STATISTICS IN MEDICINE
Statist. Med. 2007; **00**:1–18

Dose Selection in Seamless Phase II/III Clinical Trials based on Efficacy and Toxicity

Peter K. Kimani[1,*,†], Nigel Stallard[2], Jane L. Hutton[1]

[1] *Department of Statistics, University of Warwick, U.K.*
[2] *Warwick Medical School, University of Warwick, U.K.*

SUMMARY

Seamless phase II/III clinical trials are attractive in development of new drugs because they accelerate the drug development process. Seamless phase II/III trials are carried out in two stages. After stage 1 (phase II stage), an interim analysis is performed and a decision is made on whether to proceed to stage 2 (phase III stage). If the decision is to continue with further testing, some dose selection procedure is used to determine the set of doses to be tested in stage 2. In this paper we propose a dose-selection procedure for binary outcomes in adaptive seamless phase II/III clinical trials that incorporates the dose-response relationship when the experimental treatments are different dose levels of the same drug, and explicitly incorporates both efficacy and toxicity. The choice of the doses to continue to stage 2 is made by comparing the predictive power of the potential sets of doses which might continue.

KEYWORDS: Seamless phase II/III clinical trials; Closure principle; Adaptive testing; Conditional power; Predictive power.

1. Introduction

In drug development, clinical trials are categorized into three phases. Phase I is the stage where the drug is first tested in human beings and the objective is to determine the safety of the new drug. Phase I trials are small and several dose levels are generally tested. If a safe dose (or dose range) is identified, the drug is then tested for efficacy in a small clinical trial. Such a trial is referred to as a phase II clinical trial and like phase I, often more than one dose level is tested. At the end of the phase II trial, a decision has to be made on the basis of efficacy and safety data regarding which dose(s) proceeds to the next stage of testing. The last stage of drug testing in human beings before submission for regulatory approval is the phase III clinical trial which is a large confirmatory trial for efficacy.

In order to reduce the time before approval of a new drug, there has been interest in combining different phases of a clinical trial. Trials which combine phase II and phase III into a single trial with a phase II stage and phase III stage are referred to as (seamless) phase

[*] Correspondence to: Peter Kimani, Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K.
[†] p.k.kimani@warwick.ac.uk

Received
Revised

II/III trials. Such trials are conducted in two stages. In stage 1 (phase II stage) of phase II/III trials, some experimental treatments are compared with a control treatment with the aim of determining the promising treatments. Sufficiently promising treatments then continue to stage 2 (phase III stage) along with the control treatment. At the end of the trial, data from both stages are used to assess the efficacy of the selected treatment(s). In this paper we propose a procedure for selecting the set of doses that proceed to stage 2 after stage 1. Thall *et al.* [1], Schaid *et al.* [2] and Stallard & Todd [3] among others have proposed methods in which the most promising treatment is selected for further testing for experiments with more than one stage. These authors consider distinct treatments, that may be different doses of a drug, but have not considered the dose-response relationship and the safety of the experimental treatments explicitly. The procedure that we propose incorporates dose-response relationship and safety explicitly for binary outcomes that are observable rapidly after administration of a treatment.

Bretz *et al.* [4] have described a method of analysis for phase II/III trials using a single outcome. This is described in detail in Section 2. Assuming this analysis for efficacy outcome, we obtain the set of stage 2 data for which at least one of the experimental doses which proceed to stage 2 is concluded to be effective after stage 2 given the results of stage 1. After defining the distribution of stage 2 data, we then obtain the probability of this set of data. We refer to this probability as the (combined) conditional power. To incorporate the dose-response relationship, we let the parameters of stage 2 data be given by some specified parametric dose-response curves. To penalize for toxicity, we partition the conditional power into probabilities of disjoint events where at least one effective dose is rejected and multiply each component by the indicator variable that doses in that event have probability of toxicity less than some maximum accepted level. To allow the prior knowledge influence decision of which doses continue to stage 2, we use Bayesian methodology by defining some prior distributions for the dose-response curves parameters and obtain the posterior probability of concluding at least one experimental dose is effective and safe by integrating over the parameter space. We refer to this probability as penalized predictive power and dose selection is made so as to optimise the penalized predictive power among the potential doses that proceed to stage 2.

In the next section, a review of adaptive analysis of phase II/III data with multiple experimental treatments is given. In Section 3, we give the expressions for the conditional power, the probability of concluding that at least one of the experimental doses that proceed to stage 2 is effective given the stage 1 data. A possible form of the prior distributions is given in Section 4. We examine our method of dose selection using simulated examples in Section 5. A discussion is given in Section 6.

2. Adaptive analysis of phase II/III data with multiple treatments

Building on the work of Bauer & Kieser [5], Bretz *et al.* [4] consider a seamless phase II/III trial in which a control treatment is compared to more than one experimental treatment using some hypothesis tests to determine if there is an effective treatment. They focus on the case in which there is a single endpoint ([4, 5]). In this section we review the work of Bretz *et al.* [4] and ways of obtaining the p-values required in this analysis. Their notation has been maintained and will be used in subsequent sections.

Suppose in an experiment $k(>1)$ experimental treatments are to be compared with a control treatment such that k null hypotheses $H_i : \theta_i = \theta_0$, $i = 1, ..., k$ comparing each experimental dose with the control treatment are of interest where θ_i and θ_0 respectively denote the measure of effectiveness for experimental treatment i and the control treatment. In order to control the familywise error rate (FWER) associated with testing the k pairwise null hypotheses at pre-specified level α, Bretz *et al.* [4] use the closure principle (CP) of Marcus *et al.* [6]. The CP considers the set of all intersection hypotheses constructed from the initial hypotheses of interest. Marcus *et al.* [6] refer to this set, denoted by $\mathcal{H}$, as the closure set. In the closure set H_{ij} denotes $H_i \cap H_j$, H_{ijl} denotes $H_i \cap H_j \cap H_l$ and so on for $i, j, l \in \{1, ..., k\}$. Using the CP, a null hypothesis H_i is rejected at FWER α if the subset of hypotheses in $\mathcal{H}$ which are included in H_i are all rejected at level α.

To combine evidence from the two stages, Bretz *et al.* [4] use the adaptive approach as described by Bauer & Köhne [7]. In adaptive testing, data from each stage are analysed separately and in order to make a single conclusion from the two stages, p-values obtained at the end of each stage are combined into a single value. Suppose the p-value obtained from testing a (null) hypothesis H at end of stage 1 is p_1 and the corresponding p-value at stage 2 is p_2. Assuming that p_1 and p_2 have independent Uniform$[0, 1]$ distributions under the null hypothesis, several combination procedures for p_1 and p_2 into a single p-value have been proposed but none is uniformly most powerful. Zaykin *et al.* [8] have reviewed some methods of combining the p-values. A commonly used method is the weighted inverse normal method in which the combined p-value is

$$C(p_1, p_2) = 1 - \Phi[w_1\Phi^{-1}(1 - p_1) + w_2\Phi^{-1}(1 - p_2)] \tag{1}$$

where $0 < w_i < 1,\ i = 1, 2$, are arbitrary weights subject to $w_1^2 + w_2^2 = 1$ and Φ is the standard normal distribution function.

Bretz *et al.* [4] propose combining the CP and adaptive testing so that a null hypothesis H_i $(i = 1, ..., k)$ is rejected at the end of stage 2 if all the combined p-values for all the hypotheses in $\mathcal{H}$ and contained in H_i are less than the pre-specified level of testing. For example, suppose there are three experimental treatments at stage 1 and let $p_{i,j}$ denote the p-value for testing hypothesis $H_j \in \mathcal{H}$ at stage i $(i = 1, 2)$, hypothesis H_1 is rejected at the end of stage 2 at level α if

$$\max\{C(p_{1,1}, p_{2,1}),\ C(p_{1,12}, p_{2,12}),\ C(p_{1,13}, p_{2,13}),\ C(p_{1,123}, p_{2,123})\} \leq \alpha.$$

To understand what happens if some treatments are dropped after stage 1, suppose for example that treatment 3 is dropped implying no data are available for treatment 3 at stage 2. Then the tests for intersection hypotheses H_{13} and H_{123} reduce to the tests for hypotheses H_1 and H_{12} respectively so that $p_{2,13} = p_{2,1}$ and $p_{2,123} = p_{2,12}$.

Bretz *et al.* [4] do not give details of how the p-values testing the hypotheses in $\mathcal{H}$ but Westfall & Wolfinger [9] provide a simplified discussion of some methods. The pairwise hypotheses may be tested using basic tests such as the chi-squared test for binary data and the t-test for continuous data. There are several tests for the intersection hypotheses (H_{ij}, H_{ijl}, etc) but some are specific to certain forms. For example Hotelling's T^2 described by Johnson & Wichern [10] test is valid for continuous data. Flexible tests that can be used for many forms of responses (normal, poisson, etc) are Bonferroni-Holm, Šidak-Holm and Simes test. Holm's procedure for testing multiple hypotheses has been described by Westfall & Young [11]. Suppose we wish to test a hypothesis of equality of the control treatment with m $(1 < m \leq k)$ experimental treatments. The Bonferroni-Holm adjusted p-value is given by $(m \times \text{minp})$ while Šidak-Holm adjusted p-value is given by $(1 - [1 - \text{minp}]^m)$ where minp is the minimum p-value of the individual component tests. The Simes p-value is given by $\min\{\frac{m}{i}p_{(i)}\},\ i = 1, ..., m$ where $p_{(i)}$ denote the ordered p-values.

3. Conditional power

The last section reviewed the method described by Bretz *et al.* [4] for analyzing seamless phase II/III trials. Assuming this analysis for efficacious outcome (we assume safety data are not considered in hypotheses tests), the objective of this paper is to select the best set of doses for testing in stage 2. In this section we give an expression for the probability of concluding at least one of the doses in the potential set of doses tested in stage 2 is effective given the results of stage 1. We will refer to this probability as the (combined) conditional power. To give the expression for conditional power, stage 2 data distribution and set of stage 2 data for which at least one of the experimental doses will be concluded effective after stage 2 are required. Before giving stage 2 data distribution and the expressions for conditional power, we give the setting of interest while introducing more notation.

Consider an experiment with $k_1(> 1)$ experimental treatments in stage 1 of which a subset remains for testing in stage 2. Suppose the sample size for stage 1 is fixed to be $n_1(k_1 + 1)$, so that n_1 patients are randomized to receive each experimental dose and n_1 are randomized to receive the control. The data from stage 1 can be summarized by the number of observed successes, x_{1i}, and the number of observed toxicities, t_{1i}, at dose i for $i = 0, ..., k_1$, with $i = 0$ corresponding to the control treatment. At the onset of the phase II/III trial, the interest is to determine whether there is a safe treatment among the k_1 experimental treatments which is more effective than the control treatment. Thus the null hypotheses of interest are $H_1 : \theta_0 = \theta_1,\ ...,\ H_{k_1} : \theta_0 = \theta_{k_1}$ where $\theta_i,\ i \in \{0, 1, ..., k_2\}$ is a measure of the effectiveness of treatment i. Based on the efficacy data $\mathbf{x}_1$ and with the intention of using the closure principle to control the FWER, a set of p-values $p_{1,j}$ for $H_j,\ j \subseteq \{1, ..., k_1\}$ can be constructed.

Suppose that the total sample size for stage 2 is fixed. The number of patients randomized to each treatment, n_2, then depends on the number of doses that remain in the trial. Let $\mathcal{K}_2 \subseteq \{1, ..., k_1\}$ be the set of experimental doses that remain in the trial for testing in stage 2 with $k_2 = |\mathcal{K}_2|$. As any of the k_1 doses in stage 1 may continue to stage 2 there are 2^{k_1} possible sets of doses that we could choose. We will restrict these to sets of adjacent doses which means that there are $k_1(k_1 + 1)/2$ possible sets of doses, reducing the number of cases that need to be considered. Let x_{2i} and t_{2i}, $i \in \{0\} \cup \mathcal{K}_2$ with $i = 0$ corresponding to the control treatment, respectively denote the number of successes and toxicities on dose i in stage 2. At the end of stage 2, the efficacy data $\mathbf{x}_2$ can be used to construct a set of p-values $p_{2,j}$ corresponding to the closure set of p-values $p_{1,j}$ constructed using the stage 1 data.

By utilizing the method described in Section 2 the two sets of p-values from the two stages can be used to test whether there is an effective dose among the k_2 doses that proceed to the second stage. Given stage 1 data we want to determine the set $\mathcal{K}_2$ which will mostly likely lead us to finding at least one effective dose at the end of stage 2. In this section, assuming that stage 2 data have a distribution which depends on a fixed parameter vector, we obtain an expression for the probability of concluding at least one of the k_2 doses that proceed to stage 2 is effective by summing probabilities of outcomes for which we will find at least one effective dose (conditional power).

3.1. Distribution of second stage data

Let $f(\mathbf{x}_2, \mathbf{t}_2; \theta)$ denote the distribution of stage 2 data where θ is the vector of parameters giving the dose-response curves for efficacy and toxicity. Suppose a study patient is administered a dose level d, the outcome for efficacy will be either a successful treatment or a treatment failure and the probability of the successful treatment will be denoted by $p_E(d)$. The toxicity outcome will be categorized as either toxic or non-toxic and the probability of a toxic outcome will be denoted by $p_T(d)$. We propose two logistic models for the outcomes

$$p_E(d) = \frac{exp(\alpha_E + \beta_E \log\ d)}{1 + exp(\alpha_E + \beta_E \log\ d)} \tag{2}$$

and

$$p_T(d) = \frac{exp(\alpha_T + \beta_T \log\ d)}{1 + exp(\alpha_T + \beta_T \log\ d)} \tag{3}$$

such that stage 2 data $(\mathbf{x}_2, \mathbf{t}_2)$ would depend on the probability vector $\theta = (\alpha_E, \beta_E, \alpha_T, \beta_T)'$. Although we propose a logit link, other link functions may be used. A different linear predictor may also be used. Assuming the outcomes are independent, then the probability of x_{20} successes and t_{20} toxicities in the control group and x_{2i} successes and t_{2i} toxicities in the experimental dose i, $i \in \mathcal{K}_2$ is

$$f(\mathbf{x}_2, \mathbf{t}_2; \theta) = f(x_{20}; n_2, p_{E_0}) f(t_{20}; n_2, p_{T_0}) \prod_{i \in \mathcal{K}_2} f(x_{2i}; n_2, p_{E_i}) f(t_{2i}; n_2, p_{T_i})$$

where $f(x_{2i}; n_2, p_{E_i})$ and $f(t_{2i}; n_2, p_{T_i})$, $i \in \{0\} \cup \mathcal{K}_2$ are binomial mass functions with parameter vectors (n_2, p_{E_i}) and (n_2, p_{T_i}) respectively. The parameters p_{E_i} and p_{T_i}, $i \in \mathcal{K}_2$ are respectively points on the dose-response curves (2) and (3) corresponding to dose level i. If the control treatment is a dose level of the experimental drug, p_{E_0} and p_{T_0} are also points on the dose response curves (2) and (3). Otherwise their values are determined separately.

3.2. Expressions for conditional power

After obtaining the distribution of stage 2 data, the next step in obtaining the conditional power involves determining stage 2 data for which the final hypothesis will be significant given the results of stage 1. Given stage 1 data $\mathbf{x}_1$, the p-value $p_{1,j}$ corresponding to hypothesis H_j in the closure set $\mathcal{H}$ can be considered fixed. The final hypothesis test for the individual hypothesis H_j will be significant at level α if and only if $p_j = C(p_{1,j}, p_{2,j}) < \alpha$. The inequality can be rearranged to determine the minimum value of $p_{2,j}$ such that the null hypothesis H_j is rejected at the end of stage 2. For example if the combination of choice is the inverse normal combination given by equation (1), rearranging the inequality, the final hypothesis test will be significant if and only if

$$p_{2,j} < 1 - \Phi \left\{ \frac{\Phi^{-1}(1-\alpha) - w_1 \Phi^{-1}(1 - p_{1,j})}{w_2} \right\}. \tag{4}$$

Let l be the number of experimental doses in hypothesis H_j at stage 2. Then using the Bonferroni-Holm MinP, $p_{2,j} = l \times \min_{i \in j}\{p_{2,i}\}$ where $p_{2,i}$ is the p-value obtained from testing the pairwise null hypothesis H_i at the second stage. Substituting this expression of the p-value in inequality (4), then hypothesis H_j will be rejected at the end of stage 2 if and only if

$$p_{2,i} < \left(1 - \Phi\left\{\frac{\Phi^{-1}(1-\alpha) - w_1\Phi^{-1}(1-p_{1,j})}{w_2}\right\}\right)/l \quad \text{for some} \quad i \in j. \tag{5}$$

The RHS of inequality (5) could be viewed as the "level of testing" hypothesis H_j at stage 2.

For each possible x_{20}, the minimum number of successes required in either of the l doses such that inequality (5) holds can be obtained. We will denote this minimum number of successes by $B_{x_{20}}(p_{1,j})$ where the denotation reflects dependency on x_{20} and $p_{1,j}$. The next subsection is the focus of obtaining $B_{x_{20}}(p_{1,j})$. Hypothesis H_j will be rejected for the set of stage 2 data $\mathbf{x}_2$ such that $x_{2i} \geq B_{x_{20}}(p_{1,j})$ for some $i \in j$. To conclude an experimental dose i is more effective than the control treatment, we need to determine the set of stage 2 data $\mathbf{x}_2$ for which all hypotheses H_j with $i \in j$ are all rejected. We denote the set of $\mathbf{x}_2$ for which this is true by $\mathcal{R}(p_{1,i}),\ i \in \mathcal{K}_2$. The probability of concluding dose i is more effective than the control after stage 2 analysis is obtained by summing the probabilities of all outcomes in $\mathcal{R}(p_{1,i})$.

The form of $\mathcal{R}(p_{1,i})$ depends on the number of doses that continue to the second stage. For example suppose $k_1 = 4$ with a single treatment continuing, say $\mathcal{K}_2 = \{1\}$. To conclude that dose 1 is effective all the hypotheses H_{1234}, H_{123}, H_{124}, H_{134}, H_{12}, H_{13}, H_{14} and H_1 need to be rejected. Since only dose 1 proceeds to the second stage, the intersection hypotheses H_{1234}, H_{123}, H_{124}, H_{134}, H_{12}, H_{13} and H_{14} simplify to the pairwise hypothesis H_1 because no data are available for the other doses at stage 2 but the tests are carried out at different levels determined by inequality (5). The minimum x_{21} for a given x_{20} required to reject all hypotheses H_j for $j \subseteq \{1,2,3,4\}$ with $1 \in j$ could be obtained and is given by $B_{x_{20}}(\max\{p_{1,j}\})$. We take $\max\{p_{1,j}\}$ since the RHS of inequality (5) decreases when $p_{1,j}$ increases. Dose 1 would then be concluded to be more effective than the control treatment at the end of stage 2 if

$$x_{21} \geq B_{x_{20}}(\max\{p_{1,j}\})$$

for all j with $1 \in j$. The probability of concluding dose 1 is more effective than the control treatment at the end of stage 2 is then given by

$$\sum_{\mathcal{R}(p_{1,1})} f(\mathbf{x}_2;\theta)d\mathbf{x}_2 = \sum_{x_{20}=0}^{n_2}\left\{\binom{n_2}{x_{20}}p_{E_0}^{x_{20}}(1-p_{E_0})^{n_2-x_{20}}\sum_{x_{21}=B}^{n_2}\binom{n_2}{x_{21}}p_{E_1}^{x_{21}}(1-p_{E_1})^{n_2-x_{21}}\right\} \tag{6}$$

where $B = B_{x_{20}}(max\{p_{1,j}\})$ and $\mathcal{R}(p_{1,1})$ denotes the set of $\mathbf{x}_2$ for which dose 1 is rejected after stage 2.

Suppose from an initial four experimental doses at stage 1, dose 1 and dose 2 proceed to stage 2, that is, $k_1 = 4$ and $\mathcal{K}_2 = \{1,2\}$. In order to make inference on the effectiveness of dose 1 using the closure principle, the null hypotheses H_{1234}, H_{123}, H_{124}, H_{134}, H_{12}, H_{13}, H_{14} and H_1 are tested. On the other hand, the null hypotheses H_{1234}, H_{123}, H_{124}, H_{234}, H_{12}, H_{23}, H_{24} and H_2 are tested in order to make inference on dose 2. Since no data are available for doses 3 and 4, tests for hypotheses H_{134}, H_{13}, H_{14} and H_1 which are included in H_1 but not in H_2 are performed using only the test for H_1 but at different levels. The minimum x_{21} required to reject all these hypotheses which we denote by B_1 is obtained by evaluating $B_{x_{20}}(max\{p_{1,j}\})$ for $j \subseteq \{1,3,4\}$ with $1 \in j$. Similarly, only dose 2 data are available for hypotheses H_{234}, H_{23}, H_{24} and H_2 which are included in H_2 but not in H_1. The minimum x_{22} required to reject all these hypotheses which we denote by B_2 is obtained by evaluating

$B_{x_{20}}(max\{p_{1,j}\})$ for $j \subseteq \{2,3,4\}$ with $2 \in j$. On the other hand, only dose 1 and dose 2 data are available at stage 2 for hypotheses H_{1234}, H_{123}, H_{124} and H_{12} and hence their test is performed using only the test for H_{12}. The minimum number of successes required in either dose 1 or 2 to reject all these hypotheses which we denote by B_{12} is obtained by evaluating $B_{x_{20}}(max\{p_{1,j}\})$ for $j \subseteq \{1,2,3,4\}$ with $\{1,2\} \in j$.

Assuming dose 1 and dose 2 are interchangeable, there are three possible configurations for B_1, B_2 and B_{12} namely;

$$(i)\;\; B_1 < B_2 < B_{12} \qquad (ii)\;\; B_{12} < B_1 < B_2 \quad \text{and} \quad (iii)\;\; B_1 < B_{12} < B_2.$$

The expression for conditional power for each of these scenarios is different. From left to right, Figure 1 shows configurations (i) to (iii) for a given realization x_{20}. The partitions marked by 1, 2 and 12 respectively represent the realization of the number of successes in the experimental treatments for which only dose 1, only dose 2 and for which both dose 1 and 2 are concluded to be effective for a given number of successes in the control treatment. The probability of concluding at least one of the experimental doses is effective is obtained by summing all the probabilities of all outcomes in the partitions marked by 1, 2 and 12. For example, for configuration (i), the probability of concluding dose 1 or dose 2 is effective after stage 2 is

$$\sum_{\mathcal{R}(p_1)} f(\mathbf{x}_2;\theta)d\mathbf{x}_2 = \sum_{\mathcal{R}(p_{1,1})} f(\mathbf{x}_2;\theta)d\mathbf{x}_2 + \sum_{\mathcal{R}(p_{1,2})} f(\mathbf{x}_2;\theta)d\mathbf{x}_2 + \sum_{\mathcal{R}(p_{1,12})} f(\mathbf{x}_2;\theta)d\mathbf{x}_2 \qquad (7)$$

where $\mathcal{R}(p_{1,1})$, $\mathcal{R}(p_{1,2})$ and $\mathcal{R}(p_{1,12})$ respectively denote the set of stage 2 data given the stage 1 data for which after stage 2 only dose 1 would be effective, only dose 2 would be effective and when both dose 1 and 2 would be effective such that

$$\sum_{\mathcal{R}(p_{1,1})} f(\mathbf{x}_2;\theta)d\mathbf{x}_2 = \sum_{x_{20}=0}^{n_2} f(x_{20};n_2,p_{E_0}) \left\{ \sum_{x_{21}=B_{12}}^{n_2} \sum_{x_{22}=0}^{B_2} f(x_{21};n_2,p_{E_1}) f(x_{22};n_2,p_{E_2}) \right\},$$

$$\sum_{\mathcal{R}(p_{1,2})} f(\mathbf{x}_2;\theta)d\mathbf{x}_2 = \sum_{x_{20}=0}^{n_2} f(x_{20};n_2,p_{E_0}) \left\{ \sum_{x_{21}=0}^{B_1} \sum_{x_{22}=B_{12}}^{n_2} f(x_{21};n_2,p_{E_1}) f(x_{22};n_2,p_{E_2}) \right\}$$

and

$$\begin{aligned} \sum_{\mathcal{R}(p_{1,12})} f(\mathbf{x}_2;\theta)d\mathbf{x}_2 &= \sum_{x_{20}=0}^{n_2} f(x_{20};n_2,p_{E_0}) \left\{ \sum_{x_{21}=B_{12}}^{n_2} \sum_{x_{22}=B_2}^{n_2} f(x_{21};n_2,p_{E_1}) f(x_{22};n_2,p_{E_2}) \right\} \\ &+ \sum_{x_{20}=0}^{n_2} f(x_{20};n_2,p_{E_0}) \left\{ \sum_{x_{21}=B_1}^{B_{12}} \sum_{x_{22}=B_{12}}^{n_2} f(x_{21};n_2,p_{E_1}) f(x_{22};n_2,p_{E_2}) \right\} \end{aligned}$$

where $f(x_{2i};n_2,p_{E_i})$, $i = 0,1,2$, is the probability mass function of the binomial random variable X_{2i} with parameters n_2 and p_{E_i}.

Expressions (6) and (7) are respectively the combined conditional power when $\mathcal{K}_2 = \{1\}$ and $\mathcal{K}_2 = \{1,2\}$. The expressions also give the conditional power for taking $\mathcal{K}_2 = \{1\}$ or $\mathcal{K}_2 = \{1,2\}$ for any value of $k_1 \geq 2$ and similar expressions can be obtained for any $\mathcal{K}_2 = \{i\}$ and $\mathcal{K}_2 = \{i,j\}$ with $i,j \in \{1,...,k_1\}$. Bonferroni-Holm has been used to obtain the expressions for conditional power. Šidak-Holm similarly leads to simple expressions for conditional power. For Simes test,

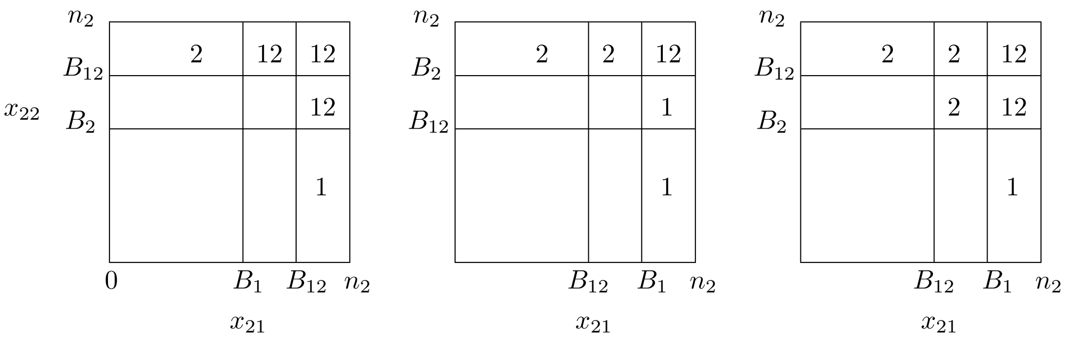

Figure 1. ***Configuration of the minimum number of successes. The x-axes are the no. of successes in dose 1 (x_{21}) and y-axes the no. of successes in dose 2 (x_{22}).***

it is not possible to obtain a single inequality such as the one resulting from Bonferroni-Holm test given by inequality (5) for composite hypotheses. However, it is still possible to obtain expressions for conditional power using this test but becomes less straightforward as the value of k_2 increases.

We have given the expressions for up to when two doses proceed to stage 2 but using the same principles expressions can be obtained for $k_2 > 2$. In practice, it would be rare to proceed to stage 2 with many experimental doses.

3.3. Obtaining the minimum number of successes

In this sub-section we illustrate how to obtain $B_{x_{20}}(p_{1,j})$, the minimum number of successes required in either of the l experimental doses in j such that the null hypothesis H_j is rejected at the end of stage 2. The left hand side of inequality (5) is the p-value from testing the null hypothesis H_i, $i \in j$ at stage 2. If a chi-squared test is used to test the null hypothesis H_i with $i \in j$, the critical chi-squared value χ_c^2 corresponding to the level of the test (RHS of inequality (5)) can be determined. The null hypothesis H_i is rejected if and only if the observed chi-square value

$$\frac{2n_2(x_{20}-x_{2i})^2}{(x_{20}+x_{2i})\{2n_2-(x_{20}+x_{2i})\}} > \chi_c^2.$$

Rearranging the expression, the null hypothesis is rejected for superiority if and only if

$$x_{2i} > \frac{U+V}{(2n_2+\chi_c^2)} = B_{x_{20}}(p_{1,j})$$

where

$$U = -\{\chi_c^2(x_{20}-n_2) - 2n_2x_{20}\}$$

and

$$V = \sqrt{\{\chi_c^2(x_{20}-n_2)-2n_2x_{20}\}^2-(2n_2+\chi_c^2)\{(2n_2+\chi_c^2)x_{20}^2-2n_2\chi_c^2x_{20}\}}.$$

Although we focus here on the χ^2 test, the value of $B_{x_{20}}(p_{1,j})$ can be evaluated for any other test statistic that can be used for making inference on binary data.

3.4. Penalizing for toxicity

Toxicity has not been incorporated in the conditional power expressions (6) and (7). Suppose a dose will be rejected for toxicity if the probability of toxicity exceeds some predetermined level γ. Then the probability that a dose is demonstrated to be both safe and effective is the product of the conditional power given by expression (6) and the indicator $I(p_{T_1} \leq \gamma)$. If more than one experimental dose proceeds to the second stage the different disjoint events for which we conclude at least one of the experimental doses in stage 2 is effective are multiplied by different indicator variables. For example if $\mathcal{K}_2 = \{1, 2\}$, there are three disjoint events for which we conclude there is an effective dose. These are; only dose 1 is effective, only dose 2 effective and both dose 1 and 2 are effective. The respective indicator variables with which the probability of these events are multiplied is $I(p_{T_1} \leq \gamma)$, $I(p_{T_2} \leq \gamma)$ and $I(p_{T_1} \leq \gamma, p_{T_2} \leq \gamma)$.

4. Predictive power

The conditional power expressions obtained in Section 3.2 assume a fixed value of the parameter vector θ. Suppose that θ is given some prior distribution with density $\pi_0(\theta)$. The posterior distribution of θ given the data observed at the end of the first stage is given by *Bayes′ theorem* to be equal to

$$\pi(\theta|\mathbf{x}_1, \mathbf{t}_1, n_1) = \frac{L(\mathbf{x}_1, \mathbf{t}_1, n_1)\pi_0(\theta)}{\int L(\mathbf{x}_1, \mathbf{t}_1, n_1)\pi_0(\theta)d\theta}$$

where $L(\mathbf{x}_1, \mathbf{t}_1, n_1)$ is the likelihood for the data from the k_1 doses of the experimental treatment observed at the end of the first stage. Assuming the observations are independent,

$$L(\mathbf{x}_1, \mathbf{t}_1, n_1) = \prod_{i=1}^{k_1} \binom{n_1}{x_{1i}} p_{E_i}^{x_{1i}} (1 - p_{E_i})^{n_1 - x_{1i}} \binom{n_1}{t_{1i}} p_{T_i}^{t_{1i}} (1 - p_{T_i})^{n_1 - t_{1i}}$$

where p_{E_i} and p_{T_i} are respectively the probabilities of success and toxicity at dose i. The predictive power is then obtained by evaluating the posterior mean of the conditional power. For example if $\mathcal{K}_2 = \{1, 2\}$, the penalized predictive power is given by

$$\int_\Theta [I(p_{T_1} \leq \gamma).A_1 + I(p_{T_2} \leq \gamma).A_2 + I(p_{T_1} \leq \gamma, p_{T_2} \leq \gamma).A_{12}]\, \pi(\theta|\mathbf{x}_1, \mathbf{t}_1, n_1)d\theta$$

where

$$A_j = \sum_{\mathcal{R}(p_{1,j})} f(\mathbf{x}_2; \theta), \quad j \in \{1, 2, 12\}$$

and $\mathcal{R}(p_{1,1})$, $\mathcal{R}(p_{1,2})$ and $\mathcal{R}(p_{1,12})$ respectively denote the set of stage 2 data given the stage 1 data for which after stage 2 only dose 1 would be effective, only dose 2 would be effective and when both dose 1 and 2 would be effective as described above.

The penalized predictive power depends on the choice of the doses selected to continue to stage 2 as these affect the number of patients per arm, n_2, the rejection region, $\mathcal{R}(p_1)$, which probabilities p_{E_i} enter the density $f(\mathbf{x}_2; \theta)$ and which probabilities p_{T_i} enter the penalty. We wish to make a choice of doses to continue on the basis of $\mathbf{x}_1$ and $\mathbf{t}_1$ to make the penalized predictive power as large as possible.

4.1. Distribution of the unknown parameters

We propose obtaining the prior beliefs on the dose-response curves for efficacy and toxicity separately using the technique of Bedrick *et al.* [12] for eliciting the priors for generalized linear models. The idea is to elicit prior belief at p locations on the dose response curve if there are p parameters in the linear predictor. For example, the dose response curve (2) for efficacy is defined by two parameters (α_E, β_E) hence we elicit priors at two dose levels, say d_{01} and d_{02}, on the dose response curve. The joint distribution at these p locations is then evaluated. The joint distribution of (α_E, β_E) is then obtained by transformation of random variables. Suppose that the probability of successful treatment $p_{E_{01}} = p_E(d_{01})$ and $p_{E_{02}} = p_E(d_{02})$ at dose levels d_{01} and d_{02} have independent beta distributions Beta(x_{01}, y_{01}) and Beta(x_{02}, y_{02}) respectively. Then assuming the dose response model (2), the prior distribution for (α_E, β_E) which has also been given by Whitehead *et al.* [13] for similar dose-response curves is

$$\pi_0(\alpha_E, \beta_E) = \prod_{i=1}^{2} \frac{p_{E_{0i}}^{x_{0i}} (1 - p_{E_{0i}})^{y_{0i}}}{B(x_{0i}, y_{0i})} \left| \log(\frac{d_{01}}{d_{02}}) \right|$$

where $p_{E_{0i}}$ is the function of $\alpha_{E_{0i}}$ and $\beta_{E_{0i}}$ given by the dose response curve (2) and B is the beta function. Similarly if the probability of toxicity at dose levels d_{01} and d_{02} have independent beta distributions Beta(t_{01}, u_{01}) and Beta(t_{02}, u_{02}) respectively, then assuming dose response (3), the prior distribution of (α_T, β_T) is

$$\pi_0(\alpha_T, \beta_T) = \prod_{i=1}^{2} \frac{p_{T_{0i}}^{t_{0i}} (1 - p_{T_{0i}})^{u_{0i}}}{B(t_{0i}, u_{0i})} \left| \log(\frac{d_{01}}{d_{02}}) \right|$$

where $p_{T_{0i}}$ is the function of $\alpha_{T_{0i}}$ and $\beta_{T_{0i}}$ given by the dose response curve (3) and B is the beta function.

The advantage of eliciting the prior distribution for $(\alpha_E, \beta_E, \alpha_T, \beta_T)$ in this way is that it is easier and more intuitive to elicit probability of efficacy (or toxicity) at a dose level than eliciting a prior for $(\alpha_E, \beta_E, \alpha_T, \beta_T)$ directly. Also for different link functions the same prior belief is used on the probability scale. The parameters for the prior beta distributions may be thought of as pseudo data. For example for Beta(x_{01}, y_{01}), x_{01} would denote the number of successfully treated patients out of $x_{01} + y_{01}$ administered dose d_{01}. To quantify the strength of the prior belief (variability) we propose examining the 90% interval of probability of efficacy (or toxicity) running from 5% to 95% as suggested by Thall & Simon [14] and checking the curve of the beta distribution as suggested by Lindley & Phillips [15].

It has been assumed that the beta distributions at doses d_{01} and d_{02} are independent. This assumption simplifies the mathematics but as Whitehead *et al.* [13] note, it has the undesired consequences that it is possible for $\beta_E < 0$ or $\beta_T < 0$ when it is believed that $\beta_E \geq 0$ and $\beta_T \geq 0$. To partly address this problem the priors are elicited at locations that are far from each other. Also like Whitehead *et al.* [13] since we are interested in the posterior means, negative parameter values for the slope parameters will not have undesired effects on the predictive power.

Let x_{1i} denote the number of successfully treated patients and $y_{1i} = n_1 - x_{1i}$ the number of patients that are not treated successfully at stage 1 at dose i $(i = 1, ..., k_1)$. Similarly let t_{1i} denote the number of patients that experience toxicity at stage 1 and $u_{1i} = n_1 - t_{1i}$ the number of patients that do not experience toxicity. After observation of the stage 1 data the updated distribution for the models of efficacy and toxicity are respectively

$$\pi_0(\alpha_E, \beta_E|\mathbf{x}_1, n_1) \propto \prod_{i=1}^{2} p_{E_{0i}}^{x_{0i}}(1 - p_{E_{0i}})^{y_{0i}} \prod_{i=1}^{k_1} p_{E_i}^{x_{1i}}(1 - p_{E_i})^{y_{1i}}$$

and

$$\pi_0(\alpha_T, \beta_T|\mathbf{t}_1, n_1) \propto \prod_{i=1}^{2} p_{T_{0i}}^{t_{0i}}(1 - p_{T_{0i}})^{u_{0i}} \prod_{i=1}^{k_1} p_{T_i}^{t_{1i}}(1 - p_{T_i})^{u_{1i}}.$$

Where the control treatment is a different drug, a beta prior Beta$(p_{E_0}; a_0, b_0)$ for the probability of successful treatment at control treatment which is conjugate for the likelihood function

$$L(\mathbf{x}_0, n_1) = \binom{n_1}{x_{10}} p_{E_0}^{x_{10}}(1 - p_{E_0})^{n_1 - x_{10}}$$

is elicited. The parameters a_0 and b_0 are elicited as explained before. The resulting posterior has a beta distribution Beta$(p; a_0 + x_{10}, b_0 + n_1 - x_{10})$.

5. Examples and simulation studies

The preceding sections have described how the doses continuing from the first stage of a seamless phase II/III clinical trial may be chosen and how a final analysis may be conducted to allow for this. In this section, the method is illustrated by simulation studies.

5.1. Simulation model parameter values and prior distributions

Following Whitehead *et al.* [13], assume that a new drug is tested at dose levels 10.5mg, 35.0mg, 87.5mg, 262.5mg, 700.0mg and 1050.0mg where each of the dose level is compared to a control. Suppose that γ, the accepted maximum proportion of toxicity, is 0.2. The control treatment is assumed to be a different drug from the experimental drug with the true probability of efficacy for the control treatment taken to be 0.3. For the dose-response curve parameters two cases are considered.

- Case 1: the true parameter values for (α_E, β_E) and (α_T, β_T) corresponding to dose-response curves (2) and (3) are assumed to be (-1.4867, 0.2720) and (-2.5782, 0.1621) respectively.
- Case 2: the true parameter vector (α_T, β_T) is equal to (-2.6728, 0.2023). The value of parameter vector (α_E, β_E) is the same as Case 1.

The dose-response curves for Case 1 and 2 are given in Figure 2. The continuous curve corresponds to the efficacy model for both Case 1 and 2, the dotted curve corresponds to the toxicity model for Case 1 and the dashed curve corresponds to the toxicity model for Case 2. In both cases the treatment is efficacious at higher dose levels. In Case 1, all the tested doses are also acceptably safe whereas in Case 2, dose 1050mg would be considered too toxic since probability of toxicity at this dose is above 0.2.

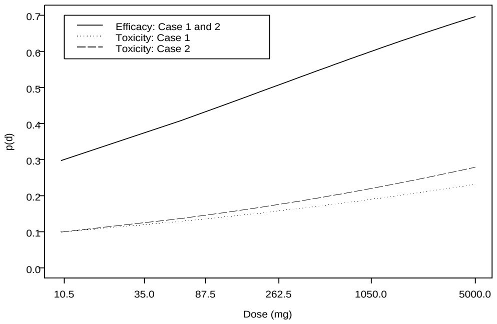

Figure 2. ***Dose-response curves.***

For all the simulation studies, it will be assumed n_1, the number of patients in each treatment at stage 1 is 20 and the total number of patients available for testing at stage 2 is 400 such that $n_2 = 400/(k_2 + 1)$. We demonstrate the method for k_2, the number of potential doses for testing in the second stage, up to 2. To obtain the predictive power, numerical quadrature is used to integrate over the parameter space and expressions (6) and (7) are evaluated using the normal approximation to the binomial distribution.

The prior distributions are elicited as described in section 4.1. The beta prior distribution for the probability of successful treatment using the control treatment is Beta(p_{E_0};12, 28). This prior belief will be used in evaluating the predictive power for all the simulation studies. Beta priors for probabilities of successful treatment and probabilities of toxicity are defined at dose levels 10.50mg and 5000mg. In order to assess the effect of strength of prior belief, three priors with equal prior means but different weights are defined for both toxicity and efficacy. Figure 3 shows the plots of the elicited prior beliefs. The first row gives the priors at dose level 10.50mg, in the second row are the priors at dose level 5000mg while in the third row are the corresponding contour plots of the resulting joint prior distributions of the intercept and slope parameters. The legends inside the Beta plots give the parameter values for the elicited beta densities. Columns 1, 3 and 5 respectively give the most informative, the middle weight and the least informative priors for efficacy. Columns 2, 4 and 6 respectively give the most informative, the middle weight and the least informative priors for toxicity. Henceforth, the most informative prior will mean most informative priors for both efficacy and toxicity (Columns 1 and 2), the middle weight prior will mean middle weight priors for both efficacy toxicity (Columns 3 and 4), and similarly least informative prior will mean the least informative priors for both efficacy and toxicity (Columns 5 and 6). For each of the three prior beliefs (most informative, middle weight and least informative) two sets of simulations (assuming Case 1 and 2) were carried out and predictive power evaluated.

5.2. Comparing results for Case 1 and Case 2

Figure 4 shows the histograms of the set of doses with the highest predictive power with each histogram based on a 1000 simulation studies. The notation di, $i \in \{1, ..., 6\}$ means the set $\mathcal{K}_2 = \{i\}$ results in the highest power while dij with $i, j \in \{1, ..., 6\}$ means the set $\mathcal{K}_2 = \{i, j\}$

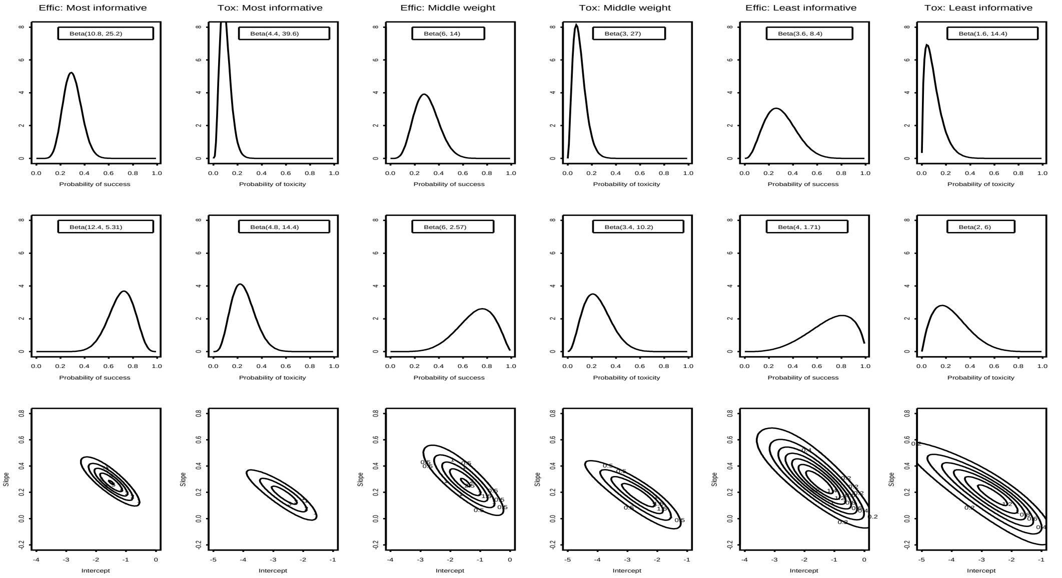

Figure 3. ***Elicited prior densities. Row 1 gives the priors at dose 10.50mg, Row 2 priors at dose 5000mg and Row 3 the resulting joint priors.***

results in the highest predictive power. The first row corresponds to the results for Case 1 and the second row the results for Case 2. From Column 1 to 3, the predictive power has been evaluated with the most informative, the middle weight and the least informative priors respectively. The bars have been partitioned into simulation studies whose maximum predictive power of potential doses is above 0.7 (shaded parts) and studies whose maximum predictive power is less than 0.7 (striped parts). The latter represent trials in which it is unlikely that any dose would continue to the second stage. To compare Case 1 and Case 2, we initially focus on results evaluated using the middle weight priors which are given by Figure 4(b) and 4(e) respectively.

For Case 1, the true probabilities of efficacy in the experimental doses are 0.30, 0.37, 0.43, 0.51, 0.57 and 0.60 while the respective probabilities of toxic outcomes are 0.10, 0.12, 0.14, 0.16, 0.18 and 0.19. Thus all the experimental doses are safe and doses 5 and 6 do not differ much in terms of efficacy. Dose 4 is considerably less efficacious than dose 5 and 6 but also considerably safer than dose 5 and 6. Based on all simulation studies (shaded and striped parts), dose 5 or 6 is in the set $\mathcal{K}_2$ with the highest predictive power in about 60% of the simulations. Dose 4 or one of the higher doses is in the set $\mathcal{K}_2$ with the highest predictive power in over 90% of the simulations. When only the simulation studies whose predictive power greater than 0.7 are considered (they were 607 out of 1000), dose 5 or 6 would be tested in stage 2 in over 65% of the simulations and dose 4 or 5 or 6 would be tested in stage 2 in over 96% of the simulations.

The true probability of toxic outcomes for Case 2 in ascending order at tested dose levels are 0.10, 0.12, 0.15, 0.18, 0.206 and 0.220. Hence the prior belief (mean) underestimates the level of toxicity. Based on all simulation studies, dose 6 alone whose true proportion of toxicity is well above 0.20, the accepted proportion of toxicity, has the highest predictive power in less

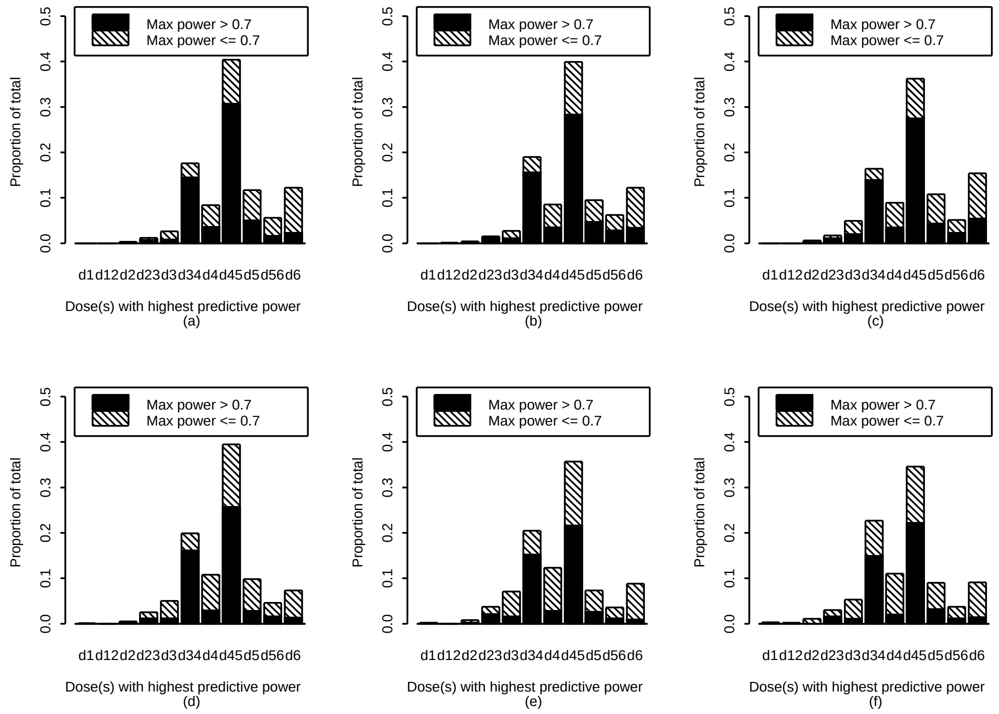

Figure 4. ***Histograms of doses with highest predictive power. Row 1 corresponds to Case 1 and row 2 to Case 2. From left to right the priors are less informative.***

than 10% of the simulations. Dose 4 which would be the desired dose for testing in stage 2 is in the set $\mathcal{K}_2$ with the highest predictive power in about 70% of the simulations. When only simulation studies whose maximum predictive power is greater than 0.7 are considered (they were 483 out of 1000), dose 6 results in the highest predictive in less than 2 % while dose 4 is one of the doses in the set with highest predictive power in over 80 % of the simulations.

Focussing on simulations with maximum predictive power above 0.7 and when $\mathcal{K}_2 = \{i\}$ $(i = 1, 2, ..., 6)$, for Case 1 the frequency increases to dose 5 and drops for dose 6. For Case 2, the frequency increases to dose 4 and drops for dose 6. These trends are what is expected because for Case 1 doses 5 and 6 do not not differ much in efficacy levels but dose 5 is safer while for Case 2 dose 5 has toxicity level slightly above the accepted level but considerably efficacious compared to dose 4. Dose 6 which is toxic results in highest predictive power with very low frequency. The same trend is observed for simulations whose set with highest predictive has two doses. The frequency increases to $\mathcal{K}_2 = \{4, 5\}$ and drops for $\mathcal{K}_2 = \{5, 6\}$ for the two cases.

5.3. Comparing results for different prior distributions

In both cases and for the three prior beliefs, $\mathcal{K}_2 = \{4, 5\}$ is the set which results in the highest predictive power more often although the relative frequency decreases as the priors become less informative. The three histograms for each of the cases show a similar trend for the set with the highest predictive power but higher doses result in highest predictive power more often for less informative priors. For example the frequency of dose 6 increases as the priors become less informative. The frequency however is contributed to more by the simulation studies whose predictive power is less than 0.7 (striped parts). This implies that small variations in the strength of the prior belief does not affect the choice of the doses much when the predictive power is high.

To further examine the results of when the drug is either not better than the control or the level of toxicity does not depend on the dose level or both, further simulations were carried out under the following scenarios (i) $(\alpha_E, \beta_E) = (-1.4867, 0.2720)$ & $(\alpha_T, \beta_T) = (-2.1972, 0)$ (ii) $(\alpha_E, \beta_E) = (-0.8473, 0)$ & $(\alpha_T, \beta_T) = (-2.5782, 0.1621)$ and (iii) $(\alpha_E, \beta_E) = (-0.8473, 0)$ & $(\alpha_T, \beta_T) = (-2.1972, 0)$. Figure 5 gives the dose-response curves and the simulation results for the three scenarios where rows 1 to 3 respectively correspond to the scenarios (i) to (iii). Column 1 gives the dose-response curves while the second to the fourth columns are respectively the histograms of doses with the highest predictive power where the predictive powers have been evaluated with most informative, middle weight and least informative priors used above.

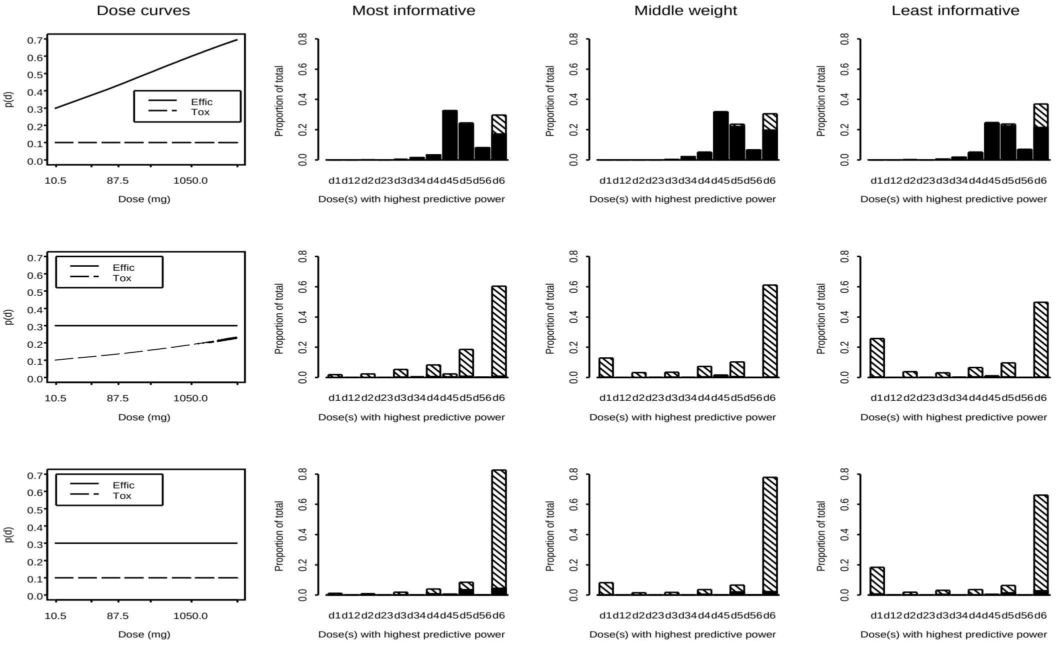

Figure 5. ***Dose curves and histograms of doses with highest predictive power when (i) toxicity is independent of dose level (Row 1) (ii) experimental doses are no better than control (Row 2) (iii) both (i) and (ii) hold (Row 3).***

For Scenario (i) efficacy improves with dose level and probabilities of toxicity at experimental doses are safe and independent of dose level. Dose 6 would be the most favorable dose but since the prior assumes the toxicity increases with dose level, like results presented in Figure 4, $\mathcal{K}_2 = \{4, 5\}$ results in the highest predictive power most often. However doses 5 or 6 are in the set with highest predictive with higher frequency compared to results of Figure 4 where simulations assumed true probability of toxicity increases with dose level. For scenarios (ii) and (iii) where the experimental doses are equally effective among themselves and no better than control, $\mathcal{K}_2 = \{6\}$ had the highest predictive power most frequently. This explains why the frequency for $\mathcal{K}_2 = \{6\}$ in Figure 4 seems not to follow the general trend. We also observe that for scenarios (ii) and (iii), $\mathcal{K}_2 = \{1\}$ results into the set with the highest predictive power more frequently especially for the least informative priors compared to results of scenario (i) and the results presented in Figure 4. This is because for scenarios (ii) and (iii) it is more likely to have simulation studies that lead to a posterior distribution with a negative slope for the efficacy so that dose 1 would be the preferred dose level.

6. Discussion

Recent methods for dose escalation/de-escalation in early clinical trials such as phase I and phase I/II trials have considered the dose-response relationship while allocating patients to the available dose levels in a trial. For example, O'Quigley *et al.* [16] proposed an exponential model for a toxicity model while Whitehead *et al.* [13] have proposed logistic models for both toxicity and therapeutic outcomes. This work generally does not focus on hypothesis testing. In phase II trials hypothesis testing is carried out and based on the results, a phase III trial is planned. In most previous work in this area a set of doses is chosen and the value of test statistics are evaluated using only the information from this set of experimental doses and the control treatment ignoring information from the other experimental doses. Incorporating the dose-response relationship such as the one proposed by O'Quigley *et al.* [16] and Whitehead *et al.* [13] helps make use of the results from the other experimental doses. Here we have proposed use of two logistic models for efficacy and toxicity in order to select the promising doses for the second stage in a seamless phase II/III clinical trials.

Both toxicity and efficacy have been considered explicitly in early clinical trials. For example Whitehead *et al.* [13] have proposed a design applicable to phase I/II clinical trials. However toxicity is often not explicitly included in the dose-selection procedure for doses to be tested in phase III. We have proposed a dose-selection procedure that incorporates both the dose-response relationship and considers toxicity explicitly. Rather than only focus on the probability that the dose will be found effective after phase III stage, the joint probability that the dose will be effective and not exceed some toxicity level is considered.

The penalty for toxicity was considered based on the distribution of proportion of toxicity rather than the distribution of the number of patients who would be experience toxicity at stage 2. This option was preferred for two reasons. If the penalty considered the probability that the number of patients treated in an experimental dose does not exceed $(\gamma \times n_2)$, larger samples will be penalized more when the true probability of toxicity is greater than γ and less when true probability of toxicity is less than γ. The second reason is that in practice adverse events are monitored as the trial continues so that the toxicity of the drug is evaluated before all patients are treated.

Finally, it has been assumed that the experimentation will continue in phase III stage when the results are promising. In some instances it may be desired to stop the trial after the phase II stage for effectiveness on the basis of a criterion determined in advance. Bauer & Köhne [7] have given an expression for the overall Type I error probability while using the Fisher's combination test. The expression while using the inverse normal method is given by

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \left(1 - \Phi\left\{[\Phi^{-1}(1-\alpha) - w_1\Phi^{-1}(1-p_1)]/w_2\right\}\right) dp_1 \tag{8}$$

where α_1 is the critical p-value for stopping at stage 1 for effectiveness and α_0 is the critical value for stopping at stage 1 with acceptance of the null hypothesis. Equating the overall Type I probability to α, numerical integration can be used to obtain α_1 for fixed value of α_0. The trial stops at stage 1 if some safe dose is concluded to be effective at level α_1. Expressions for overall Type I error probability when more than two stages of hypotheses testing are done are also obtainable so that if it is desired to have an extra evaluation of efficacy data in phase III stage, backward induction is used to obtain the predictive power. An alternative to using expression (8) in order to control the overall Type I error is to evaluate the testing boundaries as proposed by Lehmacher & Wassmer [17].

ACKNOWLEDGEMENTS

We are grateful to the Department of Statistics, University of Warwick for funding PhD studies of the first author at Warwick.

REFERENCES

1. Thall PF, Simon R, Ellenberg SS. Two-Stage Selection and Testing Designs for Comparative Clinical Trials. *Biometrika* 1988; **75**: 303–310.
2. Schaid DJ, Wieand S, Therneau TM. Sequential Designs for Phase III Clinical Trials incorporating Treatment Selection. *Biometrika* 1990; **77**: 507–513.
3. Stallard N, Todd S. Sequential Designs for Phase III Clinical Trials incorporating Treatment Selection. *Statistics in Medicine* 2003; **22**: 689–703.
4. Bretz F, Schmidli H, König F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal* 2006; **48**: 623–634.
5. Bauer P, Kieser M. Combining Different Phases in the Development of Medical Treatments within a Single Trial. *Statisitics in Medicine* 1999; **18**: 1833–1848.
6. Marcus R, Peritz E, Gabriel KR. On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance. *Biometrika* 1976; **63**: 655–660.
7. Bauer P, Köhne K. Evaluation of Experiments with Adaptive Interim Analyses. *Biometrics* 1994; **50**: 1029–1041.
8. Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. Truncated Product Method for Combining P-values. *Genetic Epidemiology* 2002; **22**: 170–185.
9. Westfall PH, Wolfinger DW. Closed Multiple Testing Procedures and PROC MULTTEST. http://support.sas.com/documentation/periodicals/obs/obswww23/ [17 September 2007].
10. Johnson RA, Wichern DW *Applied Multivariate Statistical Analysis* (5th edn). Prentice Hall: 2002.
11. Westfall PH, Young SS. *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment.* John Wiley & Sons: 1993.
12. Bedrick EJ, Christensen R, Johnson W. A New Perspective on Priors for Generalized Linear Models. *Journal of the American Statistical Association* 1996; **91**: 1450-1460.
13. Whitehead J, Zhou Y, Stevens J, Blakey G, Price J, Leadbetter J. Bayesian Decision Procedures for Dose-escalation based on Evidence of Undesirable Events and Therapeutic Benefit. *Statistics in Medicine* 2006; **25**: 37–53.

14. Thall PF, Simon R. Practical Bayesian Guidelines for Phase IIB Clinical Trials. *Biometrics* 1994; **50**: 337–349.
15. Lindley DV, Phillips LD. Inference for a Bernoulli process (A Bayesian View). *The American Statistician* 1976; **30**: 112–119.
16. O'Quigley J, Pepe M, Fisher L. Continual Reassessment Method: A Practical Design for Phase I Clinical Trials in Cancer. *Biometrics* 1990; **46**: 33–48.
17. Lehmacher W, Wassmer G. Adaptive Sample size Calculations in Group Sequentials Trials. *Biometrics* 1999; **55**: 1286–1290.