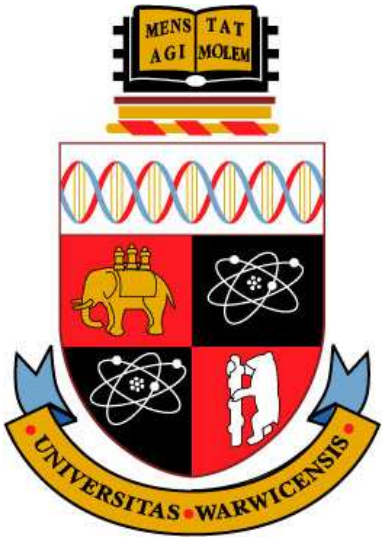

Dose Selection in Seamless Phase II/III Clinical Trials based on Efficacy and Toxicity

by

Peter Kung'u Kimani

Thesis

Submitted in partial fulfilment of the requirements

for the degree of

Doctor of Philosophy in Statistics

Department of Statistics

August 2009

THE UNIVERSITY OF WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

I would like to acknowledge various support I received from several people. Firstly, I would like to thank my supervisors Professor Jane Hutton and Professor Nigel Stallard. Throughout the preparation of this thesis, they gave me invaluable support and guidance. I would like to thank Dr. Ewart Shaw who every year gave me useful comments after reviewing my annual reports. I would also like to thank Paula and other administrative staff in the department for making sure I was comfortable.

I cannot forget the friendship of the PhD students in the department. I particularly thank Michalis and Demetris with whom from the first year until now, we have coffee breaks together at the student union. I also thank my friends at the Chaplaincy, my housemates Andrea and Flavia, and all my other friends.

To my parents and brothers; thank you for the love and the encouragement you have always offered me. I was able to pay for my expenses because of the bursary award I received from the Department of Statistics and I am very grateful for this. Lastly, I thank God for seeing me through the period of my studies.

Declarations

I declare that the work in this thesis is my own, and has not been submitted elsewhere for examination. The materials that are not my original ideas have been acknowledged by referencing. The materials in Chapters 5 and 6 expound the work by Kimani et al. (2009).

Abstract

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. Methodology exists for the analysis of such trials that allows complete flexibility of the choice of doses that continue to the second stage. There is very little work, however, on optimizing the selection of the doses. This is a challenging problem as it requires incorporation of the dose-response relationship, of the observed safety profile and of the planned analysis method. In this thesis we propose a dose-selection procedure for binary outcomes in adaptive seamless phase II/III clinical trials that incorporates the dose-response relationship, 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 to stage 2.

Chapter 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. A review of the statistical models used in design and analyses of data at each of the three phases of a clinical trial is given in Chapter 3. Chapter 2 outlines the statistical tools needed in the review of the statistical models used in each of the phases of a clinical trial.

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 II/III trials. Such trials are conducted in two stages. In stage 1 (phase II stage) of phase

II/III trials, several hypotheses, such as comparing how the drug works in different subpopulations or which doses are more efficacious than control treatment are tested. Based on stage 1 data, subpopulation(s) or dose(s) which show promising results continue to stage 2 (phase III stage) for further testing. At the end of stage 2, data from both stage 1 and stage 2 are used for the final confirmatory analysis. Although such phase II/III trials save development time, they introduce statistical complexity associated with controlling the type I error while testing multiple hypotheses and combining evidence from the two phases. In Chapter 4 we describe how to address these issues.

In addition to the issues associated with testing phase II/III clinical trials, another challenge raised by these trials is how to make the choice of the subpopulation(s) or the dose(s) to continue to stage 2 after stage 1. This is the question considered in this thesis. In Chapter 5, we develop a new method for dose selection in seamless phase II/III allowing for the final analysis that incorporates the dose response relationship, the prior knowledge and the stage 1 data. The dose selection procedure is evaluated using simulation studies in Chapter 6.

The method for dose selection developed in Chapter 5 assumes that: (1) in both stages binary outcomes are primary endpoints, (2) there is no uncertainty on the dose-response relationship, and (3) the seamless phase II/III is monitored only once and there are no opportunities for stopping early either for futility or for overwhelming evidence of efficacy. In Chapter 7, we describe how in future work, we intend to address these limitations. We end the thesis by discussing the main features of the new dose selection procedure and stating the conclusions in Chapter 8.

Chapter 2

Statistical background

In this chapter we give background on some of the statistical tools that will be needed in the rest of this thesis. The work in this thesis is based on binary outcomes, that is, occurrence or non-occurrence of an event such as toxicity or a therapeutic effect. Hence statistical tools reviewed in this chapter are demonstrated using binary outcomes. After describing the technique of transformation of random variables in Section 2.1, we will describe how to make Bayesian inference for a binary outcome parameter in Section 2.2. The chapter ends by describing Bayesian decision theoretic techniques in Section 2.3.

2.1 Transformation of random variables

In this thesis, we will occasionally need to determine the distribution of a random vector when we know the distribution of another random vector with which there is one-to-one transformation. To do this, we will use the technique of transformation of random variables that is described in several statistics text books such as in Chapter 11 of Roussas (2007). In the rest of this section, we briefly review this technique. Let $f_{X_1,...,X_n}(x_1, ..., x_n)$ be the value of the joint probability density of the continuous random vector $\mathbf{X} = (X_1, ..., X_n)'$.

Suppose the transformations $y_1 = \phi_1(x_1, ..., x_n)$, ..., $y_n = \phi_n(x_1, ..., x_n)$ are respectively partially differentiable with respect to x_1, ..., x_n and represent one-to-one transformations for all values within the range of $\mathbf{X}$ for which $f_{X_1,...,X_n}(x_1, ..., x_n) \neq 0$, then for these values of $(x_1, ..., x_n)$, the equations $y_1 = \phi_1(x_1, ..., x_n)$, ..., $y_n = \phi_n(x_1, ..., x_n)$ are uniquely solved for x_1, ..., x_n to give $x_1 = \psi_1(y_1, ..., y_n)$, ..., $x_n = \psi_n(y_1, ..., y_n)$ and for the corresponding values of $(y_1, ..., y_n)$, the joint probability density of $\mathbf{Y} = (\phi_1(X_1, ..., X_n), ..., \phi_n(X_1, ..., X_n))'$ is given by

$$f_{Y_1,...,Y_n}(y_1, ..., y_n) = f_{X_1,...,X_n}(\psi_1(y_1, ..., y_n), ..., \psi_n(y_1, ..., y_n))|J|, \tag{2.1}$$

where $|J|$ is the determinant of the Jacobian of the transformation J given by

$$J = \begin{vmatrix} \frac{d\psi_1}{dy_1} & \cdots & \frac{d\psi_1}{dy_n} \\ . & & \\ \frac{d\psi_n}{dy_1} & \cdots & \frac{d\psi_n}{dy_n} \end{vmatrix}.$$

For all the other values of $(y_1, ..., y_n)$, $f_{Y_1,...,Y_n}(y_1, ..., y_n) = 0$.

2.2 Review of Bayesian principle

As mentioned above, we will mostly focus on clinical trials with binary outcomes. Examples of binary outcomes in clinical trials are (1) after treatment is administered to a patient, the patient is successfully treated or not, and (2) after treatment is administered to a patient, the patient experiences an adverse (or toxic effect) or not. In this chapter, we will use the first example. We will assume successful treatment (success) after treatment is a Bernoulli process that occurs with probability p, that is, the probability of success is p. The number of successfully treated patients (s_n) generated using the Bernoulli process after n subjects have been entered in the trial will have a binomial distribution with parameters n and p, that is S_n is Bin(n, p). The objective of the clinical trial is to make inference on p.

A common feature of phase I and phase II trials is that they are small studies. This means that the incorporation of information from outside the trial is particularly attractive. This can be achieved by using the Bayesian principle in order to learn from previous experience. In this section, we demonstrate how to make Bayesian inference for a parameter of interest such as the binomial parameter p. In contrast to the frequentist setting where p is assumed to be fixed, in Bayesian statistics p itself is considered to be a random variable whose distribution is continually updated as more data are collected. After data **x** are collected, the updated distribution of p is referred to as the posterior distribution of p given **x**, with the density that will be denoted by $\pi(p|\mathbf{x})$. The Bayesian principle is centered around Bayes' theorem. If p is the parameter of interest and data **x** are collected, Bayes' theorem is expressed as

$$\pi(p|\mathbf{x}) = \frac{l(p|\mathbf{x}) \cdot \pi_0(p)}{\int l(p|\mathbf{x}) \cdot \pi_0(p) dp}, \tag{2.2}$$

where $l(p|\mathbf{x})$ is the likelihood function of p given the data **x** and $\pi_0(p)$ is the density of the prior distribution of p before data **x** are observed. For binary outcomes, data can be summarised by the number of successfully treated patients (s_n) and the number of patients entered in the trial (n) so that we may write $l(p|s_n, n)$ for $l(p|\mathbf{x})$. Inference on p or a function of p is then made using the posterior distribution. For example, the posterior mean can be used to estimate the probability of success p.

When a prior distribution chosen for some parameter leads to a posterior distribution of the same form as the prior distribution, the prior distribution is said to be a conjugate prior. Conjugate priors are advantageous because they may lead to integrals which can be evaluated using analytical methods. In general numerical integration techniques are required to make inference using the posterior distribution. Expression (2.2) assumes a single parameter p but this could be replaced by a vector. Gelman et al. (2004) describe making inference for several models. In this chapter, we will focus on obtaining the prior distribution for the probability of success based on a single treatment and for parameters in

a logistic regression model.

2.2.1 Eliciting beta prior distribution for a Bernoulli parameter

The beta prior distribution is a conjugate prior for a Bernoulli process parameter such as the probability of success p. The beta prior distribution for the binomial data parameter is proposed in some clinical trial designs that will be reviewed in the next chapter and is also used for research work outlined in the remainder of this thesis. For a Bernoulli process, the likelihood function of p, the probability of success, after n patients have been treated and s_n successes have been observed is given by

$$l(p|s_n, n) = \binom{n}{s_n} p^{s_n}(1-p)^{n-s_n} \quad (s_n = 0, 1, ..., n).$$

If we assume that p has a beta prior distribution with parameters $a > 0$ and $b > 0$, that is,

$$\pi_0(p) = \text{Beta}(p; a, b) = \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}, \quad 0 < p < 1,$$

where $B(a, b)$ is the beta function, then using equation (2.2), the posterior distribution of p given (s_n, n) is given by

$$\begin{aligned} \pi(p|s_n, n) &= \frac{l(p|s_n, n) \cdot \text{Beta}(p; a, b)}{\int \{l(p|s_n, n) \cdot \text{Beta}(p; a, b)\}\, dp} \\ &\propto p^{a+s_n-1}(1-p)^{b+n-s_n-1} \end{aligned}$$

which is of beta form Beta$(a+s_n, b+n-s_n)$. Hence a beta prior distribution is a conjugate prior for a Bernoulli parameter.

The prior information is elicited from investigators and quantified into a relevant distribution. For a beta prior distribution, the elicited information is quantified into a beta distribution by using the elicited information to determine the parameters a and b in Beta$(p; a, b)$. For the rest of this subsection, we describe how this may be done. In addition to being a conjugate prior for the binomial distribution parameter, a beta distribution has a

number of attractive properties which make it appealing to use it as a prior distribution for a binomial data parameter p. The domain of p in Beta$(p; a, b)$ is $[0, 1]$ which makes it sensible to use a beta distribution as a prior for a binomial distribution parameter which itself has its domain $[0, 1]$. As shown above, if a binomial distribution parameter p is assumed to have a beta distribution Beta$(p; a, b)$, the posterior distribution of p is Beta$(a + s_n, b + n - s_n)$, where s_n is the number of successfully treated patients after n patients have been administered a treatment. The mean of a random variable p that is Beta$(a + s_n, b + n - s_n)$ is

$$\frac{a + s_n}{a + b + n}. \tag{2.3}$$

If $a = b = 0$, then expression (2.3) gives the proportion of successfully treated patients after n patients have been entered in the trial. Thus the parameters of the beta prior distribution, that is a and b, may be thought of as pseudo-data elicited such that the prior belief is that if $a + b$ patients were treated, a will be successfully treated so that the proportion of successfully treated patients is $a/(a + b)$. This proportion is then updated when data (s_n, n) are collected to give expression (2.3). Figure 2.1 shows beta densities with different parameter values. The legends give the parameter values of the beta densities. For a beta density with parameter vector $(a, b) = (0.5, 10)$, most mass is at values of p close to 0 while for a beta density with parameter vector $(a, b) = (10, 0.5)$, most mass is at values of p close to 1. For a beta density with parameter vector $(a, b) = (0.5, 0.5)$, probability mass is concentrated at values of p close to 0 and 1. When p is Beta(1,1), the density is flat so that this corresponds to Uniform[0,1]. When both parameters values are greater than 1, the densities have a mode between 0 and 1. For example when p is Beta(2,8), $p = 0.1$ is the mode. When $a > 1$ and $b > 1$ the mode is

$$\frac{a - 1}{a + b - 2}$$

so that when $a \to \infty$ or $b \to \infty$, the mean is approximately equal to the mode. Hence from Figure 2.1, if the investigators do not have prior knowledge on p, the flat prior distribution

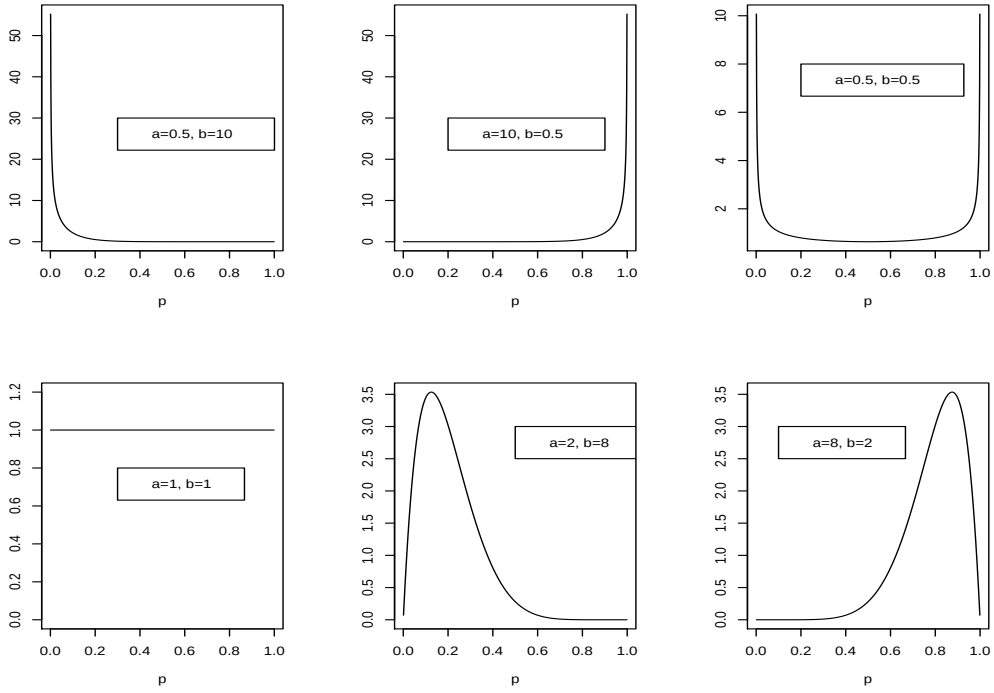

Figure 2.1: Beta densities with different parameter values. The legends give the values of the parameters.

is Beta(1,1) and prior densities with parameter value(s) less than 1 should be used with care.

In addition to the mean value of the parameter of interest, the parameter values chosen for its prior distribution should reflect the level of uncertainty (variance) associated with the parameter of interest. The variance of p which is Beta(a, b) is given by

$$\frac{ab}{(a+b)^2(a+b+1)},$$

so that different sets of a and b can result in the same mean but different variance. The variance of the probability of success p in the density curve is exhibited by the spread of the curve. Large values of variance (small values of a and b) lead to flat densities reflecting limited knowledge while small values of variance (large values of a and b) lead to curves

with high peaks at the mode reflecting more certainty on the true value of p. Accordingly, Lindley and Phillips (1976) suggest referring to curves of the beta densities plotted for different values of a and b during the elicitation process. In their paper, they give a good discussion using an example of how to elicit and quantify a beta distribution. Figure 2.2 shows curves for different values of a and b but with the same mean (0.2). As the values of a and b increase the peaks are higher and the mode moves closer to the mean 0.2. Thall and Simon (1994) refer to Lindley and Phillips (1976) for the elicitation and quantification of the beta distribution but they also introduce the idea of the width of the 90% interval (W_{90}) running from the 5% to 95% percentiles. An investigator is asked to provide the width of an interval within which he/she is 90% confident p lies. A search is then carried out to determine values of a and b such that the mean of p is $a/(a+b)$ and the difference between the 95th quantile and the 5th quantile is equal to the specified value. The shorter the width the more informative is the prior distribution since the density curves will be more peaked.

2.2.2 Prior distribution for dose-response parameters

The example given above is applicable when inference is made for the probability of success at a single treatment dose. When more than one dose of the same drug are tested, some dose-response curve may be assumed and if the Bayesian principle is used to make inference, it is necessary to give the joint prior distribution of the parameters of the dose-response curve. Prior distributions for generalized linear models parameters were proposed by Bedrick et al. (1996). The form of prior distribution for generalized linear models parameters proposed by Bedrick et al. (1996) generalizes the prior distribution proposed by Tsutakawa (1975). This form of prior distributions is used in one of the phase II design reviewed in the next chapter and will also be adopted in the dose selection procedure that we propose in Chapter 5. Rather than describe the theory given by Bedrick et al. (1996), we will demonstrate with the models which we are interested in. Whitehead (2006) has

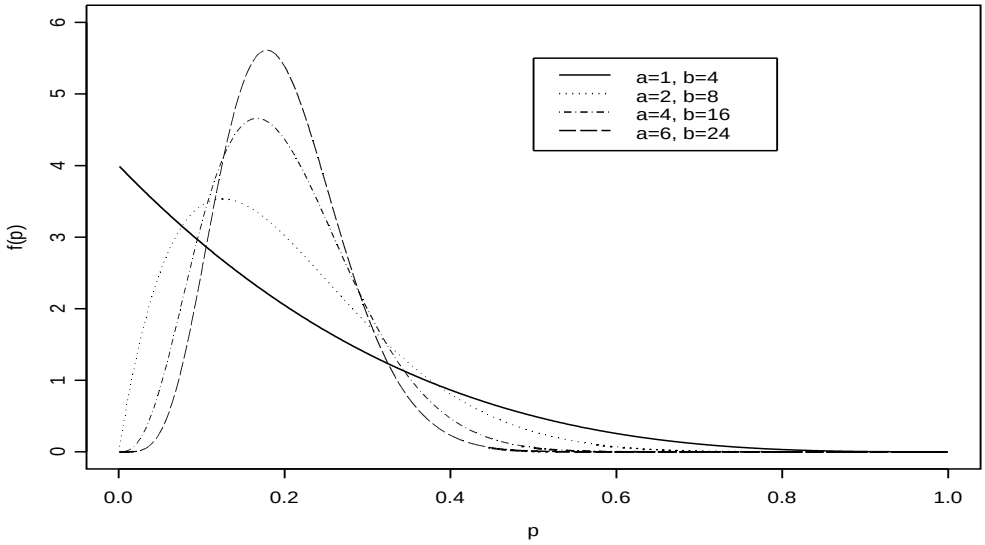

Figure 2.2: Beta curves with same mean (0.2) but different parameter values

reviewed this form of prior distribution and some of the notation used in this section is adopted from his review.

Let $p(d)$ denote the probability of success at dose level d. Further suppose that given dose, successes are independent binary outcomes with probability $p(d)$ and that $p(d)$ can be modeled by a generalized linear model. Then probabilities of success are related to the dose levels through the formula

$$g(p(d)) = \alpha + \beta\ f(d),$$

where $g(.)$ is a link function that links the probability of success ($p(d)$) to the linear predictor $\alpha + \beta\ f(d)$, where α is the intercept parameter, β is the slope parameter and $f(.)$ is some transformation of the dose such as natural log of the dose so that $f(d) = \log(d)$. Agresti (2002) describes link functions which can be used for binary outcomes such as the logit, probit and complementary log-log link functions. We will use the logit link which

links the probability of success to the linear predictor as follows

$$g(p(d)) = \text{logit}(p(d)) = \log\left(\frac{p(d)}{1-p(d)}\right) = \alpha + \beta\ f(d). \tag{2.4}$$

Using the proposal of Bedrick et al. (1996), rather than directly elicit prior distributions for the parameter vector (α, β), prior distributions for the probabilities of success are elicited at some dose levels. Because the dose-response curve (2.4) is defined by two parameters (α and β), prior distributions for probabilities of success are elicited at two dose levels. Assuming these prior distributions are independent, the joint distribution of the two probabilities of success is obtained and hence the joint distribution of the linear predictor parameters (α, β) using transformation of random variables. If there were three parameters in the linear predictor, prior distributions for probabilities of success would be elicited at three dose levels and so on. For the dose-response curve (2.4), suppose the prior distributions for probabilities of success are elicited at dose levels $d_i,\ i = -1, 0$. These dose levels do not have to be among the experimental dose levels. In this thesis, we will assume beta prior distributions Beta$(p_i; a_i, b_i)$, $i = -1, 0$ at dose i can be elicited as described above where p_i denotes the probability of success at dose i and a_i and b_i may be interpreted as pseudo-data elicited as described above. Assuming the elicited beta prior distributions at the two doses are independent, then the joint prior distribution of $p(d_{-1})$ and $p(d_0)$ is given by

$$\prod_{i=-1}^{0} \frac{p_i^{(a_i-1)}(1-p_i)^{(b_i-1)}}{B(a_i, b_i)}.$$

To obtain the joint prior distribution of α and β which we denote by $\pi_0(\alpha, \beta)$, the technique of transformation of random variables described in Section 2.1 is used. In equation (2.1), let $n = 2$, $X_1 = p(d_{-1})$, $X_2 = p(d_0)$, $y_1 = \alpha$, $y_2 = \beta$. Assuming the logit link (2.4),

$$\psi_1(\alpha, \beta) = p(d_{-1}) = p_{-1} = \frac{\exp(\alpha + \beta\ f(d_{-1}))}{1 + \exp(\alpha + \beta\ f(d_{-1}))}$$

and

$$\psi_2(\alpha,\beta) = p_(d_0) = p_0 = \frac{\exp(\alpha+\beta\ f(d_0))}{1+\exp(\alpha+\beta\ f(d_0))},$$

so that using equation (2.1),

$$\pi_0(\alpha,\beta) = \prod_{i=-1}^{0} \frac{p_i^{(a_i-1)}(1-p_i)^{(b_i-1)}}{B(a_i,b_i)}|J|, \tag{2.5}$$

where $p_i,\ (i=-1,0)$ are functions of α and β as defined above. The partial derivatives are

$$\frac{d\psi_i}{d\alpha} = p_i(1-p_i) \ \text{ and } \ \frac{d\psi_i}{d\beta} = p_i(1-p_i)g(d_i) \quad i=-1,0$$

so that

$$|J| = |g(d_{-1}) - g(d_0)| \prod_{i=-1}^{0} (p_i)(1-p_i)$$

which when substituted in equation (2.5) gives

$$\pi_0(\alpha,\beta) = |g(d_{-1}) - g(d_0)| \prod_{i=-1}^{0} \frac{p_i^{a_i}(1-p_i)^{b_i}}{B(a_i,b_i)}.$$

The transformation of the dose that we are going to use in this thesis is the natural log. Hence the joint prior density of α and β is given by,

$$\pi_0(\alpha,\beta) = \prod_{i=-1}^{0} \frac{p_i^{a_i}(1-p_i)^{b_i}}{B(a_i,b_i)} \left|\log\left(\frac{d_{-1}}{d_0}\right)\right|. \tag{2.6}$$

Suppose that a new drug is tested at k doses. Let the number of treatment successes and treatment failures at dose d_i ($i=1,...,k$) be denoted by a_i and b_i respectively. The likelihood function of (α,β) given the observed data is

$$l(\alpha,\beta|\mathbf{x}) = \prod_{i=1}^{k} \binom{n_i}{a_i} p_i^{a_i}(1-p_i)^{b_i},$$

where $n_i = a_i + b_i$ is the number of patients allocated to dose d_i so that updating the distribution of (α, β) given by equation (2.6) with these data using equation (2.2), the joint posterior density for α and β is

$$\pi(\alpha, \beta|\mathbf{x}) \propto \prod_{i=-1}^{k} p_i^{a_i}(1-p_i)^{b_i}, \tag{2.7}$$

where

$$p_i = \frac{\exp(\alpha + \beta \ \log \ d_i)}{1 + \exp(\alpha + \beta \ \log \ d_i)}, \quad i = -1, 0, 1, ..., k.$$

The form of the posterior distribution given by equation (2.7) has the same form as the prior distribution given by equation (2.6) so that this prior is a conjugate prior for (α, β). Eliciting the prior distribution for (α, β) as described in this section may also have the advantage of being easier and more intuitive since it involves elicitation of the probabilities of success at several doses from investigators rather than direct elicitation of the joint probability of (α, β).

2.3 Bayesian decision theory

Bayesian decision procedures are most common and seem appropriate in early clinical trials. For example, Stallard (1998) points out that the outcome of a phase II study is a decision of whether to continue with further evaluation or to abandon the therapy due to lack of efficacy or high toxicity or cost and hence argues for Bayesian decision techniques. Decision theory involves defining gain functions for different actions (or decisions) that can be taken and comparing the expected gain from each action. The best decision is the one with the highest expected gain. Rather than think of the gain, it is also possible to think of losses and hence take the decision with the least expected loss. Lindley (1985) gives a good introduction to the basic concepts in decision theory.

Before giving gain functions for complex decision problems, we first consider the simplest decision making problem where there are only two decisions to choose from and only two states of nature can occur. Table 2.1 summarizes this simple problem for a drug company with a capital base of $£m$ from which it can choose whether or not to invest $£c$ in a clinical trial to test efficacy of a new drug. Decision 1 is "to invest" (d_1) and decision 2 is "not to invest" (d_2). At the end of the clinical trial, the two states of nature are the new drug will be concluded to be efficacious ("drug is efficacious") and the new drug will be concluded not to be efficacious ("drug is not efficacious") with probability θ_1 and θ_2 respectively, where $\theta_1 + \theta_2 = 1$. Suppose that if the new drug is concluded to be efficacious at the end of the clinical trial, the drug company will make $£k$ from marketing the new drug. Then if the drug company decides to undertake the drug development, and the drug is concluded to be efficacious, the drug company will improve its capital to $£(m - c + k)$ while if the drug is concluded not to be efficacious, then its capital will decrease by $£c$ to $£(m - c)$. If the drug company chooses not to undertake the drug development, the drug company will neither lose nor gain anything regardless of whether the drug will have been concluded effective or not as shown in row corresponding to decision d_2. To compare decision d_1 and d_2, the expected gain function for decision d_i (i=1,2) is defined by

$$E(d_i) = \sum_{j=1}^{2} \theta_j G_i(\theta_j), \tag{2.8}$$

where $G_i(\theta_j)$ is the final capital base if state j $(j = 1, 2)$ occurs for decision i (i=1,2). The resulting expected gains from decision d_1 and d_2 are respectively

$$E(d_1) = \sum_{j=1}^{2} \theta_j G_1(\theta_j) = m + k\theta_1 - c \text{ and } E(d_2) = \sum_{j=1}^{2} \theta_j G_2(\theta_j) = m, \tag{2.9}$$

where $G_i(\theta_j)$ $(i = 1, 2)$ is as defined above. If the initial capital base ($£m$) invested is ignored so that $G_i(\theta_j)$ is the gain if state j occurs for decision i, then the expected gains

Table 2.1: Simplest decision making problem

	State of Nature	
Decision	Drug efficacious (Prob of this state is θ_1)	Drug not efficacious (Prob of this state is θ_2)
$d_1 : Invest$	$m - c + k$	$m - c$
$d_2 : Do\ not\ invest$	m	m

are evaluated as follows

$$E(d_1) = \sum_{j=1}^{2} \theta_j G_1(\theta_j) = k\theta_1 - c \text{ and } E(d_2) = \sum_{j=1}^{2} \theta_j G_2(\theta_j) = 0. \tag{2.10}$$

The two expressions (2.9 and 2.10) show that the difference in expected fortune between decisions d_1 and d_2 only depends on the amount the drug company will make from selling the new drug if it is concluded to be effective and the amount it will lose if the new drug will be concluded not to be effective. Hence the gain functions can be compared relative to any baseline.

In the example of Table 2.1, the decision is whether to invest or not to invest. A more natural decision in clinical trials is whether to proceed from one phase of a clinical trial to the next phase. For example, in a phase II study, one may want to choose between a decision to proceed from phase II to phase III (d_1) and decision to abandon drug development after the phase II study (d_2). Another example would be a phase II clinical trial that allows more than one inspection of data during the trial. Before the final inspection, one may choose to stop the phase II study and proceed to phase III study (d_1), stop phase II study and abandon drug development (d_2) or continue with the phase II study and make another inspection (d_3). Thus the number of decisions to choose from may be more than 2 but often will be finite.

Further in Table 2.1, the state of nature is that the drug is efficacious, in a clinical trial the unknown state of nature would be the probability of efficacy for an experimental drug, denoted by $p \in [0, 1]$. In Bayesian decision theory, the decision maker's prior knowledge of p is encoded by a prior distribution $\pi_0(p)$ (French and Insua, 2000). Then data are observed

that are drawn from a distribution that depends on the unknown state of nature p. These data are used to update the distribution of p using the Bayes' theorem given by equation (2.2) resulting to a posterior distribution $\pi(p|\mathbf{x})$. Then the expected gain $\mathcal{G}_a$, for action $a \in \mathcal{D}$, where $\mathcal{D}$ is the set of actions that may be chosen is given by

$$\int_0^1 G_a(p,n)\pi(p|\mathbf{x})dp, \tag{2.11}$$

where $G_a(p,n)$ is the gain associated with action a and depends on the probability of success p and the number of patients in the trial n. To give an example of the form of $G_a(p,n)$, suppose in a phase II clinical trial one of the actions that may be taken is to proceed to phase III. Suppose the average amount of money required to treat one patient in the phase II trial is k and the amount required to test a drug in a phase III trial is $m \geq 0$. After the phase III trial, the company gets a reward denoted by $l \geq 0$ which depends on the probability that the drug is concluded effective by a phase III clinical trial. This probability is given by the power function of the test denoted by $\kappa(p)$. Then the gain may be expressed as

$$-nk - m + l\kappa(p),$$

which is 0 (baseline value) less the expenses in phase II and phase III plus the reward after phase III. More gain functions are defined in the next chapter.

Chapter 3

Clinical trials

In the introduction, we mentioned that in drug development, clinical trials are categorized into three phases. In this chapter, we will first in Section 3.1 give the broader definition of a clinical trial and the definition of clinical trials in the development of a new drug and then in Sections 3.2, 3.3 and 3.4 respectively, we will give the objective and review some of the statistical models used to design and analyse clinical trials in phase I, phase II and phase III. Most of the models reviewed in this chapter will assume that the clinical trials are carried out in the traditional set-up where each phase of a clinical trial is carried out separately.

3.1 What is a clinical trial?

In this section, we define a clinical trial, describe the drug development process, and describe the different phases of a clinical trial. The section was compiled from various literature. Some of the text books used are Wang and Bakhai (2006) and Cook and DeMets (2008). The papers reviewed in later sections were also used in developing this section. A clinical trial is a research study to test how well a new intervention such as a new therapy or

a different mode of administration of an existing drug works on people. We will consider a clinical trial in the development of a new drug. The broad aim of a clinical trial in the development of a new drug is to find out whether there is a dose (or dose range) and schedule at which the drug can be shown to be simultaneously safe and effective, to the extent that the risk-benefit relationship is acceptable. The particular subjects who may benefit from the drug, and the specific indications for its use, also need to be defined.

The modern drug development process involves a series of experiments that are carried out with specific objectives. First, tests are carried out in the laboratory in isolation from living organisms. After obtaining promising results, the next step is to test the new substance in animals (animal pharmacological studies) before the testing can proceed to human beings. The testing in human beings is what is referred to as a clinical trial and is categorized into phase I clinical trials, phase II clinical trials and phase III clinical trials.

Phase I is the stage where the drug is first tested in human beings. The primary objective is to determine the safety of the new therapy. Several dose levels are made available for testing. The dose levels are determined from the animal pharmacological studies. If a safe dose (or dose range) is found, the drug is then tested for biological activity (anti-disease activity) in a small clinical trial. Such a trial is referred to as a phase II clinical trial. Before the product is released into the market, a confirmatory trial (phase III trial) has to be carried out. While phase I and II trials could include only a treatment arm, phase III trials are almost always randomized studies comparing a control (standard therapy) arm and a treatment (new drug) arm.

3.2 Phase I clinical trials

The primary objective of phase I clinical trials is to study toxicity of the new drug and determine a dose that has acceptable toxicity (tolerable dose) for further testing. In this section we give the basic set-up of phase I clinical trials and give some of the designs used

to achieve this primary objective. The basic set-up of phase I clinical trials described in Section 3.2.1 is generally adopted from combining materials in the articles later cited in this section.

3.2.1 Basic set-up of a phase I clinical trial

Phase I clinical trials are typically small, having as few as 10 participants while rarely exceeding 30 participants. Except for cancer trials (oncology), where subjects are usually patients who are at an advanced stage of the disease and/or have failed to respond to the standard therapies, healthy volunteers are used. In oncology, sick patients are used because potential cancer drugs are known to be highly toxic and it would be unethical to administer them to healthy volunteers who have not been diagnosed with cancer. These are normally patients who have not responded to existing therapies. Since in oncology the subjects are patients, it may be desired that most of the patients available are allocated to the dose that will be proposed for testing in the next phases of a clinical trial. This is to enable them have maximum benefit in case the new cancer therapy has therapeutic effect on this group of patients.

Most designs, such as those proposed by O'Quigley et al. (1990), Babb et al. (1998) and Durham et al. (1997) among others, have been developed for cancer trials but can be modified for other therapies. Suppose k different doses, $d_1 < d_2 < ... < d_k$, are chosen for consideration in an oncology trial and we wish to establish the maximum tolerated dose (MTD). We define MTD as the dose, d^*, for which the probability of a medically unacceptable dose limiting toxicity (DLT) is equal to some specified value θ. That is MTD is the dose d^* such that

$$\text{Prob}\{\text{DLT}|d^*\} = \theta.$$

The value of θ is the maximum accepted probability of a DLT and is chosen depending on the nature of the DLT and the potential benefit expected from the drug. The reason that it

is not necessarily the safest dose that is sought is because it is widely assumed that toxicity is a prerequisite for antidisease activity such as antitumor activity in cancer treatment. The MTD d^* is not necessarily one of the experimental doses $d_i's,\ i = 1, ..., k$. It is hoped that the lowest dose (d_1) is safe and that

$$d_1 \leq \text{MTD} \leq d_k.$$

Most investigators assume that there is an underlying dose-response relationship but not all phase I designs explicitly involve fitting the dose-response curve. Each dose level has a corresponding probability of DLT. The probability of DLT is assumed to be monotonic increasing in dose. Diagrammatical representation of the set-up is given by Figure 3.1. The experiment is performed using k dose levels $d_1,\ ...,\ d_k$ whose respective probabilities of DLT are $\theta_1,\ ...,\ \theta_k$. With the maximum accepted probability of DLT denoted by θ, the dose that corresponds to this value is d^* as is shown in the figure. In Figure 3.1, it has been assumed that the MTD has been captured by the experimental dose range d_1 to d_k.

For safety reasons, the available patients (volunteers) are sequentially entered into the trial in small cohorts. Each cohort usually includes at most three volunteers. The early designs are intuitive and approach the MTD conservatively from the lowest dose (d_1) while recent designs are based on statistical principles where the cohort of volunteers, for example in the Bayesian setting, are allocated to the experimental doses based on the predictive probability of toxicity at the experimental doses.

In addition to allocation based on safety reasons, ethical issues based on other factors are also considered and hence some patients may be allocated to different doses from those which the design proposes. Investigators' opinion may lead to allocation of a trial subject to a different dose from the one the design proposes. For example, for a therapy which seems to have therapeutic effect, if a design based on safety only proposes a lower dose whereas a higher dose may be fairly safe, the investigator(s) may want a very sick patient to be allocated the higher dose.

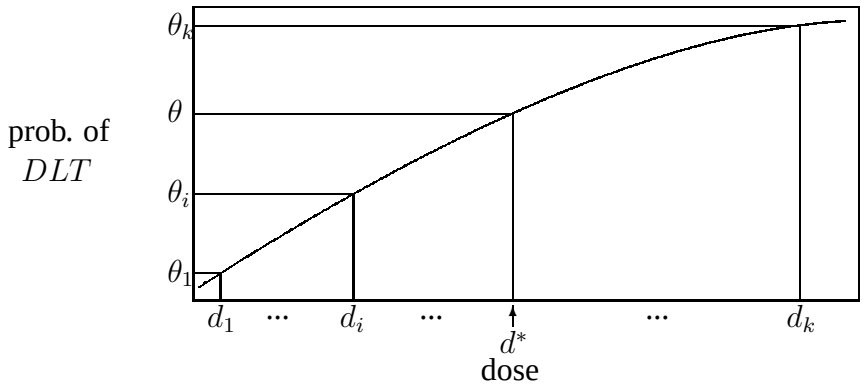

Figure 3.1: Phase I set-up

3.2.2 Early designs

Storer (1989) describes a traditional design (which he calls design A) and also defines three more designs (B, C and D). In design A, whose flowchart is given by Figure 3.2, cohorts of 3 patients are treated at a time starting from the lowest dose. The patients' responses are observed before allocating the next cohort to one of the experimental doses. If no DLT is observed in all 3 patients, escalation to the next higher dose occurs. If 2 or 3 DLTs are observed, the MTD is reached and the trial stops. If only 1 patient experiences a DLT, 3 more patients are allocated to the same dose and if no extra DLT is observed, escalation again continues; otherwise the MTD is reached and the trial stops.

Designs B, C and D are "up and down" schemes in which both escalation and de-escalation takes place. In design B, one patient is treated at a time. If a DLT occurs, the next patient is treated at the next lower dose; otherwise escalation to the next higher dose takes place. The only difference between designs C and B is the escalation rule. For design C, escalation takes place after two consecutive patients treated at the same dose do not experience a DLT. For design D, cohorts of 3 patients are treated at a time. Escalation occurs if no DLT is observed and de-escalation if more than 1 DLT is observed. If only 1 DLT is observed, then the next cohort is treated at the same dose level. A flowchart of this procedure is given in Figure 3.3. The difference between designs A and D is that D

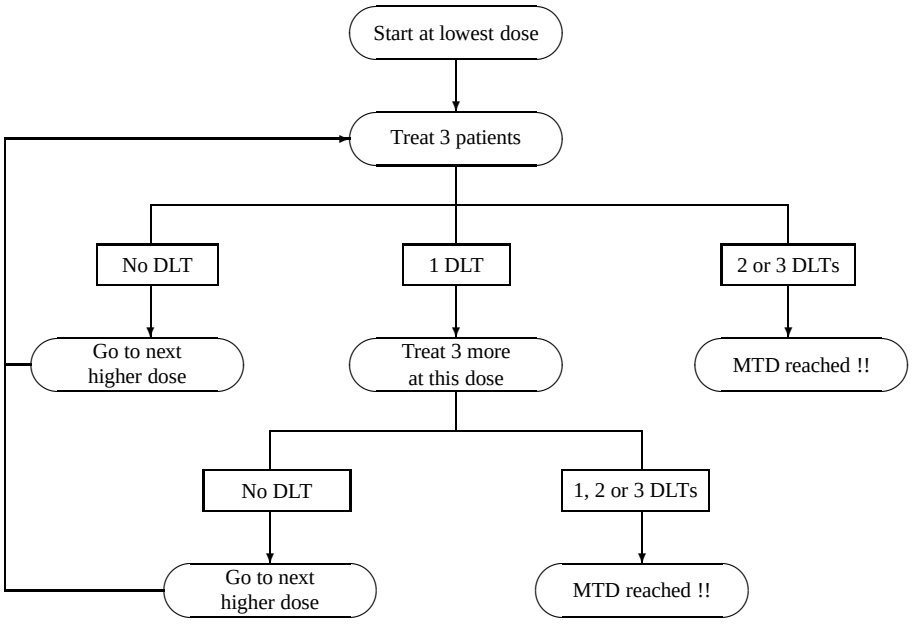

Figure 3.2: Flowcharts of the traditional design (Design A)

allows de-escalation to lower doses and all available patients are entered in the trial. For this design, there is no outcome that leads to stopping so that all available patients are tested.

Storer has proposed two two-stage designs, denoted by BC and BD which combine the single-stage designs (that is B followed by either C and D). The first stage follows design B until the first toxic response occurs. From the point at which the next patient is entered at the next lower dose level, the second stage design (C or D) is implemented. He showed the two stage designs (BC and BD) estimated the MTD with reduced bias relative to the single stage designs A, C and D.

In all these designs (A, B, C, D, BC and BD), the dose to be recommended for testing in further clinical trials depends on the results of the highest dose administered to the participants. If this highest dose administered is deemed nontoxic, it is recommended for further testing. Otherwise, the immediate lower dose is recommended.

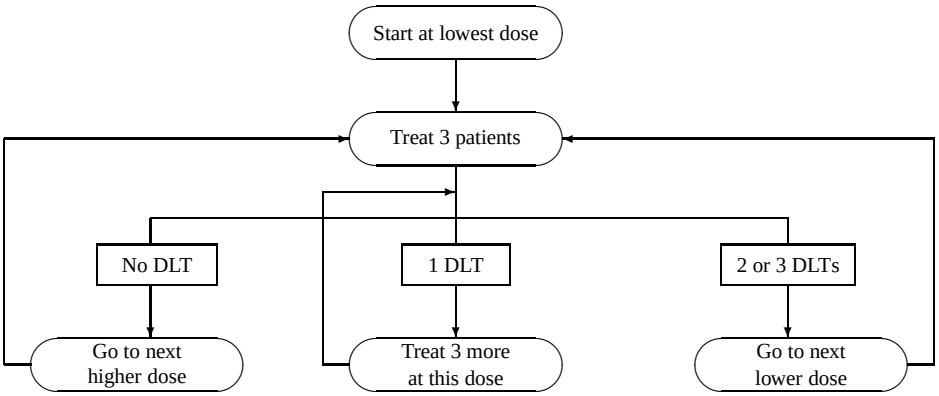

Figure 3.3: A variation of the traditional design (Design D)

3.2.3 The continual reassessment method

The Continual Reassessment Method (CRM) was developed by O'Quigley et al. (1990) for a cancer trial. Several authors, for example Babb et al. (1998), Whitehead et al. (2006), Fan and Wang (2006) and Durham et al. (1997) among others, have compared their methods' operating characteristics with those of CRM. Let X_j be a binary random variable (that is, $X_j \in \{0, 1\}$), where 1 denotes occurrence of a DLT and 0 nonoccurrence of a DLT for the j^{th} patient ($j = 1, ..., n$) entered in the trial. Further, as above, let d^* (not necessarily one of the experimental dose levels $d_1 < d_2 < ... < d_k$) be the MTD. The probability of DLT is modeled by a simple dose-response curve $\psi(d, a)$ that depends on the dose level d and a single parameter a. The dose-response function is assumed to be monotonic in d and a and that for some a, say a_0, from the set $\mathcal{A}$ of possible values of a, we have $\psi(d^*, a_0) = \theta$, where θ is the maximum accepted probability of DLT.

The version of the CRM proposed by O'Quigley et al. (1990) uses the Bayesian principle where the parameter a is considered to be a random variable. Let $(x_1, ..., x_{j-1})$, the data before experimentation on the j^{th} patient, be denoted by $\mathbf{x}_j$ and let $\pi(a|\mathbf{x}_j)$ denote the prior density of the parameter a before experimentation on the j^{th} patient. The form of $\pi(a|\mathbf{x}_j)$ is given later. When $j = 1$, $\pi_0(a) = \pi(a|\mathbf{x}_1)$ is the prior density for a before the

experimentation. They take $\mathcal{A} = (0, \infty)$ so that

$$\int_0^\infty \pi(a|\mathbf{x}_j)da = 1, \ (j = 1, ..., n).$$

Using the accumulated information on the $(j-1)$ patients responses, the probability of DLT at dose level i (denoted by θ_{ij}) is estimated by

$$\theta_{ij} = \int_0^\infty \psi(d_i, a)\pi(a|\mathbf{x}_j)da, \ \ (i = 1, ..., k). \tag{3.1}$$

This is the expected value of the probabilities over $\mathcal{A}$. As an approximation to equation (3.1), O'Quigley et al. (1990) suggest one could obtain the posterior mean of a and substitute this in the dose-response function resulting in a simple to evaluate estimate of θ_{ij} given by

$$\theta'_{ij} = \psi(d_i, \bar{a}(j)), \ \ (i = 1, ..., k), \ \ \bar{a}(j) = \int_0^\infty a\pi(a|\mathbf{x}_j)da.$$

We continue explanation of the CRM using θ'_{ij} but the same procedure would be followed if one chose to use θ_{ij}. In order to determine the best dose to allocate to the jth patient, the estimates of probabilities of DLT θ'_{ij}, ($i = 1, \ ..., \ k$) are compared with the accepted proportion of DLT θ by defining some measure of distance Δ of θ'_{ij} from θ. A commonly used choice is the absolute difference $\Delta(\theta'_{ij}, \theta) = |\theta'_{ij} - \theta|$. The j^{th} entered patient is assigned to the dose d_i such that $\Delta(\theta'_{ij}, \theta)$ is minimized.

Given the response of the j^{th} patient, which updates the knowledge about a_0, the posterior distribution $\pi(a|\mathbf{x}_{j+1})$ is obtained from $\pi(a|\mathbf{x}_j)$ using Bayes' formula given by equation (2.2). The likelihood of the outcome for the j^{th} patient is Bernoulli given by

$$\phi(d(j), x_j, a) = (\psi(d(j), a))^{x_j}\{1 - \psi(d(j), a)\}^{1-x_j},$$

where $d(j)$ is the dose allocated to the j^{th} patient and $\psi(d(j), a)$ is the probability of DLT given by the dose-response curve $\psi(d, a)$. The prior distribution of a before experimenta-

tion with j^{th} patient is $\pi(a|\mathbf{x}_j)$ so that the posterior distribution has density equal to

$$\begin{aligned}\pi(a|\mathbf{x}_{j+1}) &= \frac{\phi(d(j),x_j,a)\pi(a|\mathbf{x}_j)}{\int_0^\infty \phi(d(j),x_j,u)\pi(u|\mathbf{x}_j)du} \\ &= \frac{\pi_0(a)\prod_{l=1}^{j}\phi\{d(l),x_l,a\}}{\int_o^\infty \pi_0(u)\prod_{l=1}^{j}\phi\{d(l),x_l,u)\}du}.\end{aligned}$$

Patients are entered in this way until the results of the last patient entered are available. The recommended dose level for further testing will be the dose d_i $(i = 1, ..., k)$ such that $\Delta(\theta'_{i,n+1}, \theta)$ is minimized. As seen in the allocation of the patients to the dose levels, the design takes into consideration the large potential gain to the patients by aiming to treat as many patients as possible at the MTD. This makes it superior to the designs that begin testing at the lowest dose; these designs tend to under-treat more patients particularly if the MTD is the highest dose considered for experimentation.

O'Quigley and Shen (1996) proposed a likelihood based version of the CRM (CRML). Suppose $(j-1)$ subjects have been entered in the trial and a dose-response function is defined as before, then the likelihood is equal to

$$L(a) = \prod_{l=1}^{j-1}(\psi(d_l,a))^{x_l}\{1-\psi(d_l,a)\}^{1-x_l},$$

where $d_l \in \{d_1, ..., d_k\}$ is the dose level allocated to patient l. To obtain an estimate for a using the maximum likelihood method, the derivative of the logarithm of the above expression is obtained which results in the score function

$$U(a) = \sum_{l=1}^{j-1}\{x_l\frac{\psi'}{\psi}(d_l,a)\} - \sum_{l=1}^{j-1}\{(1-x_l)\frac{\psi'}{1-\psi}(d_l,a)\}. \tag{3.2}$$

When there is heterogeneity in the outcomes (that is, some patients with DLT and some without DLT), then the equation $U(a) = 0$ has a solution. The solution is given by $a = \hat{a}_j$, the maximum likelihood estimate of a. The maximum likelihood estimate for probability of a DLT at dose d_i for patient j is $\psi(d_i, \hat{a}_j)$, where $\hat{a}_j$ is assumed to exist. Patient j is

allocated to dose d_i such that $\Delta(\psi(d_i, \hat{a}_j), \theta)$, $i = 1, ..., k$, is minimized. The recommended dose level for further testing will be the dose d_i such that $\Delta(\psi(d_i, \hat{a}_{n+1}), \theta)$ is minimized.

Before heterogeneity, that is, when all patients experience DLTs or all patients do not experience DLTs, the equation $U(a) = 0$ has no solution so that it is not possible to obtain the maximum likelihood estimate of a and consequently the maximum likelihood estimate $\psi(d_i, \hat{a}_j)$. Before heterogeneity in results is observed, O'Quigley and Shen (1996) suggest using the Bayesian CRM or one of the early designs described above until a DLT is observed if the first outcome is a non-DLT or vice versa. This is because the early designs do not involve estimating a parameter while allocating patients to a dose and in the Bayesian CRM, a prior distribution for a is defined which is updated by the data so that we do not have problem of estimating a through maximum likelihood estimate. Comparison using different starting procedures show that the final results, that is the dose recommended for testing in the next phases of a clinical trial, are largely robust to the method used before heterogeneity is achieved. Operational characteristics would be expected to differ when the lower doses have a very low probability of DLT where starting with the traditional design, more patients are allocated to the lower dose levels. However, the probabilities of recommending the experimental doses for further testing are similar to starting with the Bayesian CRM. Comparison of likelihood CRM and Bayesian CRM using simulation studies indicated similar results.

O'Quigley and Shen (1996) performed simulation studies and compared the probabilities of recommending the experimental doses for further testing using the following three methods; (i) the Bayesian CRM, (ii) CRML while starting with traditional design until heterogeneity is observed and (iii) CRML while starting with Bayesian CRM until heterogeneity is observed. The probabilities of recommending each of the experimental doses for testing in the next phases of a clinical were similar for the three methods. Of particular interest is the scenario for which the lower doses have a very low probability of DLT, where when the traditional design is used before heterogeneity, many patients are allocated

to the lower doses. Despite the difference in operating characteristics with this scenario, the probabilities of recommending experimental doses for testing in the next phases of a clinical were similar in the three methods. The Bayesian CRM and CRML started with Bayesian CRM used the same prior distribution for a and the probabilities of recommending the experimental doses were very close. O'Quigley and Shen (1996) observed that if less informative prior distributions were used, the results of these two methods would be closer.

3.2.4 Overdose control

An attractive idea for phase I clinical trials is to impose a safety constraint in order to minimize the chance of exposing patients to dose(s) with probability of DLT above that of the MTD. This can be achieved by requiring that a dose d cannot be administered if the predictive probability of a DLT at that dose is greater than a pre-specified value given the already collected data. Whitehead et al. (2006) propose an explicit consideration which they argue is more transparent. For example, using the Bayesian CRM, safety may be incorporated by allocating the j^{th} patient to dose d_i such that $\Delta(\theta_{ij}^{'}, \theta)$ is minimized and $\theta_{ij}^{'} \leq \theta_T$, where Δ, $\theta_{ij}^{'}$ and θ are as defined above and θ_T is the probability of DLT which would be considered too high to allocate patients.

Alternatively, the constraint can be incorporated in the statistical model. Babb et al. (1998) have proposed a phase I clinical trial design that incorporates safety in the statistical model. The model selects a dose for each patient so that the predicted probability that the dose exceeds the MTD is less than or equal to some pre-specified value α. This is accomplished by also considering the MTD to be random variable with a prior distribution and then computing the posterior cumulative distribution function (CDF) of the MTD. We will only give the rule of how the patients are allocated to the doses and not the details of how to obtain the distribution of MTD (Babb et al. (1998) give an example of the distribution of

MTD for binary outcomes). For the j^{th} $(j = 1, ..., n)$ patient, if allocation is to a dose d, the probability that d exceeds the MTD is related to the posterior CDF of the MTD and is given by the function π_j defined as

$$\pi_j(d) = \text{Prob}\{\text{MTD} \leq d|\mathbf{x}_j\},$$

where $\mathbf{x}_j$ is the data at the time of j^{th} patient, that is, the responses and the dose levels administered. Hence, π_j is the conditional probability that dose d exceeds the MTD given the currently available data. Based on this criteria, the j^{th} patient is allocated to the dose level d_i such that

$$\pi_j(d_i) = \alpha.$$

That is, each patient is allocated to a dose so that the predicted probability it exceeds the MTD is equal to α. Babb et al. (1998) assume that any dose is available within the experimental dose range. If only a distinct number of doses are available, the j^{th} patient may be allocated to the highest dose level d_i such that

$$\pi_j(d_i) \leq \alpha.$$

3.3 Phase II clinical trials

The primary objective of phase II clinical trials is to study efficacy of the new drug in comparison with the standard treatment(s). Hence, although such studies can be carried out in a single arm setting, the trials are inherently comparative. In addition to efficacy, consideration of toxicity (safety) and cost of the trial may also be incorporated in a phase II trial. The trials are used to determine whether to proceed to a phase III trial depending on the efficacy level, evaluation of toxicity and cost involved in the development of the drug. Designs utilizing frequentist techniques as well as purely Bayesian and Bayesian decision

techniques have been proposed. After outlining the set-up of phase II clinical trials, we will give a review of the popular designs and the emerging new designs.

3.3.1 Set-up of phase II clinical trials

For ethical reasons, it is often important to monitor the outcomes for patients in a phase II clinical trial. For this reason, phase II trials are sometimes designed such that at least two inspections are carried out so that there are opportunities to stop early either for futility or highly promising results before all the patients available for phase II testing are entered into the trial. Suppose there are k inspections where all remaining patients are entered into the trial after the $(k-1)^{th}$ inspection. At the i^{th} inspection $(i = 1, ..., k-1)$, three actions (decisions) can be taken

- Action A: Stop the phase II study and abandon development of the drug
- Action P: Stop the phase II study and proceed to phase III study
- Action C: Continue with the phase II study and make the $(i+1)^{th}$ inspection.

Focussing on studies in which actions are based only on efficacy, Action A is taken when evidence of efficacy is below a certain level so that the new drug is not promising. The motivation for Action A is that patients should not continue to be exposed to a drug that is clearly not effective. Action P is taken when the evidence for efficacy of the new drug is high enough to mean that more evidence on efficacy from subsequent inspections in phase II is not required. The motivation for Action P is that if based on accumulated data at the i^{th} $(i = 1, ..., k-1)$ inspection, there is high probability the new drug is more beneficial compared to the standard drug, the trial should proceed to phase III stage to reduce the development duration for the new drug. Reducing the development duration avoids delay of potential benefit to the society if the new drug will be concluded better than the standard treatment after the phase III stage and saves cost for the drug company because

fewer patients are recruited and treated. Also, reducing the development duration increases profit to the drug company because lesser time of the patent life is used to develop the new drug. On the other hand Action C is taken when the drug shows evidence of efficacy but not strong enough to suggest stopping the phase II testing after the i^{th} $(i = 1, ..., k-1)$ inspection to proceed to phase III testing. At the k^{th} (last) inspection only actions A and P can be taken.

In some settings, not all three actions are considered. For example some trials allow for action P only at the k^{th} inspection; that is, they do not allow for early stopping of phase II due to highly promising results from the new drug and proceeding to phase III before all trial subjects are treated and observed. Decision (action) boundaries depend on the design being utilized. For binary data, it makes sense action P will be taken if enough successes are observed, action A will be taken if too few successes are observed and action C will be taken if the number of successes is between the number of successes required to take action A and the number of successes required to take action P. A pictorial representation of the decision boundaries is shown in Figure 3.4. In the general case, before the last inspection, if all the three actions can be taken at the i^{th} inspection ($i = 1, .., k-1$), two values U_i and L_i $(L_i < U_i)$ are predetermined. The two values are used as the decision boundaries for the action to be taken. Suppose at the i^{th} inspection the total number of treated patients in the phase II trial is n_i and s_i are treated successfully. If $s_i < L_i$, drug development is abandoned (Action A). If $s_i > U_i$, phase II trial is stopped and drug development proceeds to phase III (Action P). On the other hand, if $L_i \leq s_i \leq U_i$, more patients are treated and $(i+1)^{th}$ inspection is made (Action C). If the design does not allow Action P, no upper values $U_i's$ in Figure 3.4 are defined.

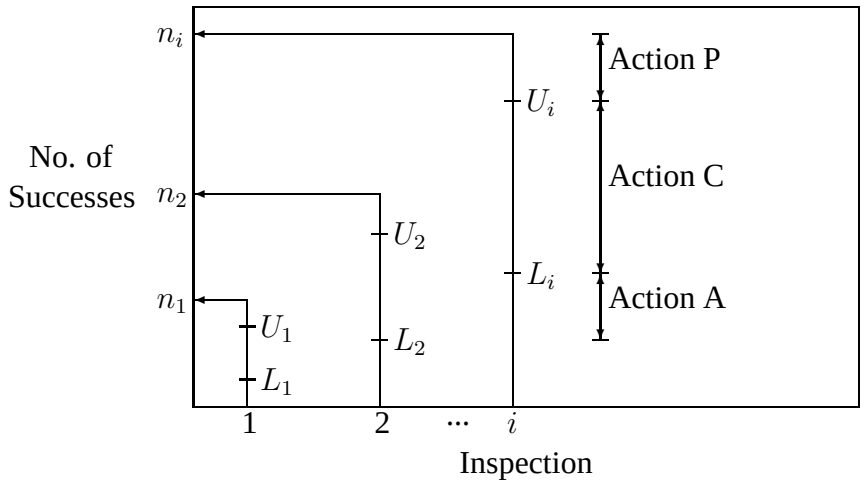

Figure 3.4: A phase II setting allowing for 3 actions at each inspection

3.3.2 Frequentist designs

Frequentist designs focus on determining decision boundaries that control the error rates. Suppose that the true probabilities of success using the standard treatment and the new drug are p_0 and p_1 respectively. Then the new drug may be considered to be sufficiently more efficacious than the control treatment if $p_1 \geq p_0 + \delta$, where $\delta > 0$ is a clinically relevant improvement of the new drug over the standard treatment. The hypothesis would then be to test $H_0 : p_1 = p_0$ Vs $H_1 : p_1 \geq p_0 + \delta$. The experiment is set up such that the error of rejecting H_0 when actually H_0 is true (type I error, usually denoted by α) and the error of concluding H_0 when in reality H_1 is the truth (type II error, commonly denoted by β) are controlled to some specified levels.

Making a type II error in a phase II clinical trial means that treatments that offer larger benefits compared to the existing treatments are rejected based on a small sample size clinical trial. Schoenfeld (1980) notes that investigators do not want to reject treatments with larger benefit on the basis of small sample size trials. Making a type I error means that a new treatment that is not better than the existing treatment is concluded to be better than the existing treatment. Schoenfeld (1980) observes that Type I error is minimized in a large

phase III clinical trial. Hence unlike in phase III clinical trials, Schoenfeld proposes that in a phase II clinical trial, preference should be given to minimizing type II error (hence increasing the power; power $= 1 - \beta$). He proposes setting type II error to less than 0.10 and type I error to less than 0.25.

Gehan (1961) proposed a design which has had considerable application in the past. The design has two stages. He stated that two decisions can be made

- Decision I: Drug is unlikely to be effective in a proportion p_1 of the patients or more
- Decision II: Drug could be effective in a proportion p_1 of patients or more.

When Decision I is made at inspection 1, Action A is taken while if Decision II is made, Action C is taken. There is no opportunity for Action P. Gehan illustrated how to determine the decision boundaries by taking $p_1 = 0.20$. With $p_1 = 0.20$, the chance of consecutive treatment failures is summarized in Table 3.1. The probability of treatment failure is $1 - p_1 = 0.8$. Assuming the observations are independent, the probability of i $(i = 1, ..., 14)$ consecutive failures is $(0.8)^i$. The chance of at least 1 success after i patients will then be given by $1 - (0.8)^i$. For example as shown in the table, the chance of 3 consecutive failures is $(0.8)^3 = 0.8 \times 0.8 \times 0.8 = 0.512$ and the chance of at least 1 success after 3 patients have been treated will be $1 - (0.8)^3 = 0.488$.

Supposing that 14 patients are inspected at the first inspection, Gehan (1961) proposed decision boundaries summarized in Table 3.2. Action A is taken if they are 14 consecutive failures while Action C is taken if 1 or more treatment successes out of the 14 patients is observed. With $p_1 = 0.2$, the probability of taking Action A is 0.044. That is, type II error is controlled at less than 5% and hence the chance of concluding the drug may be working when true probability of success is 0.2 (power) is greater than 95%. On the other hand, using the same ideas, if one is prepared to accept a type II error rate of 0.1, 11 patients are required at the first inspection where as before Action A is taken if there are 11 consecutive failures and Action C is taken otherwise.

Table 3.1: Chance of successive treatment failures when probability of success is 0.2

CONSECUTIVE PATIENTS	CHANCE OF TREATMENT FAILURE IN GIVEN CONSECUTIVE NUMBER OF PATIENTS
1	0.8
2	0.8×0.8=0.64
3	0.8×0.8×0.8=0.512
.	.
.	.
8	0.168
.	.
.	.
11	0.086
.	.
.	.
14	0.044

The number of additional subjects for the second stage is determined so that the true effectiveness of the drug is estimated with a given precision, i.e, standard error. The standard error of the estimated proportion of the treatment successes after the first sample of n_1 patients is

$$\sqrt{\frac{p(1-p)}{n_1}},$$

where p is the proportion of treatment successes in the first sample and n_1 the size of the first sample. If the proportion of successes is approximately the same for future patients, the standard error with the total number of patients is about

$$\sqrt{\frac{p(1-p)}{n_2}}, \tag{3.3}$$

where n_2 is the combined size of the first and the second stage samples and p is the same as above. The second sample number $(n_2 - n_1)$ can be determined so that approximately the required precision will result. It is hoped that p is near the true rate of treatment successes. A more conservative value of p to substitute in equation (3.3) would be the 75% confidence limit for the true rate of successes as derived from the first sample.

Table 3.2: Decision boundaries at inspection 1 using Gehan's method

TREATMENT SUCCESSES	ACTION
0	Drop drug
1	Include more patients in
.	
.	study to pinpoint
.	
14	effectiveness

The sample size for second stage using Gehan's (1961) method depends on the success rate in the first stage. Also, Gehan's design controls the error rates for the first inspection only. Simon (1989) proposed an optimal two-stage design that like Gehan's method allows for Actions A and C but the second stage sample size does not depend on first stage success rate and his design controls the error rates for the entire phase II trial. At the first inspection, the number of successes, S_1, from n_1 patients is observed. A lower bound L_1 is predetermined so that if $S_1 \leq L_1$, action A will be taken. Otherwise action C is taken, with a further $(n_2 - n_1)$ treated at the second stage. A lower bound L_2 for the second stage is also set such that if the total number of successes (in both stages) $S_2 \leq L_2$, development of the drug will be abandoned.

The probability of treatment success depends on the true probability of success, p, for the new drug. Assuming that the responses from the patients are independent and identically distributed as Bernoulli with parameter p the probability of i $(i = 0, 1, ..., n_1)$ successes in the first stage is Bin(n_1, p). Thus the probability of abandoning the drug at first stage, that is prob$(S_1 \leq L_1)$, is given by

$$\sum_{i=0}^{L_1} \binom{n_1}{i} p^i (1-p)^{n_1 - i} = F_B(L_1; p, n_1), \tag{3.4}$$

where F_B denotes the cumulative distribution function of a binomial distribution. Action A is taken at the end of the second stage if $S_1 = i$ (for $i \geq L_1 + 1$) and $(S_2 - S_1) \leq (L_2 - i)$.

Thus the probability of proceeding at stage 1 and abandoning at stage 2 is expressed as

$$\sum_{i=L_1+1}^{n_1} \text{prob}(S_1 = i \text{ and } S_2 - S_1 \leq L_2 - i; p).$$

Since $(S_2 - S_1)$ is binomial with parameter vector $(n_2 - n_1, p)$, with $(S_2 - S_1)$ independent of S_1, the above probability is

$$\sum_{i=L_1+1}^{n_1} \sum_{j=0}^{L_2-i} \binom{n_1}{i} p^i (1-p)^{n_1-i} \binom{n_2-n_1}{j} p^j (1-p)^{n_2-n_1-j}$$

$$= \sum_{i=L_1+1}^{n_1} f_B(i; n_1, p) F_B(L_2 - i; n_2 - n_1, p), \tag{3.5}$$

where f_B denotes a binomial mass function and as before F_B denotes the cumulative distribution function of a binomial distribution.

The expected sample size is $EN = n_1 + (1 - \text{PET})(n_2 - n_1)$ where PET is the probability of early termination after the first stage. Parameters p_0, p_1, α and β are specified and then the two-stage design that satisfies the error probability constraints and minimizes the expected sample size when the response probability is p_0 is determined. Optimization is taken over all values of n_1 and $(n_2 - n_1)$ as well as L_1 and L_2. This is found by searching over the range of $L_1 \in (0, n_1)$ and for each value of L_1 determine the maximum L_2 that satisfies the type II error.

3.3.3 A Bayesian design

Thall and Simon (1994) have proposed a Bayesian design for phase II clinical trials. Let E denote the new (experimental) drug and S the standard (control) drug and that all patients entered in the trial receive new drug. Further let p_E and p_S respectively denote the probabilities of success after treatment with the new drug and the standard treatment. The prior distributions for p_E and p_S are respectively denoted by $\pi_0(p_E)$ and $\pi_0(p_S)$. Because Thall and Simon (1994) assume that all patients in the phase II trial will receive the new drug, the posterior distribution of p_S after the phase II trial is equal to its prior distribution $\pi_0(p_S)$.

Let the response for the j^{th} patient in the phase II clinical trial, X_j ($j = 1, 2, ...$) take values 0 and 1 for treatment failure and successful treatment respectively. Assuming the responses $X_j^{'}s$ are independent, the total number of successes after n patients, $S_n = X_1 + X_2 + ... + X_n$, is Bin$(n, p_E)$. Suppose in an experiment after n patients, s_n are treated successfully, then the posterior distributions for p_E after observing data (s_n, n) is denoted by $\pi(p_E|s_n, n)$. Assuming an improvement of size δ is of medical significance, the objective is to determine the probability that the effect of the new treatment (p_E) is greater than the effect of the standard treatment plus δ ($p_S + \delta$) expressed as

$$\begin{aligned} \lambda(s_n, n; \pi_S, \pi_E, \delta) &= \text{Prob}(p_S + \delta < p_E | S_n = s_n \text{ out of } n) \\ &= \int_{p_S=0}^{1-\delta} \int_{p_E=p_S+\delta}^{1} \pi(p_E|s_n, n)\pi_0(p_S) dp_E dp_S. \end{aligned} \tag{3.6}$$

Figure 3.5 demonstrates the range of the parameter values used to obtain the probability. Since we want to determine the probability that the new drug is better than the control by effective size δ, p_S and p_E are integrated over values such that $(p_E - p_S) \geq \delta$. Hence parameter p_s is allowed to take values from 0 to $1 - \delta$, since beyond $1 - \delta$, $(p_E - p_S)$ will be less than δ. The parameter p_E is similarly integrated from $p_S + \delta$ to 1 to make sure that $(p_E - p_S) \geq \delta$.

Thall and Simon (1994) proposed beta prior distributions for both p_E and p_S. Suppose that $\pi_0(p_E)$ is Beta(a_E, b_E) and $\pi_0(p_s)$ is Beta(a_S, b_S). Since there is no experimentation with the control treatment, the posterior distribution of p_S is also Beta(a_S, b_S). For the new drug the likelihood is Binomial so that following the discussion of Section 2.2.1, the posterior distribution $\pi(p_E|s_n, n)$ is Beta$(a_E + s_n, b_E + n - s_n)$ and since

$$\int_{p_s+\delta}^{1} f_\beta(p_E; a_E + s_n, b_E + n - s_n) dp_E = 1 - F_\beta(p_S + \delta; a_E + s_n, b_E + n - s_n),$$

where f_β and F_β are respectively the probability density function and the cumulative distribution function of a beta distribution, then equation (3.6) simplifies to

$$\int_0^{1-\delta} \{1 - F_\beta(p_S + \delta; a_E + s_n, b_E + n - s_n)\} f_\beta(p; a_s, b_s) dp_S, \tag{3.7}$$

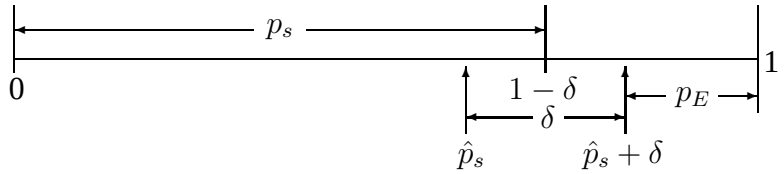

Figure 3.5: Range of the parameter values

where f_β and F_β are as defined above.

Thall and Simon assume the parameters $a_S,\ b_S,\ a_E$ and b_E can be elicited from the investigators and the parameters represent pseudo-patients. For example a_S and b_S are elicited such that if (a_S+b_S) patients are treated with the standard drug, then a_S would have successful responses to the treatment while b_S will not respond positively to the standard drug. Similarly, a_E patients would be treated successfully after $a_E + b_E$ are treated with the new drug. Thall and Simon assume an informative prior distribution $\pi_0(p_S)$ and an at most slightly informative prior distribution $\pi_0(p_E)$. They suggest eliciting and quantifying the prior distributions by setting width of the 90% interval (W_{90}) and examining the Beta curves as described in Section 2.2.1.

The design allows for the three actions (A, P and C) stated in Section 3.3.1. To determine the decision boundaries, a small value p_L such as (0.01-0.05) and a large value p_U such as (0.95-0.99) for equation (3.7) are predetermined. Let λ denote the expression (3.7) after the prior distributions ($\pi_0(p_E),\ \pi_0(p_S)$) and parameter values $s_n,\ n$ and δ are given. The lower and upper cut-offs are then given by

$$U_n = \text{smallest integer } s_n \text{ such that } \lambda(s_n, n, \pi_S, \pi_E, 0) \geq p_U$$
$$L_n = \text{Largest integer } s_n < U_n \text{ such that } \lambda(s_n, n, \pi_S, \pi_E, \delta) \leq p_L.$$

The decision rule after n patients are treated is:

if $S_n \leq L_n,$ take action A,

if $S_n \geq U_n,$ take action P, and

if $L_n < S_n < U_n$ and $n < n_{max},$ take action C,

where n_{max} is the maximum number of patients that can be entered into the phase II clinical trial.

3.3.4 A Bayesian decision design

The decision boundaries (rules) for the frequentist and Bayesian designs described in Sections 3.3.2 and 3.3.3 respectively depend only on the number of successfully treated patients. Fully Bayesian decision theory techniques can be used to define gain function which incorporate other measures such as the monetary gain for the pharmaceutical company. Stallard (1998) has proposed a method for sample size determination for phase II clinical trials using Bayesian decision theory. Here we will dwell more on Stallard's proposal for defining decision boundaries after evaluating data rather than on sample size determination. He defines a gain function that depends on the true efficacy and stage of the trial at which the decision is made. Suppose a maximum of K inspections are planned at phase II and that the i^{th} inspection ($i = 1,\ ...,\ K$) is carried out after a total of n_i patients have been entered into the trial. Further let the true probability of efficacy be denoted by p. Then the gain is a function of p and n_i and for action $a(a \in \{\text{A, P, C}\})$, it is denoted by $G_a(p, n_i)$. Actions A, P and C are as defined in Section 3.3.1. Let X_j be the indicator variable for successful treatment of patient $j,\ j = 1,\ ...,\ n_i$ and $S_{n_i} = \sum_{j=1}^{n_i} X_j$ be the number of successfully treated patients after n_i patients have been treated. After observing data $X_1 = x_1,\ X_2 = x_2, ..., X_{n_i} = x_{n_i}$ with $S_{n_i} = s_{n_i}$, using the Bayesian decision theory principles of Section 2.3, the expected utility from action a is

$$\mathcal{G}_a(s_{n_i}) = E\{G_a(p, n_i)|s_{n_i}, n_i\} = \int_0^1 G_a(p, n_i)\pi(p|s_{n_i}, n_i)dp,$$

where $\pi(p|s_{n_i}, n_i)$ is the posterior distribution of p given the data (s_{n_i}, n_i). The optimal action is the one with largest expected utility.

The baseline for the utility function defined here is 0 so that if the phase II study is assumed to have a cost $k(\geq 0)$ per patient, the utility function for abandoning the trial

(action A) at the i^{th} inspection is given by

$$G_A(p, n_i) = -n_i k$$

which is 0 (baseline value) less the number of patients entered multiplied by the cost per patient.

To proceed to phase III (action P), in addition to the cost of the phase II trial, the gain function needs to incorporate the cost of the phase III trial and the expected reward if the phase III trial shows that the new drug is efficacious. Stallard assumed that the total cost of the phase III trial is fixed and equal to some amount $m(\geq 0)$. The reward $l(\geq 0)$ is taken to depend on the speed with which the drug can be developed. Assuming the length of phase III is fixed, the variability of speed of the drug development will depend on the length of the phase II trial and hence l will be taken to be a function of n_i. Further, the reward will depend on the probability that the drug will be indicated efficacious by the phase III trial. This probability depend on p and is given by the power function of the test denoted by $\kappa(p)$. The utility function for action P at the i^{th} inspection will thus be of the form

$$G_P(p, n_i) = -n_i k - m + l(n_i)\kappa(p).$$

Expectations for the two gain functions corresponding to actions A and P are given by

$$\mathcal{G}_A(p, n_i) = E[G_A(p, n_i)] = -n_i k \tag{3.8}$$

and

$$\mathcal{G}_P(p, n_i) = E[G_P(p, n_i)] = -n_i k - m + l(n_i)E(\kappa(p)|s_{n_i}, n_i) \tag{3.9}$$

respectively, where $E(\kappa(p)|s_{n_i}, n_i)$ which we define as the predictive power in Chapter 5, is the expected value of $\kappa(p)$ obtained using the posterior distribution of p given (s_{n_i}, n_i).

At $i = K$, further continuation (that is action C) is not possible. At this inspection, G_C will be taken to be $-\infty$ so that it will have the least gain among actions A, P and

C. When $i \neq K$, the utility from action C, depends on the action that will be taken at the $(i+1)^{th}$ inspection and subsequent inspections. At the $(i+1)^{th}$ inspection, if $S_{n_{i+1}} = s_{n_{i+1}}$ and the optimal action is taken, the expected utility will be

$$\max_{a\in\{\text{A,P,C}\}}\mathcal{G}_a(s_{n_{i+1}}, n_{i+1})$$

The expected utility from action C at the i^{th} inspection can thus be given recursively by

$$\mathcal{G}_C(s_{n_i}, n_i) = \sum_{S_{n_{i+1}}=s_{n_i}}^{s_{n_i}+n_{i+1}-n_i} \max_{a\in\{\text{A,P,C}\}}\{\mathcal{G}_a(s_{n_{i+1}}, n_{i+1})\} f_{n_{i+1}}(s_{n_{i+1}}|s_{n_i}, n_i), \qquad (3.10)$$

where $f_{n_{i+1}}(s_{n_{i+1}}|s_{n_i}, n_i)$ is the density of $S_{n_{i+1}}$ given $S_{n_i} = s_{n_i}$ given by

$$f_{n_{i+1}}(s_{n_{i+1}}|s_{n_i}, n_i) = \int_0^1 g_{n_{i+1}}(s_{n_{i+1}}|s_{n_i}, p)\pi(p|s_{n_i}, n_i)dp$$

with $g_{n_{i+1}}(s_{n_{i+1}}|s_{n_i}, p)$ the density of $S_{n_{i+1}}$ given $S_{n_i} = s_{n_i}$ and the value of p.

Figure 3.6 gives all possible outcomes at stage $(i+1)$ and the probability of each possible outcome given the outcome at stage i. Suppose at inspection i, s_{n_i} successes are observed. With the $(i+1)^{th}$ inspection carried out after n_{i+1} patients have been treated, at inspection $(i+1)$ an extra $(n_{i+1} - n_i)$ patients are entered so that the extra number of successes takes values $0,\ 1,\ ...,\ (n_{i+1} - n_i)$ and consequently $S_{n_{i+1}}$ can take values $s_{n_i} + 0,\ s_{n_i} + 1,\ ...,\ (s_{n_i} + n_{i+1} - n_i)$. Thus

$$\text{Prob}(S_{n_{i+1}} = s_{n_i} + s_{(n_{i+1}-n_i)}|p) = \text{Prob}(S_{(n_{i+1}-n_i)} = s_{(n_{i+1}-n_i)}|p)$$

which is Bin$((n_{i+1} - n_i), p)$ where $S_{(n_{i+1}-n_i)}$ is the random variable denoting the extra number of successes at inspection $(i+1)$. This is the distribution of $g_{n_{i+1}}(s_{n_{i+1}}|s_{n_i}, p)$.

If K is finite, that is for a truncated test, equation (3.10) can be solved using equations (3.8) and (3.9) using backward induction starting at $i = K$. It is thus possible to compare the utilities for the three actions A, P and C given respectively by equations (3.8), (3.9) and (3.10) and choose the optimal action. For a truncated test, it can be shown that

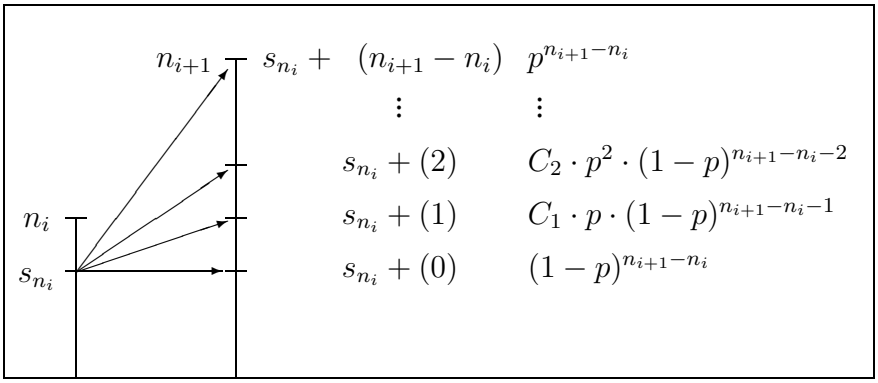

Figure 3.6: Possible outcomes at stage $(i+1)$

functions c and d exist that determine decision boundaries so that

$$\max_{a\in\{\text{A,P,C}\}}\mathcal{G}_a(s_{n_i}, n_i) = \begin{cases} \mathcal{G}_A(s_{n_1}, n_i), & s_{n_i} < c(i) \\ \mathcal{G}_C(s_{n_i}, n_i), & c(i) \le s_{n_i} < d(i) \\ \mathcal{G}_P(s_{n_i}, n_i), & d(i) \le s_{n_i}. \end{cases}$$

3.3.5 Phase II studies based on therapeutic benefit and toxicity

The phase II designs described above focussed only on efficacy data. However, it may be desirable to make the decision on which doses to consider for further testing based on both efficacy and safety data. Both frequestists and Bayesian methods that use both efficacy and safety are available. We will mention several methods but we will describe in detail one frequentist method and two Bayesian methods.

A frequentist method

In the frequentist setting, if both efficacy and safety are considered, the type I error needs to be controlled at some level α. Let X_1 and X_2 denote the outcome variables for efficacy and toxicity (DLT) respectively. Following Pocock et al. (1987), one possible solution is to consider X_1 as the primary endpoint whose p-value for treatment difference is used for the

formal test of hypothesis and toxicity as a subsidiary endpoint requiring exploratory rather than formal interpretation. However, this may sometimes not be desirable so that X_2 may also be used in the hypothesis testing. Because a family of hypotheses (hypothesis testing X_1 and hypothesis testing X_2) are tested, procedures that control the type I familywise error rate (FWER), the probability of rejecting at least one true null hypothesis in the family under any configuration, need to be employed. By under any configuration, we mean when only one null hypothesis is true or both the null hypotheses are true. One possible method as pointed by Geller and Pocock (1987) and Pocock et al. (1987) is the Bonferroni correction, where to control the type I FWER, each variable $X_i,\ (i = 1, 2)$ is tested at level $\alpha/2$. Other methods that may be used to control the type I FWER by adjusting the level of the tests are Šidák's method and Holm's procedure among others. These methods are described in detail in Section 4.3. In this section we will describe the method proposed by Jennison and Turnbull (1993, 2000).

Jennison and Turnbull (1993, 2000) consider pairs $X = (X_1, X_2)$ that have bivariate normal distributions with mean $\mu = (\mu_1, \mu_2)$, correlation ρ and known variances which by appropriate re-scaling are such that var$(X_1) =$ var$(X_2) = 1$. They also assume that X_1 and X_2 are defined such that higher values of μ_1 and μ_2 are desirable. Jennison and Turnbull (1993) further assume that with regard to $X_i\ (i = 1, 2)$, there are constants $\varepsilon_i < \Delta_i$ such that the new drug is preferred if $\mu_i > \Delta_i$ and is unacceptable if $\mu_i \leq \varepsilon_i$, but the region with $\varepsilon_i < \mu_i \leq \Delta_i$ is a region of indifference so that the parameter space for μ is divided into nine preference regions as shown in Figure 3.7 (a). In the pairs, the first position corresponds to X_1 and the second position to X_2. The symbols $-$, 0 and + respectively indicate that the new drug is unacceptable, new drug is indifferent to the standard treatment and the new drug is preferred. After a trial, the objective is to decide whether to accept (A) or reject (R) the new drug so that for each of the nine regions, the investigators will either accept the new drug or drop (reject) it. We have given one example of collapsing the nine regions in Figure 3.7 (b). In this case, the investigators would be interested in a new drug

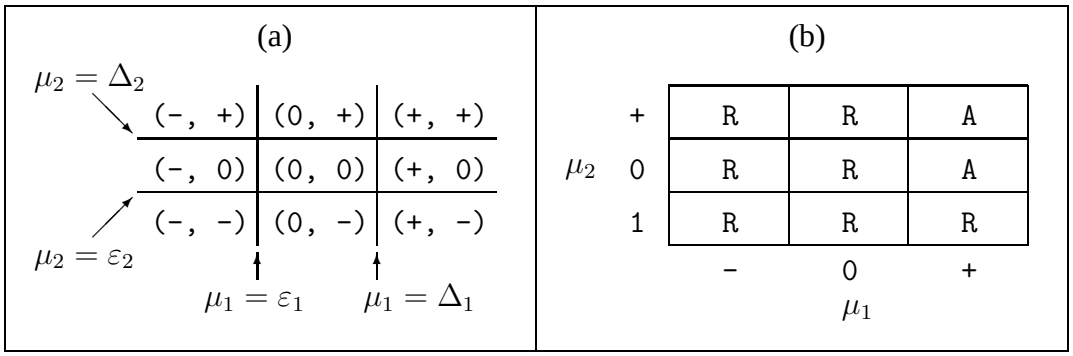

Figure 3.7: (a) The preference regions for $\mu = (\mu_1, \mu_2)$. The first position in the pair correspond to X_1 (efficacy) and the second position to X_2 (toxicity). The symbols +, 0 and - respectively indicates new drug is preferred, considered equivalent and unacceptable. (b) An example of appropriate actions for specified values of μ; R (reject new drug) and A (accept new drug)

that is more efficacious than the standard drug while it is at least as safe as the standard drug. Hence the new drug is preferred if μ_1 and μ_2 are either in the region (+, +) or (+, 0). More examples of categorization are presented in Jennison and Turnbull (1993).

Jennison and Turnbull (1993, 2000) propose a unified method for different categorization based on preferences for the new drug which is achieved by shifting X. For the example presented in Figure 3.7 (b), the shifted random vector is $X - (\Delta_1, \varepsilon_2)$. With the transformation, there is a single region over which to control type I FWER. The type I FWER is controlled at level α if

$$\max\{\mathcal{P}_A(\mu_1, \mu_2); \mu_1 \leq 0 \text{ or } \mu_2 \leq 0\} \leq \alpha,$$

where $\mathcal{P}_A(\mu_1, \mu_2)$ is the probability of concluding that the new drug is preferred when $\mu = (\mu_1, \mu_2)$. Suppose after n patients data $X_{ij};\ i = 1, 2;\ j = 1, ..., n$ are taken. Let $\bar{X}_i = n^{-1}(X_{i1} + ... + X_{in}),\ i = 1, 2$ to be the sample means with the standardized values $Z_i = \bar{X}_i\sqrt{n},\ i = 1, 2$. The decision rule is:

- If $\min(Z_1, Z_2) > \Phi^{-1}(1 - \alpha)$, accept new drug;
- Otherwise, reject new drug.

Jennison and Turnbull (1993, 2000) show that with this decision rule, the type I FWER is controlled at level α. To show this result, they first show that $\mathcal{P}_A(\cdot,\cdot)$ is monotone in both directions for this decision rule. This is accomplished by comparing the probability of concluding that a new drug is preferred based on two sets of bivariate random vectors Z and Z'. Let the first random vector $Z = (Z_1, Z_2)$ be bivariate normal with mean $(\mu_1\sqrt{n},\ \mu_2\sqrt{n})$, var$(Z_1)$ = var$(Z_2) = 1$ and $corr(Z_1, Z_2) = \rho$. Let the second random vector $Z' = (Z_1', Z_2') = Z + (v_1\sqrt{n}, v_2\sqrt{n})$, where $v_1 > 0$ and $v_2 > 0$ so that Z' is distributed as Z except that its mean is greater than the mean of Z and is equal to $([\mu_1 + v_1]\sqrt{n},\ [\mu_2 + v_2]\sqrt{n})$. Since $Z_1' > Z_1$ and $Z_2' > Z_2$ for all values of Z_1 and Z_2,

$$\begin{aligned} \mathcal{P}_A(\mu_1 + v_1, \mu_2 + v_2) &= \text{Prob}\{\min(Z_1', Z_2') > \Phi^{-1}(1-\alpha)\} \qquad (3.11) \\ &\geq \text{Prob}\{\min(Z_1, Z_2) > \Phi^{-1}(1-\alpha) = \mathcal{P}_A(\mu_1, \mu_2)\}. \end{aligned}$$

The equality (3.11) holds from the definition of $\mathcal{P}_A(\cdot,\cdot)$ and the decision rule. The above inequality indicates that $\mathcal{P}_A(\mu_1, \mu_2)$ is monotone increasing in both arguments. Hence, for any value of ρ,

$$\begin{aligned} \max\{\mathcal{P}_A(\mu_1, \mu_2); \mu_1 \leq 0 \text{ or } \mu_2 \leq 0\} &\leq \max\{\mathcal{P}_A(0,\infty), \mathcal{P}_A(\infty, 0)\} \qquad (3.12) \\ &= \max\{\text{Prob}[Z_1 > \Phi^{-1}(1-\alpha)|\mu_1 = 0], \\ &\quad \text{Prob}[Z_2 > \Phi^{-1}(1-\alpha)|\mu_2 = 0]\} \\ &= \alpha, \end{aligned}$$

since Prob$[Z_1 > \Phi^{-1}(1-\alpha)|\mu_1 = 0]$ = Prob$[Z_2 > \Phi^{-1}(1-\alpha)|\mu_2 = 0] = \alpha$. The right hand side of inequality (3.12), by monotonicity of $\mathcal{P}_A(\mu_1, \mu_2)$, represents scenarios where the probability of making type I error is highest. Although Jennison and Turnbull (1993, 2000) consider normally distributed random variables, using the central limit theorem, the method can be used for binary outcomes if the number of patients is large enough. Further, Jennison and Turnbull (1993) have extended the method to allow for more than one inspection.

Bayesian methods

In Bayesian setting the possible outcomes (based on efficacy and toxicity) are assigned some utility values. While focussing on efficacy only, for a binary outcome, binomial models are reasonable. The same is true for toxicity (DLT). When interest is in both efficacy and toxicity, two categories are no longer adequate. Combinations of efficacy and toxicity outcomes will result in more than two categories. Suppose we consider the simplest case where both efficacy and toxicity are binary outcomes. The possible outcomes are shown in Table 3.3. Administration of a drug to a patient will either result in efficacy and DLT (YY), DLT without efficacy (NY), efficacy without toxicity (YN) or neither efficacy nor toxicity (NN). Let l denote the possible categories based on efficacy and toxicity. Loke et al. (2006) propose a method that models the probability of the four outcomes so that $l = 4$. Whitehead et al. (2006) give priority to avoiding a DLT such that the four outcomes YY, NY, YN and NN reduce to $*Y$, YN and NN where $*Y$ means either NY or YY so that $l = 3$. Stallard et al. (1999) noted that these two cases of $l = 3$ and $l = 4$ encompass a very large proportion of phase II clinical trials.

The method by Loke et al. (2006) was intended for phase I trials but like the method proposed by Whitehead et al. (2006), it could be used in early phase II clinical trials where another separate phase II trial is expected to be carried out. In the two methods, all patients are allocated to the experimental treatment. The method by Stallard et al. (1999) includes a control arm and is applicable to the late phase II clinical trials. Loke et al. (2006) and Stallard et al. (1999) assume the outcomes in Table 3.3 have a multinomial density and Dirichlet prior distribution can be elicited. To compare the outcomes, utilities are assigned to the different possible outcomes. Whitehead et al. (2006) model two dose-response curves to estimate the probabilities of the three outcomes $*Y$, YN and NN. In this subsection, we will describe the method by Stallard et al. (1999) because it is different from the other two methods in that it has a control arm. We will also describe the work of Whitehead et al.

Table 3.3: Cross tabulation of toxicity and efficacy

	Toxicity (X_2)	
Efficacy (X_1)	$Yes\ (Y)$	$No\ (N)$
$Yes\ (N)$	YY	YN
$No\ (N)$	NY	NN

(2006) because later in this thesis we borrow ideas from this method.

Stallard's method

The phase II design proposed by Stallard et al. (1999) is a decision theoretic method which allows for more than one data inspection and at any data inspection, the decision theoretic design they propose is flexible enough to allow the three actions (A, P and C) described in Section 3.3.1 or only actions A and C. In their design, they assume a maximum sample size of M patients is available for testing in phase II. In order to determine whether to accept or not to accept early stopping in phase II in favour of the new drug, some $M_1 \leq M$ is pre-determined such that at least M_1 patients are treated before action P can be allowed. $M_1 = 0$ allows proceeding to phase III at any data inspection while $M_1 = M$ only allows proceeding to phase III when all the available patients have been treated.

The possible actions (A, P and C) may be compared using gain functions. The gain functions Stallard et al. (1999) proposed are similar to the gain functions given in Section 3.3.4. As before, let E denote the new drug and S the standard drug and that the probability of outcome $i,\ i = 1, ..., l$ ($l \leq 4$) for treatment $t,\ t \in \{E, S\}$ be denoted by θ_{ti} such that $\theta_{t1} + ... + \theta_{tl} = 1$. Further let the probability vectors $(\theta_{E1}, ..., \theta_{El})'$ for the new drug and $(\theta_{S1}, ..., \theta_{Sl})'$ for the standard drug be denoted by θ_E and θ_S respectively. The addition in the gain functions of Section 3.3.4 is the patient gain which we denote by some function $g(\theta_E, \theta_S)$ for patient treated with E under the pair (θ_E, θ_S). To specify the form of this function utilities are assigned to the l possible outcomes such that the expected utility when a patient is treated with the new drug E is $\mathbf{u}'\theta_E = u_1\theta_{E1} + ... + u_l\theta_{El}$. The corresponding

expected utility for a patient treated with the standard drug is $u^{'}\theta_S = u_1\theta_{S1} + ... + u_l\theta_{Sl}$ and the patient gain could be defined as

$$g(\theta_E, \theta_S) = u^{'}(\theta_E - \theta_S). \tag{3.13}$$

The utilities $u_1, ..., u_l$ may be elicited from the investigators. One way is to assign the best outcome utility value +1 and the worst outcome utility value -1. Other outcomes are elicited such that they take values in the interval $[-1, +1]$.

In Section 3.3.4, the gain function for action A at the i^{th} inspection was defined as $G_A(p, n_i) = -n_i k$ where p was the parameter of interest, n_i the number of patients treated at the i^{th} inspection and k the cost of treating 1 patient. Now when the patient gain is included and p replaced with the new parameter vector (θ_E, θ_S) the gain function becomes

$$\begin{aligned} G_A(\theta_E, \theta_S, n_i) &= n_i g(\theta_E, \theta_S) - n_i k \\ &= n_i\{u^{'}(\theta_E - \theta_S) - k\}. \end{aligned} \tag{3.14}$$

The gain function for action P in Section 3.3.4 was given as $G_P(p, n_i) = -n_i k + l(n_i)\kappa(p) - m$, where m is the total cost of the phase III clinical trial. By including the patient gain for patients treated at the end of i^{th} inspection, the term $-n_i k$, as in equation (3.14), is replaced by $n_i\{u^{'}(\theta_E - \theta_S) - k\}$. Stallard et al. (1999) take the benefit to future patients to be $\Pi g(\theta_E, \theta_S)$ for some $\Pi > 0$. Because $\Pi g(\theta_E, \theta_S)$ is in the same scale as the patient gain, Π may be interpreted as the number of future patients to benefit from treatment with E. The values of Π might also reflect the gains to the clinicians or pharmaceutical companies and thus need not to be equated to a number of potential patients. Thus the term $l(n_i)\kappa(p)$ is replaced by $\Pi g(\theta_E, \theta_S)$ so that

$$\begin{aligned} G_P(\theta_E, \theta_S, n_i) &= n_i\{u^{'}(\theta_E - \theta_S) - k\} + \Pi g(\theta_E, \theta_S) - m \\ &= n_i\{u^{'}(\theta_E - \theta_S) - k\} + \Pi u^{'}(\theta_E - \theta_S) - m. \end{aligned} \tag{3.15}$$

To compare the actions, expectations of equations (3.14) and (3.15) are evaluated. The

expected gain function for action C at i^{th} inspection depends on the action that will taken at $(i+1)^{th}$ inspection and is obtained as was explained in Section 3.3.4.

Whitehead's method

As mentioned before, Whitehead et al. (2006) give priority to avoiding a DLT so that the four outcomes in Table 3.3, YY, NY, YN and NN reduce to $*Y$, YN and NN where $*Y$ means either NY or YY. The probabilities for the three outcomes are respectively denoted by $p_{*Y}(d), p_{YN}(d)$ and $p_{NN}(d)$ and the conditional probability of DO, given no DLT is denoted by $p_{Y|N}(d)$. Two logistic models are used to describe the probabilities;

$$p_{*Y}(d) = \frac{\exp(\alpha_{*Y} + \beta_{*Y} \log\ d)}{1 + \exp(\alpha_{*Y} + \beta_{*Y} \log\ d)} \tag{3.16}$$

$$p_{Y|N}(d) = \frac{\exp(\alpha_{Y|N} + \beta_{Y|N} \log\ d)}{1 + \exp(\alpha_{Y|N} + \beta_{Y|N} \log\ d)}. \tag{3.17}$$

The advantage of modeling the conditional probability ($p_{Y|N}(d)$) is that this does not require modeling the association between DO and DLT. Using the multiplicity probability law $p(A \cap B) = p(B|A) \times p(A)$, we have

$$\begin{aligned} p_{YN}(d) &= p_{Y|N}(d) \times (p_{*Y})^c = p_{Y|N}(d) \times (1 - p_{*Y}) \\ &= \frac{\exp(\alpha_{Y|N} + \beta_{Y|N} \log\ d)}{\{1 + \exp(\alpha_{*Y} + \beta_{*Y} \log\ d)\}\{1 + \exp(\alpha_{Y|N} + \beta_{Y|N} \log\ d)\}}. \end{aligned}$$

Using the law of probability $p_{*Y}(d) + p_{YN}(d) + p_{NN}(d) = 1$, then

$$\begin{aligned} p_{NN}(d) &= 1 - p_{*Y}(d) - p_{YN}(d) \\ &= \frac{1}{\{1 + \exp(\alpha_{*Y} + \beta_{*Y} \log\ d)\}\{1 + \exp(\alpha_{Y|N} + \beta_{Y|N} \log\ d)\}}. \end{aligned}$$

The Whitehead et al. (2006) method for recommending the doses to which the next cohort of patients should be allocated uses the Bayesian principle. The joint distribution

for $(\alpha_{*Y}, \beta_{*Y})$ in model (3.16) and the joint distribution for $(\alpha_{Y|N}, \beta_{Y|N})$ in model (3.17) are elicited separately. Figure 3.8 shows the curves of the prior means for different dose response relationships using the prior distributions that Whitehead et al. (2006) use in their illustrating example. Using the elicited prior distributions, the dose-response curves based on prior means indicate the probability of a DLT (p_{*Y}) and the probability of DO given no DLT ($p_{Y|N}$) increase with dose level. The other curves are derived from these two models. The probability that the administration of a drug to a patient results in neither a therapeutic effect nor a toxic effect (p_{NN}) decreases with the dose level. The highest dose is not necessarily the best choice as the curve of $p_{YN}(d)$ shows. The posterior probability of a therapeutic effect and no toxic outcome increases to some dose level and then decreases. Thus if the investigators' objective is to identify a dose that has the highest chance of therapeutic effect but no DLT, this dose is not necessarily the highest experimental dose level although the probability of DO given no DLT ($p_{Y|N}$) increases with dose level. This is similar to the objective of the dose selection procedure we propose in Chapter 5. We will aim to select the dose that is more efficacious compared to the control treatment and has the probability of DLT less than a specified value.

The joint prior distributions for the parameter vectors $(\alpha_{*Y}, \beta_{*Y})$ and $(\alpha_{Y|N}, \beta_{Y|N})$ are obtained as described in Section 2.2.2. For model (3.16), pseudo-data are used to define the prior distributions at two dose levels d_{i1} $(i = -1, 0)$. These consist of $n_{i1} = a_i + b_i$ pseudo-subjects treated at dose d_{i1}, of whom a_i suffer DLTs. The second subscript on dose d_{i1}, that is 1, is an indicator for model (3.16). Thus assuming the prior distribution of the form (2.6), the prior distribution for $(\alpha_{*Y}, \beta_{*Y})$,

$$\pi_{01}(\alpha_{*Y}, \beta_{*Y}) = \prod_{i=-1}^{0} \frac{p_{i1}^{a_i}(1-p_{i1})^{b_i}}{B(a_i, b_i)} \left|\log\left(\frac{d_{-11}}{d_{01}}\right)\right|, \tag{3.18}$$

where

$$p_{i1} = \frac{\exp(\alpha_{*Y} + \beta_{*Y} \log\ d_{i1})}{1 + \exp(\alpha_{*Y} + \beta_{*Y} \log\ d_{i1})}, \quad i = -1, 0.$$

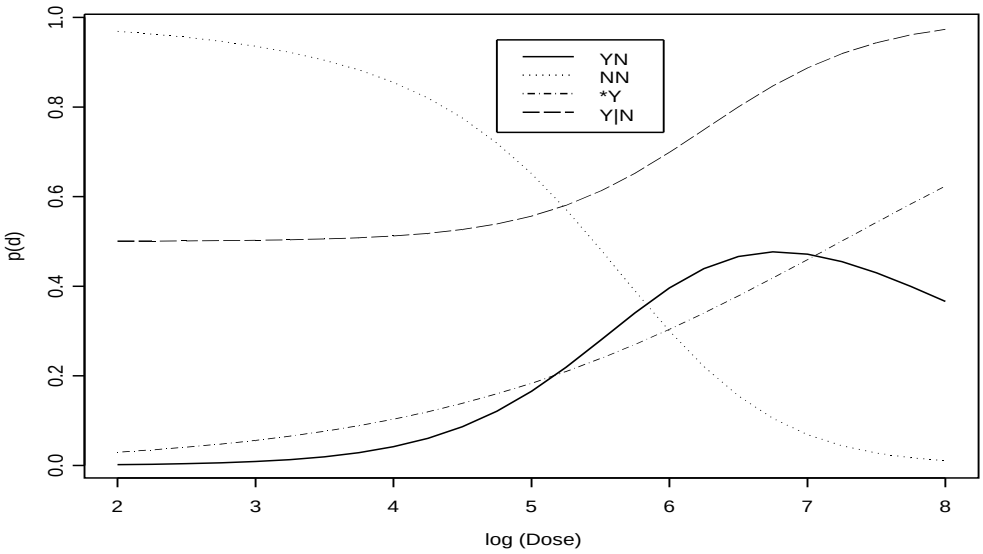

Figure 3.8: Dose response curves using the prior means

After the trial starts, suppose that n_{i1} subjects have been treated with the experimental dose d_{i1} of whom m_i have shown no response, t_i have exhibited a DO without a DLT, and a_i have suffered a DLT so that $m_i + t_i + a_i = n_{i1}$ for $i = 1, ..., k$. Denoting these observed data by $\mathbf{x}$, the posterior distribution will be of the form given by (2.7) so that

$$\pi(\alpha_{*Y}, \beta_{*Y}|\mathbf{x}) \propto \prod_{i=-1}^{k} p_{i1}^{a_i}(1 - p_{i1})^{b_i}, \tag{3.19}$$

where

$$p_{i1} = \frac{\exp(\alpha_{*Y} + \beta_{*Y} \log\ d_{i1})}{1 + \exp(\alpha_{*Y} + \beta_{*Y} \log\ d_{i1})}, \quad i = -1, 0, 1, ..., k.$$

Similarly, for model (3.17), to define the prior distribution for $\alpha_{Y|N}$ and $\beta_{Y|N}$, let $n_{i2} = t_i + u_i$ pseudo-subjects treated at dose d_{i2}, all of whom have no DLT, and t_i of whom have a DO. The second subscript on dose d_{i2}, that is 2, is an indicator for model (3.17). Thus

assuming the prior distribution of the form (2.6), the prior distribution for $(\alpha_{Y|N}, \beta_{Y|N})$,

$$\pi_{02}(\alpha_{Y|N}, \beta_{Y|N}) = \prod_{i=-1}^{0} \frac{p_{i2}^{t_i}(1-p_{i2})^{u_i}}{B(t_i, u_i)} \left| \log\left(\frac{d_{-12}}{d_{02}}\right) \right|,$$

and the joint posterior density for $\alpha_{Y|N}$ and $\beta_{Y|N}$ is

$$\pi(\alpha_{Y|N}, \beta_{Y|N}|\mathbf{x}) \propto \prod_{i=-1}^{k} p_{i2}^{t_i}(1-p_{i2})^{u_i}, \tag{3.20}$$

where

$$p_{i2} = \frac{\exp(\alpha_{Y|N} + \beta_{Y|N} \log\ d_{i2})}{1 + \exp(\alpha_{Y|N} + \beta_{Y|N} \log\ d_{i2})}, \quad i = -1, 0, 1, ..., k,$$

and $u_i = m_i$, $n_{i2} = u_i + t_i$ with $d_{i1} = d_{i2}$ for $i = 1, ..., k$.

Let us define the therapeutic window as the interval (d_L, d_U) for which $p_{NN}(d_L) = c_L$ and $p_{*Y}(d_U) = c_U$ (c_L and c_U small values such as 0.2). Whitehead et al. (2006) propose dose allocation so as to maximize the inverse of the sum of the variances of the boundaries of the therapeutic window defined as

$$G(\theta) = \{w_L Var(d_L|\mathbf{x}_a) + w_U Var(d_U|\mathbf{x}_a)\}^{-1},$$

where θ is a vector of parameters $\alpha_{*Y}, \beta_{*Y}, \alpha_{Y|N}$ and $\beta_{Y|N}$, $\mathbf{x}_a$ denotes the current data $\mathbf{x}$ augmented with the data that will be observed on the next cohort of subjects, and w_L and w_U are appropriate weights. The number of inspections with this method depend on the total sample size for the whole trial and the cohort sizes.

3.3.6 Phase II studies with several doses

In the late phase II studies, the investigators while doing tests to decide whether it is worth continuing to phase III studies, may still be uncertain as to which is the best potential dose of the new drug to test in the phase II clinical trial. To overcome this difficulty, several doses of the new drug may be compared to the standard treatment. This results to testing

multiple hypotheses comparing the standard treatment to many doses of the new drug. In frequentist testing, if it is desired to control the type I FWER associated with comparing several doses to a control treatment at some level α, then the pairwise p-value comparing each dose to the control needs to be adjusted. These methods are discussed extensively in Section 4.3. Bayesian methods are also available. For example, the ideas in the works by Stallard et al. (1999) and Loke et al. (2006) could be combined to develop a method that allows for several doses. The method by Loke et al. (2006) uses several doses and Stallard et al. (1999) define gain functions which incorporate whether it is worth continuing to the phase III stage.

3.4 Phase III clinical trials

Phase III trials are typically large confirmatory trials for efficacy. The main focus is placed on efficacy but safety is also monitored. The new drug is compared with a commonly used drug (the control or the standard drug) usually in a randomized trial. The trial subjects are allocated randomly to the new drug treatment arm and the standard drug treatment arm and the measure of efficacy, side effects and all information that will allow the new treatment to be used safely are examined. Evidence of efficacy is usually assessed by testing hypotheses usually using frequentist methods.

3.4.1 Sample size calculation in fixed sample trials

The statistical aspects involved in designing a phase III clinical trial include determining whether there will be interim analyses or not and calculating the sample size. Adopting the definition of Whitehead (1997), we refer to clinical trials where analysis is carried out after all patients have been entered in the trial and the outcomes observed as fixed sample clinical trials. In this subsection, we briefly describe the rationale for sample size formulae

for a fixed sample clinical trial. More detail of the rationale for sample size formulae are given by Friedman et al. (1998), Machin et al. (2009) and Cleophas et al. (2009) among others.

For simplicity, suppose a new drug is being tested for superiority. Before a new drug is accepted for use by the regulatory authorities, the investigators must demonstrate clearly that the new drug is better than the standard drug. For this reason, the probability of concluding that the new drug is better than the standard drug while the truth is that the new drug is not better than the standard drug, is often set to a maximum of 2.5% (0.025). This probability is referred to as the type I error and is usually denoted by α. On the other hand it is essential to have a clinical trial with sufficient statistical power to detect a difference between the new drug and the standard drug when it truly exists. The probability of concluding that the new drug is better than the standard drug when the new drug is truly better than the standard drug by some specified amount is called the power and is usually denoted by $(1-\beta)$, where β denotes the type II error. Type II error is the probability of failing to reject the null hypothesis that the new drug is not better than the standard drug when the truth is that the new drug is better than the standard drug by the specified amount. The danger of conducting a clinical trial with low power is that new treatments that are beneficial are discarded without adequate testing and may never be considered in future (Friedman et al., 1998). In addition to the monetary loss the drug company will incur, this leads to loss to society associated with the lack of effective therapies. In practice trials are normally designed to have power of between 0.8 and 0.95 so that the probability of a type II error is controlled at between 0.05 and 0.2.

To plan a trial with the desired statistical power and control type I error, sample size calculation is based on α and β. To determine the power of a test the effectiveness of the new and standard drug are required. In sample size calculation, the hypothetical effectiveness of the standard is determined and the effectiveness of the new drug is taken as the sum of the effectiveness of the standard drug and a difference of medical relevance.

For example for binary outcomes, suppose the probability of successful treatment with the standard drug is p_0 and with the new drug is $p_1 = p_0 + \delta$, where δ is a difference of medical relevance, then one of the sample size approximation for a one-sided test at level α is given by

$$2N = \frac{2\left\{Z_\alpha\sqrt{2\bar{p}(1-\bar{p})} + Z_\beta\sqrt{p_0(1-p_0)+p_1(1-p_1)}\right\}^2}{(p_0-p_1)^2},$$

where $\bar{p} = (p_0 + p_1)/2$ and N is the number of patients in each treatment arm, and Z_α and Z_β are standard normal values such that $\Phi(Z_\alpha) = 1 - \alpha$ and $\Phi(Z_\beta) = 1 - \beta$, where Φ as before is the standard normal distribution function. Alternative sample size formulae are given for example by Friedman et al. (1998) and Machin et al. (2009) but they point out that these formulae give similar results to the above sample size formula. Sample size formulae for other outcome variables such as survival outcomes, continuous outcomes are available in most clinical trials books such as the ones cited at the beginning of this subsection. Due to loss in follow-up visits, some investigators increase the calculated sample size by some factor.

3.4.2 Sequential investigations

Recruitment of patients in a clinical trial occurs gradually during the course of the trial which can extend to years depending on the prevalence of the targeted disease and the size of targeted population. This feature opens the possibility of stopping the trial earlier based on the emerging evidence (Armitage, 1975; Whitehead, 1997). In clinical trials based on efficacy, on ethical grounds, it may be desirable to stop the trial if there is a clear advantage of either drug (new or standard) over the other thereby avoiding the allocation of more patients to the less efficacious drug. The pharmaceutical company may also wish to save costs by stopping a trial early for a drug which appears to have little chance of demonstrating improved efficacy. In principle sequential investigation may be carried out after the outcome

of each patient has been observed but for practical reasons, sequential investigations are performed either at some pre-specified times or after a pre-specified number of patients (or pre-specified number of events such as number of deaths for survival outcomes) have been observed.

In sequential trials, some method for combining evidence from the interim analyses is required. In this thesis, we will describe two methods. In the first method, raw data or sufficient statistics are merged to make the final analysis. We will refer to this method of combining evidence as the group sequential technique. In the second method, data from each interim analysis are analysed separately and some combination function is used to combine the p-values. We will refer to this method of combining evidence using the p-values as the method of combination tests. In Chapter 4, we will introduce seamless phase II/III clinical trials which are trials that combine phase II and phase III into single trial. Analysing these trials require combining evidence from phase II stage and the phase III stage which can be done using the group sequential technique or by combination tests so that these methods of combining evidence will be described in detail in Chapter 4.

In the next chapter, while demonstrating how evidence from the phase II and phase III stages may be combined using the group sequential technique and by use of combination tests, we will assume that there will be no opportunity to stop the trial after the phase II stage for futility or for strong evidence against the null hypothesis that the new drug is not better than the control treatment. Phase III clinical trials with sequential investigation (testing) would allow for early stopping either for futility or overwhelming evidence that the new drug is better than the control treatment. Suppose in the entire trial we aim to control the type I error at level α. Armitage (1975), by use of examples based on binary and continuous data, has shown that if at each stage of the investigation hypothesis testing is carried out at level α, the overall type I error is inflated above α. Accordingly, methods for analysing data sequentially without inflating overall type I error have been developed. Pocock (1977), O'Brien and Fleming (1979), and Lan and DeMets (1983) have developed

methods based on group sequential testing that allow for early stopping while controlling overall type I error rate α and Brannath et al. (2002) have developed methods for adjusting for early stopping without inflating overall rate α based on combination tests.

Chapter 4

Seamless phase II/III clinical trials

In the last chapter, we have stated the objective of each phase of a clinical trial, reviewed some methods used to design trials in each phase, and discussed how conclusions are made from these trials. In the methods described, the conclusion from a trial did not include evidence from the previous trials. In this chapter, we introduce seamless phase II/III clinical trials, which are trials that combine phase II and phase III clinical trials into a single trial. These trials are attractive because data from both phase II and phase III are used in the final confirmatory analysis. The combination of phases II and III does, however, introduce complexity in analysis. The analysis poses a challenge of how to combine evidence from the phase II stage and phase III stage without inflating the type I error rate. Further, if multiple hypotheses are tested, the analysis poses a second challenge of how to adjust for multiple testing associated with testing several hypotheses. In the next section, we describe the testing process in phase II/III clinical trials and explicitly describe the challenges posed by these trials. In Section 4.2, we review some methods that can be used to combine evidence from the phase II stage and phase III stage. In Section 4.3, we review some methods that can be used to adjust for multiple testing. In Section 4.4, we describe the proposal by Bauer and Kieser (1999), Hommel (2001) and Bretz et al. (2006) to analyse

phase II/III clinical trials data that address both the challenge of combining evidence and testing multiple hypotheses without inflating the type I error rate. The notation given in Bretz et al. (2006) is used. In Section 4.5, we review some of the existing methods for treatment selection in phase II/III clinical trials.

4.1 The testing process and challenges in phase II/III clinical trials

In most of the designs that we reviewed in Chapter 3, it was assumed that the testing of the new drug takes place in the traditional way: each phase is carried out as a separate trial. Furthermore, it may appear as if in each phase, only one trial is required. However, for example in phase II, two or more trials may be carried out. If two trials are carried out in phase II, the first trial may be a proof of concept trial, where some dose-response modelling is done with the intention of identifying the minimum effective dose (MED). The second trial would then be a phase IIb trial, which may involve testing of hypotheses, where for example, several doses of a new drug (of higher efficacy level than MED) are compared to the standard treatment. Each clinical trial requires careful planning which means considerable time may be required to plan a trial. Thus the traditional procedure for testing a new drug, with many trials to be carried out, is very time consuming. Secondly, in the traditional procedure, data from the previous trials are not used in the analysis of the current trial data. This means to achieve adequate power, more patients are required, hence prolonging the recruitment time.

When a drug company starts testing a new product, the product is registered and the company is given a patent period during which no other company is allowed to test or produce that product. The patent period includes the development process time; hence there is a financial benefit to a drug company if the development time is reduced, increasing

the period the drug company will have a monopoly to produce and market the new drug. In addition to the financial benefit to the drug company, accelerated drug development avoids delay in potential benefits to the society. Hence any procedure or technique that will reduce the duration of drug development while maintaining the trial integrity is welcome.

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 II/III clinical trials. Such trials are conducted in two stages. In stage 1 (phase II stage) of phase II/III clinical trials, usually several hypotheses are of interest. For example, a new drug may be tested in different sub-populations such as different age-groups or groups based on a set of biomarkers which could affect sensitivity to the new drug, with the aim of identifying the sub-populations that respond favorably to the new drug. Another example is that in stage 1, a control treatment is compared to different experimental treatments, which could be different doses of a new drug, with the aim of identifying promising new treatments. In the case of sub-population selection, sub-populations that show promising results continue to stage 2 (phase III stage). Similarly, in the case of treatment selection, sufficiently promising treatments continue to stage 2 along with the control treatment. After stage 2 results, at the end of the phase II/III clinical trial, data from both stages are used to test the hypotheses of interest. In both the examples that we have given above, two issues arise while analysing data generated from such a phase II/III clinical trials, namely: (i) how to combine the evidence from the two stages without inflating the type I error rate, and (ii) how to control the type I familywise error rate (FWER) associated with testing multiple hypotheses.

The work in this thesis is based on the second example above, where in the phase II stage, the objective is to identify promising treatments that will continue for testing in the phase III stage. Specifically, in the phase II stage, we will assume that several dose levels of a new drug are compared to the control treatment, and a subset of the dose levels

tested in phase II stage continue with the control treatment to the phase III stage. Planning such a phase II/III clinical trial presents two challenges: (i) how to perform analysis that controls the error rates, and (ii) how to choose which doses should continue to the phase III stage after the phase II stage. Methods that control error rates in the analysis of phase II/III clinical trials with flexible choice of doses exist but there is very little work to guide the choice of doses. An example of an analysis that allows flexible choice of doses to test in stage 2 is given in Section 4.4. The objective of this thesis is to provide a solution for the second challenge by developing a new dose selection procedure. This procedure is described in the next chapter. In order to point out the differences between this procedure and the existing methods that can be used to make a choice of the doses to test in stage 2, in Section 4.5 we review some methods available in literature that may be used to select the doses that proceed to stage 2.

4.2 Combining evidence from two stages

In Section 3.4.2, we explained that a phase III trial could include one or more interim analyses. We mentioned two techniques of including evidence from interim analyses in the final analysis: the group sequential method, and the use of combination tests. The same techniques would apply in a phase II/III clinical trial, where the phase II stage could be viewed as being equivalent to an interim analysis. In this Section, we describe how the two methods may be used to test data from a phase II/III clinical trial when we assume there is no stopping after stage 1 (phase II stage).

4.2.1 Combining evidence using group sequential technique

Using the group sequential techniques, data from stage 1 and stage 2 are merged and an analysis is carried on the merged data set. Alternatively, as has been the case while design-

ing group sequential clinical trials, some sufficient statistics could be used to combine the evidence from the two stages. We demonstrate combining evidence in a two-stage group sequential using the efficient score statistics described by Whitehead (1997). Suppose that the new drug and the control are compared using some parameter θ, which is a measure of the treatment difference between the new and the control drug. For example if the outcome of interest is continuous and normally distributed, the parameter θ could be given by the standardized mean difference

$$\theta = \frac{\mu_E - \mu_C}{\sigma},$$

where μ_E and μ_C are treatment means for the new drug and the control drug respectively and σ is the population standard deviation for patients treated using the new drug and the control drug. The inference on θ using the data from stage s alone $(s = 1, 2)$, is based on a statistic Z_s. The statistic Z_s is the efficient score and is asymptotically normally distributed with mean θV_s and variance V_s, where V_s is the Fisher's information about θ contained in Z_s, that is, Z_s will be taken to be normally distributed $N(\theta V_s, V_s)$. As Z_1 and Z_2 are calculated from data from separate stages, they are independent. This notation is used to facilitate comparison with the method described in Section 4.2.2. Notation used by Whitehead (1997) is different with Z_2 being the efficient score based on both stage 1 and stage 2 data and V_2 the accumulated Fisher's information so that $Z_2 - Z_1 \sim N(\theta(V_2 - V_1), V_2 - V_1)$ and is independent of Z_1.

To describe how the statistics Z_s and V_s are derived, we for the moment ignore the subscript denoting the stage for which the statistics are based so that we describe using notation Z and V. The statistics Z and V can be derived from appropriate likelihood functions. If there are unknown nuisance parameters, the profile likelihood is used and this guarantees

$$l\{\theta, \hat{\phi}(\theta)\} = \text{const} + \theta Z - \frac{1}{2}\theta^2 V + O(\theta^3),$$

where $\hat{\phi}(\theta)$ is the maximum likelihood estimate of the nuisance parameter vector $\hat{\phi}$ given the value of θ. In the absence of nuisance parameters,

$$Z = l_\theta(0)$$

and

$$V = -l_{\theta\theta}(0),$$

where $l_\theta(0)$ and $l_{\theta\theta}(0)$ denote respectively the first and second derivatives of $l(\theta)$ evaluated with respect to θ, evaluated at $\theta = 0$. To illustrate with inference for mean of normal data with known variance, let a sample $x_1, ..., x_n$ be observations from a normal population distributed as $N(\mu, 1)$ so that the likelihood of these data is given by

$$\begin{aligned} L(\mu) &= \prod_{i=1}^{n} (2\pi)^{-1/2} \text{exp}\left(\frac{-(x_i - \mu)^2}{2}\right) \\ &= (2\pi)^{-n/2} \text{exp}\left(\frac{-(\sum_i^n x_i^2 - 2\mu \sum_i^n x_i + n\mu^2)}{2}\right). \end{aligned}$$

The corresponding log-likelihood has the form

$$l(\mu) = \text{const} + \mu S_n - \frac{1}{2}\mu^2 n,$$

where $S_n = \sum_{i=1}^n x_i$ so that $Z = S_n$ and $V = n$ and inference on μ is made from the statistic S_n which is such that $S_n \sim N(\mu n, n)$. Whitehead (1997) also gives forms for Z and V for comparative studies.

Reverting to Z_s and V_s to denote the statistics at stage s ($s = 1, 2$), the statistic

$$Z = (Z_1 + Z_2) \sim N(\theta(V_1 + V_2), (V_1 + V_2))$$

is the efficient score statistic based on data from both stage 1 and stage 2. Suppose that after collecting data in stage 1 and stage 2, the realizations for the efficient scores Z_1 and

Z_2 are z_1 and z_2 respectively. Then, using the group sequential approach, the p-value from the two stages is given by

$$\begin{aligned}
\text{Prob}(Z \geq z_1 + z_2|\theta_0) &= 1 - \text{Prob}(Z \leq z_1 + z_2|\theta_0) \\
&= 1 - \text{Prob}\left(\frac{Z - \theta_0(V_1+V_2)}{\sqrt{V_1+V_2}} \leq \frac{z_1 + z_2 - \theta_0(V_1+V_2)}{\sqrt{V_1+V_2}}\right) \\
&= 1 - \text{Prob}\left(Z^* \leq \frac{z_1 - \theta_0 V_1}{\sqrt{V_1+V_2}} + \frac{z_2 - \theta_0 V_2}{\sqrt{V_1+V_2}}\right) \\
&= 1 - \Phi\left\{\frac{z_1 - \theta_0 V_1}{\sqrt{V_1+V_2}} + \frac{z_2 - \theta_0 V_2}{\sqrt{V_1+V_2}}\right\}, \qquad (4.1)
\end{aligned}$$

where $Z^* \sim N(0,1)$ and θ_0 is the value of the parameter θ under the null hypothesis.

4.2.2 Combining evidence using combination tests

Bretz et al. (2006) use the combination test as described by Bauer and Köhne (1994). Using the combination test, data from each stage are analysed separately. In order to make a single conclusion from the two stages, p-values obtained at the end of each stage are combined using some function C into a single p-value. Bauer and Köhne (1994) implement combination tests in adaptive clinical trials but the technique of combining evidence using combinations tests had been proposed by Fisher (1932) to address the need to combine results from a number of independent tests used to test a common hypothesis. Suppose a null hypothesis H (notice here we do not use the conventional notation H_0) is tested at stage 1 and stage 2 obtaining the p-value p_s at stage s ($s = 1, 2$). Further, let the combined p-value be denoted by $C(p_1, p_2)$. Zaykin et al. (2002) have reviewed some methods of combining the p-values. Two of the commonly used methods are the Fisher's combination method and the weighted inverse normal method.

For uniformly distributed p_1 and p_2, the functions $-2\ \log\ p_s\ (s = 1, 2)$, where $\log$ is to base e, have a chi-square distribution with two degrees of freedom. Using the fact that the sum of random variables that are χ^2-distributed has a χ^2-distribution with degrees

of freedom equal to the sum of the degrees of freedom of the summed random variables, Fisher (1970) noted that

$$T = -2\sum_{s=1}^{2} \log\ p_s = -2\log\prod_{s=1}^{2} p_s$$

has a χ^2-distribution with 4 degrees of freedom when the null hypothesis H is true and the p-values p_1 and p_2 are independent. Therefore, the p-value for testing the null hypothesis H using the evidence from the 2 stages is the probability of a χ^2_4 variable being greater or equal to the observed value T^* of T so that, using the Fisher's combination method, the combined p-value

$$C(p_1, p_2) = 1 - F_{\chi^2_4}(-2\log\prod_{s=1}^{2} p_s), \tag{4.2}$$

where $F_{\chi^2_4}$ is the distribution function of a chi-square distribution with 4 degrees of freedom.

The inverse normal procedure uses the normal-transformed p-values. Let X be a normally distributed random variable with mean 0 and variance 1, that is, $X \sim N(0,1)$. Further, let the distribution function of X be denoted by $\Phi(x)$ and suppose that Pr$(X \leq x) = \Phi(x) = c$. Because $X \leq x$ is equivalent to $\Phi(X) \leq \Phi(x)$, then

$$\begin{aligned} \text{Prob}(\Phi(X) \leq c) &= \text{Prob}\left(\Phi(X) \leq \Phi(x)\right) \\ &= \text{Prob}(X \leq x) = \Phi(x) = c. \end{aligned}$$

Hence Prob$(\Phi(X) \leq c) = c$, which implies the distribution function of a standard normal random variable is Uniform[0,1] so that the p-value p_s for hypothesis H at stage s ($s = 1, 2$) can be transformed into standard normal score when the hypothesis H is true by taking

$$z_s = \Phi^{-1}(1 - p_s), \quad s = 1, 2.$$

Let $X_j,\ j = 1, 2, ..., n$ be distributed $N(0,1)$ and $\alpha_1, ..., \alpha_n$ be constants such that $\sum_j \alpha_j^2 = 1$, then the linear combination $Y = \sum_j \alpha_j X_j$ is distributed $N(0,1)$. Using this standard

result, then

$$z = \sum_{s=1}^{2} \frac{z_s}{\sqrt{2}}$$

is standard normal and the combined p-value may be given by

$$C(p_1, p_2) = 1 - \Phi\left(\frac{1}{\sqrt{2}} \sum_{s=1}^{2} \Phi^{-1}(1 - p_s)\right).$$

Other weights w_1 and w_2 which satisfy $w_1^2 + w_2^2 = 1$ can be used in place of the 2 equal weights $1/\sqrt{2}$ so that using the weighted inverse normal method for combining evidence in a phase II/III clinical trial, the combined p-value is given by

$$C(p_1, p_2) = 1 - \Phi[w_1 \Phi^{-1}(1 - p_1) + w_2 \Phi^{-1}(1 - p_2)], \tag{4.3}$$

where $0 < w_s < 1,\ s = 1, 2$, are arbitrary weights subject to $w_1^2 + w_2^2 = 1$. Suppose the efficient scores given in Section 4.2.1 are used to obtain the p-values at each stage, then

$$\begin{aligned} 1 - p_1 &= 1 - \text{Prob}(Z_1 \geq z_1 | \theta_0) = \text{Prob}(Z_1 \leq z_1 | \theta_0) \\ &= \text{Prob}\left(\frac{Z_1 - \theta_0 V_1}{\sqrt{V_1}} \leq \frac{z_1 - \theta_0 V_1}{\sqrt{V_1}}\right) \\ &= \Phi\left(\frac{z_1 - \theta_0 V_1}{\sqrt{V_1}}\right). \end{aligned} \tag{4.4}$$

Equivalently,

$$1 - p_2 = \Phi\left(\frac{z_2 - \theta_0 V_2}{\sqrt{V_2}}\right). \tag{4.5}$$

Substituting the expressions (4.4) and (4.5) in the expression (4.3) for inverse normal method combined p-value and letting $w_s = \sqrt{V_s/(V_1 + V_2)}\ \ (s = 1, 2)$, then

$$\begin{aligned} C(p_1, p_2) &= 1 - \Phi\left\{\sum_{s=1}^{2} \sqrt{\frac{V_s}{V_1 + V_2}} \Phi^{-1}\left\{\Phi\left(\frac{z_s - \theta_0 V_s}{\sqrt{V_s}}\right)\right\}\right\} \\ &= 1 - \Phi\left\{\sum_{s=1}^{2} \left(\frac{z_s - \theta_0 V_s}{\sqrt{V_1 + V_2}}\right)\right\} \\ &= 1 - \Phi\left\{\frac{z_1 - \theta_0 V_1}{\sqrt{V_1 + V_2}} + \frac{z_2 - \theta_0 V_2}{\sqrt{V_1 + V_2}}\right\}. \end{aligned} \tag{4.6}$$

The combined p-value obtained from expression (4.6) and the p-value obtained using the group sequential in expression (4.1) are equal. Hence, if the weights $w_s \quad (s = 1, 2)$ for the inverse combination method are appropriately chosen, this combination function corresponds to the two-stage group sequential test. Choosing the weights proportional to the sample size, that is taking $w_s = \sqrt{n_s/(n_1 + n_2)}$ ($s = 1, 2$), where n_1 and n_2 are the stage 1 and stage 2 sample sizes, achieves this since V_1 and V_2 are approximately proportional to the respective sample sizes.

Next we give the expressions for the type I error and the critical p-value for testing hypothesis H such that the type I error is not inflated. Suppose there is opportunity to stop the trial early after stage 1 for overwhelming evidence against the null hypothesis, that is when $p_1 \leq \alpha_1$ ($\alpha_1 \leq \alpha$) or for futility, that is when $p_1 > \alpha_0$ ($\alpha_0 > \alpha$). Then, the type I error is the probability that, under the null hypothesis H, either $p_1 \leq \alpha_1$ or $\alpha_1 < p_1 \leq \alpha_0$ and the combined p-value $C(p_1, p_2) \leq c$, that is

$$\text{Prob}_H[p_1 \leq \alpha_1] + \text{Prob}_H[C(p_1, p_2) \leq c,\ \alpha_1 < p_1 \leq \alpha_0],$$

where c is the combined critical p-value and is obtained by equating the above equation to overall type I error α and solving for c. Assuming that the p-values p_1 and p_2 have independent Uniform$[0, 1]$ distributions under the null hypothesis, then the overall type I error is given by

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{[C(p_1,p_2)\leq c]} dp_2 dp_1, \tag{4.7}$$

where $\mathbf{1}_{[C(p_1,p_2)\leq c]}$ equals 1 if $C(p_1, p_2) \leq c$ and 0 otherwise. For the first part of our work, we assume we do not stop for overwhelming evidence and stopping for unpromising results in stage 1 does not depend on the observed stage 1 p-value p_1. Note in this case we do not make any type I error at stage 1 so that equation (4.7), if the trial proceeds to stage 2, simplifies to

$$\int_0^1 \int_0^1 \mathbf{1}_{[C(p_1,p_2)\leq c]} dp_2 dp_1. \tag{4.8}$$

In expression (4.8), if the overall type I error is controlled at level α, we have $c = \alpha$, so that the combined p-values defined by equations (4.2) and (4.3) control error rate at level α.

The p-values defined by (4.2) and (4.3) clearly control the type I error rate because the p-values p_1 and p_2 are assumed independent under H. Brannath et al. (2002) explain this is a strong requirement. The only requirement needed in order for the p-values defined by equations (4.2) and (4.3) to control the type I error rate is that the distribution of the p-values p_1 and p_2 under H to satisfy

$$\Pr_H(p_1 \leq \alpha) \leq \alpha \text{ and } \Pr_H(p_2 \leq \alpha | p_1) \leq \alpha \text{ for all } 0 \leq \alpha \leq 1. \tag{4.9}$$

Brannath et al. (2002) refer to this property of the distribution of p-values p_1 and p_2 as "p clud".

4.3 Controlling familywise error rate in multiple hypotheses testing

Suppose in an experiment $k(> 1)$ experimental treatments are to be compared with a control treatment such that k null hypotheses $H_j : \theta_j = \theta_0$, $j = 1, ..., k$ comparing each experimental dose with the control treatment are of interest, where θ_j and θ_0 respectively denote the measure of effectiveness for experimental treatment j and the control treatment. Without loss of generality, suppose the first k_1 ($k_1 \leq k$) null hypotheses are true. Let E_j be the event that the null hypothesis H_j ($j = 1, ..., k$) is rejected, then if no multiple testing adjustment is made, the overall (type I) FWER associated with testing the k null hypotheses is

$$1 - \text{Prob}(\cap_{j=1}^{k_1} E_j^c \mid H_{0k_1}),$$

where the notation "$\mid H_{0k_1}$" means that given the null hypotheses $H_j,\ j = 1, ..., k_1$ are true. If the events $E_j's$, $j = 1, ..., k_1$ are independent the above expression reduces to

$$1 - \prod_{j}^{k_1}[1 - \text{Prob}(E_j \mid H_{0k_1})].$$

For example, if $k_1 = 2$ and each hypothesis is tested at level $\alpha = 0.05$ and the hypothesis H_1 and H_2 are independent, then the unadjusted type I error is

$$1 - (1 - 0.95)^2 = 0.0975$$

so that the FWER is almost double the individual type I errors associated with testing H_1 and H_2. Indeed, one is almost certain to make a type I error when the number of true null hypotheses to be tested becomes large (Hochberg and Tamhane, 1987). Thus, for a credible analysis, methods are required to control the FWER associated with testing the k pairwise null hypotheses at the pre-specified level α.

There are several testing procedures that can be used to test the multiple hypotheses so that the FWER is controlled at the desired level α. Hochberg and Tamhane (1987) explain that the FWER may be strongly or weakly controlled. The FWER is strongly controlled if

$$\left[1 - \text{Prob}(\cap_{j=1}^{k_1} E_j^c \mid H_{0k_1})\right] \leq \alpha, \text{ for all } k_1 \leq k,$$

and the FWER is controlled weakly if

$$\left[1 - \text{Prob}(\cap_{j=1}^{k_1} E_j^c \mid H_{0k_1})\right] \leq \alpha, \text{ only when } k_1 = k,$$

that is, when all the tested null hypotheses are true. An example of a test that controls FWER weakly is due to Fisher (1935). In this test, individual pairwise hypotheses are tested only when the global null hypothesis of no difference among all the $k + 1$ treatments (that is the k experimental and the control treatment are all equal) is tested and rejected.

This test controls the FWER strongly only if $k = 2$. For $k \geq 3$, the test controls the FWER weakly. All the other procedures we describe later in this section strongly control the FWER.

Westfall and Young (1993) have reviewed some methods that control the FWER. The simplest is the Bonferroni method in which each of the k null hypothesis is rejected when the observed p-value is less or equal to α/k, which leads to the Bonferroni adjusted p-value $\tilde{p}_j = \min(kp_j, 1)$, where p_j is the unadjusted p-value obtained from testing the null hypothesis H_j, $j = 1, ..., k$. The FWER is protected since

$$\begin{aligned} \text{Prob(Reject at least one } H_j \mid H_0) &= \text{Prob}\left(\min_{1\leq j\leq k} p_j \leq \alpha/k \mid H_0\right) \\ &\leq \sum_{j=1}^{k} \text{Prob}(p_j \leq \alpha/k \mid H_0) = \alpha. \end{aligned}$$

A similar adjustment is by use of the Šidák method, which rejects each of the k null hypothesis when the observed p-value is less than $1 - (1 - \alpha)^{1/k}$. This leads to the Šidák adjusted p-value $\tilde{p}_j = 1 - (1 - p_j)^k$. This method is less conservative than the Bonferroni correction and is exact for protecting FWER if all p-values are independent since

$$\begin{aligned} \text{Prob(Reject at least one } & H_j \mid H_0) \\ &= \text{Prob}\left(\min_{1\leq j\leq k}\{1 - (1 - p_j)^k\} \leq \alpha \mid H_0\right) \\ &= 1 - \text{Prob}\left(\min_{1\leq j\leq k}\{1 - (1 - p_j)^k\} > \alpha \mid H_0\right) \\ &= 1 - \text{Prob}\left(p_j > \{1 - (1 - \alpha)^{1/k}\} \text{ for all } j \mid H_0\right) \\ &= 1 - \prod_{j=1}^{k} \text{Prob}\left(p_j > \{1 - (1 - \alpha)^{1/k}\} \mid H_0\right) \qquad (4.10) \\ &= 1 - \{(1 - \alpha)^{1/k}\}^k = \alpha. \end{aligned}$$

The equality in step 4.10 holds assuming independence and the final results holds assuming $p_j,\ j = 1, ..., k$ are Uniform[0,1].

The Bonferroni and the Šidák methods described above are single step procedures. Holm (1979) introduced a sequentially rejective algorithm to test multiple hypotheses. The algorithm is based on the ordered p-values, $p_{(1)} \leq ... \leq p_{(k)}$, corresponding to hypotheses $H_{(1)}, ..., H_{(k)}$. The reasoning is that once $H_{(1)}$ has been rejected using for example the Bonferroni critical value α/k, we should believe that $H_{(1)}$ is false. Thus, there are only $k-1$ hypotheses which might still be true, implying the critical value $\alpha/(k-1)$ should be used for $H_{(2)}$ and so on. Holm's Sequentially rejective algorithm is given below.

- Step 1: If $p_{(1)} > \alpha/k$, then accept all hypotheses $H_{(1)}, ..., H_{(k)}$ and stop; otherwise, reject $H_{(1)}$ and continue.

  .

- Step j: If $p_{(j)} > \alpha/(k-j+1)$, then accept all hypotheses $H_{(j)}, ..., H_{(k)}$ and stop; otherwise, reject $H_{(j)}$ and continue.

  .

- Step k: If $p_{(k)} > \alpha$, then accept hypothesis $H_{(k)}$; otherwise, reject $H_{(k)}$.

The adjusted p-values of this algorithm are $\tilde{p}_{(1)} = \max\{kp_{(1)}, 1\}$, $\tilde{p}_{(2)} = \max\{(k-1)p_{(2)}, 1\}$, ..., $\tilde{p}_{(k)} = p_{(k)}$.

Westfall and Young (1993) introduce the Bootstrap adjustments which have the advantage of capturing the correlation structure. There are also other adjustments methods in literature. In order to control the FWER associated with testing the k pairwise null hypotheses at pre-specified level α, Bretz et al. (2006) use the closure principle (CP) of Marcus et al. (1976). The CP considers the set of all intersection hypotheses $H_J = \cap_{j \in J} H_j,\ J \subseteq \{1, ..., k\}$ constructed from the initial hypotheses of interest. Marcus et al. (1976) refer to this set, denoted by $\mathcal{H}$, as the closure set. Using the CP, a null hypothesis $H_j,\ j = 1, ..., k$ is rejected at FWER α if all hypotheses $H_J,\ J \subseteq \{1, ..., k\}$ with $j \in J$ are rejected at level α. Consider Figure 4.1 when $k = 3$. The closure set $\mathcal{H}$ equals

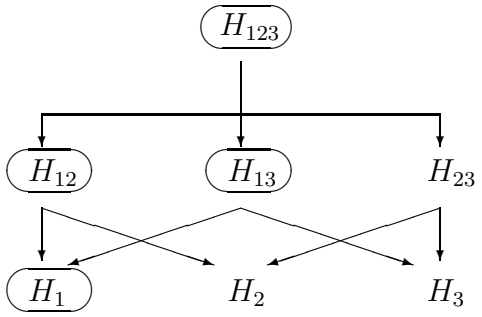

Figure 4.1: Closure set with 3 treatments. The hypotheses contained in H_1 are circled.

$\{H_1, H_2, H_3, H_{12}, H_{13}, H_{23}, H_{123}\}$. Hypothesis H_1 is rejected if the circled hypotheses $H_{123},\ H_{12},\ H_{13},$ and H_1 are all rejected each at level α.

Marcus et al. (1976) have explained how the type I error for this procedure is at most α. Let X be a random variable with distribution depending on a parameter $\theta \in \Omega$ such that $\mathcal{H}$, the set of null hypotheses defined above, is a set of subsets of Ω. For each $H_J \in \mathcal{H}$, let $\phi_J(X)$ be a level α test, that is, $\text{Prob}_\theta\{\phi_J(X) = 1\} \leq \alpha$ for all $\theta \in H_J$ where $\phi_J(X)$ is an indicator variable for rejecting H_J. As detailed in the steps for the closure principle above, any null hypothesis H_J is rejected by means of $\phi_J(X)$ if and only if all hypotheses H that are included in H_J ($H \subset H_J$) and belonging to $\mathcal{H}$ ($H \in \mathcal{H}$) have been tested and rejected. A type I error is committed if and only if the intersection of all true hypotheses, H_τ say, is tested and rejected by means of $\phi_\tau(X)$; in other words, if we denote by A the event that any true H_J is rejected, and by B the event that $\phi_\tau(X) = 1$, then

$$\text{Prob}(A \cap B) = \text{Prob}(B)\text{Prob}(A|B) \leq \alpha,$$

since ϕ_τ is a level α test hence $\text{Prob}(B) \leq \alpha$ and $\text{Prob}(A|B) \leq 1$. However, since $A \cap B = A$, $\text{Prob}(A \cap B) = \text{Prob}(A)$ and hence $\text{Prob}(A) \leq \alpha$. The probability of making no type I error with this procedure is thus at least $1 - \alpha$.

4.4 Analysing data from a phase II/III clinical trial

Bauer and Kieser (1999), Hommel (2001) and Bretz et al. (2006) propose using the combination tests and the CP for the analysis of phase II/III clinical trials data. In their proposal, a null hypothesis H_j ($j = 1, ..., k$) is rejected at the end of stage 2 if all the combined p-values for all the hypotheses H_J, $J \subseteq \{1, ..., k\}$ with $j \in J$ are less than the pre-specified level of testing. For example, suppose there are three experimental treatments at stage 1 and let $p_{s,J}$ denote the p-value for testing hypothesis $H_J,\ J \subseteq \{1, 2, 3\}$ at stage s $(s = 1, 2)$. Then hypothesis H_1 is rejected at the end of stage 2 at level α if

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

Figure 4.2 gives the flow chart of this example. Panel (a) gives the stage 1 p-values corresponding to the hypotheses given in Figure 4.1. The p-values for hypotheses contained in H_1 are circled. On the other hand, panel (b) gives stage 2 p-values corresponding to the hypotheses given in Figure 4.1 and once again, p-values for hypotheses contained in H_1 are circled. Panel (c) gives the combined p-values. The combined p-values for hypotheses contained in H_1 are circled and they must all be rejected for hypothesis H_1 to be rejected after stage 2.

To illustrate what happens if some treatments are dropped after stage 1, suppose for example that treatment 3 is dropped after stage 1, so that no data are available for treatment 3 at stage 2. The stage 2 p-values for this scenario are given in Figure 4.3. The tests for intersection hypotheses H_{13}, H_{23} and H_{123} respectively reduce to the tests for hypotheses H_1, H_2 and H_{12} so that $p_{2,13} = p_{2,1}$, $p_{2,23} = p_{2,2}$ and $p_{2,123} = p_{2,12}$. If treatment 3 is tested for efficacy after stage 2, $p_{2,3}$ may be fixed to 1. This follows the proposal by Posch et al. (2005) where stage 2 p-values for hypotheses that do not have stage 2 data are fixed to 1, which lead to conservative final tests for the hypotheses in the closure set.

To show tests for H_{13}, H_{23} and H_{123} respectively using $p_{2,1}$, $p_{2,2}$ and $p_{2,12}$ at stage

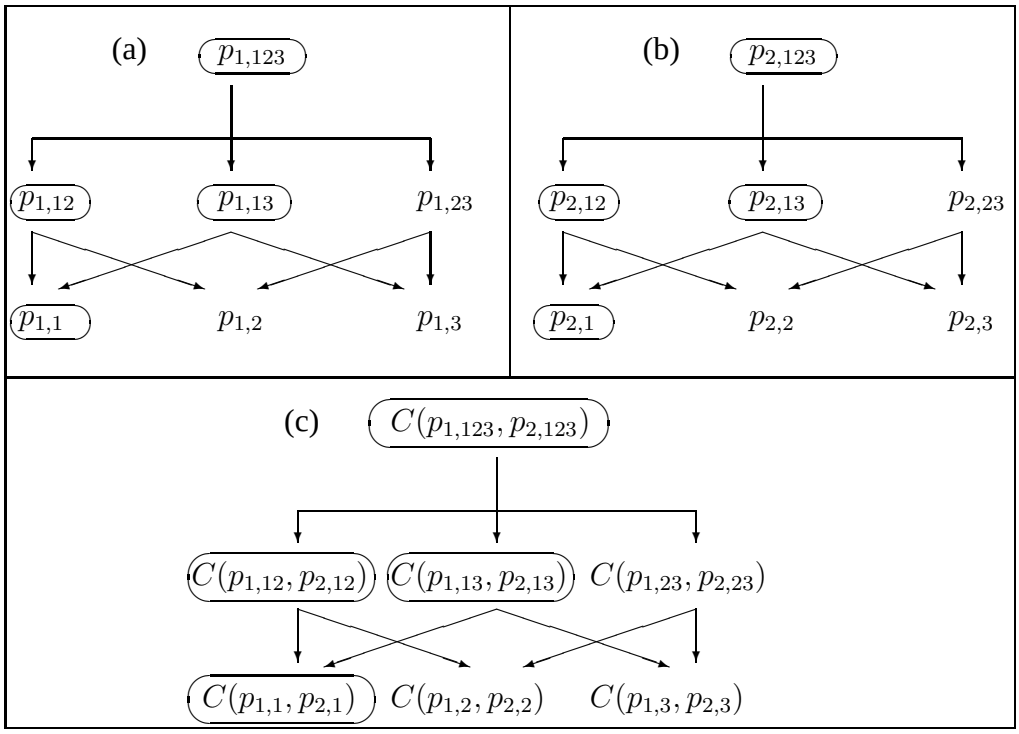

Figure 4.2: P-values required to test 3 elementary hypotheses. Panels (a) and (b) respectively give stage 1 and stage 2 p-values corresponding to hypotheses given in Figure 4.1. Panel (c) gives the combined p-values for these hypotheses.

2 are level α tests when dose 3 is not tested in stage 2, we use a general case. Suppose we wish to test a hypothesis H_J using the p-value $p_{2,J'}$ for hypothesis $H_{J'}$ with $H_J \subseteq H_{J'}$ (that is $J' \subseteq J$). Since $H_J \subseteq H_{J'}$, under H_J, $H_{J'}$ is also true so that $p_{2,J'} \sim U[0,1]$. Hence testing H_J using $p_{2,J'}$ provides a level α test.

For the test described above to control the type I error rates strongly, the p-values $p_{1,J}$ and $p_{2,J}$ should satisfy the "p clud" condition given by expression (4.9). In this thesis, we will consider p-values $p_{1,J}$ and $p_{2,J}$ obtained using separate data, that is, p-value $p_{s,J}$ ($s = 1, 2$) will be obtained using stage s data only. For now, we also assume appropriate level α tests are used so that for all hypotheses H_J, $\text{Prob}_{H_J}(p_{s,J} \leq \alpha) \leq \alpha$ ($s = 1, 2$). Hence, when no treatments are dropped, the p-values $p_{1,J}$ and $p_{2,J}$ are independent so that under H_J, the distribution of $p_{1,J}$ and $p_{2,J}$ satisfy the "p clud" condition. If some fixed

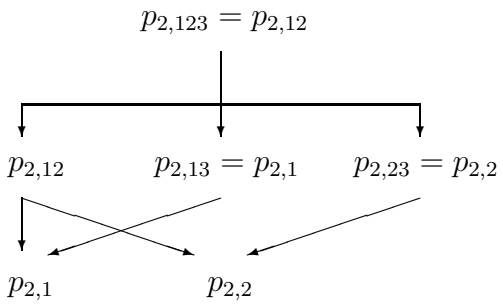

Figure 4.3: Stage 2 p-values when treatment 3 is dropped

treatments are dropped after stage 1, as demonstrated above H_J is tested at stage 2 using the p-value $p_{2,J'}$ for hypothesis $H_{J'}$ with $H_J \subseteq H_{J'}$ (that is $J' \subseteq J$). We need to show that Prob$_{H_J}(p_{2,J'} \leq \alpha | p_{1,J}) \leq \alpha$. From above a test for H_J using $p_{2,J'}$ is an α test, and since $p_{1,J}$ and $p_{2,J'}$ are independent for a fixed $H_{J'}$, then Prob$_{H_J}(p_{2,J'} \leq \alpha | p_{1,J}) \leq \alpha$ so that the p-values $p_{1,J}$ and $p_{2,J'}$ satisfy the "p clud" condition. If the dropped treatments are tested for efficacy after stage 2, for some hypotheses H_J, $J' = \emptyset$. If following Posch et al. (2005), the p-values $p_{2,J'}$ for these hypotheses are set to 1, that is $p_{2,J'} = 1$ for all $J' = \emptyset$, the p-values for hypotheses with $J' = \emptyset$ satisfy the "p clud" condition since Prob$_{H_J}(p_{2,J'} \leq \alpha \mid p_{1,J}) = 0 \leq \alpha$, so that the type I error rate is maintained.

In the next chapter, we develop a new procedure to select the treatments to test in stage 2 and the selection procedure depends on stage 1 p-values $p_{1,J}$ ($J \subseteq \{1, ..., k\}$). We will set a rule that the intersection hypothesis H_J will be tested using a test for the smallest intersection hypothesis $H_{J'}$ ($H_J \subseteq H_{J'}$) that can be constructed from all the experimental treatments that are selected for testing in stage 2. For example, if only experimental treatment j ($j = 1, 2, 3$) is tested in stage 2, hypothesis H_{123} is tested using $p_{2,j}$ and if we test treatments i and j ($i, j \in \{1, 2, 3\}$) at stage 2, H_{123} will be tested using $p_{2,ij}$ so that $H_{J'}$ used in test for H_J at stage 2 is random. We need to show that Prob$_{H_J}(p_{2,J'} \leq \alpha | p_{1,J}) \leq \alpha$, where J' is random. Under H_J, all hypotheses $H_{J'}$ with $H_J \subseteq H_{J'}$ (that is $J' \subseteq J$) are also true. Hence a test for H_J using $p_{2,J'}$ using hypothesis $H_{J'}$ defined using the above rule

is a level α test so that $\text{Prob}_{H_J}(p_{2,J'} \leq \alpha | p_{1,J}) \leq \alpha$. Hence $p_{1,J}$ and $p_{2,J'}$ are "p clud".

In the above discussion, while showing the p-values satisfy the "p clud" condition, we have assumed appropriate level α tests are used for all the hypotheses H_J ($J \subseteq \{1, ..., k\}$). We will illustrate the dose selection procedure developed in the next chapter using chi-squared tests for the pairwise hypotheses H_j ($j = 1, ..., k$). Asymptotically, the chi-square test provides a test with the type I error rate close to the desired nominal level α. We will assume that sufficiently large samples will be available at stage 1 for each treatment arm so that the type I error rate will be close to the nominal value for hypotheses H_j ($j = 1, ..., k$). At stage 2, generally large samples are available so that chi-square tests for hypotheses H_j ($j = 1, ..., k$) control type I error rates close to nominal value α. With treatment selection, at stage 2, the number of experimental treatments to be tested will vary so that the sample available to test hypotheses H_j ($j = 1, ..., k$) will vary with more data available if few treatments are tested in stage 2. However, since large samples are available at stage 2, the chi-square test will still be adequate to control the type I error rates when number of treatments to test in stage 2 vary. In the next paragraph, we describe how the p-values for the hypotheses H_J with $|J| \geq 2$ may be obtained. The tests described are conservative. Hence, all the p-values used to analyse the seamless phase II/III clinical trial incorporating the treatment selection will be asymptotically "p clud" (Zuber et al., 2006).

Bauer and Kieser (1999), Hommel (2001) and Bretz et al. (2006) do not give details of how the p-values testing the hypotheses in $\mathcal{H}$ should be calculated but Westfall and Wolfinger (2000) 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 or the t-test for continuous data. There are several tests for the intersection hypotheses $H_J,\ J \subseteq \{1, ..., k\}$ with $|J| \geq 2$ but some are specific to certain forms. For example, Hotelling's T^2 test described by Johnson and Wichern (2002) is valid for continuous data. Flexible tests that can be used for many forms of responses (normal, Poisson, etc) are Bonferroni, Šidak and Simes tests. Suppose we wish to test a hypothesis of equality of the

control treatment with m $(1 < m \leq k)$ experimental treatments. The Bonferroni adjusted p-value is given by $\min\{1, (m \times \text{minp})\}$, while Šidak 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 adjusted p-value is given by $\min\{\frac{m}{i} p_{(i)}\},\ i = 1, ..., m$ where $p_{(i)}$ denote the ordered p-values.

When several treatments are compared to a control treatment using the same control group, it would be desired rather than assume the pairwise tests are independent, to utilize correlation in the comparisons because of the pairwise comparisons versus the same control group. Dunnett (1955) proposed a multiple comparison procedure that makes use of the correlation associated with comparing several treatments to the same control group for continuous normal data. This test can be used for all the intersection hypotheses in the closure set. Let Z_j ($j = 0, 1, ..., k$) be the standardized response from treatment j with $j = 0$ corresponding to the control treatment. For hypothesis H_J ($J \subseteq \{1, ..., k\}$), let $Z_J^{\max} = \max_{j \in J} Z_j$ and define

$$F_{Z_J^{\max}}(z) = \int_{-\infty}^{\infty} [\Phi(\sqrt{2}z + z_0]^{|J|} \phi(z_0) dz_0,$$

where as before $\phi(.)$ and $\Phi(.)$ respectively denote the density and the distribution function of a standard normal distribution. For observed $Z_J^{\max} = z$, the p-value for hypothesis H_J is given by $1 - F_{Z_J^{\max}}(z)$. The test can be used for other outcomes such as binary data by applying the central limit theorem.

To illustrate hypotheses testing in a phase II/III clinical trial with an example, we assume a new drug is tested at three doses; dose 1, dose 2 and dose 3. The primary hypotheses of interest are $H_j,\ j = 1, 2, 3$, where H_j is the null hypothesis comparing dose j to the control treatment. Suppose that using the stage 1 data, as in the data used to demonstrate CP in Westfall and Wolfinger (2000), $p_{1,1} = 0.0982$, $p_{1,2} = 0.0262$ and $p_{1,3} = 0.0067$. Using the Bonferroni adjusted p-values for the intersection hypotheses gives $p_{1,123} = \min\{1, 3 \times \min\{p_{1,1}, p_{1,2}, p_{1,3}\}\} = 0.0201$, $p_{1,12} = \min\{1, 2 \times \min\{p_{1,1}, p_{1,2}\}\} = 0.0524$, $p_{1,13} =$

$\min\{1, 2 \times \min\{p_{1,1}, p_{1,3}\}\} = 0.0134$ and $p_{1,23} = \min\{1, 2 \times \min\{p_{1,2}, p_{1,3}\}\} = 0.0134$. Quite naively, suppose the stage 2 data result in similar p-values, that is, $p_{2,1} = p_{1,1} = 0.0982$, $p_{2,2} = p_{1,2} = 0.0262$ and $p_{2,3} = p_{1,3} = 0.0067$ so that $p_{2,123} = p_{1,123} = 0.0201$, $p_{2,12} = p_{1,12} = 0.0524$, $p_{2,13} = p_{1,13} = 0.0134$ and $p_{2,23} = p_{1,23} = 0.0134$.

Let us assume that the total sample sizes at stage 1 and stage 2 are respectively 120 and 400. Then we could choose the weights proportional to the sample sizes in each treatment arm so that the stage 1 weight $w_1 = \sqrt{30/130}$ and stage 2 weight $w_2 = \sqrt{100/130}$. Using the inverse normal method,

$$\begin{aligned} C(p_{1,123}, p_{2,123}) &= 1 - \Phi\{w_1\Phi^{-1}(1 - p_{1,123}) + w_2\Phi^{-1}(1 - p_{2,123})\} \\ &= 1 - \Phi\left\{\sqrt{\frac{30}{130}}\Phi^{-1}(1 - 0.0201) + \sqrt{\frac{100}{130}}\Phi^{-1}(1 - 0.0201)\right\} \\ &= 0.0027. \end{aligned}$$

Similarly, $C(p_{1,12}, p_{2,12}) = 0.0138$, $C(p_{1,13}, p_{2,13}) = 0.0013$, $C(p_{1,23}, p_{2,23}) = 0.0013$, $C(p_{1,1}, p_{2,1}) = 0.0397$, $C(p_{1,2}, p_{2,2}) = 0.0042$ and $C(p_{1,3}, p_{2,3}) = 0.0004$ so that for each dose j $(j = 1, 2, 3)$,

$$\max\{C(p_{1,J}, p_{2,J})\} \leq 0.05 \text{ for } J \subseteq \{1, 2, 3\} \text{ with } j \in J.$$

Hence at the end of phase II/III clinical trial, all doses are concluded to be more effective than the control treatment.

4.5 Treatment selection in phase II/III clinical trials

Methods that can be used (or adapted) to select the most promising treatment(s) after stage 1 for testing in stage 2 have been developed by Thall et al. (1988), Schaid et al. (1990), Stallard and Todd (2003), Schmidli et al. (2007), and Zuber et al. (2006). Thall et al. (1988) consider binary outcomes and select the most promising treatment if the global null

hypothesis is not accepted at stage 1. Evidence from stage 1 and stage 2 is combined in a test similar to the combination test given by equation (4.3). Schaid et al. (1990) consider survival outcomes and their method allows for stopping after stage 1 results either for futility or overwhelming evidence. The method allows to continue with more than one experimental treatment and multiple testing is adjusted for using the Bonferroni correction. The methods by Thall et al. (1988) and Schaid et al. (1990) are respectively specific to binary and survival outcomes because of the statistics used. Stallard and Todd (2003) method generalizes these two methods because it assumes using Statistics introduced in Section 4.2.1 which can be derived for many outcomes. The most promising treatment is selected for further testing. These authors consider distinct treatments that may be different doses of the same drug but have not considered the dose-response relationship. The method we develop in the next chapter is for binary outcomes and we consider a phase II/III trial where in stage 1 several doses of the same drug are compared to a control treatment so that we model the dose-response relationship while making the choice of the dose(s) to test in stage 2.

Like Schmidli et al. (2007) and Zuber et al. (2006), we will assume the analysis will be conducted as described in Section 4.4. Given the stage 1 data, for each candidate set of the treatments (or subgroups) to be tested in stage 2, Schmidli et al. (2007) and Zuber et al. (2006) obtain the expression for the probability of all stage 2 data for which the null hypothesis will be rejected after stage 2. They use the Bayesian tools so that the expected value of this expression, which is referred to as the predictive power, is obtained. The treatment (or subgroup) that results in highest predictive is proposed for testing in stage 2. Schmidli et al. (2007) and Zuber et al. (2006) consider survival outcomes. In our proposed method, we use the same ideas but for binary outcomes. In addition, because we consider experimental treatments that are different dose levels of the same drug, we incorporate dose-response relationship. Also, we explicitly include safety data in the dose selection procedure.

Chapter 5

Dose selection in phase II/III trials

In Chapter 4, we reviewed methods of how phase II/III clinical data may be analysed. We also briefly described procedures that may be used to select experimental dose(s) to test at stage 2 after the stage 1 results. In this chapter, we expound a new procedure we have proposed (Kimani et al., 2009) for selecting the doses to test in stage 2 based on stage 1 data and prior knowledge. The selection procedure is different from the methods described in Section 4.5 in at least one of the following characteristics of our new procedure; the outcomes of interest are binary, more than one experimental dose may be selected to continue to stage 2, dose-response relationships are incorporated in the dose selection procedure and safety data is considered explicitly to make the choice of the doses to test in stage 2. This selection procedure assumes the efficacy data will be analysed as described in Section 4.4, where the closure principle is used to control the type I FWER associated with comparing the control treatment to several experimental doses and the combination tests are used to combine evidence from stage 1 and stage 2.

In the next section, we explicitly describe the setting of interest while introducing the notation that we will use to develop the dose selection procedure. In Section 5.2, we develop expressions for the probability that at least one of the candidate set of doses that

may continue to stage 2 is concluded to be effective and safe. We refer to this expression as the (penalized) conditional power. In Section 5.3, we propose a prior distribution for the parameters in the penalized conditional power which is updated by the stage 2 data to obtain the posterior distribution. We define the expected value of the penalized conditional power using the posterior distribution as the (penalized) predictive power. We propose to test at stage 2 the set of doses that has the highest predictive power. We summarise the dose selection procedure developed in this chapter in Section 5.4. In Section 5.5, we compare the new dose selection procedure to the some of the selection procedures in literature. The chapter ends by remarks describing how various associations are modeled in Section 5.6.

5.1 Setting of interest

Consider an experiment with $k_1(> 1)$ experimental doses 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_{1j}, and the number of observed toxicities, t_{1j}, at dose j for $j = 0, ..., k_1$, with $j = 0$ corresponding to the control treatment. At the onset of the phase II/III trial, the interest is to determine whether there is a safe dose among the k_1 experimental doses 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 θ_j, $j \in \{0, 1, ..., k_1\}$ is a measure of the effectiveness of treatment j. Based on the efficacy data $\mathbf{x}_1 = \{x_{10}, x_{11}, ..., x_{1k_1}\}$ 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 dose, 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 $k_2 = |\mathcal{K}_2|$. The selection procedure we propose in this chapter allows considering any of the k_1 doses in stage 1 to continue to stage 2 so that there are 2^{k_1} possible sets of doses that we could choose. To reduce the set of doses to be considered, the search may be restricted to sets of adjacent doses. Also, in practice at the phase III stage, the number of experimental doses is fewer so that the possible set of doses to be considered could be lower. Let x_{2j} and t_{2j}, $j \in \{0\} \cup \mathcal{K}_2$ with $j = 0$ corresponding to the control treatment, respectively denote the number of successes and toxicities on dose j in stage 2. At the end of stage 2, the efficacy data $\mathbf{x}_2 = (\{x_{2j}\})$, $j \in \{0\} \cup \mathcal{K}_2$ 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 4.4, 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 be most likely to lead us to finding at least one effective and safe dose at the end of stage 2 using the predictive power. In the next section, given stage 1 data, for each potential set of doses $\mathcal{K}_2$ to test in stage 2, we develop an expression for the probability at least one of the doses in $\mathcal{K}_2$ will be concluded effective and safe after stage 2 (conditional power). This probability is the sum of probabilities of different stage 2 outcomes for which at least one dose will be concluded effective and safe. The predictive power, which is the expected value of the conditional power, is given in Section 5.3.

5.2 Conditional power

As described above, in this section, assuming that stage 2 data have a distribution which depends on a fixed parameter vector, we develop an expression for the probability of concluding at least one of the k_2 doses in the potential set of doses $\mathcal{K}_2$ to be tested in stage 2 is effective given the results of stage 1. The expression is obtained by summing probabilities

of outcomes for which we will find at least one effective dose after stage 2 given stage 1 data. To incorporate the safety measure, we multiply this probability by an indicator variable that doses that are effective are safe. We will refer to this probability as the penalized (combined) conditional power. Since the conditional power is a summation of probabilities of stage 2 data, we need to determine the distribution of stage 2 data. The distribution of stage 2 data is given in the next subsection. In Section 5.2.2, we give the probability of stage 2 data for which at least one dose will be concluded effective. This probability is penalized for toxicity in Section 5.2.4.

5.2.1 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. To give the form of θ, 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;

$$p_E(d) = \frac{\exp(\alpha_E + \beta_E \log\ d)}{1 + \exp(\alpha_E + \beta_E \log\ d)} \tag{5.1}$$

and

$$p_T(d) = \frac{\exp(\alpha_T + \beta_T \log\ d)}{1 + \exp(\alpha_T + \beta_T \log\ d)} \tag{5.2}$$

so 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. For the logistic dose-response models (5.1) and (5.2), we have taken the dose in the log scale as in common in drug development but 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_{2j} successes and t_{2j} toxicities in the experimental dose $j,\ j \in \mathcal{K}_2$ is

$$f(\mathbf{x}_2, \mathbf{t}_2; \theta) = f_B(x_{20}; n_2, p_{E_0}) f_B(t_{20}; n_2, p_{T_0}) \prod_{j \in \mathcal{K}_2} f_B(x_{2j}; n_2, p_{E_j}) f_B(t_{2j}; n_2, p_{T_j}),$$

where $f_B(x_{2j}; n_2, p_{E_j})$ and $f_B(t_{2j}; n_2, p_{T_j})$, $j \in \{0\} \cup \mathcal{K}_2$ are binomial mass functions with parameter vectors (n_2, p_{E_j}) and (n_2, p_{T_j}) respectively. The parameters p_{E_j} and p_{T_j}, $j \in \mathcal{K}_2$ are respectively points on the dose-response curves (5.1) and (5.2) corresponding to dose level j. 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 (5.1) and (5.2). Otherwise, for example, an estimate of p_{E_0} is obtained by maximizing the likelihood

$$l(p_{E_0}|x_{10}, n_1) = \binom{n_1}{x_{10}} p_{E_0}^{x_{10}} (1 - p_{E_0})^{n_1 - x_{10}}.$$

5.2.2 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 an intersection hypothesis H_J in the closure set $\mathcal{H}$ can be considered fixed. The final hypothesis test for the intersection hypothesis H_J will be significant at level α if and only if $C(p_{1,J}, p_{2,J}) \leq \alpha$. The inequality can be rearranged to determine the maximum 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 test of choice is the inverse normal combination given by equation (4.3), rearranging the inequality, the final hypothesis test will be significant if and only if

$$p_{2,J} \leq 1 - \Phi \left\{ \frac{\Phi^{-1}(1-\alpha) - w_1 \Phi^{-1}(1 - p_{1,J})}{w_2} \right\}. \tag{5.3}$$

Let $l \leq |J|$ be the number of experimental doses in hypothesis H_J at stage 2. Then using the Bonferroni adjusted p-value, $p_{2,J} = \min(1, l \times \min_{j \in J}\{p_{2,j}\})$, where $p_{2,j}$ is the p-

value obtained from testing the pairwise null hypothesis H_j at the second stage. Since $p_{2,J} = \min(1, l \times \min_{j\in J}\{p_{2,j}\})$, inequality (5.3) holds if and only if

$$l \times \min_{j\in J}\{p_{2,j}\} \leq 1 - \Phi\left\{\frac{\Phi^{-1}(1-\alpha) - w_1\Phi^{-1}(1-p_{1,J})}{w_2}\right\}$$

since, as the right hand side (RHS) is less than 1, we so cannot have $1 \leq$ RHS. Dividing both sides of the above inequality by l, then hypothesis H_J is rejected after stage 2 if and only if

$$\min_{j\in J}\{p_{2,j}\} \leq \left(1 - \Phi\left\{\frac{\Phi^{-1}(1-\alpha) - w_1\Phi^{-1}(1-p_{1,J})}{w_2}\right\}\right)\Big/ l. \tag{5.4}$$

Note that if inequality (5.4) holds, then it means that for some pairwise hypothesis H_j ($j \in J$), $p_{2,j}$ is less than the RHS of inequality (5.4). Thus hypothesis H_J will be rejected at the end stage 2 if and only if

$$p_{2,j} \leq \left(1 - \Phi\left\{\frac{\Phi^{-1}(1-\alpha) - w_1\Phi^{-1}(1-p_{1,J})}{w_2}\right\}\right)\Big/ l \quad \text{for some} \quad j \in J. \tag{5.5}$$

The RHS of inequality (5.5) could be viewed as the "level of testing" for hypothesis H_J at stage 2.

For each possible number of successes in the control treatment (x_{20}), the minimum number of successes required in either of the l doses such that inequality (5.5) holds can be obtained. We will denote this minimum number of successes by $B_{x_{20}}(p_{1,J})$ where the notation reflects dependency on x_{20} and the stage 1 p-value $p_{1,J}$ for the intersection hypothesis H_J. The next subsection focusses on 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 $x_{2j} \geq B_{x_{20}}(p_{1,J})$ for some $j \in J$. To conclude that an experimental dose j 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 $j \in J$ are all rejected. We denote the set of $\mathbf{x}_2$ for which this is true by $\mathcal{R}(p_{1,j}),\ j \in \mathcal{K}_2$. The probability of concluding dose j is more effective than the control after stage 2 analysis is obtained by summing the probabilities of all outcomes in $\mathcal{R}(p_{1,j})$.

The form of $\mathcal{R}(p_{1,j})$ 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 (5.5). The minimum number of successes at dose 1 (x_{21}) for a given number of successes in the control treatment (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 (5.5) 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

$$\sum_{\mathcal{R}(p_{1,1})} f(\mathbf{x}_2; \theta) = \sum_{x_{20}=0}^{n_2} \left\{ f_B(x_{20}; n_2, p_{E_0}) \sum_{x_{21}=B}^{n_2} f_B(x_{21}; n_2, p_{E_1}) \right\}, \tag{5.6}$$

where $f_B(x_{2j}; n_2, p_{E_j})$ ($j = 0, 1$) is the probability mass function of the binomial random variable X_{2j} with parameter vector (n_2, p_{E_j}), $B = B_{x_{20}}(\max\{p_{1,J}\})$ 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 5.1 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 doses 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

$$\sum_{\mathcal{R}(p_{1,1})} f(\mathbf{x}_2;\theta) \ + \sum_{\mathcal{R}(p_{1,2})} f(\mathbf{x}_2;\theta) \ + \sum_{\mathcal{R}(p_{1,12})} f(\mathbf{x}_2;\theta), \tag{5.7}$$

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

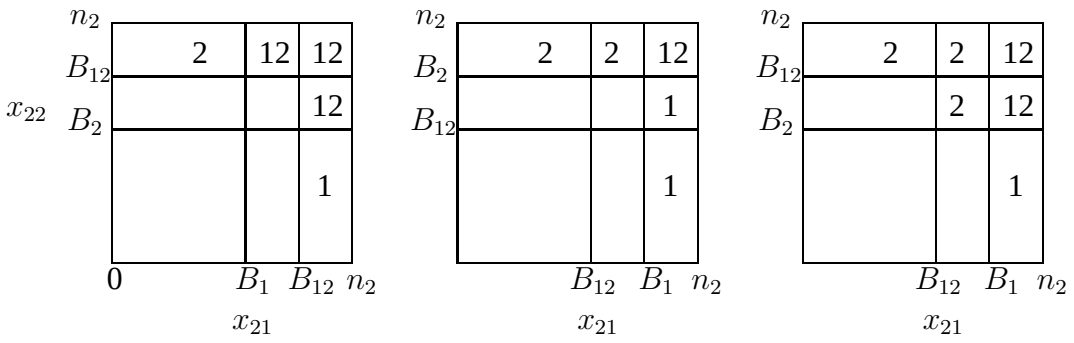

Figure 5.1: Configuration of the minimum number of successes. The x-axes are the number of successes in dose 1 (x_{21}) and y-axes the number of successes in dose 2 (x_{22}).

effective and when both dose 1 and 2 would be effective so that

$$\sum_{\mathcal{R}(p_{1,1})} f(\mathbf{x_2};\theta) = \sum_{x_{20}=0}^{n_2} f_B(x_{20};n_2,p_{E_0}) \left\{ \sum_{x_{21}=B_{12}}^{n_2} \sum_{x_{22}=0}^{B_2} f_B(x_{21};n_2,p_{E_1}) f_B(x_{22};n_2,p_{E_2}) \right\},$$

$$\sum_{\mathcal{R}(p_{1,2})} f(\mathbf{x_2};\theta) = \sum_{x_{20}=0}^{n_2} f_B(x_{20};n_2,p_{E_0}) \left\{ \sum_{x_{21}=0}^{B_1} \sum_{x_{22}=B_{12}}^{n_2} f_B(x_{21};n_2,p_{E_1}) f_B(x_{22};n_2,p_{E_2}) \right\}$$

and

$$\begin{aligned} \sum_{\mathcal{R}(p_{1,12})} f(\mathbf{x_2};\theta) &= \sum_{x_{20}=0}^{n_2} f_B(x_{20};n_2,p_{E_0}) \left\{ \sum_{x_{21}=B_{12}}^{n_2} \sum_{x_{22}=B_2}^{n_2} f_B(x_{21};n_2,p_{E_1}) f_B(x_{22};n_2,p_{E_2}) \right\} \\ &+ \sum_{x_{20}=0}^{n_2} f_B(x_{20};n_2,p_{E_0}) \left\{ \sum_{x_{21}=B_1}^{B_{12}} \sum_{x_{22}=B_{12}}^{n_2} f_B(x_{21};n_2,p_{E_1}) f_B(x_{22};n_2,p_{E_2}) \right\}, \end{aligned}$$

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

Expressions (5.6) and (5.7) 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 $\mathcal{K}_2=\{i,j\}$ for any $i,j\in\{1,...,k_1\}$. The Bonferroni adjusted p-values have been used to obtain the expressions for conditional power. The Šidak adjusted p-values similarly lead to simple expressions for conditional power. For Simes adjusted

p-values, it is not possible to obtain a single inequality such as the one resulting from Bonferroni adjusted p-values given by inequality (5.5) for composite hypotheses. However, it is still possible to obtain expressions for conditional power using this test but this becomes less straightforward as the value of k_2 increases.

We have given the expressions for when up to 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.

5.2.3 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 (5.5) is the p-value from testing the null hypothesis H_j, $j \in J$ at stage 2. If a chi-squared test is used to test the null hypothesis H_j with $j \in J$, the critical chi-squared value χ_c^2 corresponding to the level of the test (RHS of inequality (5.5)) can be determined. The null hypothesis H_j is rejected if and only if the observed chi-square value

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

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

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

where

$$U = -\{\chi_c^2(x_{20} - n_2) - 2n_2x_{20}\} \quad \text{and} \quad V = \sqrt{n_2\chi_c^2\{n_2\chi_c^2 + 8x_{20}(n_2 - x_{20})\}}.$$

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.

5.2.4 Penalizing for toxicity

Toxicity has not been incorporated in the conditional power expressions (5.6) and (5.7). 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 (5.6) 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 indicators. 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 indicators with which the probability of these events are multiplied are $I(p_{T_1} \leq \gamma)$, $I(p_{T_2} \leq \gamma)$ and $I(p_{T_1} \leq \gamma, p_{T_2} \leq \gamma)$.

5.3 Predictive power

The conditional power expressions obtained in Section 5.2 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

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

where $l(\theta|\mathbf{x}_1, \mathbf{t}_1, n_1)$ is the likelihood function of θ given the observed data $(\mathbf{x}_1, \mathbf{t}_1, n_1)$ from the k_1 doses of the experimental treatment observed at the end of the first stage. Assuming the number of successes and toxicities at each dose level are independent,

$$l(\theta|\mathbf{x}_1, \mathbf{t}_1, n_1) = \prod_{j=1}^{k_1} \binom{n_1}{x_{1j}} p_{E_j}^{x_{1j}} (1 - p_{E_j})^{n_1 - x_{1j}} \binom{n_1}{t_{1j}} p_{T_j}^{t_{1j}} (1 - p_{T_j})^{n_1 - t_{1j}},$$

where p_{E_j} and p_{T_j} are respectively the probabilities of success and toxicity at dose j. The predictive power is then obtained by evaluating the posterior mean of the conditional power, that is, the predictive power is given by

$$\int_{\Theta} (\text{CP}_\theta)\pi(\theta|\mathbf{x}_1, \mathbf{t}_1, n_1)d\theta, \tag{5.8}$$

where CP_θ denotes the conditional power. For example if $\mathcal{K}_2 = \{1, 2\}$, the penalized predictive power is given by

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

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_j} enter the density $f(\mathbf{x}_2; \theta)$ and which probabilities p_{T_j} 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.

5.3.1 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. (1996) as was described in Section 2.2.2. This requires eliciting beta prior distributions at two dose levels d_{-1} and d_0 for each dose-response curve since each dose-response curve is defined by two parameters. We assume

the beta prior distributions at each dose level can be elicited as was described in Section 2.2.1 using the contribution by Thall and Simon (1994) and Lindley and Phillips (1976). Suppose for the probability of success, $p_{E_j} = p_E(d_j)$ at dose j ($j = -1, 0$), the elicited prior distribution is Beta(x_{1j}, y_{1j}). Then assuming that the probabilities of success are related to the dose levels according to the logistic model (5.1) and using the prior distribution given by equation (2.6), the prior distribution for (α_E, β_E)

$$\pi_0(\alpha_E, \beta_E) = \prod_{j=-1}^{0} \frac{p_{E_j}^{x_{1j}}(1-p_{E_j})^{y_{1j}}}{B(x_{1j}, y_{1j})} \left|\log\left(\frac{d_{-1}}{d_0}\right)\right|, \tag{5.9}$$

where B is the beta function and

$$p_{E_j} = \frac{\exp(\alpha_E + \beta_E \ \log \ d_j)}{1 + \exp(\alpha_E + \beta_E \ \log \ d_j)}, \quad j = -1, 0.$$

Similarly suppose a beta prior distribution Beta$(p_{T_j}; t_{1j}, u_{1j})$ is elicited for the probability of toxicity, $p_{T_j} = p_T(d_j)$ at dose j ($j = -1, 0$), then assuming logistic dose-response (5.2) for the probabilities of toxicity, the prior distribution of (α_T, β_T)

$$\pi_0(\alpha_T, \beta_T) = \prod_{j=-1}^{0} \frac{p_{T_j}^{t_{1j}}(1-p_{T_j})^{u_{1j}}}{B(t_{1j}, u_{1j})} \left|\log\left(\frac{d_{-1}}{d_0}\right)\right|, \tag{5.10}$$

where B is the beta function and

$$p_{T_j} = \frac{\exp(\alpha_T + \beta_T \ \log \ d_j)}{1 + \exp(\alpha_T + \beta_T \ \log \ d_j)}, \quad j = -1, 0.$$

As in Section 5.1, let x_{1j} denote the number of successfully treated patients and $y_{1j} = n_1 - x_{1j}$ the number of patients that are not treated successfully at stage 1 after treatment with dose j $(j = 1, ..., k_1)$. After observation of the stage 1 data, using equation (2.7), the updated distribution (posterior distribution) of (α_E, β_E) is

$$\pi(\alpha_E, \beta_E | \mathbf{x}_1, n_1) \propto \prod_{j=-1}^{k_1} p_{E_j}^{x_{1j}}(1-p_{E_j})^{y_{1j}}, \tag{5.11}$$

where

$$p_{E_j} = \frac{\exp(\alpha_E + \beta_E \ \log \ d_j)}{1 + \exp(\alpha_E + \beta_E \ \log \ d_j)}, \quad j = -1, 0, 1, ..., k_1.$$

Similarly let t_{1j} denote the number of patients that experience toxicity at stage 1 and $u_{1j} = n_1 - t_{1j}$ the number of patients that do not experience toxicity, then the posterior distribution of

$$\pi(\alpha_T, \beta_T | \mathbf{t}_1, n_1) \propto \prod_{j=-1}^{k_1} p_{T_j}^{t_{1j}} (1 - p_{T_j})^{u_{1j}}, \tag{5.12}$$

where

$$p_{T_j} = \frac{\exp(\alpha_T + \beta_T \ \log \ d_j)}{1 + \exp(\alpha_T + \beta_T \ \log \ d_j)}, \quad j = -1, 0, 1, ..., k_1.$$

If the control treatment is a lower dose of the same drug as the experimental treatments, data from the control group are used in updating the prior distributions of (α_E, β_E) and (α_T, β_T). If it is a different drug, a beta prior distribution Beta$(p_{E_0}; a_0, b_0)$ for the probability of successful treatment at control treatment which is conjugate for the likelihood function

$$l(p_{E_0} | x_{10}, n_1) = \binom{n_1}{x_{10}} p_{E_0}^{x_{10}} (1 - p_{E_0})^{n_1 - x_{10}}$$

is elicited. The parameters a_0 and b_0 are elicited as explained in Section 2.2.1. The resulting posterior has a beta distribution Beta$(p_{E_0}; a_0 + x_{10}, b_0 + n_1 - x_{10})$.

5.4 Summarizing the dose selection procedure

In Chapter 4, we introduced seamless phase II/III clinical trials and described the challenges in these trials. One of the challenges of seamless phase II/III clinical trials is how to analyse these trials without inflating type I error rates. In Chapter 4, we described in detail an analysis method given in Bauer and Kieser (1999), Hommel (2001) and Bretz et al. (2006).

When several doses are to be tested with the possibility of dropping some doses after stage 1, this analysis allows selection of any combination of doses to continue to stage 2 with a final analysis that strongly controls the FWER.

The second challenge of seamless phase II/III is how best to select the doses that are tested in stage 2. In the previous sections of this chapter, we have proposed a procedure for selecting the doses to test in stage 2 of a seamless phase II/III clinical trials assuming the analysis described in Bauer and Kieser (1999), Hommel (2001) and Bretz et al. (2006). The doses are selected by calculating the predictive power of each set of doses that may be tested in stage 2 using the knowledge of this analysis and stage 1 while also incorporating the prior knowledge through prior distributions, and proposing to test in stage 2 the set of doses with the highest predictive power.

In more detail, before the seamless phase II/III clinical trial, prior distributions given by equations (5.9) and (5.10) for the parameters that define the dose-response curves for the probability efficacy and for the probability of toxicity are elicited. After stage 1, stage 1 efficacy and toxicity data are used to update prior distributions of the parameters for the efficacy and toxicity dose-response curves using equations (5.11) and (5.12). In addition to using the stage I data to update prior distribution, stage 1 efficacy data are used to obtain all intersection hypotheses p-values p_J ($J \subseteq \{1, ..., k_1\}$), where k_1 is the number of experimental doses in stage 1.

The stage 1 p-values are required to obtain expressions for conditional power, the conditional probability of concluding at least one of the experimental doses that is tested in stage 2 is efficacious and safe given stage 1 data. The expressions for conditional power are obtained for each potential set of doses that may be tested in stage 2. For example, if the number of experimental doses at stage 1 is 3, and can only proceed with a single dose or consecutive pairs of doses, then 5 expressions for conditional power are obtained. The 5 expressions for conditional power correspond to the single experimental doses 1, 2 and 3, and the consecutive pairs doses 1 and 2 and doses 2 and 3. For proceeding with the

single doses, the conditional powers have the form given by expression (5.6) multiplied by probability of safety described in Section 5.2.4. For continuing with two experimental doses, the conditional powers have the form given by expression (5.7) multiplied by probabilities of safety as described in Section 5.2.4. Expressions (5.6) and (5.7) are expressions for probability of all data for which we conclude at least one of the experimental doses is more efficacious than control. Stage 1 p-values enter these expressions by determining the minimum successes using equation (5.5). Equation (5.5) also requires the weights w_1 and w_2, and the level of the test α to be pre-defined. A typical level for one-sided tests is 0.025. The weights w_1 and w_2 could be chosen proportional to the pairwise comparisons sample sizes (that is patients treated in each treatment arm) as was demonstrated using the example given at the end of Section 4.4.

The posterior distribution is then used to obtain the expected value of the conditional powers using equation (5.8). This is the predictive power. For example, using the example given in the previous paragraph, 5 predictive power values corresponding to proceeding to single doses 1, 2 and 3 and the consecutive pairs doses 1 and 2 and doses 2 and 3 will be obtained. The predictive power values are compared to choose the set of doses to test in stage 2. The set of doses with the highest predictive power is chosen for testing in stage 2.

The final analysis does not include the prior knowledge. The prior distributions are only used to plan stage 2.

5.5 Comparing the dose selection procedure with existing methods

In this section, we explicitly discuss the features of the dose selection developed above in comparison with the existing methods. The features described in detail in this section are generally attractive. Before describing the attractive features, in this paragraph, we briefly

mention the features of the dose selection developed in this chapter that may be improved while also giving the thesis sections in which these features are described in more detail. The first feature that may be improved concerns how the association between efficacy and toxicity is modeled. The association between efficacy and toxicity is not modeled explicitly and the details of how the association between efficacy and toxicity are given in Section 5.6. Also, we have assumed some known dose-response curves (5.1) and (5.2) respectively for efficacy and toxicity. However, there may be uncertainty in the form of the dose-response curves and it would be desirable to include this uncertainty in the dose selection procedure. Bretz et al. (2005) and Klingenberg (2009) have proposed methods that include model uncertainty while estimating the minimum effective dose. In Section 7.2, we have described how the ideas in Klingenberg (2009) could be borrowed to include model uncertainty in the dose selection procedure.

Some of the key attractive features of this procedure are that it: (1) allows for the dose-response relationships by using the logistic models (5.1) and (5.2) respectively to model the probability of efficacy and toxicity, (2) as described in Section 5.3, uses the Bayesian tools by defining some prior distributions for the parameters that enter the probability of concluding at least one of the doses that is tested in stage 2 is effective and safe, and (3) explicitly includes safety by including the probability of concluding that the doses that are concluded effective after stage 2 are also safe as described in Section 5.2.4. In the light of these features, we explain why this new procedure is different from the existing methods and why the new procedure is expected to perform better.

The method proposed by Stallard and Todd (2003) selects among the experimental treatments (doses) tested in stage 1, the best performing treatment (dose) in terms of its efficacious level compared to the control treatment. If more than one dose is selected, it is not guaranteed the type I error rates are controlled at the desired level. This restriction also applies to the method proposed by Thall et al. (1988). Our procedure does not have this restriction because the analysis assumed allows for any decision rule without inflating

type I error rates. The simulation results presented in the next chapter show under some scenarios, it is often better to choose two doses so that it is better to have a procedure that allows continuing with more than a single dose. Further, unlike Stallard and Todd (2003) and Thall et al. (1988), with this procedure the prior knowledge about the experimental doses and the control treatment is formally used in planning stage 2.

Stallard and Todd (2003), Thall et al. (1988), Schmidli et al. (2007) and Zuber et al. (2006) would be adequate if the experimental treatments are distinct treatments. However, when the treatments are different doses of the same drug, then these methods do not exploit the dose-response relationship which is expected when the treatments are different doses of the same drug. The simulation results presented in the next chapter show the dose selection developed in this chapter capture the dose-response relationships.

Finally, we have explicitly included safety in the dose selection procedure. It is likely the methods proposed by Thall et al. (1988), Schaid et al. (1990) and Stallard and Todd (2003) require a lot effort to include safety. It may be easy to incorporate safety in the methods proposed by Schmidli et al. (2007) and Zuber et al. (2006) but these authors have not done this.

5.6 Remarks on the dose selection procedure

The method of Bedrick et al. (1996) of eliciting prior distributions for the dose-response curves parameters assumes that the beta prior distributions elicited at dose levels d_{-1} and d_0 are independent. This assumption simplifies the mathematics but as noted in Whitehead et al. (2006), 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 $\beta_T \geq 0$. This is because assuming the elicited beta distributions at dose levels d_{-1} and d_0 are independent, for example implies that the probability that the probability of efficacy at dose level d_{-1} is higher than the probability of efficacy at dose level d_0 for $d_{-1} < d_0$ is not zero even when it is believed efficacy improves

with dose level. This in turn means it is possible to have $\beta_E < 0$ when it is believed $\beta_E \geq 0$. To partly address this problem the beta prior distributions are elicited at locations that are far from each other. Also as in Whitehead et al. (2006), since we are interested in the posterior means of the conditional powers associated with continuing with different set of doses, negative parameter values for the slope parameters will not have undesired effects on the predictive power. Further, since we obtain the posterior distributions by updating the prior distributions using all the phase II clinical trial data, for the posterior distributions, the slope parameters are unlikely to be negative when the slope parameters are actually positive.

The use of conditional efficacy and toxicity models (5.1) and (5.2) may raise concern about the association between efficacy and safety. At each dose level, we are assuming independence between the probabilities of efficacy and toxicity to obtain the predictive power. However, because we are using more than one experimental dose, this does not imply marginal independence between efficacy and toxicity. To demonstrate this, using odds ratio as a measure of the association, first we give the expression for the odds ratio and then give the implied odds ratio for some scenarios. As above, let p_{E_j} and p_{T_j} respectively denote the probability of efficacy and of toxicity at dose j $(j = 1, ..., k_1)$. Further let p_{R_j} denote the probability of a patient being randomized to dose j $(j = 1, ..., k_1)$. Using law of total probability, the marginal probabilities of efficacy (p_E) and toxicity (p_T) and the probability of efficacy and toxicity (p_{ET}) assuming independence of safety and efficacy at each dose level are expressed as:

$$p_T = \sum_{j=1}^{k_1} p_{T_j} p_{R_j}, \qquad p_E = \sum_{j=1}^{k_1} p_{E_j} p_{R_j} \quad \text{and} \quad p_{ET} = \sum_{j=1}^{k_1} p_{E_j} p_{T_j} \cdot p_{R_j}$$

so that the marginal odds ratio is given by

$$\frac{p_{ET}(1 - p_E - p_T + p_{ET})}{(p_T - p_{ET})(p_E - p_{ET})}. \tag{5.13}$$

To give examples of some implied odds ratio, we use the three scenarios used to assess the

effect probability of efficacy in the next chapter. We refer to the three scenarios as the reference scenario, Scenario 2 and Scenario 3. The dose-response curves for the three scenarios are given in Figure 5.2. In the three scenarios, the dose-response curve for the probability of toxicity is the same with $(\alpha_T, \beta_T) = (-2.5782, 0.1621)$ and is given by the continuous line (—). The three scenarios differ in the parameter vector (α_E, β_E). For the reference scenario, the parameter vector $(\alpha_E, \beta_E) = (-1.4867, 0.2720)$ and the dose-response curve is given by the dashed line (- - -). In Scenario 2, $(\alpha_E, \beta_E) = (-2.6226, 0.3187)$ and the dose-response curve is given by the dotted line ($\cdots$). For Scenario 3, $(\alpha_E, \beta_E) = (-0.8473, 0)$ and the dose-response curve is given by the dashed and dotted line ($\cdot$ - $\cdot$ - $\cdot$).

Assuming a new drug is tested at the marked dose levels on the x-axis of Figure 5.2, that is dose levels 10.5mg, 35mg, 87.5mg, 262.5mg, 700.0mg and 1050.0mg, the marginal odds ratios for reference scenario, Scenario 2 and Scenario 3 respectively are 1.13, 1.14 and 1.0. In Scenario 3 the probability of toxicity increases with dose level and probability of efficacy does not change with the dose level so that an odds ratio of 1 would not be a bad assumption. Scenario 2 has a higher odds ratio than the reference scenario which is what we would desire. This is made possible since we assume some dose-response curves. The marginal odds ratio expression (5.13) holds even when probabilities of efficacy and toxicity are not modelled using some dose-response curves. Modelling the probabilities at each dose level independently may result in instances where the marginal odds ratio for Scenario 2 is less than the odds ratio for the reference scenario. By using different dose levels, as would be expected, we observed that the modelled odds ratios for the reference scenario and Scenario 2 are higher when: (1) patients are allocated to more dose levels, and (2) the experimental dose level are further apart.

To conclude, by modelling the probabilities of efficacy and the probabilities of toxicity as described above, we assume that the probability of efficacy is independent of the probability of toxicity given dose subject to a given marginal odds ratio. The marginal odds ratio is induced by assuming some dose-response curves for the probabilities of ef-

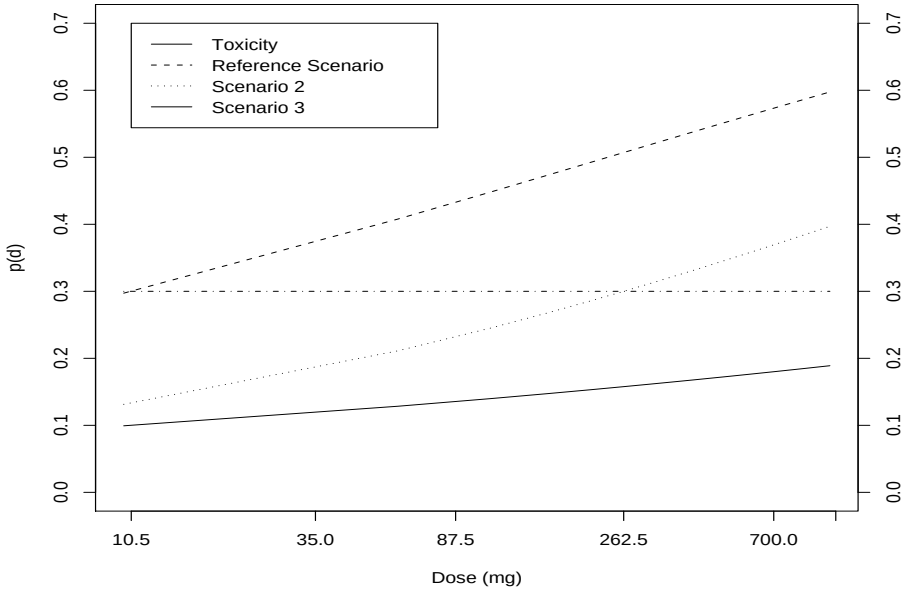

Figure 5.2: Different scenarios of dose response curves used to give examples of implied marginal associations.

ficacy and of toxicity. Thus, although we assume independence at each dose level, there is a restriction of the values the probabilities of efficacy and of toxicity can take. If there is correlation between efficacy and toxicity, we reduce the set of values probabilities of efficacy and of toxicity can take at each dose level so that the independence assumption is less strong compared to modelling outcomes (efficacy and toxicity) at each dose level independent and also outcomes at a dose level independent of the outcomes in other dose levels.

Chapter 6

Simulation studies

In Chapter 5, we 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 without inflating type I error rates. In this chapter, the performance of the selection procedure is investigated using simulation studies. Different scenarios for the underlying true probabilities of toxicity and efficacy using the different doses are considered. For each of these scenarios, 1000 studies were simulated in order to obtain the probabilities of continuing to stage 2 with each of the potential doses.

6.1 Simulation model parameter values

The simulation studies are based on the trial described by Whitehead et al. (2006). We assume that the new drug is tested at dose levels 10.5mg, 35.0mg, 87.5mg, 262.5mg, 700.0mg and 1050.0mg plus a control. To conform with the previous chapters, we simply refer to the experimental doses in increasing dose levels as dose 1, dose 2, dose 3, dose 4, dose 5 and dose 6. Further, in all simulation studies, we assume that γ, the accepted maximum probability 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, the true parameter values for (α_E, β_E) and (α_T, β_T) corresponding to dose-response curves (5.1) and (5.2) are assumed to be (-1.4867, 0.2720) and (-2.5782, 0.1621) respectively. We refer to this set of parameter values for (α_E, β_E) and (α_T, β_T) as the reference scenario. With these parameter values, all doses are acceptably safe and dose 1 is as efficacious as the control treatment while all the other experimental doses are more efficacious than the control treatment.

To explore the effect of efficacy, two more scenarios are compared to the reference scenario. We will refer to them as efficacy Scenario 2 and efficacy Scenario 3. The two scenarios have the same value for the true parameter vector (α_T, β_T) as the reference scenario but differ from the reference scenario in the value of the parameter vector (α_E, β_E). The probabilities of efficacy and toxicity at each experimental dose level for the two scenarios and the reference scenario are given in Table 6.1 while the dose-response curves for the two scenarios and the reference scenario are given in the left panel of Figure 6.1. The marked points on the x-axis correspond to the experimental doses. As the linear predictor of the dose-response curves are on the natural log dose scale, doses are plotted on the log scale in Figure 6.1, so that the higher doses are closer to each other on the x-axis. The continuous line (—) shows the toxicity dose-response curve for the reference scenario. The same toxicity dose-response curve will be used for the efficacy Scenario 2 and efficacy Scenario 3. As already described above, the dose-response curves shows that all the experimental doses are acceptably safe since the probability of toxicity in all cases is less than 0.2. The dashed line (- - -) gives the efficacy dose-response curve for the reference scenario. In this scenario, dose 1 is as efficacious as the control while doses 2 to 6 are more efficacious than the control. The dotted line ($\cdot\cdot\cdot$) gives the efficacy dose-response curve for the efficacy Scenario 2 for which $(\alpha_E, \beta_E) = (-2.6226, 0.3187)$. In this scenario only doses 5 and 6 are more efficacious compared to control. The dashed and dotted line ($\cdot$ - $\cdot$ - $\cdot$) gives the efficacy Scenario 3 with $(\alpha_E, \beta_E) = (-0.8473, 0)$. In this scenario, all the experimental

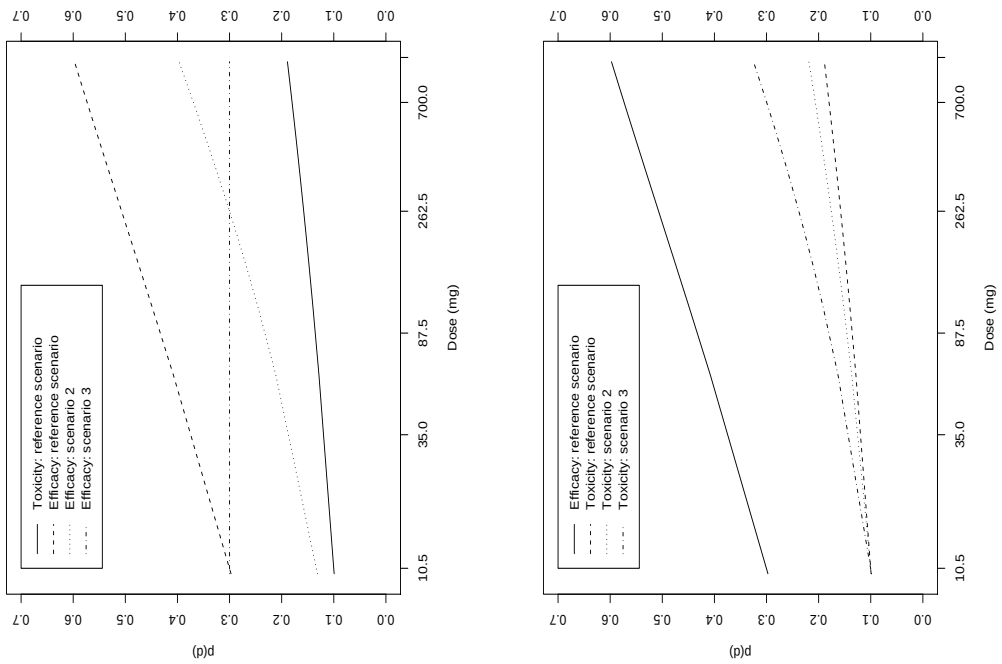

Figure 6.1: Underlying true dose-response curves. The left panel shows different scenarios for efficacy while the right panel shows different scenarios for toxicity.

doses have the same efficacy level as the control treatment. For each of these three scenarios, 1000 studies were simulated and in each case predictive power calculated to determine the dose(s) that continue to stage 2. The simulation results are given in Section 6.5.

We also wish to explore the effect of true toxicity on the proposed dose selection procedure. To do this, two more scenarios that will be compared to the reference scenario will be considered. We will refer to them as toxicity Scenario 2 and toxicity Scenario 3. The dose-response curves for the probability of efficacy for the toxicity Scenario 2 and the toxicity Scenario 3 have the same parameter vector (α_E, β_E) as the reference scenario. However, the dose-response curves for the probability of toxicities for the two scenarios are different from the reference scenario. Table 6.2 gives the probabilities of efficacy and toxicity at each experimental dose level for the reference scenario and the toxicity Scenario 2 and the toxicity Scenario 3 while Figure 6.1 (right panel) gives the dose-response curves

Table 6.1: Probabilities of efficacy and toxicity at tested dose levels for the three scenarios used to assess effect of efficacy

	Dose levels (mg)					
Scenario (Outcome)	10.5	35.0	87.5	262.5	700.0	1050
All Scenarios (Toxicity)	0.10	0.12	0.14	0.16	0.18	0.19
Reference Scenario (Efficacy)	0.30	0.37	0.43	0.51	0.57	0.60
Scenario 2 (Efficacy)	0.13	0.18	0.23	0.23	0.37	0.40
Scenario 3 (Efficacy)	0.30	0.30	0.30	0.30	0.30	0.30

for the reference scenario and the toxicity Scenario 2 and the toxicity Scenario 3. The continuous line (—) shows the efficacy dose-response curve for the reference scenario. The same dose-response curve for efficacy will be used to simulate studies for toxicity Scenario 2 and 3. The dashed line (- - -) gives the toxicity dose-response curve for the reference scenario for which all doses are acceptably safe. The dotted line ($\cdot\cdot\cdot$) and the dashed and dotted line ($\cdot$ - $\cdot$ - $\cdot$) respectively give the toxicity dose-response curves for the toxicity Scenario 2 and toxicity Scenario 3. For the toxicity Scenario 2 ($\cdot\cdot\cdot$), $(\alpha_T, \beta_T) = (-2.6728, 0.2023)$. In this scenario dose 5 is nearly safe (probability of toxicity is 0.206) and dose 6 is unacceptably toxic while the other doses are acceptably safe. On the other hand, for the toxicity Scenario 3 ($\cdot$ - $\cdot$ - $\cdot$), $(\alpha_T, \beta_T) = (-2.9523, 0.3211)$ so that doses 1 to 3 are acceptably safe while the other experimental doses would be considered too toxic. A further two sets of 1000 studies for each of these two new scenarios were simulated and predictive powers evaluated to determine the dose(s) that continue to stage 2. The simulation results for these scenarios are also given in Section 6.5.

In all the simulation studies, it will be assumed that n_1, the number of patients tested at each dose at stage 1 is 20 and the total number of patients available for testing at stage 2 is 400 such that n_2, the number of patients allocated to each treatment arm at stage 2 is $400/(k_2 + 1)$, where k_2 is the number of the number of doses chosen to be tested in the second stage. We will demonstrate the method for clinical trials in which up to 2 experimental doses are included with the control in stage 2, that is $k_2 = 1$ or $k_2 = 2$. We

Table 6.2: Probabilities of efficacy and toxicity at tested dose levels for the three scenarios used to assess effect of toxicity

	Dose levels (mg)					
Scenario (Outcome)	10.5	35.0	87.5	262.5	700.0	1050
All Scenarios (Efficacy)	0.30	0.37	0.43	0.51	0.57	0.60
Reference Scenario (Toxicity)	0.10	0.12	0.14	0.16	0.18	0.19
Scenario 2 (Toxicity)	0.10	0.12	0.15	0.18	0.206	0.22
Scenario 3 (Toxicity)	0.10	0.14	0.18	0.24	0.30	0.33

will also restrict testing consecutive experimental doses so that we will not consider for example a stage 2 trial with dose 1 and dose 3. The restriction to consecutive experimental and considering $k_2 \leq 2$ is not a limitation of the selection procedure. The selection procedure can be extended to consider $k_2 > 2$ but the expressions for the conditional power would involve summing over more dimensions which is computational expensive. Further, including non-consecutive experimental doses increases the sets to be compared increasing the computation time.

6.2 Prior distributions

As described in Section 5.3, to evaluate the predictive power, beta prior distributions for the probability of efficacy using the control treatment and for the probabilities of efficacy and toxicities at two dose levels of the new drug are required. To determine the parameter values for the beta prior distributions, we examined the beta curves and the 90% credible interval width as described in Section 2.2 while also considering what would be typical of the prior distributions elicited in practice. In all the simulation studies, the predictive power is evaluated assuming a beta prior distribution Beta(12, 28) for the probability of efficacy using the control treatment.

Beta prior distributions for probabilities of successful treatment and probabilities of toxicity are defined at dose levels 10.50mg and 5000mg. Figure 6.2 shows the densities

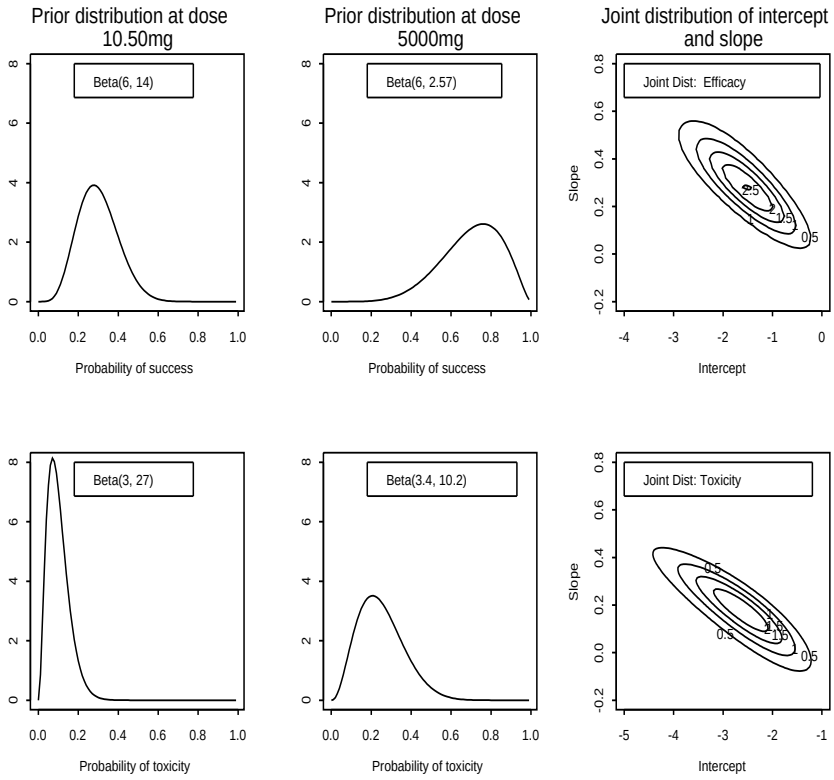

Figure 6.2: Elicited prior densities. Row 1 and 2 give the prior distributions for efficacy and toxicity respectively. Columns 1 and 2 correspond to prior distributions at dose 10.50mg and 5000mg respectively. Column 3 gives the resulting joint prior distributions.

of the elicited prior distributions. Rows 1 and 2 respectively give the prior distributions for the efficacy and toxicity models. Columns 1 and 2 respectively give the beta prior distributions at dose levels 10.50mg and 5000mg whose parameter values are given in the legends. Column 3 is the resulting joint prior distributions of the intercept and slope parameters obtained as described in Section 2.2.2.

6.3 Computational details

The predictive power was defined in Section 5.3 as the expected value of the conditional power using the posterior distribution. Evaluating this expectation has two complexities: (1) the integral in equation (5.8) cannot be given in closed form, and (2) the expression for CP in equation (5.7) requires calculation of an expectation that involves summing over more than one dimension. To overcome the first complexity, we used numerical quadrature to integrate over the parameter space. In the rest of this section, we describe how we avoided more than one summation using the normal approximation to the binomial distribution.

Examples of the form of the conditional power are given by expressions (5.6) and (5.7). These expressions entail summing over possible values in the control treatment and over outcomes in the experimental doses for which at least one dose is concluded to be better than control. These summations are computationally expensive so that to make it feasible to evaluate the predictive power for the simulation studies, some approximations are needed. Expression (5.6) is given by

$$\sum_{x_{20}=0}^{n_2} \left\{ \binom{n_2}{x_{20}} p_{E_0}^{x_{20}} (1-p_{E_0})^{n_2-x_{20}} \sum_{x_{21}=B}^{n_2} \binom{n_2}{x_{21}} p_{E_1}^{x_{21}} (1-p_{E_1})^{n_2-x_{21}} \right\}, \tag{6.1}$$

where $B = B_{x_{20}}(\max\{p_{1,J}\})$ is the minimum number of successes in dose 1 required to conclude that dose 1 is better than control after stage 2. This expression requires summing over a grid of possible number of successes in the control treatment (x_{20}) and for each possible value of x_{20}, summing over the number of successes in dose 1 (x_{21}) from $B_{x_{20}}(\max\{p_{1,J}\})$ to n_2. To reduce the computation time, the latter summation is approximated using the normal approximation to the binomial distribution. Suppose the probability of successful treatment with dose 1 is p_{E_1} so that the number of successes at stage 2 (x_{21}) is Bin(n_2, p_{E_1}). The number of successes (x_{21}) is approximately normal

$N(n_2 p_{E_1}, n_2 p_{E_1}(1-p_{E_1}))$ so that using the properties of the normal distribution

$$\begin{aligned} \text{Prob}(X_{21} \geq x_{21}) &= 1 - \text{Prob}(X_{21} < x_{21}) \\ &\simeq 1 - \Phi\left\{\frac{x_{21} - n_2 p_{E_1}}{\sqrt{n_2 p_{E_1}(1-p_{E_1})}}\right\}, \end{aligned}$$

where Φ is the standard normal distribution function. Hence, to reduce the computation time for evaluating expression (6.1), the following approximation is used

$$\sum_{x_{21}=B}^{n_2} \binom{n_2}{x_{21}} p_{E_1}^{x_{21}} (1-p_{E_1})^{n_2 - x_{21}} = 1 - \Phi\left\{\frac{B - n_2 p_{E_1}}{\sqrt{n_2 p_{E_1}(1-p_{E_1})}}\right\}.$$

On the other hand, expression (5.7) has summations of the form

$$\sum_{x_{20}=0}^{n_2} \left\{ f_B(x_{20}; n_2, p_{E_0}) \left\{ \sum_{x_{21}=B_{12}}^{n_2} \sum_{x_{22}=0}^{B_2} f_B(x_{21}; n_2, p_{E_1}) f_B(x_{22}; n_2, p_{E_2}) \right\} \right\}, \qquad (6.2)$$

where $f_B(x_{20}; n_2, p_{E_0})$, $f_B(x_{21}; n_2, p_{E_1})$ and $f_B(x_{22}; n_2, p_{E_2})$ are respectively probability mass functions of $\text{Bin}(x_{20}; n_2, p_{E_0})$, $\text{Bin}(x_{21}; n_2, p_{E_1})$ and $\text{Bin}(x_{22}; n_2, p_{E_2})$ and B_{12} and B_2 are quantities that depend on x_{20} and stage 1 results. The normal approximation to the binomial distribution in expression (6.2) is used for the term

$$\left\{ \sum_{x_{21}=B_{12}}^{n_2} \sum_{x_{22}=0}^{B_2} f_B(x_{21}; n_2, p_{E_1}) f_B(x_{22}; n_2, p_{E_2}) \right\}.$$

Since conditional on p_{E_1} and p_{E_2}, the number of successes in dose 1 (x_{21}) and dose 2 (x_{22}) are independent, then the above expression could be re-expressed as

$$\left\{ \sum_{x_{21}=B_{12}}^{n_2} f_B(x_{21}; n_2, p_{E_1}) \sum_{x_{22}=0}^{B_2} f_B(x_{22}; n_2, p_{E_2}) \right\},$$

which using the normal approximation to the binomial distribution as described above can be approximated by

$$\left(1 - \Phi\left\{\frac{B_{12} - n_2 p_{E_1}}{\sqrt{n_2 p_{E_1}(1-p_{E_1})}}\right\}\right) \left(\Phi\left\{\frac{B_2 - n_2 p_{E_2}}{\sqrt{n_2 p_{E_2}(1-p_{E_2})}}\right\}\right).$$

These approximations reduced the computation time for evaluating the predictive powers associated with continuing to stage 2 with a single dose and two consecutive doses in a single simulation study from a few days to a few minutes on a personal computer using the R package. For each scenario, 1000 simulation studies were used and it takes about three days to simulate and evaluate predictive powers for the 1000 simulation studies.

6.4 Explanation of how results will be obtained

In the previous sections of this chapter, we have described the scenarios (based on true parameter values) we will use to investigate the operating characteristics of the dose selection procedure, the total sample sizes at each stage that will be used in the simulation studies (120 for stage 1 and 400 for stage 2), prior distributions that will be used to calculate the predictive powers and the computational details. Also, we explained that in the simulation results we present in this chapter, we can proceed to stage 2 with single experimental doses 1, 2, 3, 4, 5 and 6 or pairs of consecutive experimental doses, that is, doses 1 and 2, doses 2 and 3, doses 3 and 4, doses 4 and 5, and doses 5 and 6. In this section, we describe how we obtained the simulation results given in this chapter.

For each scenario, 1000 simulation studies will be carried out. In each simulation study, stage 1 data are simulated using the underlying true parameter values. The stage 1 data consist of the number of successes from 20 simulated patients per experimental dose and the control treatment. Using the stage 1 data, stage 1 p-values for all intersection hypotheses $p_{1,J}$, $J \subseteq \{1, ..., 6\}$ are obtained. Given these p-values, for each potential set of doses that may be tested in stage 2, that is, the single doses and consecutive pairs of consecutive doses, we obtain the expressions for conditional power as described in Section 5.2. Thus there will be 11 separate expressions for conditional power corresponding to the 6 experimental doses and 5 pairs of consecutive doses. For the single doses, the conditional powers have the form given by expression (5.6) multiplied by probability of safety

described in Section 5.2.4. For pairs of consecutive doses, the conditional powers have the form given by expression (5.7) multiplied by probabilities of safety as described in Section 5.2.4. To obtain the expressions for conditional power, we need to define the weights and level of the tests given in inequality (5.3). The hypotheses are tested at level 2.5% and the squares of the weights are proportional to the total sample sizes, that is, $w_1 = \sqrt{120/520}$ and $w_2 = \sqrt{400/520}$.

The next step in the dose selection is obtaining the predictive power. To obtain the predictive power, we need the distribution of the parameters in the expressions for conditional power. The parameters are given by the dose response curves and the prior distributions of these parameters are given in Section 6.2. The prior distribution are updated using the stage 1 data to obtain the posterior distributions of the dose response curves parameters using equations (5.11) and (5.12). The posterior distributions are used to obtain the expected value of each of the 11 expressions for conditional power described in the previous paragraph. This is the predictive power. The 11 predictive power values are compared to choose the set of doses to test in stage 2. The set of doses with the highest predictive power is proposed for testing in stage 2. This is repeated for the 1000 simulated studies. To obtain the simulated probability of selecting each potential set of doses, the number of times this set is proposed for testing in stage 2 is divided by 1000.

6.5 Comparing results for different scenarios

Figure 6.3 shows histograms of the simulated probabilities of continuing to stage 2 with each dose and consecutive pair of doses. Each histogram corresponds to one of the scenarios described in Section 6.1 and is based on a 1000 simulation studies. On the x-axis, the notation di, $i \in \{1, ..., 6\}$ means dose i is selected for testing at stage 2 while dij with $i, j \in \{1, ..., 6\}$ means both doses i and j are selected for testing at stage 2. The selected set of doses has the highest predictive power among the potential doses or pair of doses

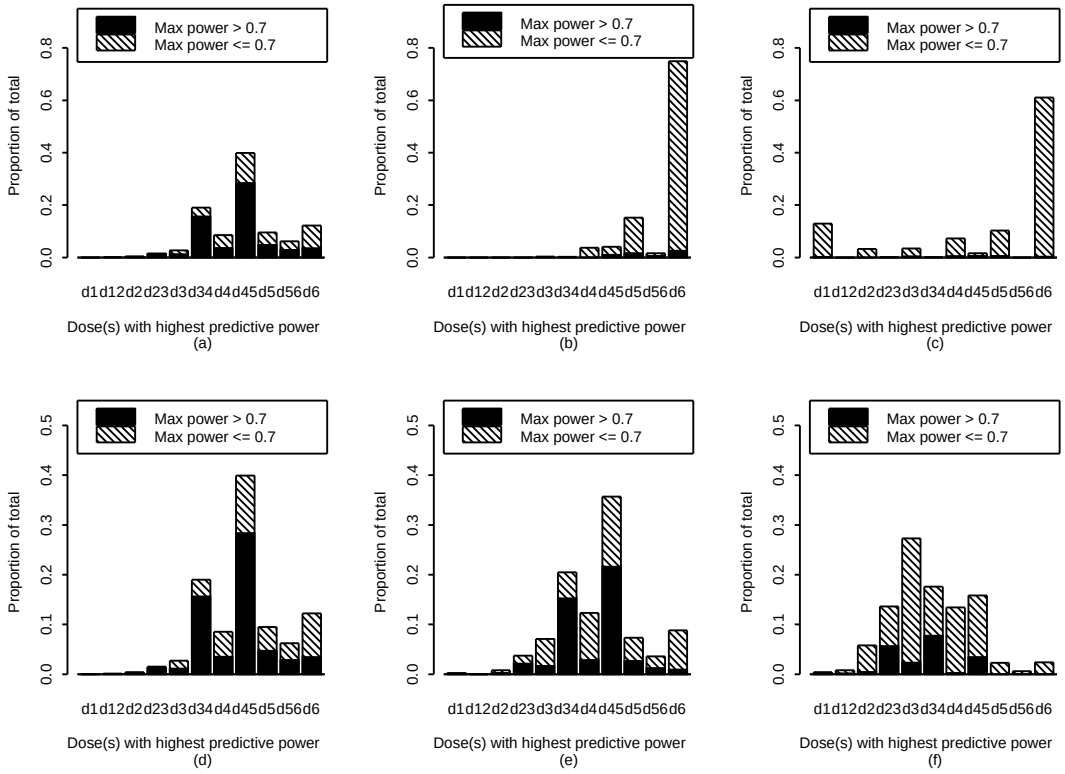

Figure 6.3: Histograms of set of doses with highest predictive power. Row 1 explores different scenarios for efficacy. In (a), only dose 1 is ineffective, in (b) only doses 5 and 6 are effective and in (c), all doses are ineffective. Row 2 explores different scenarios for toxicity. In (d), all doses are safe, in (e) dose 6 is toxic and in (f), doses 4 to 6 are toxic.

considered for testing at stage 2. The y-axis gives the proportion for which the doses on the x-axis are selected out of the total 1000 simulation studies. The bars have been partitioned into simulation studies whose maximum predictive power of potential doses to test in stage 2 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 because the probability of a successful trial, that is, a trial that is going to identify a safe and efficacious dose, is low.

Row 1 of Figure 6.3 corresponds to the results comparing different scenarios for efficacy (the dose-response curves for these scenarios were given in Figure 6.1, left panel). The same true dose-response curve for probabilities of toxicity (reference scenario toxicity dose-response curve) is used in these scenarios. As shown in Table 6.1 Row 3, the respective true probabilities of toxic outcomes for these scenarios at doses 1 to 6 are 0.10, 0.12, 0.14, 0.16, 0.18 and 0.19. In panel (a), the reference scenario for dose-response for efficacy is used. The respective true probabilities of efficacy at doses 1 to 6 for this scenario are given in Table 6.1 Row 4 and are 0.30, 0.37, 0.43, 0.51, 0.57 and 0.60. 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 doses 5 and 6 but also considerably safer than doses 5 and 6. Based on all simulation studies (shaded and striped parts), dose 5 or 6 is selected for testing at stage 2 with probability of about 0.6. Dose 4 or one of the higher doses is among the selected doses for testing at stage 2 with probability of over 0.9. When only the simulation studies whose predictive power greater than 0.7 are considered (607 studies out of 1000), dose 5 or 6 would be tested in stage 2 with probability above 0.65 and dose 4 or 5 or 6 would be tested in stage 2 with probability above 0.96. Panel (b) gives results for efficacy Scenario 2. The respective probabilities of efficacy for this scenario are given in Table 6.1 Row 5 and are 0.13, 0.18, 0.23, 0.30, 0.37 and 0.40 so that in comparison to the control treatment, dose 4 is not better, dose 5 and 6 are more efficacious while doses 1 to 3 are less efficacious. In this scenario, the desired dose for testing at stage 2 would be dose 6. In the simulation studies, this dose alone is selected with probability above 0.75. Doses 5 or 6, which are the only dose levels efficacious than the control, are selected for testing at stage 2 with probability of above 0.90. Sets which include only doses that are less efficacious than the control treatment, that is doses 1 to 3, are selected with probability of about 0.01. In panel (c), the probabilities of efficacy are the same among all the experimental doses and the control treatment. As would be desired the predictive power for almost all the simulated studies is less than or equal to 0.7. In this scenario dose 1 would be the

most desired because it is the least toxic. However dose 6 alone has the highest probability of being tested in stage 2. The selection procedure may be driven more by the prior distribution since the prior belief is that the efficacy improves with dose level. In the next section, more simulations based on the same scenario for dose response curves but different prior distributions support this view. With stronger prior belief, dose 1 is selected with less probability and when the prior belief is weak, dose 1 is selected with higher probability.

A distinctive difference between results in panel (a) and panels (b) and (c) is that in (a) continuing with more than one dose has higher probability than in (b) and (c). This may be explained by the true probabilities of efficacy of the experimental doses. In panels (b) and (c), the probability of efficacy is low and hence the (predictive) power of potential doses will be lower so that it would be preferable to allocate the available patients to the control and to only one dose of the new drug. Further comparing results in panel (a) to the results in panel (b), in panel (a), the probabilities of selecting the doses increase to selecting both dose 4 and 5 and then drops for continuing with dose 5 or 6 or both dose 5 and 6. On the other hand, in panel (b), the probability of selecting doses increase with dose level with dose 6 selected with the highest probability. This may be explained by the probabilities of efficacy and toxicity. In panel (a), the probabilities of efficacy are high so that the probability of obtaining a significant result in a trial that includes dose 4 and dose 5 is as high as a trial that includes either dose 5 only, dose 6 only or both dose 5 and dose 6. However, since testing both dose 4 and 5 involves testing safer doses, this set is selected with higher probability. In panel (b), the probabilities of efficacy are low so that the probability of obtaining a significant result in a trial with higher dose levels is higher so that dose 6 is selected with highest probability since it is also safe.

Row 2 of Figure 6.3 corresponds to the results comparing different scenarios for toxicity (the dose-response curves for these scenarios were given in Figure 6.1, right panel). As shown in Table 6.2 Row 3, for all these scenarios the respective true probabilities of efficacy at doses 1 to 6 are 0.30, 0.37, 0.43, 0.51, 0.57 and 0.60. Results in panel (d)

are the same as in panel (a) but on a different vertical scale to facilitate comparison of results in panels (e) and (f). Panel (e) gives results of toxicity Scenario 2 and as shown in Table 6.2 Row 5, the true probabilities of toxicity for this scenario at tested dose levels in increasing order 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, is selected for testing at stage 2 with probability less than 0.10. Dose 4 which would be the desired dose for testing in stage 2 is among the selected doses with probability of about 0.70. When only simulation studies with maximum predictive power greater than 0.7 are considered (483 studies out of 1000), dose 6 is selected with probability less than 0.02 while dose 4 is among the selected doses with probability above 0.80. Dose 5 which is nearly safe is selected with probability 0.05. Both dose 5 (nearly safe) and dose 6 (toxic) are selected for testing in stage 2 with probability 0.08. Panel (f) gives results for toxicity Scenario 3 and as shown in Table 6.2 Row 6, the true probabilities of toxicity for this scenario at the tested dose levels in increasing order are 0.10, 0.14, 0.18, 0.24, 0.30 and 0.33. The desired dose is dose 3. When only simulation studies with predictive power greater than 0.7 are considered, dose 3 is among the set selected for testing in stage 2 with probability 0.79. Dose 4 or both dose 4 and 5 which are all toxic are selected with a high probability of 0.18. We could not find an explanation to this high probability rather than chance.

Comparing the results in Row 2, we observe that the proportion of simulation studies with predictive power above 0.7 decreases from panel (d) to panel (f). This is because the probabilities of toxicity for doses 2 to 6 which are more efficacious than the control treatment increase from panel (d) to (f) so that from panel (d) to (f) lower doses which are less effective than the higher doses would be desired for testing in stage 2. The difference in panels (d) and (e) is particularly interesting. Although in both panels testing both dose 4 and dose 5 has the highest probability, testing either dose 5, dose 6 or both dose 5 and dose 6 is selected with lower probability in panel (e). When only studies with predictive

power above 0.7 are considered, the probability of continuing with either of these doses is even lower. Further we observe that dose 1 is selected with very low probability in the three scenarios so that even though higher doses are selected with lower probability when they are considered too toxic, the selection procedure still does not favour dose 1 which is not efficacious.

6.6 Comparing results for different prior distributions

The results discussed in the last section were obtained using the prior distributions presented in Figure 6.2. We refer to these sets of prior distributions as the middle weight prior belief. In order to assess the effect of prior distribution weight, we consider two more sets of prior distributions. In the second set of the prior distributions, Beta prior distributions at dose levels 10.50mg and 5000mg for the efficacy model have parameter vectors (18, 42) and (18, 7.71) respectively. For the toxicity model the Beta prior distributions have parameter vectors (9, 81) and (10.2, 30.6) at dose levels 10.50mg and 5000mg respectively. These beta distributions have the same prior means as the middle weight belief but smaller variance. From left, the first and the second contour plots in Figure 6.4 respectively give the resulting joint prior distribution of the slope and intercept for the efficacy and toxicity model. We refer to this set of prior belief as the most informative prior belief. The third set of the prior distribution is less informative and we refer to this set as the least informative prior belief. For the least informative prior belief, for both efficacy and toxicity the beta distributions at both dose 10.50mg and 5000mg are assigned parameter vector (1, 1) (that is Beta(1, 1) which is equivalent to the Uniform(0,1)). The resulting joint prior distribution for the intercept and the slope parameters is given by the contour in the right panel of Figure 6.4. Note that the scale for this contour plot is different from the scales of the other contour plots in Figure 6.4 and in Figure 6.2 with its contours spread wider along the intercept and slope parameters.

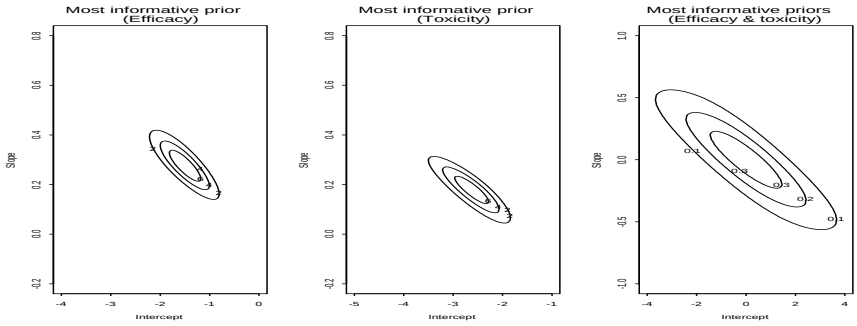

Figure 6.4: Contour plots for more informative and less informative prior densities.

We chose to demonstrate the effect of the prior distributions using the reference scenario (all doses safe, dose 1 as efficacious as control treatment and doses 2 to 6 more efficacious than the control) and efficacy Scenario 2 (all doses safe and equally as efficacious as the control treatment). Figure 6.5 gives the results using the different sets of prior distributions. Row 1 and Row 2 respectively correspond to the reference scenario and efficacy Scenario 2. Columns 1 to 3 respectively correspond to the most informative, middle weight and least informative prior distributions. For the reference scenario (Row 1), using the three sets of the prior distribution, probability of testing both dose 4 and dose 5 at stage 2 is highest. However the relative frequency decreases as the prior distributions become less informative. As the prior belief becomes less informative lower, higher doses are selected with higher probability. For example, the frequency of dose 6 increases as the prior distributions become less informative. The frequency, however, has a larger contribution from the simulation studies whose predictive power is less than 0.7 (striped parts).

For efficacy Scenario 2 (Row 2), when the prior distributions have higher weight (panel (d)), higher doses are selected for testing at stage 2 with higher probability with dose 6 selected with the highest probability. This is driven by the prior distribution in which higher doses are considered more efficacious compared to the lower doses. In panel (e), the prior distributions have lesser weight compared to panel (d) but the prior belief

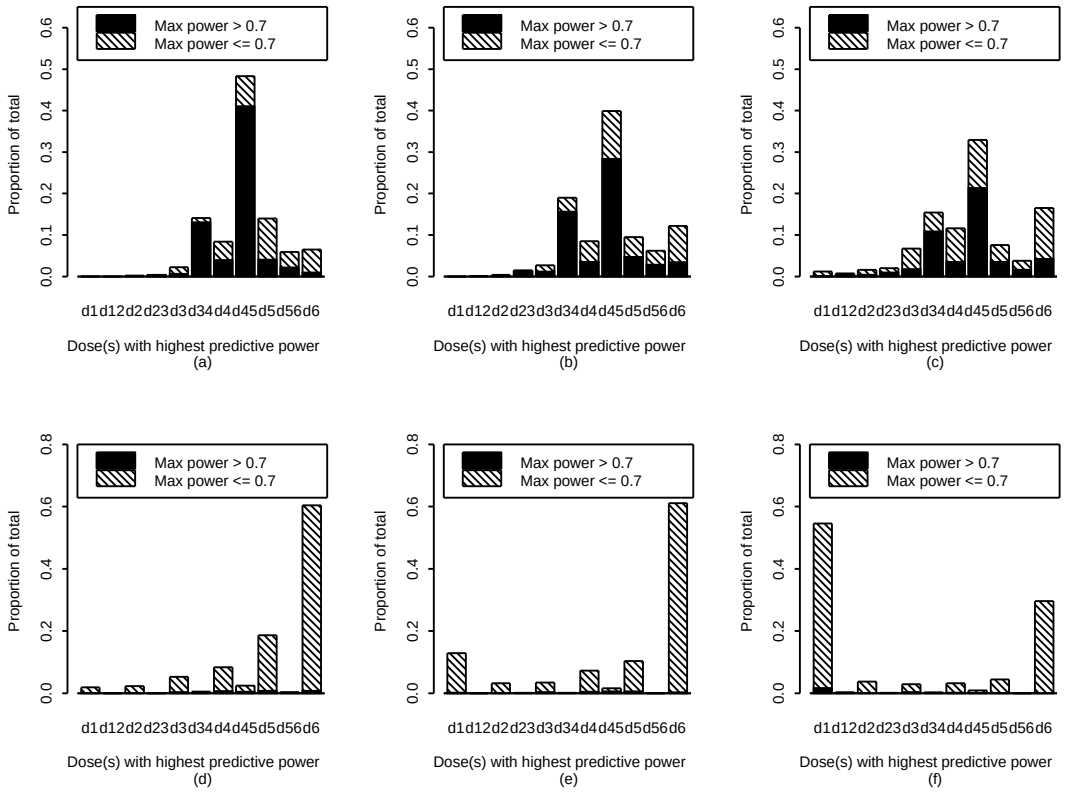

Figure 6.5: Histograms of set of doses with highest predictive power. From left to right the prior beliefs are less informative.

is still that higher doses are more efficacious so that dose 6 is selected with the highest probability. However, in panel (e), dose 1 is selected with higher probability compared to panel (d) as would be expected. In panel (f), the prior distributions used in obtaining the predictive power, assumed there is no difference between lower and higher doses. Further, the prior distribution had least weight with beta parameters having values (1,1) so that the selection probability compared to results in panels (e) and (f), puts more weight on stage 1 data. As expected, dose 1 is selected for testing at stage 2 with the highest probability because it is the safest and is not inferior compared to doses 2 to 6. Dose 6 has the second highest probability of selection. This may be explained by the stage 1 efficacy data. The

stage 1 data may suggest either a positive or a negative slope for probability of efficacy. A positive slope will favour dose 6 resulting into dose 6 selected more compared to doses 2 to 5.

6.7 Summary findings from the simulation results

In the previous sections of this chapter, we have presented simulation results under several scenarios. In this section, we summarize the key characteristics from these results while also pointing the advantage of the dose selection procedure as exhibited by the results. First, we observed that, as desired, when the second stage is not adequately powered in terms of probabilities of efficacy, the candidate sets with a single treatment are selected for testing at stage 2 with higher probability compared to candidate sets with two experimental doses. This is the case of the scenario whose results are given in Figure 6.3 (b). However, under scenarios where some experimental doses are highly efficacious and safe, such as the scenario whose results are given in Figure 6.3 (d), it is often better to test two experimental doses at stage 2. Thus the new dose selection is advantageous than the methods that allow only to proceed with one experimental dose.

Secondly, when all the experimental doses are acceptably safe but toxicity increases with dose level and the second stage sample size is such that it powers (based on probabilities of efficacy) the lower doses enough so that the difference in power of highest dose levels and some lower dose levels is not large, lower dose levels are selected with higher probability because they are safer to administer to patients. The results given in Figure 6.3 (d) are from such a scenario. Also, the inclusion of safety in the dose selection procedure avoids selecting unsafe doses although they may be more efficacious. For example, for the results given in Figure 6.3 (e), dose 6 is unsafe and thus compared to results in Figure 6.3 (d) in which all experimental doses are safe, it is selected with lower probability and when it is selected, the predictive power is low as in most cases the predictive is less than

0.70. The same trend, as shown in Figure 6.3 (f), is observed when doses 4 to 6 are unsafe. Thus new dose selection procedure is advantageous than the previous procedures that do not explicitly include safety.

The results for all scenarios captured the dose-response relationships. For example, in Figure 6.3 (b), the probability of selection increases with dose. The same is observed from Figure 6.3 (f) where the probability of selection increases with dose and then decreases with dose when doses 4 or higher, which are unsafe and which become less safe with increase in dose, are selected for testing in stage 2. The use of dose-response also captures the probabilities of efficacy and toxicity well. For example, in Figure 6.3 (d), there is huge increase in probability of selection from when only dose 3 is selected to when both doses 3 and 4 are selected. Probability of efficacy in control 0.3 while probabilities of efficacy at doses 3 and 4 are respectively 0.37 and 0.43 so that it may be that it is only when dose 4 or higher doses are selected that the probability of concluding efficacy is high.

Finally, the results for a scenario where the experimental doses of the new drug are not better than the control treatment, that is for the scenario whose results are given in 6.3 (c), the selected doses to test in stage follow the profile of the prior distribution but the predictive power is low so that the trial is unlikely to proceed to stage 2.

Chapter 7

Further work

In Chapter 5, we have developed a new method for selecting doses that continue from a phase II stage to a phase III stage of a seamless phase II/III clinical trial. The work focussed on binary outcomes at both stages, assumes the data are generated using some known generalized linear model and that there are only two stages. These features lead to three new research questions that can extend the work described in Chapter 5: (1) can the work be extended to include more than 2 stages with monitoring boundaries, (2) can the dose selection procedure be extended to include model uncertainty and can the outcomes be modelled using models that are not in the family of generalized linear models, and (3) can the work be extended to include other endpoints or a change of endpoint. In this chapter, we describe the direction we are taking to answer these questions. For the first question, we will describe ongoing recent work. For the second and third questions, we will summarise some works that have answered similar questions in different contexts that may be relevant.

7.1 Extending to more than two stages with monitoring

In Chapter 5, we assumed that there is no early stopping at end of phase II stage and that in the phase III stage, there will be no monitoring. However, as described in Sections 3.3.1 and 3.4.2, the investigators may wish stop the trial after the phase II stage and to also monitor the phase III data for among other reasons, ethical considerations. In this section, we give the progress made on exploring how a seamless phase II/III clinical trials with a single monitoring in the phase III stage so that the seamless phase II/III clinical trial will have 3 stages; the phase II stage and two phase III stages, may be planned. We will assume that there are no opportunities for adaptation in the phase III stage and we present the simple case of proceeding with one experimental dose into the phase III stage. In Section 3.3.1, we explained that clinical trials may stopped early either for overwhelming evidence of efficacy, futility or both. In Section 7.1.1, we will describe an example of how p-values using the combination tests may be adjusted when there are opportunities to stop early for overwhelming evidence of efficacy or for futility. We will set the notation of the seamless phase II/III clinical trial of interest in Section 7.1.2. In Section 7.1.3, given the stage 1 data, we will describe how to obtain the expression for the conditional power, the probability of concluding at least one of the experimental doses that continue to the phase III stage is effective after stage 2 or after stage 3. Finally, in Section 7.1.4, we define the predictive power for a trial with opportunities to stop early.

7.1.1 Combined p-value with opportunity to stop early

Let H be a null hypothesis of interest that is tested at two stages. Let p_1 and p_2 respectively denote stage 1 and stage 2 p-values. Suppose that at stage 1 the null hypothesis H will be rejected if $p_1 \leq \alpha_1$ and is accepted if $p_1 > \alpha_0$, where $0 \leq \alpha_1 < \alpha < \alpha_0 \leq 1$. Assuming p_1 and p_2 are independent and uniform[0,1] under H, in Section 4.2.2, we gave the expression

for type I error rate as

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{[C(p_1,p_2)\leq c]} dp_2 dp_1, \tag{7.1}$$

where $C(p_1, p_2)$ is some combination function of the p-values such as the Fisher's combination function given by equation (4.2) and the inverse normal function given by equation (4.3), c is the stage 2 critical value obtained by solving expression (7.1) for overall type I error rate α, and $\mathbf{1}_{[C(p_1,p_2)\leq c]}$ equals 1 if $C(p_1, p_2) \leq c$ and equals 0 otherwise. In the previous chapters, we assumed that we do not have opportunities to stop early so that expression (7.1) simplifies to

$$\int_0^1 \int_0^1 \mathbf{1}_{[C(p_1,p_2)\leq c]} dp_2 dp_1. \tag{7.2}$$

In this subsection, we consider seamless phase II/III clinical trials with opportunities to stop early for overwhelming evidence or futility. The values of α_1 and α_0 can be determined using the group sequential methods (Brannath et al., 2002) such as the alpha spending function of Lan and DeMets (1983).

Let $q(p_1, p_2)$ denote the adjusted combined p-value. For expression (7.2), since there are no opportunities to stop early, $q(p_1, p_2) = C(p_1, p_2)$. To control type I error rate defined by expression (7.1), Brannath et al. (2002) propose adjusting the combined p-values as follows

$$q(p_1, p_2) = \begin{cases} p_1 & \text{if } p_1 \leq \alpha_1 \text{ or } p_1 > \alpha_0 \\ \alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{[C(x,y)\leq C(p_1,p_2)]} dy dx & \text{otherwise.} \end{cases} \tag{7.3}$$

For the Fisher's combination function given by expression (4.2), solving expression (7.3) gives the combined p-value as

$$q(p_1, p_2) = \begin{cases} p_1 & \text{if } p_1 \leq \alpha_1 \text{ or } p_1 > \alpha_0 \\ \alpha_1 + p_1.p_2[\ln \alpha_0 - \ln \alpha_1] & \text{if } p_1 \in (\alpha_1, \alpha_0] \text{ and } p_1.p_2 \leq \alpha_1 \\ p_1.p_2 + p_1.p_2[\ln \alpha_0 - \ln(p_1.p_2)] & \text{if } p_1 \in (\alpha_1, \alpha_0] \text{ and } p_1.p_2 \geq \alpha_1. \end{cases} \tag{7.4}$$

For the normal combination method, there is no closed solution to expression (7.3) so that numerical integration methods are required.

The type I error rate expression (7.1) was obtained by evaluating the following definition of type I error for a two stage trial

$$\text{Prob}_H[p_1 \leq \alpha_1] + \text{Prob}_H[C(p_1, p_2) \leq c,\ \alpha_1 < p_1 \leq \alpha_0].$$

Extending this to a trial with three stages (3 stage trial), let α_{s0} and α_{s1} ($s = 1, 2$) respectively denote the futility boundary and rejection boundary at stage s so that the trial stops at stage s for futility if the adjusted p-value is greater than α_{s0} and for efficacy if the adjusted combined p-value is less or equal to α_{s1}. Let p_s ($s = 1, 2, 3$) denote the stage s p-value, then for a 3 stage trial, the type I error is given by

$$\begin{aligned}&\text{Prob}_H[p_1 \leq \alpha_{11}] + \text{Prob}_H[C(p_1, p_2) \leq c_{\alpha_{21}},\ \alpha_{11} < p_1 \leq \alpha_{10}] + \\ &\text{Prob}_H[c(p_1, p_2, p_3) \leq c, \alpha_{11} < p_1 \leq \alpha_{10}, c_{\alpha_{21}} < C(p_1, p_2) \leq c_{\alpha_{20}}],\end{aligned}$$

where $c_{\alpha_{21}}$ and $c_{\alpha_{20}}$ are respectively the critical values for efficacy and futility. Following Brannath et al. (2002), the combined p-value for a 3 stage seamless phase II/III clinical trial may be given by

$$q(p_1, p_2, p_3) = \begin{cases} p_1, & \text{if trial stops at stage 1} \\ \alpha_{11} + \int_{\alpha_{11}}^{\alpha_{10}} \int_0^1 \mathbf{1}_{[C(x,y) \leq C(p_1,p_2)]} dy dx, & \text{if trial stops at stage 2} \\ \alpha_{21} + \int_{\alpha_{11}}^{\alpha_{10}} \int_0^1 \int_0^1 \mathbf{1}_{[c_{\alpha_{21}} < C(x,y) \leq c_{\alpha_{20}}]} \mathbf{1}_{[C(x,y,z) \leq C(p_1,p_2,p_3)]} dz dy dx, & \text{otherwise.} \end{cases} \quad (7.5)$$

In equation (7.5), the trial stops in stage 1 if $p_1 \leq \alpha_{11}$ or $p_1 > \alpha_{10}$ and in stage 2 if $C(p_1, p_2) \leq c_{\alpha_{21}}$ or $C(p_1, p_2) > c_{\alpha_{20}}$. In general, numerical methods may be required to evaluate equation (7.5).

7.1.2 Notation and setting of interest

As in Chapter 5, suppose k_1 experimental doses are compared to the control treatment at stage 1 so that the null hypotheses of interest are $H_1 : \theta_0 = \theta_1,\ ...,\ H_{k_1} : \theta_0 =$

θ_{k_1} where θ_j, $j \in \{0, 1, ..., k_1\}$ is a measure of the effectiveness of dose j, with $j = 0$ corresponding to the control treatment. If the closure principle is to be used to control the FWER associated with comparing the k_1 experimental doses to the control treatment, all the intersection hypotheses H_J, $J \subseteq \{1, ..., k_1\}$ are constructed. We assume n_1 patients are allocated to each dose j, $j \in \{0, 1, ..., k_1\}$ at stage 1 and we respectively denote the number of successes and toxicities in dose j after stage 1 by x_{1j} and t_{1j}. The probabilities of efficacy and toxicity at dose j will respectively be denoted by p_{E_j} and p_{T_j}. The efficacy data $\mathbf{x}_1 = (x_{10}, x_{11}, ..., x_{1k_1})$ can be used to obtain the stage 1 p-values $p_{1,J}$ ($J \subseteq \{1, ..., k_1\}$ corresponding to the constructed intersection hypotheses H_J.

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 $k_2 = |\mathcal{K}_2|$. We assume that there is no adaptation after stage 2 so that if the trial does not stop after stage 2, all the experimental doses tested at stage 2 and control continue to stage 3. Let n_2 and n_3 respectively denote the total sample sizes at stage 2 and stage 3. We assume that the total stage 2 and stage 3 sample sizes are fixed so that the number of patients allocated to dose j, $j \in \{0\} \cup \mathcal{K}_2$ with $j = 0$ corresponding to the control treatment at stage s ($s = 2, 3$) is n_s/k_2. At stage s ($s = 2, 3$), let x_{sj} and t_{sj}, $j \in \{0\} \cup \mathcal{K}_2$ with $j = 0$ corresponding to the control treatment, respectively denote the number of successes and toxicities on dose j. At the end of stage s ($s = 2, 3$), the efficacy data $\mathbf{x}_s = (\{x_{sj}\})$ ($j \in \{0\} \cup \mathcal{K}_2$) can be used to construct a set of p-values $p_{s,J}$ corresponding to stage 1 p-values $p_{1,J}$ constructed using the stage 1 data.

7.1.3 Conditional power

From the setting above, $p_{s,J}$ denotes stage s ($s = 1, 2, 3$) p-value obtained from testing hypothesis H_J, $J \subseteq \{1, ..., k_1\}$. Given the stage 1 p-value $p_{1,J}$ for hypothesis H_J, as in inequality (5.3), using the adjusted p-value $q(p_{1,J}, p_{2,J}, p_{3,J})$ given by equation (7.5), we can obtain the minimum value of stage 2 p-value $p_{2,J}$ so that hypothesis H_J is rejected at

the end of stage 2. If a Bonferroni adjustment is used to obtain the p-value for hypothesis H_J, the minimum p-value testing the pairwise hypothesis comparing the control treatment to the experimental doses contained in the hypothesis H_J can be obtained as described for inequality (5.5). Subsequently, as was described in detail in Section 5.2.2, for each possible number of successes in the control arm at stage 2, x_{20}, we can obtain the configurations of data such that at least one experimental dose is concluded effective after stage 2.

Suppose we continue to stage 2 with a single experimental dose i, $i \in \{1, ..., k_1\}$. In Section 5.2.2, for each possible x_{20}, we denoted the minimum number of successes required at experimental dose i at stage 2, x_{2i} so that dose i is concluded effective at stage 2 by $B_{x_{20}}(\max\{p_{1,J}\})$ for all J with $i \in J$. For a 3 stage seamless phase II/III clinical trial, for each x_{20}, we are also interested in determining the minimum number of successes required in dose i so that the trial does not stop at stage 2 with acceptance of the null hypothesis of no treatment difference. This number is obtained similarly to $B_{x_{20}}(\max\{p_{1,J}\})$. To differentiate between the notation for the two numbers, we denote the minimum number of successes required to stop the trial at stage 2 for efficacy by $B^{\mathcal{R}}_{x_{20}}(\max\{p_{1,J}\})$ and the minimum number of successes required to avoid stopping the trial at stage 2 for acceptance of the null hypothesis by $B^{\mathcal{A}}_{x_{20}}(\max\{p_{1,J}\})$. Thus the expression for the probability of concluding dose i ($i \in \{1, ..., , k_1\}$) is more effective than the control treatment at stage 2 is

$$\sum_{x_{20}=0}^{n_2} \left\{ f_B(x_{20}; n_2, p_{E_0}) \sum_{x_{2i}=B}^{n_2} f_B(x_{2i}; n_2, p_{E_i}) \right\}, \tag{7.6}$$

where $f_B(x_{2j}; n_2, p_{E_j})$ ($j = 0, i$) is the probability mass function of the binomial random variable X_{2j} with parameter vector (n_2, p_{E_j}), $B = B^{\mathcal{R}}_{x_{20}}(\max\{p_{1,J}\})$ for all $J \subseteq \{1, ..., k_1\}$ with $i \in J$.

If the trial does not stop at stage 2, for all possible data in stage 2 for which we do not stop at stage 2 and for each possible number of successes with control treatment at stage 3, x_{30} we can obtain the minimum number of successes required in dose i at stage 3

so that dose i is concluded effective after stage 3. Let us denote this minimum number of successes by $B_{x30}(\mathbf{x}_2, \max\{p_{1,J}\})$, where notation reflects number of successes in control treatment x_{30}, the stage 2 data $\mathbf{x}_2$ and stage 1 data $\max\{p_{1,J}\}$. Then the probability that dose i is concluded not futile after stage 2 and effective after stage 3 is given by

$$\sum_{x_{20}=0}^{n_2}\left\{f_B(x_{20};n_2,p_{E_0})\sum_{x_{2i}=A}^{B-1}\left(f_B(x_{2i};n_2,p_{E_i})\sum_{x_{30}=0}^{n_3}\left[f_B(x_{30};n_3,p_{E_0})\sum_{x_{3i}=C}^{n_3}f_B(x_{3i};n_3,p_{E_i})\right]\right)\right\}, \quad (7.7)$$

where $f_B(x_{2j};n_2,p_{E_j})$ $(j=0,i)$ is as defined above, $f_B(x_{3j};n_3,p_{E_j})$ $(j=0,i)$ is the probability mass function of the binomial random variable X_{3j} with parameter vector (n_3,p_{E_j}), $A = B^{\mathcal{A}}_{x_{20}}(\max\{p_{1,J}\})$, $B = B^{\mathcal{R}}_{x_{20}}(\max\{p_{1,J}\})$ and $C = B_{x_{30}}(\mathbf{x}_2,\max\{p_{1,J}\})$. Expression (7.7) contains many summations which would make its evaluation computationally expensive so that some approximation may be required. As an alternative, we will explore whether it is easier to use the efficient score statistics described in Section 4.2.1.

The conditional power, the expression for probability of concluding dose i is effective at stage 2 or at stage 3 given stage 1 data is given by summing expressions (7.6) and (7.7). The probability of concluding dose i is effective and safe is obtained by multiplying expressions (7.6) and (7.7) by the indicator that dose i is safe. For example, if an experimental dose is considered safe if its probability of toxicity is less than or equal to some value γ, expressions (7.6) and (7.7) are multiplied by the indicator $I(p_{T_i} \leq \gamma)$. We will denote the conditional power by CP_θ where θ is a vector of parameters in the conditional power.

7.1.4 Predictive power

The probabilities of efficacy and toxicity, p_{E_j} and p_{T_j} that enter the conditional power are respectively given by the dose-response models (5.1) and (5.2) so that $\theta = (\alpha_E, \beta_E, \alpha_T, \beta_T)'$. As in Section 5.3, the predictive power may be obtained by evaluating the posterior mean

of the conditional power, that is, the predictive power is given by

$$\int_{\Theta} (\text{CP}_{\theta}) \pi(\theta | \mathbf{x}_1, \mathbf{t}_1, n_1) d\theta, \tag{7.8}$$

where $\pi(\theta|\mathbf{x}_1, \mathbf{t}_1, n_1)$ is the posterior mean of θ given the stage 1 data $(\mathbf{x}_1, \mathbf{t}_1, n_1)$ obtained as described in Section 5.3.1. To improve the predictive power given by expression (7.8), we will investigate how feasible it is to update the distribution of θ with second stage data $(\mathbf{x}_2, \mathbf{t}_2)$ so that the parameter values that enter part of the conditional power that includes stage 3 data reflect the knowledge gained in stage 2. This is unlikely to be an easy task but has the advantage of reducing the uncertainty (variance) of the parameter values.

7.2 Uncertainty in the dose-response curves

In Chapter 5, we assumed that the probabilities of efficacy and of toxicity can be modelled by generalized linear models of known form. However, the investigators may be uncertain about the dose-response curve so that they would want to consider several models. Klingenberg (2009) has proposed a method for estimating the maximum estimated dose (MED) that incorporates model uncertainty. He proposes including all dose-response curves that significantly reflect the dose-response signal from the data to estimate the MED by getting the weighted average of the MEDs from these dose-response models. In future work, we intend to consider including model uncertainty by obtaining the predictive power possibly by averaging the predictive power obtained using different models similar to the proposal of Klingenberg (2009) of estimating the MED.

Klingenberg (2009) gives several dose-response curves in the generalized linear model family that result into different dose-response shapes such as probabilities increasing and then dropping. However, these models may not capture the dose-response profile adequately so that further work needs to develop models that are outside the generalized linear models family. Yin et al. (2006) have proposed models that do not specify any parametric

form for the dose-response curve. Rather they model probabilities so that it is possible to induce some relationship such as probability of toxicity increasing with dose level. In particular, they assume that the probability of toxicity increases with dose because most investigators assume a monotonically increasing relationship between toxicity and dose. They do not enforce any ordering for the probability of efficacy since for certain therapy, probability of efficacy may decrease with dose.

In detail, let p_{T_j} be the probability of toxicity at experimental dose j ($j = 1, ..., k_1$), Yin et al. (2006) model the probabilities of toxicity as follows

$$\phi_j = \begin{cases} \log \frac{p_{T_j}}{1-p_{T_j}} & \text{if} \quad j = 1 \\ \log\left(\frac{p_{T_j}}{1-p_{T_j}} - \frac{p_{T_{j-1}}}{1-p_{T_{j-1}}}\right) & \text{for} \quad j = 2, ..., k_1 \end{cases}$$

so that

$$p_{T_j} = \begin{cases} \frac{\exp(\phi_j)}{1+\exp(\phi_j)} & \text{if} \quad j = 1 \\ \frac{\exp(\phi_1)+...+\exp(\phi_j)}{1+\exp(\phi_1)+...+\exp(\phi_j)} & \text{for} \quad j = 2, ..., k_1. \end{cases} \tag{7.9}$$

Modelling probability of toxicity using equation (7.9) ensures that

$$p_{j-1} < p_j, \quad j = 2, ..., k_1,$$

that is, probability of toxicity increases with dose as required because the terms $\exp(\phi_j)$ are positive. To define the model of efficacy, let p_{E_j} be the probability of efficacy at experimental dose j ($j = 1, ..., k_1$). Yin et al. (2006) model the probabilities of efficacy as follows

$$\psi_j = \begin{cases} \log \frac{p_{E_j}}{1-p_{E_j}} & \text{if} \quad j = 1 \\ \log\left(\frac{p_{E_j}}{1-p_{E_j}}\right) - \log\left(\frac{p_{E_{j-1}}}{1-p_{E_{j-1}}}\right) & \text{for} \quad j = 2, ..., k_1 \end{cases}$$

so that

$$p_{E_j} = \begin{cases} \frac{\exp(\psi_1)}{1+\exp(\psi_1)} & \text{if} \quad j = 1 \\ \frac{\exp(\psi_1+...+\psi_j)}{1+\exp(\psi_1+...+\psi_j)} & \text{for} \quad j = 2, ..., k_1. \end{cases} \tag{7.10}$$

Expression (7.10) does not impose any order for the probability of efficacy. For further work on model uncertainty, we will study characteristics of fitting models given by expressions (7.9) and (7.10) and consider including these models in the candidate set of models so as to capture dose-response relationships not described by the generalized linear models.

7.3 Change of endpoints

The dose selection procedure developed in Chapter 5 focussed on binary outcomes both at phase II and phase III stage. Posch et al. (2005) present such a trial where in both stage 1 and stage 2 of a seamless phase II/III clinical trial, binary outcomes are considered. However, for some therapies as Inoue et al. (2002) explain, the primary outcome in the phase II study is a binary outcome which is used to plan a phase III study whose primary outcome is a time to an event outcome. Hence, it would be practically important to design seamless phase II/III clinical trials with change of primary endpoint from a binary endpoint to a survival outcome.

Inoue et al. (2002) propose a fully Bayesian seamless phase II/III clinical trial by using Bayesian tools to plan a trial and make inference from the observed data. Here we just describe how the binary outcomes are included in the inference. Let Y and T respectively denote the binary outcome and the survival time. Inoue et al. (2002) assume there will be censoring after time U and that the binary outcome is observed after time $t_0 < U$. If the survival time for a patient is less than t_0, Y is not observed, that is, Y is observed if $T^* \geq t_0$, where $T^* = \min(T, U)$. Using the law of total probability, they write the distribution of T as follows

$$f(t) = f(t|T < t_0)\text{Prob}(T < t_0) + f(t|T \geq t_0)\text{Prob}(T \geq t_0). \tag{7.11}$$

Since Y is observed for $T \geq t_0$, Inoue et al. (2002) use it to learn about the distribution of T in the second part of equation (7.11). Thus they incorporate Y in the second piece of the

distribution of T, that is in the second part of RHS of equation (7.11), using the law of total probability to result in the following mixture distribution

$$f(t) = f(t|T < t_0)\text{Prob}(T < t_0) + \sum_{y=0}^{1} f(t|T \geq t_0, Y = y)\pi_y \text{Prob}(T \geq t_0), \quad (7.12)$$

where $\pi_y = \text{Prob}(Y = y|T \geq t_0)$. Let W be the indicator that Y is observed or equivalently the indicator $T \geq t_0$, Inoue et al. (2002) assume $T = T_0(1 - W) + (T_1 + t_0)W$, where T_0 and T_1 are latent survival times with T_0 following a distribution f_0 not depending on Y and T_1 following the mixture distribution $\sum_{y=0}^{1} f_y \pi_y$. Hence expression (7.12) may be written as

$$\begin{aligned} f(t) &= \{f_0(t)\}^{1-W} \left\{ \mathcal{F}_0(t_0) \sum_{y=0}^{1} f_y(t - t_0)\pi_y \right\}^W \\ &= \begin{cases} f_0(t) & \text{if } t < t_0 \\ \mathcal{F}_0(t_0) \sum_{y=0}^{1} f_y(t - t_0)\pi_y & \text{if } t \geq t_0, \end{cases} \end{aligned}$$

where $\mathcal{F}_0(t) = \text{Prob}(T_0 > t)$ is the survival function corresponding to f_0. The inference is then made using the distribution of T defined above.

Schmidli et al. (2007) do not consider change of endpoint but they propose a seamless phase II/III design that utilize binary and survival outcomes techniques. Schmidli et al. (2007) consider survival outcomes with right censoring. At both stage 1 and stage 2, survival or censoring time and the binary outcome that denotes whether the outcome of interest (the outcome that gives survival time) are recorded. Suppose the right censoring time is U. The analysis is based on the binary outcomes, that is, comparing the number of events that occur within the censoring time U. To determine the distribution of stage 2 data, the probability of an event of interest is considered Bernoulli$(1 - S(U; \theta_j))$, where θ_j are the parameters of the survival function $S(t)$ which enter the stage 2 data distribution when treatment (dose) j is considered for testing at stage 2. At the time of planning the phase III stage, some of the phase II stage patients will not have an event and will not have

reached the censoring time U. If the phase III stage planning is done at time $t_0 < U$ for a patient, the probability of an event for this patient in the remaining follow-up time is $(1 - S(U; \theta_j)) - (1 - S(t_0; \theta_j)) = S(t_0; \theta_j) - S(U; \theta_j)$.

Both the work of Schmidli et al. (2007) and Inoue et al. (2002) consider survival outcome up to some censoring time U. Schmidli et al. (2007) do not incorporate the survival time in the analysis but incorporate it the planning by assuming that the probability of a survival event between time 0 and censoring time U depends on some survival functions. Inoue et al. (2002) use the survival outcome to make inference but the distribution of the survival time T is a mixture of distributions for different outcomes of the binary outcomes. The two approaches underline the complexity of having two endpoints. However, the ideas in these articles in which the authors use one outcome to determine another outcome and the technique of surrogate endpoints which we have not yet studied, may offer a starting point for further work on designing seamless phase II/III clinical trials with change of endpoints.

Chapter 8

Discussion and conclusions

The work in this thesis was based on seamless phase II/III clinical trials. Seamless phase II/III clinical trials are carried out in two stages; the phase II stage (stage 1) and the phase III stage (stage 2). After collecting stage 1 data, an interim analysis is done so that there are opportunities to adapt the trial based on stage 1 results. Possible adaptations in clinical trials are sample size re-estimation for example as described by Friede and Kieser (2006), sub-population selection for example as proposed by Zuber et al. (2006) and treatment selection for example as proposed by Thall et al. (1988), Schaid et al. (1990), Stallard and Todd (2003), and Schmidli et al. (2007). Our work also focussed on treatment selection in seamless phase II/III clinical trials but include features that are not included in the above works on treatment selection.

Most work for treatment selection in phase II/III trials consider candidate treatments which are distinct treatments. In this thesis, we have considered candidate treatments which are different doses of the same drug. To incorporate this, we have proposed some dose-response curves to estimate the probabilities of efficacy and toxicity at the experimental doses to inform the planning of the phase III stage so that dose selection is based on data observed over the entire experimental dose range in the phase II stage. This is similar to

dose-finding phase I studies where it is common to allocate patients to the experimental doses assuming some dose-response curves. For example, O'Quigley et al. (1990) proposed an exponential curve for a safety model while Whitehead et al. (2006) have proposed logistic models for both safety and efficacy outcomes. The work in dose-finding phase I studies generally does not focus on hypotheses testing. However, in phase II/III clinical trials hypotheses testing is done and the dose selection procedure should be made such that the type I error is not inflated. The flexible two stage hypothesis tests we assume will be used to analyze data from the two stages of the seamless phase II/III trials allow the use of the dose-response curves and also the prior knowledge about dose-response curves without inflating type I error rates.

Both efficacy and toxicity have been considered explicitly in early clinical trials. For example Whitehead et al. (2006) have proposed a design applicable to phase I/II clinical trials. However, safety is often not explicitly included in the dose-selection procedure for doses to be tested in phase III. For example, at the planning stage, the dose selection procedure may determine the (predictive) probability that the candidate sets will be concluded effective after stage 2 and a separate decision is made on whether the promising doses are safe for further experimentation. The dose-selection procedure that we have proposed considers both the efficacy and the safety of potential doses explicitly. Rather than only focus on the probability that the dose will be concluded effective after phase III stage, the procedure uses the joint probability that the dose will be concluded effective and safe by not exceeding some threshold safety level.

The penalty for safety was considered based on the distribution of probability of toxicity rather than the distribution of the number of patients who would experience a toxic outcome 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 $(\gamma \times n_2)$, where γ is the maximum probability of toxicity that can be tolerated and n_2 is the number of patients randomized to each treatment arm in stage 2, then larger samples

will be penalized more when the true probability of toxicity is greater than γ and less when the true probability of toxicity is less than γ. The second reason is that, in practice, safety data are monitored as the trial continues so that the safety of the drug is evaluated before all patients have been treated.

In Chapter 6, to study the characteristics of the dose selection procedure, we restricted the doses to be considered for testing at phase III stage to consecutive doses. However, as described in Sections 5.2.2 and 6.1, this does not reflect a limitation to the dose selection procedure developed in Chapter 5. The restriction reduces the sets of doses to be considered for testing in the phase III stage and hence reduces the computation time. Also continuing with nonconsecutive doses seems practically implausible. However, in some scenarios it may be reasonable to consider nonconsecutive doses. To demonstrate when it may be desirable to consider proceeding with non-consecutive doses, consider three doses, say dose 1, dose 2 and dose 3. Suppose the efficacy dose response curve is such that the probabilities of efficacy at dose 1 and dose 2 differ very little and the probability of efficacy at dose 3 is considerably higher than at dose 2. Then if the safety dose response curve is such that the probabilities of toxicity at consecutive doses are considerably different, then it would be desirable to proceed to phase III stage with dose 1 and dose 3 rather than with doses 2 and 3.

We have assumed that efficacy and toxicity are independent given the dose level. In Section 5.6, we showed by using the conditional efficacy and toxicity models (5.1) and (5.2), we do not assume marginal independence and that the modelled association between probabilities of efficacy and toxicity are reasonable. Alternatively the association between efficacy and safety may be modelled explicitly by introducing a parameter for association. For example, Yin et al. (2006) include parameters for the odds ratio at each dose to model the association between efficacy and toxicity. This is likely to capture association better but would introduce complexity in obtaining the joint distribution of the parameters in the model and increase the computation time, and we do not think this would make much

difference on the choice of doses to continue to stage 2.

To summarise, in this thesis, we have proposed a new method for dose selection in seamless phase II/III clinical trials. The method enables rational choice of doses to continue to stage 2 while: (1) allowing for the final analysis, (2) incorporating the stage 1 data profile, and (3) incorporating the prior knowledge.

Bibliography

Agresti, A. (2002). *Categorical Data Analysis* (2 ed.). New Jersey: John Wiley & Sons, Inc.

Armitage, P. (1975). *Sequential Medical Trials* (2 ed.). Oxford: Blackwell Scientific Publications.

Babb, J., A. Rogatko, and S. Zacks (1998). Cancer phase I clinical trials: Efficient dose escalation with overdose control. *Statistics in Medicine* **17**, 1103–1120.

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

Bauer, P. and Köhne (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50**, 1029–1041.

Bedrick, E. J., R. Christensen, and W. Johnson (1996). A new perspective on priors for generalized linear models. *Journal of the American Statistical Association* **91**, 1450–1460.

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

Bretz, F., , F. König, A. Racine, and W. Maurer (2006). Confirmatory seamless phase

II/III clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal* **48**, 623–634.

Bretz, F., J. C. Pinheiro, and M. Branson (2005). Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics* **61**, 738–748.

Cleophas, T. J., A. H. Zwinderman, T. F. Cleophas, and E. P. Cleophas (2009). *Statistics Applied to Clinical Trials* (4 ed.). Springer.

Cook, T. D. and D. L. DeMets (2008). *Introduction to Statistical Methods for Clinical Trials* (1 ed.). USA: Chapman & Hall/CRC.

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

Durham, S. D., F. N., and W. F. Rosenberger (1997). A random walk rule for phase I clinical trials. *Biometrics* **53**, 745–760.

Fan, S. K. and Y.-G. Wang (2006). Decision-theoretic designs for dose-finding clinical trials with multiple outcomes. *Statistics in Medicine* **25**, 1699–1714.

Fisher, R. A. (1932). *Statistical Methods for Research Workers* (4 ed.). London: Oliver & Boyd.

Fisher, R. A. (1935). *The Design of Experiments* (1 ed.). Edinburgh and London: Oliver & Boyd.

Fisher, R. A. (1970). *Statistical Methods for Research Workers* (14 ed.). Edinburgh: Oliver & Boyd.

French, S. and D. R. Insua (2000). *Statistical Decision Theory* (1 ed.). London: Arnold.

Friede, T. and M. Kieser (2006). Sample size recalculation in internal pilot study designs: A review. *Biometrical Journal* **48**, 537–555.

Friedman, L. M., C. D. Furberg, and D. L. DeMets (1998). *Fundamentals of Clinical Trials* (3 ed.). Newyork: Springer-Verlag, Inc.

Gehan, E. A. (1961). The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic Diseases* **13**, 346–353.

Geller, N. L. and S. J. Pocock (1987). Interim analyses in randomized clinical trials: Ramifications and guidelines for practitioners. *Biometrics* **43**, 213–223.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004). *Bayesian Data Analysis* (2 ed.). USA: Chapman & Hall/CRC.

Hochberg, Y. and A. C. Tamhane (1987). *Multiple Comparison Methods* (1 ed.). New York: John Wiley & Sons Ltd.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.

Hommel, G. (2001). Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal* **43**, 581–589.

Inoue, L. Y. T., P. F. Thall, and D. A. Berry (2002). Seamlessly expanding a randomized phase `II` trial to phase `III`. *Biometrics* **58**, 823–831.

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

Jennison, C. and B. W. Turnbull (2000). *Group Sequential Methods with Applications to Clinical Trials* (1 ed.). USA: Chapman & Hall/CRC.

Johnson, R. A. and D. W. Wichern (2002). *Applied Multivariate Statistical Analysis* (5 ed.). New Jersey: Prentice Hall, Inc.

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

Klingenberg, B. (2009). Proof of concept and dose estimation with binary responses under model uncertainity. *Statistics in Medicine* **28**, 274–292.

Lan, K. K. G. and D. L. DeMets (1983). Discrete group sequential boundaries for clinical trials. *Biometrika* **70**, 659–663.

Lindley, D. V. (1985). *Making Decisions* (2 ed.). Britain: John Wiley & Sons.

Lindley, D. V. and L. D. Phillips (1976). Inference for a Bernoulli process (a Bayesian view). *The American Statistician* **30**, 112–119.

Loke, Y.-C., S.-B. Tan, Y. Cai, and D. Machin (2006). A Bayesian dose finding design for dual endpoint phase I trials. *Statistics in Medicine* **25**, 3–22.

Machin, D., M. J. Campbell, S. B. Tan, and S. H. Tan (2009). *Sample Size Tables for Clinical Studies* (3 ed.). UK: Wiley-Blackwell.

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

O'Brien, P. C. and T. R. Fleming (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549–556.

O'Quigley, J., M. Pepe, and L. Fisher (1990). Continual reassessment method: A practical design for phase I clinical trials in cancer. *Biometrics* **46**, 33–48.

O'Quigley, J. and L. Z. Shen (1996). Continual reassessment method: A likelihood approach. *Biometrics* **52**, 673–684.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191–199.

Pocock, S. J., N. L. Geller, and A. A. Tsiatis (1987). The analysis of multiple endpoints in clinical trials. *Biometrics* **43**, 487–498.

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

Roussas, G. (2007). *Introduction to Probability* (1 ed.). USA: Elsevier Academic Press.

Schaid, D. J., S. Wieand, and T. M. Therneau (1990). Optimal two-stage screening designs for survival comparison. *Biometrika* **77**, 507–513.

Schmidli, H., F. Bretz, and A. Racine-Poon (2007). Bayesian predictive power for interim adaptation in seamless phase II/III trials where the endpoint is survival up to some specified timepoint. *Statistics in Medicine* **26**, 4925–4938.

Schoenfeld, D. (1980). Statistical consideration for pilot studies. *International Journal of Radiation Oncology, Biology and Physics* **6**, 371–374.

Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* **10**, 1–10.

Stallard, N. (1998). Sample size determination for phase II clinical trials based on decision theory. *Biometrics* **54**, 279–294.

Stallard, N., P. F. Thall, and J. Whitehead (1999). Decision theoretic designs for phase II clinical trials with multipe outcomes. *Biometrics* **55**, 971–977.

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

Storer, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics* **45**, 925–937.

Thall, P. F. and R. Simon (1994). Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics* **50**, 337–349.

Thall, P. F., R. Simon, and S. S. Ellenberg (1988). Two-stage selection and testing designs for comparative clinical trials. *Biometrika* **75**, 303–310.

Tsutakawa, R. K. (1975). Bayesian inference for bioassay. *Mathematical Sciences, University of Missouri, Columbia*, Tech. Rep **52**.

Wang, D. and A. Bakhai (2006). *Clinical Trials: A Practical Guide to Design, Analysis, and Reporting* (1 ed.). UK: Remidica.

Westfall, P. H. and D. W. Wolfinger (2000). Closed multiple testing procedures and proc multtest. *http://ftp.sas.com/techsup/download/observations/obswww23/obswww23.pdf [23 Jan 2009]*.

Westfall, P. H. and S. S. Young (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*. New Jersey: John Wiley & Sons.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials* (2 ed.). England: John Wiley & Sons Ltd.

Whitehead, J. (2006). *Using Bayesian Decision Theory in Dose-escalation studies, Statistical Methods for Dose-Finding Experiments*. Chevret, S (1 ed.). England: John Wiley & Sons Ltd.

Whitehead, J., Y. Zhou, J. Stevens, G. Blakey, and J. Price (2006). Bayesian decision procedures for dose-escalation based on evidence of undesirable events and therapeutic benefit. *Statistics in Medicine* **25**, 37–53.

Yin, G., Y. Li, and Y. Ji (2006). Bayesian dose-finding in phase I/II clinical trials using toxicity and efficacy odds ratios. *Biometrics* **62**, 777–787.

Zaykin, D., L. A. Zhivotovsky, P. Westfall, and B. Weir (2002). Truncated product method for combining p-values. *Genetic Epidemiology* **22**, 170–185.

Zuber, E., W. Brannath, M. Branson, F. Bretz, P. Gallo, M. Posch, and A. Racine-Poon (2006). Phase II/III seamless adaptive designs with Bayesian decision tools for an efficient development of a targeted therapy in oncology. *http://tr.dac.univie.ac.at/tr200605.pdf [23 Jan 2009]*.