**:623–634.
4. Schmidli H, Bretz F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: applications and practical considerations. *Biometrical Journal* 2006; **48**:635–643.
5. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
6. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994; **50**:1029–1041.
7. Jennison C, Turnbull BW. Discussion of papers on "Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim". *Biometrical Journal* 2006; **48**:650–655.
8. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 1999; **18**:1833–1848.
9. Fisher RA. *Statistical Methods for Research Workers* 4th ed. Oliver and Boyd: London, 1932.
10. Mosteller F, Bush RR. Selected quantitative techniques. In *Handbook of Social Psychology*, Lindsey G (ed.), Vol. 1. Addison-Wesley: Cambridge, MA, 1954; 289–334.
11. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics* 1999; **55**:1286–1290.
12. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **73**:751–754.
13. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* 1955; **50**:1096–1121.
14. Sarkar SK, Chang CK. The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* 1997; **92**:1601–1608.
15. Liu Q, Pledger GW. Phase 2 and 3 combination designs to accelerate drug development. *Journal of the American Statistical Association* 2005; **100**:493–502.
16. Jennison C, Turnbull BW. Adaptive seamless designs: selection and prospective testing of hypotheses. *Journal of Biopharmaceutical Statistics* 2007; **17**:1135–1161.
17. Eales JD, Jennison C. An improved method for deriving optimal one-sided group sequential tests. *Biometrika* 1992; **79**:13–24.
18. Barber S, Jennison C. Optimal asymmetric one-sided group sequential tests. *Biometrika* 2002; **89**:49–60.
19. Banerjee A, Tsiatis AA. Adaptive two-stage designs in phase II clinical trials. *Statistics in Medicine* 2006; **25**:3382–3395.
20. Öhrn F, Jennison C. Optimal group sequential designs for simultaneous testing of superiority and non-inferiority. *Statistics in Medicine* 2010; **29**:743–759.
21. Hampson LV, Jennison C. Group sequential tests for delayed responses (with discussion). *Journal of the Royal Statistical Society B* 2013; **75**:3–54.
22. Robbins H, Monro S. A stochastic approximation method. *The Annals of Mathematical Statistics* 1951; **22**:400–407.
23. Bretz F, Pinheiro JC, Branson M. Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics* 2005; **61**:738–748.
24. Bretz F, Strassburger K, Maurer W. Dicussion of "Drop-the-losers design: Normal case". *Biometrical Journal* 2005; **3**:274–275.
25. Bischoff W, Miller F. Adaptive two-stage test procedures to find the best treatment in clinical trials. *Biometrika* 2005; **92**:197–212.
26. Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine* 2003; **22**:689–703.
27. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC: Boca Raton, 2000.
28. 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.
29. Stallard N, Friede T. A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine* 2008; **27**:6209–6227.

30. Stallard N. A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Statistics in Medicine* 2010; **29**:959–971.
31. Magirr D, Jaki T, Whitehead J. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika* 2012; **99**:494–501.
32. Wason J, Jaki T. Optimal design of multi-arm multi-stage trials. *Statistics in Medicine* 2012; **31**:4269–4279.
33. Di Scala L, Glimm E. Time-to-event analysis with treatment arm selection at interim. *Statistics in Medicine* 2011; **30**: 3067–3081.
34. Bauer P, Posch M. Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections (letter to the editor). *Statistics in Medicine* 2004; **23**:1333–1335.
35. Jenkins M, Stone A, Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics* 2011; **10**:347–356.
36. Irle S, Schäfer H. Interim design modifications in time-to-event studies. *Journal of the American Statistical Association* 2012; **107**:341–348.
37. König F, Bauer P, Brannath W. An adaptive hierarchical test procedure for selecting safe and efficient treatments. *Biometrical Journal* 2006; **48**:663–678.
38. Kimani PK, Stallard N, Hutton JL. Dose selection in seamless phase II/III clinical trials based on efficacy and safety. *Statistics in Medicine* 2009; **28**:917–936.
39. Patel N, Bolognese J, Chuang-Stein C, Hewitt D, Gammaitoni A, Pinheiro J. Designing phase 2 trials based on program-level considerations: a case study for neuropathic pain. *Drug Information Journal* 2012; **46**:439–454.
40. Marchenko O, Miller J, Parke T, Perovozskaya I, Quian J, Wang Y. Improving oncology clinical program by use of innovative designs and comparing them by simulations. *Therapeutic Innovation and Regulatory Science* 2013; **47**:602–612.

Supporting information

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