A STUDY OF STAGEWISE PHASE II AND PHASE II/III DESIGNS FOR CLINICAL TRIALS

By GAOHONG DONG

A Dissertation submitted to

The School of Public Health

University of Medicine and Density of New Jersey

and

The Graduate School – New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

UMDNJ – School of Public Health

Awarded jointly by these institutions

Written under the direction of

Professor Weichung Joe Shih, PhD

and approved by

____

____

____

____

New Brunswick, New Jersey

October 2010

ABSTRACT OF THE DISSERTATION

A STUDY OF STAGEWISAE PHASE II AND PHASE II/III DESIGNS FOR CLINCIAL TRIALS

By GAOHONG DONG

Dissertation Director: Professor Weichung Joe Shih

Clinical trials play vital roles in drug development. Traditionally, phase II and phase III studies are conducted separately. However, in the pharmaceutical industry there is a recent trend toward combining phase II and phase III in a seamless fashion (a so-called phase II/III clinical trial). To first understand traditional phase II clinical trial designs, we develop two two-stage single-arm phase II clinical trial designs: a Bayesian-frequentist design and a Bayes factor-based design. Both designs control frequentist Type I and Type II error rates. Then we develop a varying-stage adaptive phase II/III clinical trial design. In this design, in addition to traditional initial learning stage (phase II) and final confirmatory stage (phase III), we also consider whether there is a need to have an intermediate stage to obtain more data, so that a more informative decision can be made to advance the trial to the final confirmatory stage. With respect to adaptations, we consider dropping dose arm(s), changing the primary study endpoint, determining sample size, and early stopping for futility. We use an adaptive combination test to perform final statistical analyses. Under conditional distribution of p-values or combined p-values, we derive Type I error rate and statistical power for each decision path. By applying closed testing procedure, we control family-wise Type I error rate at nominal level of α in the strong sense.

Preface

This dissertation consists of the following two parts:

- Two-stage single-arm phase II clinical trial designs: a Bayesian-frequentist design and a Bayes factor-based design
- A varying-stage adaptive phase II/III clinical trial design

An overview of this dissertation by chapters is presented as below.

Chapter 1 describes the background for our proposed two-stage single-arm phrase II clinical trial designs and the varying-stage adaptive phase II/III clinical trial design. The objectives of this dissertation are also addressed.

Chapter 2 reviews the literature for existing two-stage single-arm phrase II clinical trial designs including frequentist designs, Byesian designs and Bayesian clinical trial monitoring.

Chapter 3 develops a Bayesian-frequentist two-stage single-arm phase II clinical trial design. This design allows both early acceptance and rejection of the null hypothesis. The frequentist setting is very similar to Fleming (1982), Chang et al (1987) and Shuster's design (2002). Equivalently, upper and lower boundaries are determined with predictive probability of trial success outcome. With respect to other Bayesian settings, given a beta prior and a sample size for stage I, based on marginal distribution of the responses at stage I, Bayesian Type I and Type II error rates are derived. Our design controls both frequentist and Bayesian error rates. The properties of this design are demonstrated with examples and comparisons with other frequentist and Bayesian designs.

Chapter 4 develops a Bayes factor-based two-stage single-arm phase II clinical trial design using an iMOM prior. The property of prior distribution mis-specification and the property of iMOM prior are discussed. This new design is also demonstrated with examples and comparisons with other frequentist and Bayesian designs.

Chapter 5 reviews literature for adaptive phase II/III clinical trial designs.

Chapter 6 introduces the main concept of our varying-stage adaptive phase II/III clinical trial design. Under this new design, an intermediate stage can be added if the data from the first stage are not sufficient to make informative decisions. By deriving the distributions of p-values or combined p-values conditional to a trial decision path, the method of the final analysis is proposed and type I error control is proved.

Chapter 7 discusses dose selection and multiple comparisons on the primary endpoint of our varying-stage adaptive phase II/III clinical trial design. In our design, the closed testing procedure (Marcus et al, 1976) is used to protect Type I error rate control.

Chapter 8 addresses statistical power and sample size determination of our varying-stage adaptive phase II/III clinical trial design. The statistical power for each decision path is derived based on the distribution of p-values under the alternative hypothesis (Hung, O'Neil, Bauer and Kohne, 1997). The sample size for the final stage is determined based on conditional power (Shih et al, 2004).

Chapter 9 demonstrates our varying-stage adaptive phase II/III clinical trial design with an illustration and simulations. The simulations are carried out for the two study endpoints that are assumed following normal distributions, thus, Dunnett test is used to perform many-to-one comparisons. However, our design and analysis approach are not limited to normally distributed study endpoints. In addition, a special case of our design is discussed.

Chapter 10 discusses practical issues and trial implementation for our design.

Acknowledgements

First, I would like to express my sincere appreciation to my dissertation advisor Dr. Shih. My special thanks to him for his guidance and encouragement in my doctoral research. Although I have been working in the pharmaceutical industry for a few years, the area of adaptive design is new to me. Thanks to Dr. Shih for inspiring me to study this interesting and cutting-edge field. My doctoral research under his direction is the solid foundation for my future career development.

I would like to thank my other dissertation committee members for their thoughtful comments and support in my doctoral research. Thanks to Dr. Quan for his critical thoughts and time in fully reviewing my dissertation within his busy schedule. I also greatly appreciate Dr. Moore for his direction to my early research on phase II clinical trial designs and his constructive comments on my dissertation. My gratitude is further extended to Dr. Marcella who helped me expand my thinking on applications of my dissertation work.

Thanks to my other UMDNJ faculty members for their teaching, encouragement and advice, and thanks to the department staff for their administrative support.

In addition, I would like to take this opportunity to express special thanks to my management at Novartis for their financial support during my Ph.D. study. Thanks for their understanding and support for my time off and flexible work schedule as needed for my course work and doctoral research. I also greatly appreciate my Novartis colleagues for their help and fruitful discussions.

Furthermore, I am grateful to many professionals outside of UMDNJ and Novartis for their insights to my questions during my doctoral research.

Last, but not least, thanks to my family and friends. Their support and encouragement made this sustained effort possible.

Table of contents

List of tables

List of figures

1 Introduction

Clinical trials play vital roles in drug development. Phase II studies are the basis for planning of phase III clinical trials. Traditionally, these two phases are conducted separately. However, in the pharmaceutical industry there is a recent trend towards combining phase II and phase III in a seamless fashion (a so-called phase II/III clinical trial). In order to develop a phase II/III clinical trial design, which is more complex, it is important to first understand traditional phase II clinical trial designs. In this dissertation, we discuss traditional phase II clinical trial designs first, then the newer phase II/III clinical trial design. The first part of this dissertation considers two two-stage single-arm phase II clinical trial designs (Bayesian-frequentist design and Bayes factor-based design), and the second part develops a varying-stage adaptive phase II/III clinical trial design.

1.1 Two-stage designs for single-arm phase II clinical trials

It is well known that frequently trial designers are uncertain about the initial estimates of variation, treatment effect, recruitment pattern, patient compliance, etc (Shih, 2001) since limited information regarding the new therapy is available at the time of clinical trial planning, particularly for phase I and phase II clinical trials. Under a Bayesian framework, uncertainty of initial estimate of clinical trial design parameters can be considered, and the outcome of a clinical trial at the end of the study can be predicted based on accumulated interim data. Due to these advantages, the Bayesian approach has gained popularity during the past decades. However, there are some obstacles of Bayesian approach being widely used including: (1) obtaining prior information for some situations; (2) less familiarity to investigators on Bayesian trial design and Bayesian data analysis (Tan, Machin, 2002); and (3) potential resistance by regulatory agencies (Berry, 2006).

In contrast, conventional clinical trial designs have their own advantages, such as Type I and Type II error rate control, familiarity to investigators, etc. In addition, preventing drug approval with false positive results is always a concern for regulators. For Bayesian medical device trials, the FDA requires evaluations of trial operating characteristics including frequentist Type I and Type II error rates (FDA, 2010).

Therefore, in this dissertation, we develop two new two-stage designs for single-arm phase II clinical trials. These two proposed new designs have both Bayesian and frequentist properties. In fact, Bayesian-frequentist approach has been used in various areas. For example, dose response trials (Chang and Chow, 2005), treatment selection strategy (Thall et al, 2007) and trial reproducibility and power evaluation (Shao, Mukhi and Goldberg, 2008).

Our first proposal is a strict Bayesian-frequentist two-stage design. The second design is a Bayes factor-based two-stage design using an inverse moment (iMOM) prior. The main feature of the second design is that a Bayes factor is used to derive posterior probabilities, which are used for constructing the stopping rules. By using a Bayes factor, mis-specification of prior densities for the trial design parameters of interest from an alternative model in the single-arm phase II clinical trial setting can only decrease the expected weight of evidence in favor of the alternative model (Johnson and Cook, 2009). Hence the more severely the alternative model deviates from the true parameter model, the more penalty Bayes factor-based hypothesis testing would pay, no matter whether the prior density model is optimistic or skeptical as long as it is mis-specified.

The Table 1-1 summarizes what have been done in single-arm phase II clinical trial designs and what are to improve in this dissertation.

Table 1-1 What have been done in single-arm phase II clinical trial designs and what are to improve in this dissertation

Characteristic	**What have been done**	**What are to improve in this dissertation**
Bayesian Type I and Type II error rate	Derived based on the prior probability of the null hypothesis being true (Lee & Zelen, 2000). Bayesian version of Simon's two-stage design (Wang, et al, 2005) following Lee & Zelen's derivation (2000) on Bayesian Type I and Type II error rate.	To derive based on the marginal probability of the number of responses s_1 at stage I. s_1 follows a Dirichlet-multinomial distribution given a beta prior for the parameter of response rate. [Section 3.4]
Stopping rules based on Bayesian predictive probability	Threshold (boundary) probabilities are obtained mathematically from search space (Lee & Liu, 2008). The boundary predictive probability may not be practical. e.g. the low boundary is 0.001 in one of their examples of trial designs.	To pre-specify practical threshold (boundary) predictive probabilities to construct stopping rules. e.g. set the low boundary probability = 0.5 to accept the null hypothesis of the treatment response rate is equal to the maximum uninteresting response rate. [Section 3.2]
Stopping rules based on Bayes factor	Continuously monitor a phase II trial (Johnson & Cook, 2009)	To derive for two-stage design. [Section 4.3]
Control Bayesian and frequentist error rates	None.	To derive. [Section 3.6 and 4.3]

The objectives of this dissertation on single-arm two-stage phase II clinical trial design are:

1. To develop new designs for single-arm phase II clinical trials.
    - Bayesian-frequentist two-stage design.
    - Bayes factor-based two-stage design using an iMOM prior.
2. To characterize the new designs.
    a. To derive Bayesian Type I and Type II error rates.
    b. To demonstrate Bayesian and frequentist properties, particularly for the 2nd design to theoretically and/or numerically demonstrate Bayes factor's property of mis-specification for the trial design parameters and the property of iMOM as a non-local alternative prior.
    c. To control both Bayesian and frequentist Type I and Type II error rates for the Bayesian-frequentist, and control frequentist Type I and Type II error rates for the Bayes factor-based design.
    d. To establish an algorithm to find an optimal design (minimax design).
3. To demonstrate the new designs with numerical examples.
4. To compare the new designs with typical Bayesian and frequentist designs.

1.2 A varying-stage adaptive phase II/III clinical trial design

Currently, adaptive phase II/III clinical trials are typically carried out with a strict two-stage design. In general, the first stage is a learning stage as phase II, and the second stage is a confirmatory stage as phase III. During interim analysis, inefficacious or harmful dose arms are dropped, then one or two promising dose arms are selected for the second stage. Based upon interim results, other adaptations, such as change of primary study endpoint and/or primary hypothesis, adjustment of sample size, etc, could be applied.

Frequently there are some situations, in which researchers are in dilemma to make “go or no-go” decision and/or to select “best” dose arm(s), since interim data from the first stage may not provide sufficient data for their decision making. In this case, it is challenging to follow a strict two-stage plan. Therefore, we propose a varying-stage adaptive phase II/III clinical trial design, in which we also consider whether there is a need to have an intermediate stage to obtain more data, so that a more informative decision could be made regarding whether the trial can be advanced to the final confirmatory stage. Hence, the number of further investigational stages in our design is determined based upon data accumulated up to the current interim analysis.

During the past two decades, adaptive designs have been well studied by many researchers; however, existing adaptive phase II/III designs only discuss one or two aspects of adaptations, and their focus is stagewise design, in which the number of stages is fixed in order to construct decision rules, control Type I error rate and test hypotheses. In contrast to a conventional design with a fixed number of stages, under the framework of our varying-stage adaptive phase II/III clinical trial design, we consider the adaptations of dropping dose arm(s), changing primary study endpoint, adjusting sample size, and early stopping for futility. In our design, two study endpoints are considered. The endpoint 1 is initially designated as the primary study endpoint. The endpoint 2 can be switched as the primary study endpoint if the endpoint 1 does not seem sensitive to show treatment effect whereas the endpoint 2 appears a better measure of clinical benefit for the study treatment. We use an adaptive combination test to perform final statistical analyses.

Table 1-2 provides a summary of what have been done in adaptive phase II/III clinical trial designs and what are to improve in this dissertation.

Table 1-2 What have been done in adaptive phase II/III clinical trial design and what are to improve in this dissertation

Characteristic	**What have been done**	**What are to improve in this dissertation**
Flexible number of stages	Typically strict two-stage design. The first stage is a learning stage (phase II), and the 2nd (final) stage is for a confirmatory stage (phase III).	To add an intermediate stage if the data from the first stage is not sufficient to make decisions, such as selecting doses, advancing the trial to the final confirmatory stage, etc. [Section 6.1, 6.2, 6.3 and 6.4]
Number of adaptations	Consider one or two adaptations.	To consider more adaptations including dropping inefficacious/harmful dose arm(s), changing primary study endpoint, adjusting sample size, and early stopping for futility. [Section 6.1, 6.2, 6.3, 6.4, 8.3.4 and 8.3.5]
Type I error rate control	The secondary endpoints are tested only if the primary endpoint achieves statistical significance (Hung, Wang, O'Neill, 2007).	To control Type I error rate by considering both study endpoints. [Section 6.5, 7.2]

The objectives of this dissertation on adaptive phase II/III design are:

1. To propose a new design of varying-stage adaptive phase II/III clinical trial design.
2. To characterize the proposed design.
    a. To consider multiple adaptations including dropping inefficacious/harmful dose arm(s), changing primary study endpoint, adjusting sample size, and early stopping for futility.
    b. To define decision paths regarding primary study endpoint change and other adaptations.
    c. To develop final analysis methods with adaptive p-value combination.
    d. To derive Type I error rate for each decision path, and to prove Type I error rate control.
    e. To derive statistical power for each decision path.
    f. To derive algorithms to determine the sample size for the next stage.
3. To demonstrate the proposed design with illustrations/simulations.

The remainder of this dissertation is organized as follows. In Chapter 2, we will review the literature for existing single-arm phase II clinical trial designs. In Chapter 3, we will present our new design – Bayesian-frequentist single-arm phase II clinical trial design. A second new design – Bayes factor-based phase II design will be presented in Chapter 4. The literature review for phase II/III clinical trial designs is provided in Chapter 5. In Chapter 6, we will discuss our varying-stage adaptive phase

II/III clinical trial design. The other topics of the varying-stage adaptive phase II/III clinical trial design include dose selection and multiple comparisons on the primary endpoint, statistical power and sample size, simulations and a special case of the proposed design, and practical issues and trial implementation. These topics are discussed in Chapter 7, Chapter 8, Chapter 9 and Chapter 10.

2 Literature review for single-arm phase II clinical trial designs

In a phase II clinical trial, several doses of the new therapy are considered. In some diseases or medical conditions, phase II trial could be conducted without an active control treatment arm when the standard therapy is not established. In addition, placebo control might not be feasible due to ethical considerations. For example, single-arm phase II clinical trials are sometimes conducted in cancer research. In this chapter, based on our literature review, we introduce frequentist and Bayesian two-stage designs and continuous monitoring for single-arm phase II clinical trials.

2.1 Frequentist two-stage design

A typical frequentist single-arm phase II clinical trial design is Simon's two-stage design (1989). This design is widely used in cancer research. Under a frequentist framework, a two-stage single-arm phase II trial is designed to test the following hypotheses.

$$H_0: \theta \leq \theta_0, \text{ vs. } H_1: \theta \geq \theta_1 \qquad (2.1)$$

Where θ is the unknown response rate of the new therapy, θ_0 and θ_1 are the maximum uninteresting response rate and the minimum response rate of interest, respectively. The acceptance boundaries are determined under the constraints of Type I and Type II errors. By allowing trial early termination at stage I due to insufficient responses, Simon's design is flexible and requires less expected sample size compared to a single stage design. Simon developed two two-stage designs: the optimal design that minimizes the expected sample size under the null hypothesis, and the minimax design that minimizes the maximal sample size. Simon's two-stage designs have been extended by many researchers, including admissible two-stage design (Jung et al, 2004), two-stage design minimizing median sample size (Hanfelt, Slack and Gehan, 1999) and others.

Early acceptance of the new therapy is appropriate for situations where patients are very limited or the new drug is very expensive. Fleming (1982) applied O'Brien-Fleming bounds in his design, in which early rejection of null hypothesis occurs only when the interim results are quite extreme. Chang et al (1987) proposed a multi-stage phase II clinical trial design that minimizes the average of the expected sample size under null and alternative hypotheses with sample size in multiples of five. Shuster's minimax two-stage design (2002) has the smallest globally maximized expected sample size.

In practice, it is difficult to have a trial conducted exactly as initially planned. Green and Dahlberg (1992) investigated planned vs attained designs in Phase II clinical trials. Wu and Shih (2008) constructed ways to handle four different scenarios of deviation or interruptions from original Simon's two-stage design. Instead of rigid two-stage designs, some flexible or adaptive two-stage designs have been developed during the past decades. Typically, Chen and Ng (1998) proposed flexible optimal and minimax two-stage designs as a collection of two-stage designs such that the sample size and boundary for each stage are a set of consecutive values. Lin and Shih (2004) pointed out that there is a very high probability to reject a promising new treatment if the initial expectation in Simon's two-stage design is set too high. Lin & Shih's adaptive two-stage design (2004) allows both low and high expected response rate be considered. Other adaptive two-stage designs include the designs by Johns and Andersen (1999), and by Banerjee and Tsiatis (2006).

The conventional clinical trial designs described above are typical frequentist phase II clinical trial designs, which use information of treatment effect (target response rate of the new therapy vs. uninteresting response rate) to determine the sample size with sufficient statistical power and appropriate Type I error rate control. A common drawback of these designs is that the treatment effect is considered as a fixed value. With the constraints of controlling Type I and Type II error rates, frequentist clinical trial designs are rigid and not optimal in terms of sample size saving and decision making due to the lack of consideration on uncertainty, utility or loss function. In contrast, The Bayesian approach has the advantages of flexibility and accounting for uncertainties, utility or loss function.

2.2 Bayesian two-stage design

Under a Bayesian framework, Tan and Machin (2002) published single threshold design (STD) and dual threshold design (DTD). These two types of designs are based on evaluation of whether the posterior probability of the response rate exceeding a threshold response rate plus a pre-specified small value ε (e.g ε=0.05) is ≥ a threshold posterior probability (λ). Although Tan and Machin’s STD has been extended by some researchers including Mayo and Gajewski (2004), Gajewski and Mayo (2006), and Sambucini (2008), like Tan and Machin’s designs, error rate control is not considered.

Lee and Zelen (2000) argued that conventional type I error rate $\alpha = 0.05$ may result in an excessive number of false positive trial outcomes in clinical trial practice. They proposed a method to calculate posterior false positive and false negative error rates conditional on the trial outcome. These two probabilities are considered as Bayesian Type I and Type II error rates. The Bayesian Type I and Type II error rates Lee and Zelen (2000) constructed are based on the prior probability of the null hypothesis being true. Simon (2000) and Bryant and Day (2000) criticized this method for ignoring observed data and hence violating likelihood principles. Technically, as Lee and Zelen (2000) described, this prior probability can be estimated as the ratio of the number of clinical trials with positive outcome among historical trials. Obviously, this is a difficult task in some clinical practices. Moreover, this prior (especially, if it is based on subjective assessment in favor of the difference between treatments) needs to be updated with the new trial data under general consideration of the Bayesian framework and likelihood principle, in contrast to designing a trial by controlling Type I and Type II error rates directly defined based on this prior.

Wang, Leung, Li and Tan (2005) introduced a Bayesian version of Simon's two-stage design, which inherits the features of Simon's design but also possesses attractive Bayesian attributes. However, following Lee and Zelen's derivation on Bayesian Type I and Type II error rate (2000), Wang, Leung, Li and Tan (2005) used the prior probability that the study therapy is promising; therefore, Bayesian and frequentist error rates are related in their designs.

2.3 Bayesian clinical trial monitoring

To take account of the uncertainty of future data, the Bayesian predictive probability approach has been used by many researchers for clinical trial monitoring and designs (e.g. Herson, 1979; Grieve, 1991; Johns, Anderson, 1999; Dmitrienko, Wang, 2006; Lee and Liu, 2008; Sambucini, 2008). The predictive approach for clinical trial designs was first introduced by Herson (1979). Lee and Liu (2008) developed a predictive probability (PP) design, such that if $PP < P_L$ (lower boundary), then the trial is stopped and the new therapy is rejected; if $PP > P_U$(upper boundary), then the trial is stopped and the new therapy is accepted; otherwise the trial is continued to the next stage. Unlike other Bayesian designs, Lee and Liu's approach (2008) controls strict frequentist Type I and Type II error rates using the recursive method from Schultz et al (1973). Through a three-dimensional search, a minmax design can be obtained with optimal parameters of P_L, P_U, N_{max} and P_T (threshold value for posterior probability). However, this search is performed mathematically. For the two examples in their paper, the optimal P_L (lower boundary of predictive probability to reject the test treatment) was 0.001 and 0.075-0.079. For such very low boundary, it is almost impossible to reject the new therapy even if it is not promising for further investigation.

Johnson and Cook (2009) developed a Bayes factor-based, continuous monitoring approach for single-arm phase II studies. In their approach, a non-local prior for the alternative hypothesis was used such that no mass was assigned to parameter values that were consistent to the null hypothesis. Furthermore, they pointed out that mis-specification of a prior density on the treatment effect can not increase the expected weight of evidence; therefore, Bayes factor-based design can reduce potential bias from a prior density.

3 A Bayesian-frequentist two-stage single-arm phase II clinical trial design

As set in Section 2.1, we assume that the response from each patient follows a Bernoulli distribution with parameter θ, which is the unknown response rate of the new therapy. Let θ_0 and θ_1 denote the maximum uninteresting response rate and the minimum response rate of interest, respectively. Under a frequentist framework, the single-arm phase II clinical trial is designed to test the null hypothesis H_0: $\theta \leq \theta_0$, versus the alternative hypothesis H_1: $\theta \geq \theta_1$. If the null hypothesis H_0 is rejected at the pre-specified significance level α, then the test treatment is accepted for further investigation; otherwise, the new therapy is concluded to be an unpromising treatment.

In this section, we propose a new design – Bayesian-frequentist two-stage single-arm phase II clinical trial design. This design consists of two components: frequentist setting and Bayesian setting.

3.1 Frequentist setting

a. Two-stage design

Unlike Simon's design (1989), we propose a two-stage design allowing early stopping due to either futility or efficacy. Following Fleming (1982), Chang et al (1987) and Shuster's design (2002), a two-stage clinical trial can be designed as follows.

a.1 Stage I

For the first stage, n_1 patients are enrolled in the study. If the number of responses s_1 at stage I is greater than or equal to the upper boundary r_1 $(s_1 \geq r_1)$, then reject the null hypothesis $H_0 : \theta \le \theta_0$ and stop the trial due to efficacy. The probability of the rejecting this null hypothesis is

$$R_1(\theta) = P(S_1 \ge r_1 | \theta) = 1 - Bin(\theta, r_1 - 1, n_1) \qquad (3.1)$$

where *Bin* denotes cumulative binomial distribution function.

If $s_1 \le a_1$ (the lower boundary), then accept the null hypothesis H_0: $\theta \le \theta_0$ and stop the trial due to futility. The probability of accepting the null hypothesis is

$$A_1(\theta) = P(S_1 \le a_1 | \theta) = Bin(\theta, a_1, n_1) \qquad (3.2)$$

Otherwise, if s_1 is between r_1 and a_1 ($a_1 < s_1 < r_1$), then continue the trial and enroll n_2 patients into stage II.

a.2 Stage II

If the cumulative number of the responses s at the end of stage II is greater than or equal to r ($s \ge r$), then reject the null hypothesis H_0: $\theta \le \theta_0$ and claim that further investigation of the study therapy is warranted. The probability of rejecting the null hypothesis H_0: $\theta \le \theta_0$ at the stage II is

$$R_2(\theta) = P(a_1 < S_1 < r_1, S \ge r \mid \theta) = \sum_{s_1=a_1+1}^{r_1-1} bin(\theta, s_1, n_1)[1 - Bin(\theta, r - s_1 - 1, n_2)] \qquad (3.3)$$

where bin is the probability density function of binomial distribution.

Otherwise, the test treatment is rejected and no further investigation is warranted.

b. Measures of frequentist two-stage design

One of the main advantages of having a two-stage design is to have a smaller average sample size compared to a single stage design. For the two-stage design described above, the probability of early termination is the sum of the probabilities of stopping a trial at stage I. This probability can be expressed as

$$PET_f(\theta) = R_1(\theta) + A_1(\theta) \qquad (3.4)$$

The expected sample size is

$$E_f(n|\ \theta) = n_1 + [1\text{-}PET_f(\theta)]*n_2 \qquad (3.5)$$

Here the subscript f denotes frequentist properties. In the late sections, we will use the superscript f to denote frequentist properties as well. In contrast, we will use subscript or superscript B to denote Bayesian properties.

For the frequentist setting, it is important to control Type I error rate and maintain study power. The Type I error rate is the probability of stopping the trial due to efficacy given the null hypothesis $H_0 : \theta \leq \theta_0$ is true, while power is the probability of claiming the test treatment being promising given the alternative hypothesis $H_1: \theta \geq \theta_1$ is true.

$$\text{Type I error rate: } \alpha^f = R_1(\theta_0) + R_2(\theta_0) \qquad (3.6a)$$

$$Power{:}\ 1 - \beta^f = R_1(\theta_1) + R_2(\theta_1) \qquad (3.6b)$$

where α^f and β^f are the frequentist Type I and Type II error rate, respectively.

3.2 Bayesian setting

At stage I, suppose that s_1 responses among total number of n_1 patients are obtained. Given the data (s_1, n_1) at stage I, the posterior distribution of the response rate (θ) of the study therapy is a beta distribution as (3.7) if a conjugate beta prior - beta(a, b) is applied.

$$P(\theta|(s_1, n_1)) = beta(a+s_1, b+ n_1 - s_1) \qquad (3.7)$$

where beta is the probability density function of beta distribution.

Let s_2 denote the number of responses in future n_2 patients in stage II, whose predictive distribution is a beta-binomial distribution. Let s denote total number of responses ($s = s_1+ s_2$) and n be total number of patients enrolled in the whole study ($n = n_1+ n_2$), then success of the trial can be measured by the posterior probability $P(\theta > \theta_0|(s, n)) = P(\theta > \theta_0|(s_1, n_1), (s_2, n_2))$ exceeding a threshold probability P_T, namely,

$P(\theta > \theta_0|(s, n)) > P_T$. This probability is a non-decreasing function with respect to number of responses s_1 at stage I (see Appendix A.4 for proof).

Given the hypothetical stage I data (s_1, n_1) and sample size n_2 for stage II, the posterior probability of the trial outcome can be predicted with the predictive probability defined as follows.

$$PP(TS|s_1,n_1,n_2) = \sum_{s_2=0}^{n_2}\{P(s_2|(s_1,n_1,n_2)) * I(P(\theta > \theta_0|(s_1,n_1),(s_2,n_2)) > P_T)\} \quad (3.8)$$

where TS denotes trial success outcome measured by $P(\theta > \theta_0 | (s_1, n_1), (s_2, n_2)) > P_T$. For convenience, we have abbreviated the notation $P(S_2=s_2)$ by $P(s_2)$ in (3.8). We will continue to use the similar abbreviations in the sequel.

The probability defined above is the predictive probability of success of a trial given the number of responses (s_1) at stage I and sample size in stage I (n_1) and stage II (n_2). In this thesis, we use the simple term 'predictive probability' (PP) for the predictive probability of trial success as defined in (3.8), unless otherwise indicated.

Lee and Liu (2008) proposed a predictive probability design; we use similar decision rules as follows:

- If $PP \leq P_L$, then stop the trial for futility and claim the study therapy is not promising. Since PP defined in (3.8) is a non-decreasing function of number of responses in stage II (s_1), determine if $PP \leq P_L$ is equivalent to evaluating if $s_1 \leq a_1$ (lower boundary for futility).
- If $PP \geq P_U$, then stop the trial for efficacy and claim the study therapy is promising. Similarly, this criteria can be equivalently assessed with if $s_1 \geq r_1$ (upper boundary for efficacy).

- Otherwise if $P_L < PP < P_U$, (or equivalently $a_1 < s_1 < r_1$), then continue the trial to stage II with additional n_2 patients.

In Lee and Liu (2008)'s work, they searched for the optimal P_L (threshold predictive probability to reject the test treatment) mathematically. For the two examples in their paper, optimal P_L is 0.001 and 0.075-0.079. For such very low predictive probabilities, it is almost impossible to reject the study therapy even if it is not promising for further investigation. In our research, we pre-specify the lower and upper boundaries of predictive probabilities P_L and P_U from a real clinical practice point of view, e.g $P_L = 0.5$ and $P_U = 0.95$.

3.3 Probability of early termination and expected sample size under Bayesian framework

Given a beta(a, b) prior and sample size n_1 for stage I, the number of responses s_1 from stage I follows a Dirichlet–multinomial distribution.

$$P(s_1 \mid n_1) = \binom{n_1}{s_1} \frac{B(a+s_1, b+n_1-s_1)}{B(a,b)} \tag{3.9}$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$, which is a beta function.

Therefore, the probability of early termination (PET_B) under Bayesian setting is as follows.

$$PET_B = \sum_{s_1=0}^{a_1} P(s_1 \mid n_1) + \sum_{s_1=r_1}^{n_1} P(s_1 \mid n_1) \tag{3.10}$$

The expected sample EN_B under the Bayesian framework is

$$EN_B = n_1 + (1 - PET_B) * n_2 \tag{3.11}$$

3.4 Bayesian error rates

As described in Section 2.2, Lee and Zalen (2000) constructed Bayesian error rates based on the prior probability of the null hypothesis being true. Simon (2000) and Bryant and Day (2000) criticized this method for ignoring observed data and hence violating likelihood principles. Unlike Lee and Zalen (2000), we define Bayesian Type I and Type II error rates based on marginal probability of s_1 as derived in (3.9).

Let A denote acceptance of the null hypothesis H_0: $\theta \le \theta_0$, and R denote rejection of the null hypothesis H_0: $\theta \le \theta_0$. Namely A and R are for claiming negative trial outcome and positive trial outcome respectively. Bayesian error rates can be defined as follows:

$$\alpha^B = P(\theta \le \theta_0 | R) \qquad (3.12a)$$

$$\beta^B = P(\theta \ge \theta_1 | A) \qquad (3.12b)$$

In order to derive the Bayesian error rates α^B and β^B, we first derive the distributions of $\theta \le \theta_0 \cap R$ and $\theta \ge \theta_1 \cap A$. At stage I, the distribution of $\theta \le \theta_0$ and (s_1, n_1) is.

$$P(\theta \le \theta_0 \ and \ (s_1, n_1)) = P(\theta \le \theta_0 \mid (s_1, n_1)) P(s_1 \mid n_1) \qquad (3.13)$$

The number of responses s_1 given n_1 follows *a* Dirichlet–multinomial distribution as derived in (3.9), hence**,**

$$P(\theta \le \theta_0 \ and \ (s_1, n_1)) = \binom{n_1}{s_1} \frac{B(a+s_1, b+n_1-s_1)}{B(a,b)} Beta(\theta_0, a+s_1, b+n_1-s_1) \qquad (3.14)$$

where *Beta(θ,a,b)* is the cumulative beta distribution function.

And

$$P(\theta \le \theta_0 \cap R) = P(\theta \le \theta_0 \ and \ (s_1 \ge r_1))$$

$$=\sum_{s_1=r_1}^{n_1}\left\{\binom{n_1}{s_1}\frac{B(a+s_1,b+n_1-s_1)}{B(a,b)}Beta(\theta_0,a+s_1,b+n_1-s_1)\right\} \qquad (3.15)$$

Therefore, $\alpha_1^B = P(\theta \le \theta_0 \mid R) = \frac{P(\theta \le \theta_0 \cap R)}{P(R)}$

$$=\frac{\sum_{s_1=r_1}^{n_1}\left\{\binom{n_1}{s_1}\frac{B(a+s_1,b+n_1-s_1)}{B(a,b)}Beta(\theta_0,a+s_1,b+n_1-s_1)\right\}}{\sum_{s_1=r_1}^{n_1}\binom{n_1}{s_1}B(a+s_1,b+n_1-s_1)/B(a,b)} \qquad (3.16)$$

Assume that responses s_1 and s_2 are independent, the distribution of $\theta \le \theta_0$ and (s_1,n_1,s_2,n_2) is

$$P(\theta \le \theta_0 \ and \ (s_1,n_1,s_2,n_2) = P(\theta \le \theta_0 \mid (s_1,s_2))P(s_1 \mid n_1)P(s_2 \mid n_2) \qquad (3.17)$$

$$where \ P(s_2 \mid n_2) = \binom{n_2}{s_2}\frac{B(a+s_1+s_2,b+n_1-s_1+n_2-s_2)}{B(a+s_1,b+n_1-s_1)}$$

$$=\binom{n_2}{s_2}\frac{B(a+s,b+n-s)}{B(a+s_1,b+n_1-s_1)} \qquad (3.18)$$

Hence, $P(\theta \le \theta_0 \cap R) = P(\theta \le \theta_0 \ and \ (a_1 < s_1 < r_1, s_2 \ge r - s_1)$

$$=\sum_{s_1=a_1+1}^{r_1-1}\sum_{s_2=r-s_1}^{n_2}\left\{\binom{n_1}{s_1}\binom{n_2}{s_2}\frac{B(a+s,b+n-s)}{B(a,b)}Beta(\theta_0,a+s,b+n-s)\right\} \qquad (3.19)$$

and $\alpha_2^B = P(\theta \le \theta_0 \mid R) = \frac{P(\theta \le \theta_0 \cap R)}{P(R)}$

$$=\frac{\sum_{s_1=a_1+1}^{r_1-1}\sum_{s_2=r-s_1}^{n_2}\left\{\binom{n_1}{s_1}\binom{n_2}{s_2}\frac{B(a+s,b+n-s)}{B(a,b)}Beta(\theta_0,a+s,b+n-s)\right\}}{\sum_{s_1=a_1+1}^{r_1-1}\sum_{s_2=r-s_1}^{n_2}\binom{n_1}{s_1}\binom{n_2}{s_2}B(a+s,b+n-s)/B(a,b)} \qquad (3.20)$$

Therefore, the Bayesian Type I error rate is

$$\alpha^B = \alpha_1^B + \alpha_2^B$$

$$
= \frac{\sum_{s_1=r_1}^{n_1} \left\{ \binom{n_1}{s_1} \frac{B(a+s_1, b+n_1-s_1)}{B(a,b)} Beta(\theta_0, a+s_1, b+n_1-s_1) \right\}}{\sum_{s_1=r_1}^{n_1} \binom{n_1}{s_1} B(a+s_1, b+n_1-s_1) / B(a,b)}
$$

$$
+ \frac{\sum_{s_1=a_1+1}^{r_1-1} \sum_{s_2=r-s_1}^{n_2} \left\{ \binom{n_1}{s_1} \binom{n_2}{s_2} \frac{B(a+s, b+n-s)}{B(a,b)} Beta(\theta_0, a+s, b+n-s) \right\}}{\sum_{s_1=a_1+1}^{r_1-1} \sum_{s_2=r-s_1}^{n_2} \binom{n_1}{s_1} \binom{n_2}{s_2} B(a+s, b+n-s) / B(a,b)}
$$

(3.21a)

Similarly, the Bayesian Type II error rate is

$$
\beta^B = P(\theta \geq \theta_1 \mid A) = \frac{P(\theta \geq \theta_1 \cap A)}{P(A)} = \beta_1^B + \beta_2^B
$$

$$
= \frac{\sum_{s_1=0}^{a_1} \left\{ \binom{n_1}{s_1} \frac{B(a+s_1, b+n_1-s_1)}{B(a,b)} \left[1 - Beta(\theta_1, a+s_1, b+n_1-s_1) \right] \right\}}{\sum_{s_1=0}^{a_1} \binom{n_1}{s_1} B(a+s_1, b+n_1-s_1) / B(a,b)}
$$

$$
+ \frac{\sum_{s_1=a_1+1}^{r_1-1} \sum_{s_2=0}^{r-s_1-1} \left\{ \binom{n_1}{s_1} \binom{n_2}{s_2} \frac{B(a+s, b+n-s)}{B(a,b)} \left[1 - Beta(\theta_1, a+s, b+n-s) \right] \right\}}{\sum_{s_1=a_1+1}^{r_1-1} \sum_{s_2=0}^{r-s_1-1} \binom{n_1}{s_1} \binom{n_2}{s_2} B(a+s, b+n-s) / B(a,b)}
$$

(3.21b)

3.5 Beta prior

Single-arm phase II clinical trials are typically conducted in cancer research. At the time of phase II trial initiation, usually, there are limited data available from historical clinical trials; therefore, priors for the clinical trial design parameters have to be elicited from experts. For phase II cancer trials with a binary response as the primary study endpoint, conjugate beta priors are widely used for Bayesian inferences. Thall and Simon (1994) used W_{90} for beta prior elicitation, which is the width of the 90% probability interval running from 5th to 95th percentiles.

For a beta prior beta(a, b), $a + b - 2$ can be interpreted as the prior sample size, $a-1$ as prior successes and $b-1$ as prior failures (Gelman, 2004). This interpretation is intuitive and easily understood by clinical trial practitioners. Further, Hetjan (1997) expressed beta priors with the parameters $a = n_0\pi_0 + 1,$ and $b = n_0(1- \pi_0) + 1$. The hyper parameters n_0 and π_0 can be interpreted as the prior sample size and the prior mode for a beta prior. When $n_0 = 1$, the prior distribution is very flat, hence it provides less prior information. As n_0 increases, the prior distribution becomes more concentrated at the prior mode π_0. When n_0 is equal to infinity (∞), the prior density is completely condensed at π_0. Tan and Machin (2002) used prior sample size n_0=1 in their Bayesian two-stage designs (STD and DTD). Sambucini (2008) used this beta prior as well for his predictive two-stage design. Mayo and Gaewski (2004) compared the non-informative prior with n_0=1, the informative prior with n_0>1, beta priors elicited by median with W_{90}, and by mean with W_{90} for Bayesian sample size calculation.

Wu, Shih and Moore (2008) provided methods of eliciting beta priors from clinical information. It is straightforward to elicit prior response rate of the test treatment from medical investigators. One elicitation question could be like 'the response rate that is most likely to occur". In fact, this elicited response rate is the mode θ_1 of a prior beta distribution. Another elicitation question could be for tail probability of $P(\theta \leq \theta_0)$. With the mode θ_1 and tail probability $P(\theta \leq \theta_0)$, the parameters a and b can be determined for a beta prior. In this thesis, we use this elicitation method.

3.6 Algorithm to find optimal design

In single-arm phase II studies, many researchers choose a design by minimizing the expected sample size under the frequentist null or alternative hypothesis. For example, one with skeptical view on the study therapy may minimize expected sample size under the null hypothesis, whereas it is also reasonable to choose a design by minimizing the expected sample size under the alternative hypothesis if the study therapy is anticipated promising. In this thesis, our optimal design is defined as the design with minimal expected sample size under the Bayesian framework. This approach can avoid arguing on minimization of expected sample size under the null vs the alternative hypothesis.

To find the optimal design, the following parameters need to be given.

- Maximum uninteresting response rate and minimum response rate of interest: θ_0 and θ_1;
- Beta prior beta(a,b) for θ;
- Lower and upper boundaries of predictive probability: P_L, P_U;
- Threshold posterior probability: P_T;
- Type I and Type II error rates: α and β.
- Maximum sample size Nmax: set Nmax = 1.5* size for a single stage trial

The algorithm to numerically find the optimal design is described as follows

(1) For each possible n ranging from 15 to Nmax,

(2) Set n_1 ranging from max(5, n/3) to n -1

(3) Determine the boundaries r_1 and a_1

(3.1) Calculate predictive probability (PP) from (3.8) corresponding to possible number of responses at stage I (s_1), which ranges from 0: n_1

(3.2) Determine a_1 and r_1 by comparing PP against the boundaries P_L and P_U.

(4) Search r ranging from r_1+1 to n -1.

(4.1) Calculate Bayesian and frequentist type I and Type II error rates

(4.2) If all the calculated error rates < pre-specified thresholds α and β, then a design of (a_1, r_1, n_1, r, n) is obtained.

(4.3) For each obtained design, determine Bayesian and frequntist properties (e.g probability of early termination, expected sample size, and Bayesian posterior probabilities).

(5) Choose the design with the minimum expected sample size under the Bayesian setting from (3.11) as the optimal design.

3.7 Some properties of Bayesian-frequentist two-stage single-arm phase II clinical trial design

3.7.1 Posterior probability of $\theta > \theta_i$ |(s,n)

Definition 3.1: optimistic (enthusiastic) and pessimistic (skeptical) prior

Many researchers, for example, Heitjan (1997), use a prior with information centered at the minimum response rate of interest (θ_1) as an optimistic (enthusiastic) prior. Similarly, a prior with information centered at the maximum uninteresting response rate (θ_0) is used as a pessimistic (skeptical) prior. In addition, a prior centering information at the half-way between θ_0 and θ_1 is used as an 'indifference' prior (Cronin, 1999). The degree or strength of enthusiasm or pessimism can be measured by the prior size $n_0 = a + b$ of a prior beta(a, b).

Definition 3.2: more optimistic (enthusiastic) prior

In general, under two-stage single-arm phase II clinical trial setting, a beta prior expecting a higher mean response rate or a higher mode of response rate of the study therapy is considered as a more optimistic (enthusiastic) prior compared to another beta prior that result in a lower mean response rate or a lower mode of response rate. Let's consider two beta priors: beta(a_1, b_1) and beta(a_2, b_2) with the same size $a_1 + b_1 = a_2 + b_2 = n_0$. If $a_1 > a_2$, then the prior beta(a_1, b_1) is considered as the more optimistic (enthusiastic) prior than the prior beta(a_2, b_2) since a higher mean response rate or a higher mode of response rate of the study therapy is expected from the prior beta(a_1, b_1). In other word, the prior beta(a_2, b_2) is considered as the less optimistic or the more pessimistic (skeptical) prior compared to the prior beta(a_1, b_1).

Please note that a "more optimistic" prior defined here is not necessarily an optimistic prior, which centers the prior information at the minimum response rate of interest (θ_1) or above, but rather a prior that presents more optimistic belief compared to another prior.

Proposition 3.1: A more optimistic prior results in a higher posterior probability of $\theta>\theta_i \mid (s, n)$ ($i = 0, 1$) compared to another beta prior with the same size of $a + b = n_0$.

Proof: As shown in Appendix A.1, the data (s_1, n_1) obtained from the first stage does not affect the posterior distribution of $\theta|(s,n)$. $\theta|(s,n)$ follows a beta distribution if conjugate beta prior - $beta(a, b)$ is used.

$$P(\theta|(s,n))= beta(a + s, b + n - s)$$

The posterior probability $\theta > \theta_i \,|(s,n)$ is

$$P(\theta > \theta_i \,|(s,n)) = 1 - Beta(\theta_i, a + s, b + n - s) \qquad (3.22)$$

Appendix A.5.1 provides the relation of beta and binomial probability calculation. Apply (A.5.5a) to the right side of (3.22), (3.22) can be written in binomial form as:

$$\begin{aligned} P(\theta > \theta_i \,|(s,n)) &= 1 - Beta(\theta_i, a + s, b + n - s) \\ &= Bin(\theta_i, a + s - 1, n + a + b - 1) \qquad (3.23) \end{aligned}$$

For the designs using the two priors given previously in the definition 3.2: beta(a_1, b_1) and beta(a_2, b_2) where $a_1>a_2$ and $a_1 + b_1 = a_2 + b_2 = n_0$:

$$P_1(\theta > \theta_i \,|(s,n)) =Bin(\theta_i, a_1 + s - 1, n + n_0 - 1) \qquad (3.24a)$$

and

$$P_2(\theta > \theta_i \mid (s,n)) = Bin(\theta_i, a_2 + s - 1, n + n_0 - 1) \qquad (3.24b)$$

Given $a_1 > a_2$, apparently, $P_1(\theta > \theta_i \mid (s,n)) > P_2(\theta > \theta_i \mid (s,n)$. This proves that for the design with a more optimistic prior, the posterior probability of $\theta > \theta_i \mid (s, n)$ (for example, $\theta > \theta_0$ *or* $\theta > \theta_1 \mid (s, n)$) is higher than another design with the same size $a + b$ of prior *beta(a, b)*.

Definition 3.3: stronger beta prior

A beta prior with a larger size of a + b is considered as a stronger prior. Let's consider two beta priors: beta(a_1, b_1) and beta(a_2, b_2), the prior sizes are $n_{01} = a_1 + b_1$ for beta(a_1, b_1) and $n_{02} = a_2 + b_2$ for beta(a_2, b_2). If $n_{01} > n_{02}$, then the prior beta(a_1, b_1) is considered as the stronger prior compared to the prior beta(a_2, b_2).

Proposition 3.2: For the optimistic beta priors with the same mean response rate or mode of response rate, a stronger optimistic prior results in a higher posterior probability of $\theta > \theta_i \mid (s, n)$ $(i = 0, 1)$ if $s < n\theta_i$.

Proof: Given the two priors described in the definition 3.3: *beta*(a_1, b_1) and *beta*(a_2, b_2) where $n_{01} = a_1 + b_1$, $n_{02} = a_2 + b_2$, and $n_{01} > n_{02}$, the posterior probabilities of $\theta > \theta_1 \mid (s, n)$ are

$$P_1(\theta > \theta_1 \mid (s,n)) = \mathrm{Bin}(\theta_1, a_1 + s - 1, n + n_{01} - 1) \qquad (3.25a)$$

and

$$P_2(\theta > \theta_1 \mid (s,n)) = \mathrm{Bin}(\theta_1, a_2 + s - 1, n + n_{02} - 1) \qquad (3.25b)$$

Suppose the two priors beta(a_1, b_1) and beta(a_2, b_2) have the same mode as

$$\frac{a_1 - 1}{n_{01} - 2} = \frac{a_2 - 1}{n_{02} - 2} = \theta_1$$

then (3.25a) and (3.25b) can be written as

$$P_1(\theta > \theta_1 \,|(s,n)) = \text{Bin}(\theta_1, s + (n_{01}-2)\,\theta_1, n + n_{01} - 1) \qquad (3.26a)$$

$$P_2(\theta > \theta_1 \,|(s,n)) = \text{Bin}(\theta_1, s + (n_{02}-2)\,\theta_1, n + n_{02} - 1) \qquad (3.26b)$$

Apply the incremental property of binomial distribution as proved in Appendix A.5.7,

$$\begin{aligned}\text{Bin}(\theta_1, s, n) &< \text{Bin}(\theta_1, s + (n_{02}-2)\,\theta_1, n + n_{02} - 2) \text{ if } s < n\theta_1 \\ &< \text{Bin}(\theta_1, s + (n_{01}-2)\,\theta_1, n + n_{01} - 2) \text{ per Appendix A.5.4} \\ &< \text{Bin}(\theta_1, s + (n_{01}-2)\,\theta_1, n + n_{01} - 1) \qquad (3.27)\end{aligned}$$

Therefore, $P_1(\theta > \theta_1 \,|(s,n)) > P_2(\theta > \theta_1 \,|(s,n))$ and the Proposition 3.2 holds if $s < n\theta_1$. Appendix A.5.5 shows that there exist the most probable number of responses k such that $(n+1)\theta-1<k\leq(n+1)\theta$. $n\theta_1$ is approximately equal to k, therefore, the condition $s < n\theta_1$ indicates that Proposition 3.2 holds when s is less than the most probable number of responses k.

Similarly, the Proposition 3.2 can be proved for the posterior probability of $\theta>\theta_0 \mid (s, n)$.

3.7.2 Sufficient prior conditions satisfying Bayesian error rates < corresponding frequentist error rates in stage I

3.7.2.1 Sufficient prior condition satisfying Bayesian Type I error rate < frequentist Type I error rate in stage I

The aim of this section is to find a condition on beta priors to ensure that the Bayesian Type I error rate is les than the frequentist Type I error rate in the 1st stage as expressed below.

$$\alpha_1^B \leq R_1(\theta_0) \qquad (3.28)$$

Namely

$$\frac{\sum_{s_1=r_1}^{n_1}\left\{\binom{n_1}{s_1}\frac{B(a+s_1,b+n_1-s_1)}{B(a,b)}Beta(\theta_0,a+s_1,b+n_1-s_1)\right\}}{\sum_{s_1=r_1}^{n_1}\binom{n_1}{s_1}B(a+s_1,b+n_1-s_1)/B(a,b)} < 1 - Bin(\theta_0, r_1 - 1, n_1) \tag{3.29}$$

Base on the Proposition A3 in Appendix A.5.2, Beta(θ_0, $a+s_1$, $b+n_1-s_1$) is a monotonic decreasing function as s_1 increases given $a+b+n_1$, which is sum of prior size $n_0= a+b$ and sample size n_1 for the 1st stage. Therefore, to show (3.29), we can show the following (3.30). One should note that this consideration is conservative by reducing the right side of (3.29) to Beta(θ_0, $a+r_1$, $b+n_1-r_1$) as the right side of (3.30), which is the largest value of Beta(θ_0, $a+s_1$, $b+n_1-s_1$) when $s_1=r_1$ given $a+b+n_1$.

$$\text{Beta}(\theta_0, a+r_1, b+n_1-r_1) < 1 - Bin(\theta_0, r_1 - 1, n_1) \tag{3.30}$$

Following (A.5.5a) per the relation between beta and binomial probability calculation provided in Appendix A.5.1, (3.30) can be written in binomial form as

$$1-Bin(\theta_0, r_1 -1+a, n_1+a+b-1) < 1- Bin(\theta_0, r_1 - 1, n_1)$$

Simplify further,

$$Bin(\theta_0, r_1 -1+a, n_1+a+b-1) > Bin(\theta_0, r_1 - 1, n_1) \tag{3.31}$$

If n_1 is not small, and θ_0 is not close to 0 or 1, apply normal approximation to the binomial distributions, (3.31) can be written as

$$\frac{r_1-1+a-(n_1+a+b-1)\theta_0}{\sqrt{(n_1+a+b-1)\theta_0(1-\theta_0)}} > \frac{r_1-1-n_1\theta_0}{\sqrt{n_1\theta_0(1-\theta_0)}} \tag{3.32}$$

Further simplify,

$$\frac{r_1-1+a}{\sqrt{n_1+a+b-1}} - \theta_0\sqrt{n_1+a+b-1} > \frac{r_1-1}{\sqrt{n_1}} - \theta_0\sqrt{n_1}$$

$$\frac{r_1-1+a}{\sqrt{n_1+a+b-1}} - \frac{r_1-1}{\sqrt{n_1}} > (\sqrt{n_1+a+b-1} - \sqrt{n_1})\theta_0$$

Divide by $\sqrt{n_1(n_1+a+b-1)}$

$$\frac{1}{\sqrt{n_1}}\frac{r_1-1+a}{n_1+a+b-1}-\frac{1}{\sqrt{n_1+a+b-1}}\frac{r_1-1}{n_1}>(\frac{1}{\sqrt{n_1}}-\frac{1}{\sqrt{n_1+a+b-1}})\theta_0 \qquad (3.33)$$

In practice,

$$\frac{r_1-1}{n_1}>\theta_0 \qquad (3.34)$$

To satisfy (3.33), we need to have

$$\frac{r_1-1+a}{n_1+a+b-1}>\frac{r_1-1}{n_1}$$

$$(r_1-1+a)n_1>(n_1+a+b-1)(r_1-1)$$

$$a>\frac{(r_1-1)(n_0-1)}{n_1} \qquad (3.35)$$

In practice, $\frac{r_1-1}{n_1}$ is approximately equal to θ_1, then we have

$$a > (n_0-1)\,\theta_1 \qquad (3.36)$$

The (3.36) is a sufficient condition to (3.31), hence it is a sufficient condition to (3.28). But it is not a necessary condition to (3.28) since for some $a \leq (n_0-1)\,\theta_1$, (3.28) is still true. For example, when $n_0=12$, $\theta_0=0.2$, $\theta_1=0.4$ and $n_1=22$, $a1=6$, $r1=10$, if $a=3 < (n_0-1)\,\theta_1=4$, (3.28) is still true. This design is included in Section 3.9 with the prior beta(3,9).

3.7.2.2 Sufficient prior condition satisfying Bayesian Type II error rate < frequentist Type II error rate in stage I

The aim of this section is to find a condition on beta priors to ensure that the Bayesian Type II error rate is les than the frequentist Type II error rate in the 1st stage as expressed in (3.37) below.

$$\beta_1^B<A_1(\theta_1) \qquad (3.37)$$

Namely

$$\frac{\sum_{s_1=0}^{a_1}\left\{\binom{n_1}{s_1}\frac{B(a+s_1,b+n_1-s_1)}{B(a,b)}\left[1-Beta(\theta_1,a+s_1,b+n_1-s_1)\right]\right\}}{\sum_{s_1=0}^{a_1}\binom{n_1}{s_1}B(a+s_1,b+n_1-s_1)/B(a,b)} < Bin(a_1,\ n_1,\ \theta_1)$$

(3.38)

Following similar process as Section 3.8.2.1, conservatively, (3.38) can be written in binomial form as

$$Bin(a_1+a-1,\ n_1+a+b-1,\ \theta_1) < Bin(a_1,\ n_1,\ \theta_1) \tag{3.39}$$

If n_1 is not small, and θ_1 is not close to 0 or 1, apply normal approximation to the binomial distributions in (3.39),

$$\frac{a_1+a-1-(n_1+a+b-1)\theta_1}{\sqrt{(n_1+a+b-1)\theta_1(1-\theta_1)}} < \frac{a_1-n_1\theta_1}{\sqrt{n_1\theta_1(1-\theta_1)}} \tag{3.40}$$

(3.40) can be simplified as follows

$$\frac{a_1-1+a}{\sqrt{n_1+a+b-1}}-\theta_1\sqrt{n_1+a+b-1} < \frac{a_1}{\sqrt{n_1}}-\theta_1\sqrt{n_1}$$

$$\frac{a_1-1+a}{\sqrt{n_1+a+b-1}}-\frac{a_1}{\sqrt{n_1}} < (\sqrt{n_1+a+b-1}-\sqrt{n_1})\theta_1$$

Divide by $\sqrt{n_1(n_1+a+b-1)}$

$$\frac{1}{\sqrt{n_1}}\frac{a_1-1+a}{n_1+a+b-1}-\frac{1}{\sqrt{n_1+a+b-1}}\frac{a_1}{n_1} < \left(\frac{1}{\sqrt{n_1}}-\frac{1}{\sqrt{n_1+a+b-1}}\right)\theta_1 \tag{3.41}$$

In practice,

$$\frac{a_1}{n_1}\approx\theta_0 \tag{3.42}$$

The following conservative condition satisfies (3.41)

$$\frac{a_1-1+a}{n_1+a+b-1} < \frac{a_1}{n_1}$$

This formula can be simplified as

$$(a_1 - 1 + a)n_1 < (n_1 + a + b - 1)a_1$$

$$a < \frac{a_1(n_0 - 1)}{n_1} + 1 \qquad (3.43)$$

Apply (3.42) into (3.43), the condition is

$$a < (n_0-1)\theta_0 + 1 \qquad (3.44)$$

The condition (3.44) is a sufficient condition to (3.39), hence it is a sufficient condition to (3.37). But it is not a necessary condition to (3.37) since for some $a \geq (n_0-1)\theta_0 + 1$, (3.37) is still true. For example, when $n_0=12$, $\theta_0=0.2$, $\theta_1=0.4$ and $n_1=24$, $a1=6$, $r1=9$, if $a = 5 > (n_0-1)\theta_0 + 1 = 3$, (3.37) is still true. This design is included in Section 3.9 with the prior beta(5,7).

3.7.2.3 Existence of a sufficient prior condition satisfying Bayesian error rates < frequentist error rate in stage I

Comparing the sufficient conditions (3.36) and (3.44), they conflict to each other. One requires the prior parameter $a > (n_0-1)\theta_1$, but the other requires $a < (n_0-1)\theta_0 + 1$. This conflict may be due to too conservative consideration in simplifying the Bayesian Type I and Type II error rates in binomial form as the right side of (3.30) and (3.39).

3.8 Examples

Based on the algorithm described in Section 3.6, we developed R programs to perform the Bayesian-frequentist two-stage design for single-arm phase II clinical trials. This section presents numerical examples with various priors under the design parameters $\theta_0 = 0.2$, $\theta_1 = 0.4$, $P_L = 0.5$, $P_U = 0.90$, $P_T = 0.95$, frequentist Type I error rate $\alpha = 0.05$ and Type II error rate $\beta = 0.2$.

We also use these examples to investigate the roles of a beta prior to a two-stage clinical trial design. Two groups of priors are considered: one group is for pessimistic priors that set the mode or mean of beta prior equal to the maximum uninteresting response rate of θ_0; another group for optimistic priors, which set the minimum response rate of interest θ_1 as the prior mode or mean. For each group, two priors are used, one is for a weak beta prior with parameters $a + b = 1$, another one for a relatively strong beta prior with $a + b = 12$. That is, the following 4 beta priors are under considerations.

- Weak pessimistic prior: beta (0.2, 0.8) with mean = 0.2, variance = 0.08; $P(\theta \leq \theta_0) = 0.68$ and $P(\theta \geq \theta_1) = 0.21$
- Strong pessimistic prior: beta (3, 9) with mode = 0.2, variance = 0.014; $P(\theta \leq \theta_0) = 0.38$ and $P(\theta \geq \theta_1) = 0.12$
- Weak optimistic prior: beta (0.4, 0.6) with mean = 0.4, variance = 0.12; $P(\theta \leq \theta_0)=0.41$ and $P(\theta \geq \theta_1)=0.44$
- Strong optimistic prior: beta (5, 7) with mode = 0.4, variance = 0.018; $P(\theta \leq \theta_0)=0.05$ and $P(\theta \geq \theta_1)=0.53$

In each situation, we identify an optimal design that gives the smallest expected sample size under the Bayesian framework ($EN._b$ as denoted in the tables given below).

a) Examples

a.1) Designs obtained with beta prior (0.2, 0.8)

```
     a1 r1 n1  r  n Type1.f Type2.f Type1.b Type2.b PET(θ0)  EN(θ0)  PET(θ1)  EN(θ1)  Post.θ0 Post.θ1  PET.b  EN.b

[1,]  8 11 27 12 39   0.044   0.197   0.023   0.046   0.937    27.8    0.725    30.3   0.0646  0.0996   0.966  27.4

[2,]  7 10 25 13 40   0.038   0.193   0.017   0.054   0.908    26.4    0.729    29.1   0.0371  0.1430   0.962  25.6

[3,]  5  8 19 13 41   0.048   0.199   0.017   0.037   0.860    22.1    0.675    26.1   0.0450  0.1180   0.950  20.1

[4,]  8 11 27 13 42   0.038   0.199   0.018   0.042   0.937    27.9    0.725    31.1   0.0540  0.0967   0.966  27.5
     a1 r1 n1  r  n Type1.f Type2.f Type1.b Type2.b PET(θ0)  EN(θ0)  PET(θ1)  EN(θ1)  Post.θ0 Post.θ1  PET.b  EN.b

[5,]  7 10 25 13 43   0.050   0.170   0.022   0.030   0.908    26.7    0.729    29.9   0.0641  0.0786   0.962  25.7

[6,]  5  8 19 14 44   0.043   0.198   0.014   0.034   0.860    22.5    0.675    27.1   0.0377  0.1139   0.950  20.2
```

[7,]	5	8	19	14	45	0.047	0.191	0.015	0.028	0.860	22.6	0.675	27.4	0.0453	0.0938	0.950	20.3
[8,]	8	11	27	13	46	0.049	0.188	0.025	0.019	0.937	28.2	0.725	32.2	0.1017	0.0407	0.966	27.6
[9,]	6	9	22	15	47	0.036	0.191	0.011	0.034	0.887	24.8	0.704	29.4	0.0317	0.1099	0.957	23.1

Note: The bolded in blue is the optimal design.

Notation:

- Type1.f, Type2.f: Frequentist Type I and Type II error rate.
- Type1.b, Type2.b: Bayesian Type I and Type II error rate.
- PET(θ_0), PET(θ_1): Frequentist PET (probability of early termination) under null and alternative hypothesis.
- EN(θ_0), EN(θ_1): Frequentist expected sample size under null and alternative hypothesis.
- Post.θ_0, Post.θ_1: Posterior probability of $\theta<\theta_0$ and $\theta>\theta_1$ given the trial data (s,n).
- PET.b: Bayesian PET
- EN.b: Bayesian expected sample size.

a.2) Designs obtained with beta prior (3, 9)

	a1	r1	n1	r	n	Type1.f	Type2.f	Type1.b	Type2.b	PET(θ_0)	EN(θ_0)	PET(θ_1)	EN(θ_1)	Post.θ_0	Post.θ_1	PET.b	EN.b
[1,]	8	11	27	12	39	0.044	0.197	0.030	0.035	0.937	27.8	0.725	30.3	0.0607	0.0540	0.869	28.6
[2,]	8	11	27	12	40	0.047	0.193	0.032	0.030	0.937	27.8	0.725	30.6	0.0707	0.0436	0.869	28.7
[3,]	7	10	25	13	41	0.042	0.183	0.027	0.033	0.908	26.5	0.729	29.3	0.0436	0.0648	0.852	27.4
[4,]	9	12	30	13	42	0.037	0.189	0.025	0.034	0.948	30.6	0.745	33.1	0.0512	0.0528	0.880	31.4
[5,]	8	11	27	13	43	0.041	0.195	0.027	0.029	0.937	28.0	0.725	31.4	0.0597	0.0429	0.869	29.1
[6,]	7	11	25	14	44	0.032	0.187	0.016	0.032	0.896	27.0	0.568	33.2	0.0367	0.0631	0.798	28.8
[7,]	**6**	**10**	**22**	**14**	**45**	**0.036**	**0.187**	**0.018**	**0.027**	**0.873**	**24.9**	**0.534**	**32.7**	**0.0432**	**0.0517**	**0.777**	**27.1**
[8,]	8	11	27	13	46	0.049	0.188	0.031	0.021	0.937	28.2	0.725	32.2	0.0908	0.0222	0.869	29.5
[9,]	8	11	27	14	47	0.038	0.193	0.024	0.025	0.937	28.3	0.725	32.5	0.0585	0.0341	0.869	29.6

Note: The bolded in blue is the optimal design.

a.3) Designs obtained with beta prior (0.4, 0.6)

	a1	r1	n1	r	n	Type1.f	Type2.f	Type1.b	Type2.b	PET(θ_0)	EN(θ_0)	PET(θ_1)	EN(θ_1)	Post.θ_0	Post.θ_1	PET.b	EN.b
[1,]	7	10	25	12	37	0.043	0.193	0.017	0.075	0.908	26.1	0.729	28.3	0.0379	0.1654	0.946	25.6
[2,]	7	10	25	12	38	0.048	0.183	0.019	0.062	0.908	26.2	0.729	28.5	0.0463	0.1366	0.946	25.7
[3,]	8	11	27	12	39	0.044	0.197	0.019	0.060	0.937	27.8	0.725	30.3	0.0559	0.1120	0.951	27.6
[4,]	7	10	25	13	40	0.038	0.193	0.013	0.068	0.908	26.4	0.729	29.1	0.0317	0.1585	0.946	25.8
[5,]	**5**	**8**	**19**	**13**	**41**	**0.048**	**0.199**	**0.013**	**0.050**	**0.860**	**22.1**	**0.675**	**26.1**	**0.0388**	**0.1315**	**0.930**	**20.5**
[6,]	6	9	22	13	42	0.049	0.183	0.015	0.045	0.887	24.3	0.704	27.9	0.0468	0.1083	0.939	23.2
[7,]	7	10	25	13	43	0.050	0.170	0.018	0.040	0.908	26.7	0.729	29.9	0.0559	0.0886	0.946	26.0
[8,]	5	8	19	14	44	0.043	0.198	0.010	0.046	0.860	22.5	0.675	27.1	0.0325	0.1266	0.930	20.8
[9,]	5	8	19	14	45	0.047	0.191	0.012	0.040	0.860	22.6	0.675	27.4	0.0392	0.1048	0.930	20.8
[10,]	5	8	20	15	46	0.046	0.174	0.009	0.045	0.836	24.3	0.710	27.5	0.0224	0.1458	0.932	21.8

Note: The bolded in blue is the optimal design.

a.4) Designs obtained with beta prior (5, 7)

	a1	r1	n1	r	n	Type1.f	Type2.f	Type1.b	Type2.b	PET(θ_0)	EN(θ_0)	PET(θ_1)	EN(θ_1)	Post.θ_0	Post.θ_1	PET.b	EN.b
[1,]	**6**	**9**	**24**	**13**	**38**	**0.047**	**0.181**	**0.004**	**0.172**	**0.847**	**26.1**	**0.768**	**27.2**	**0.0049**	**0.2724**	**0.832**	**26.4**
[2,]	7	10	29	15	39	0.050	0.195	0.002	0.182	0.840	30.6	0.842	30.6	0.0009	0.4465	0.865	30.4
[3,]	6	9	24	14	42	0.048	0.160	0.003	0.147	0.847	26.7	0.768	28.2	0.0053	0.2259	0.832	27.0
[4,]	7	10	29	17	43	0.049	0.197	0.002	0.181	0.840	31.2	0.842	31.2	0.0004	0.4927	0.865	30.9

```
 [5,]  5  9 25 16 44  0.049  0.177  0.002  0.125  0.663  31.4  0.756  29.6  0.0014 0.3429  0.781  29.2
 [6,]  3  7 17 16 46  0.045  0.172  0.002  0.128  0.587  29.0  0.599  28.6  0.0024 0.2691  0.682  26.7
 [7,]  4  8 21 17 47  0.046  0.185  0.002  0.128  0.629  30.6  0.687  29.1  0.0012 0.3271  0.740  27.8
 [8,]  5  9 25 17 48  0.050  0.152  0.002  0.107  0.663  32.7  0.756  30.6  0.0016 0.2907  0.781  30.0
 [9,]  4  8 21 18 51  0.046  0.160  0.002  0.110  0.629  32.1  0.687  30.4  0.0014 0.2776  0.740  28.8
[10,]  5  9 25 19 52  0.048  0.167  0.002  0.112  0.663  34.1  0.756  31.6  0.0007 0.3337  0.781  30.9
```

Note: The bolded in blue is the optimal design.

b) Summary of characteristics of the optimal Bayesian-frequentist two-stage designs

As described in Section 3.7, in this thesis, the optimal Bayesian-frequentist two-stage design is defined as the design with minimal expected sample size under the Bayesian framework. The following is the summary of characteristics of the four aforementioned optimal designs.

Table 3-1 Characteristics of the optimal Bayesian-frequentist two-stage designs

Beta prior	a1	r1	n1	r	n	Type1.f	Type2.f	Type1.b	Type2.b	PET(θ_0)	EN(θ_0)	PET(θ_1)	EN(θ_1)	Post.θ_0	Post.θ_1	PET.b	EN.b
Beta(0.2, 0.8)	5	8	19	13	41	0.048	0.199	0.017	0.037	0.860	22.1	0.675	26.1	0.0450	0.1180	0.950	20.1
Beta(3, 9)	6	10	22	14	45	0.036	0.187	0.018	0.027	0.873	24.9	0.534	32.7	0.0432	0.0517	0.777	27.1
Beta(0.4, 0.6)	5	8	19	13	41	0.048	0.199	0.013	0.050	0.860	22.1	0.675	26.1	0.0388	0.1315	0.930	20.5
Beta(5, 7)	6	9	24	13	38	0.047	0.181	0.004	0.172	0.847	26.1	0.768	27.2	0.0049	0.2724	0.832	26.4

b.1) Frequentist properties

Total sample size

For the four optimal designs under the different beta priors: weak pessimistic prior - beta(0.2, 0.8), strong pessimistic prior – beta(3,9), weak optimistic prior – beta (0.4, 0.6), and strong optimistic prior – beta(5, 7), the respective total sample size is 41, 45, 41 and 38. Although the differences among these total sample sizes are not large, this example shows the roles of a beta prior to a two-stage sing-arm clinical trial design: 1) under weak priors, the total sample sizes are very similar. For the examples presented here, both weak priors beta(0.2, 0.8) and beta(0.4, 0.6) result in the same total sample size 41; 2) under the strong pessimistic prior beta(3, 9), the total sample size is 45, which is the largest; 3) under the strong optimistic prior beta(5, 7), the total sample size is 38, which is the smallest.

The above findings from the examples are intuitive. Weak priors result in the similar total sample sizes since these priors provide less prior information on the response rate of the study therapy. When a strong pessimistic prior is used, in order to claim efficaciousness of the study therapy, more data need to be obtained to overweigh the skeptical opinion from the pessimistic prior. On the other hand, when a strong optimistic prior is used, less total sample size is required.

Expect sample size and early termination probability (PET)

For the four optimal designs presented in Table 3-1, although the weak priors lead to the same expected sample sizes EN(θ_0) and EN(θ_1), in general, the sizes should be similar since weak priors only provide little prior information for a trial design.

Under strong beta priors, the strong optimistic prior beta(5, 7) result in a larger expected sample size of 26.1 compared to the size of 24.9 by the strong pessimistic prior beta (3, 9) under the null hypothesis. This is due to less likely the trial is to be terminated early under the null hypothesis if the optimistic prior belief is true (PET = 0.847 vs 0.873 for beta(5, 7) vs beta(3, 9)). Under the alternative hypothesis, it is more likely that the trial could be terminated early for efficacy if the optimistic prior belief is true (PET = 0.768 vs 0.543 for beta(5, 7) vs beta(3, 9)), therefore the strong optimistic prior beta (5,7) result in a smaller expected sample size of 27.2 compare to the size of 32.7 by the strong pessimistic prior beta (3, 9).

Summary of the examples

In summary, the trial design characteristics under weak priors are very similar. Compared to a strong pessimistic prior, a strong optimistic prior with same “strength” (e.g a + b is same for both priors) result in a smaller expected sample size under the alternative hypothesis, a larger expected size under the null hypothesis, a smaller total sample size, a lower probability of early termination under the null hypothesis, and a higher probability of early termination under the alternative hypothesis.

b.2) Bayesian properties

Bayesian error rates

The four single-arm phase II clinical trial designs summarized in Table 3-1 are controlled under frequentist Type I error rate $\alpha = 0.05$ and Type II error rate $\beta = 0.2$. For these four designs, the Bayesian Type I error rate ranges from 0.004 to 0.018, and the Bayesian Type II error rate ranges from 0.027 to 0.172. Both Bayesian Type I and Type II error rates are lower than the corresponding frequentist Type I and Type II error rates. However, as discussed in Section 3.7.2, we have not found a sufficient condition for setting a beta prior so that the Bayesian Type I and Type II error rates are less than the corresponding frequentist Type I and Type II error rates.

Posterior probability

For the four designs, posterior probability of $\theta<\theta_0$ given hypothetical trial data (s, n) range from 0.0049 to 0.0450. With respect to posterior probability of $\theta>\theta_1$ given hypothetical trial data (s, n), the optimistic priors beta (0.4, 0.6) and beta(5,7) result in the posterior probabilities of 0.1315 and 0.2724, which are higher than the posterior probability of 0.1180 obtained with the weak pessimistic prior beta(0.2, 0.8), and much higher than the posterior probability of 0.0517 from the strong pessimistic prior beta(3, 9). This finding is consistent with what theoretically demonstrated in Section 3.7.1.

Summary of the examples

In summary of the examples, for the clinical trials designed by controlling frequentist Type I and Type II error rates, an optimistic prior result in a higher poster probability of $\theta>\theta_0$ or $\theta>\theta_1$ than a pessimistic prior with the same size of a + b; and the stronger the optimistic prior, the higher the posterior probability is.

3.9 Comparisons with typical frequentist and Bayesian single-arm phase II clinical trial designs

To demonstrate the performance of our Bayesian-frequentist design, in this section, we compare our Bayesian-frequentist two-stage design with the following typical frequentist and Bayesian single-arm phase II clinical trial designs.

a. Bayesian predictive probability continuous monitoring design with early acceptance boundaries (Lee & Liu, 2008)

b. Frequentist two-stage design with early rejection and acceptance boundaries

    - Fleming's design (1982)
    - Chang's design (1987)
    - Shuster's design (2002)

c. Frequentist two-stage design with an early acceptance boundary

    - Simon's optimal design (1989)
    - Simon's minimax design (1989)

d. Frequentist single stage design

In order to have fair comparisons, we evaluate these designs under the same design parameters: $\theta_0 = 0.2$, $\theta_1 = 0.4$, Type I error rate $\alpha = 0.05$ and Type II error rate $\beta = 0.2$. For Bayesian designs, the four beta priors as described in Section 3.8 are used. The operating characteristics of these designs are summarized in Table 3-2. For

Lee & Liu's Bayesian predictive probability continuous monitoring design (Lee & Liu, 2008), the trial early acceptance boundaries are only listed for the first stage and the last stage. Figure 3-1 though Figure 3-5 present probability of rejecting H_0, probability of accepting H_0, probability of early termination (PET), expected sample size, and expected sample size per correct decision for each design if applicable.

Table 3-2 **Comparisons of single-arm phase II designs for θ_0=0.2, θ_1 = 0.4, α = 0.05 and β = 0.2**

Bayesian-frequentist two-stage design (BF) **P_L= 0.5, P_U = 0.90, P_T = 0.95**	a1	r1	n1	r	n
Prior beta(0.2, 0.8)	5	8	19	13	41
Prior beta(3, 9)	6	10	22	14	45
Prior beta(0.4, 0.6)	5	8	19	13	41
Prior beta(5, 7)	6	9	24	13	38
Bayesian predictive probability continuous monitoring design with early acceptance boundary Lee & Liu, 2008) **P_L=0.01, P_U=1, P_T=0.95**					
Prior beta(0.2, 0.8)	1	11	NA	13	36
Prior beta(3, 9)	1	12	NA	13	36
Prior beta(0.4, 0.6)	1	11	NA	13	36
Prior beta(5, 7)	0	13	NA	13	36
Frequentist two-stage design with early an acceptance boundary					
Simon's optimal design	3	NA	13	13	43
Simon's minimax design	4	NA	18	11	33
Frequentist two-stage design with early acceptance and rejection boundaries					
Fleming's design	4	9	20	12	35
Chang's design	7	9	25	17	50
Shuster's design	5	8	20	13	39
Single-arm				12	35

Probability of rejecting and accepting H_0

As shown in Figure 3-1, the probability of rejecting H_0 is almost same for all these designs. This is due to these designs are performed by controlling Type I error rates at the same level of α = 0.05. Similarly, the probability of accepting H_0 (Figure 3-2) is almost same for all these designs.

Probability of early termination

With respect to the probability of early termination shown in Figure 3-3, Chang's design has highest early termination probabilities for futility or efficacy. This is due to the fact that Chang's design requires the largest total samples size and the largest number of responses to stop a trial early for futility, and requires the smallest number of responses to stop a trial early for efficacy. The requirement on number of responses for early termination by Fleming's design is exactly the opposite to Chang's design, hence, Fleming's design has the lowest early termination probabilities for futility or efficacy. For our Bayesian-frequentist design, in general under a weak prior, its probability of early termination is comparable to Shuster's design, which is lower than Chang's design, but higher then Fleming's design. Under a strong optimistic prior, the probability of early termination for efficacy by our design is higher than Shuster's design, but lower than Shuster's design. This is due to the fact that our design under a strong optimistic prior requires less responses to stop a trial early for efficacy.

For the designs with early acceptance boundary, Lee & Liu's design seems not sensitive to priors. Under the four priors, the design's operating characteristics are almost same. However, in general, its probability of early termination for futility is higher than Simon's design. This result is expected since Lee & Liu's design monitors the trial for each new patient, therefore, there are more chances to terminate a trial with Lee & Liu's design than a two-stage trial.

Expected sample size

As shown in Figure 3-4, Fleming's design and Chang's design require larger expected sample size than Shuster's design. This is due to Shuster's design is a minimax design, which has the smallest globally minimized expected sample size under the frequentist framework. Our Bayesian-frequentist design under a weak prior is comparable to Shuster's design. However, under a strong optimistic prior, the expected sample size around θ_1 (for the example, around $\theta = 0.35 \sim 0.55$) is lower than that for a design under a strong pessimistic prior.

Not surprisingly, expected sample size from Simon's optimal design is the lowest around θ_0, but is the largest when $\theta>\theta_1$. For Lee & Liu's design, a prior does not have much impact on the expected sample size when $\theta>\theta_1$, but the expected size is larger around θ_0 when a strong optimistic prior is used.

Expected sample size per correct decision

Following Johnson and Cook (2009), we define expected sample size per correct decision as the ratio of the expected sample size to the probability of correctly stopping the trial. In these examples, we assume that it is correct to accept H_0 when $\theta \leq 0.3$, which is $(\theta_0 + \theta_1)/2$. As shown in Figure 3-5, our Bayesian-frequentist design under a weak prior has the smallest expect sample size per correct decision, and is comparable to Shuster's design. When a strong optimistic prior is used, lower expected sample size per correct decision is needed compared to the design with a strong pessimistic prior when $\theta = 0.3$ to 0.55. Whereas, when $\theta \leq \theta_0$, slightly higher expected sample size per correct decision is required for the design with a strong optimistic prior compared to the design with a strong pessimistic prior. These results explicitly imply the impact of a prior to decision-making.

Lee & Liu's design has highest expected sample size per correct decision, and there is no much impact from a prior. This is not surprising as explained in the previous sections regarding expected sample size and probability of early termination. When response rate is around $\theta>\theta_1$, the designs with only early acceptance boundary including Simon's design and Lee & Liu's design, their expected sample size per correct decision are larger than that by designs with both early rejection and acceptance boundaries. This is due to the fact that the decision of rejecting H_0 is more likely correct when true θ is great than θ_1.

Summary of the comparisons

In summary, we have observed the following from the comparisons between our Bayesian-frequentist design and typical Bayesian and frequentist designs with the examples described previously: 1) Our two-stage design for single-arm phase II clinical trials is comparable to Shuster's minimax design when a weak prior is used; 2) In our design, a strong optimistic prior results in a smaller total sample size and higher posterior probabilities. In contrast, Lee & Liu's predictive probability continuous monitoring design has limited advantages of incorporating prior information.

Figure 3-1 Probability of rejecting H_0

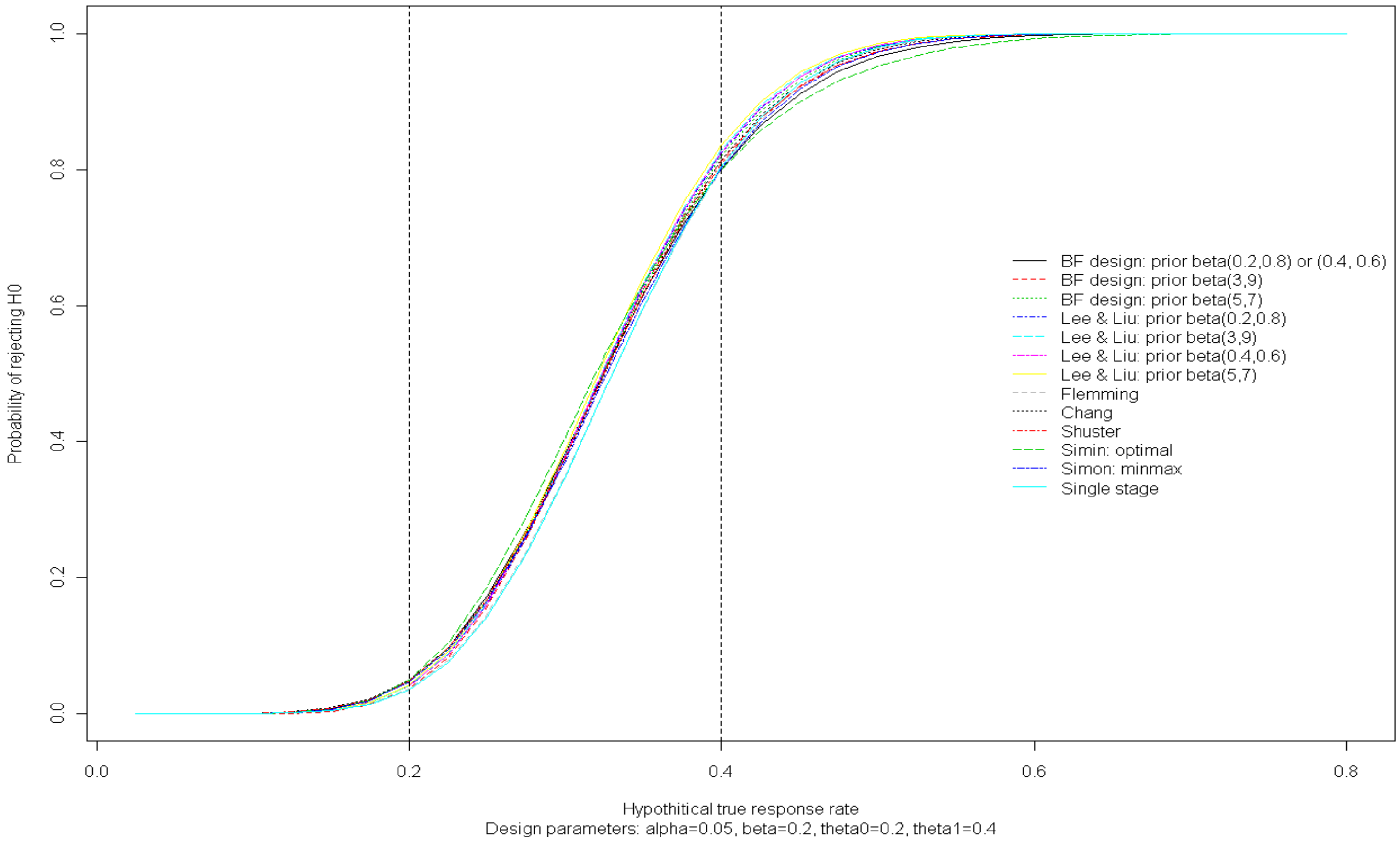


Figure 3-2 Probability of accepting H_0

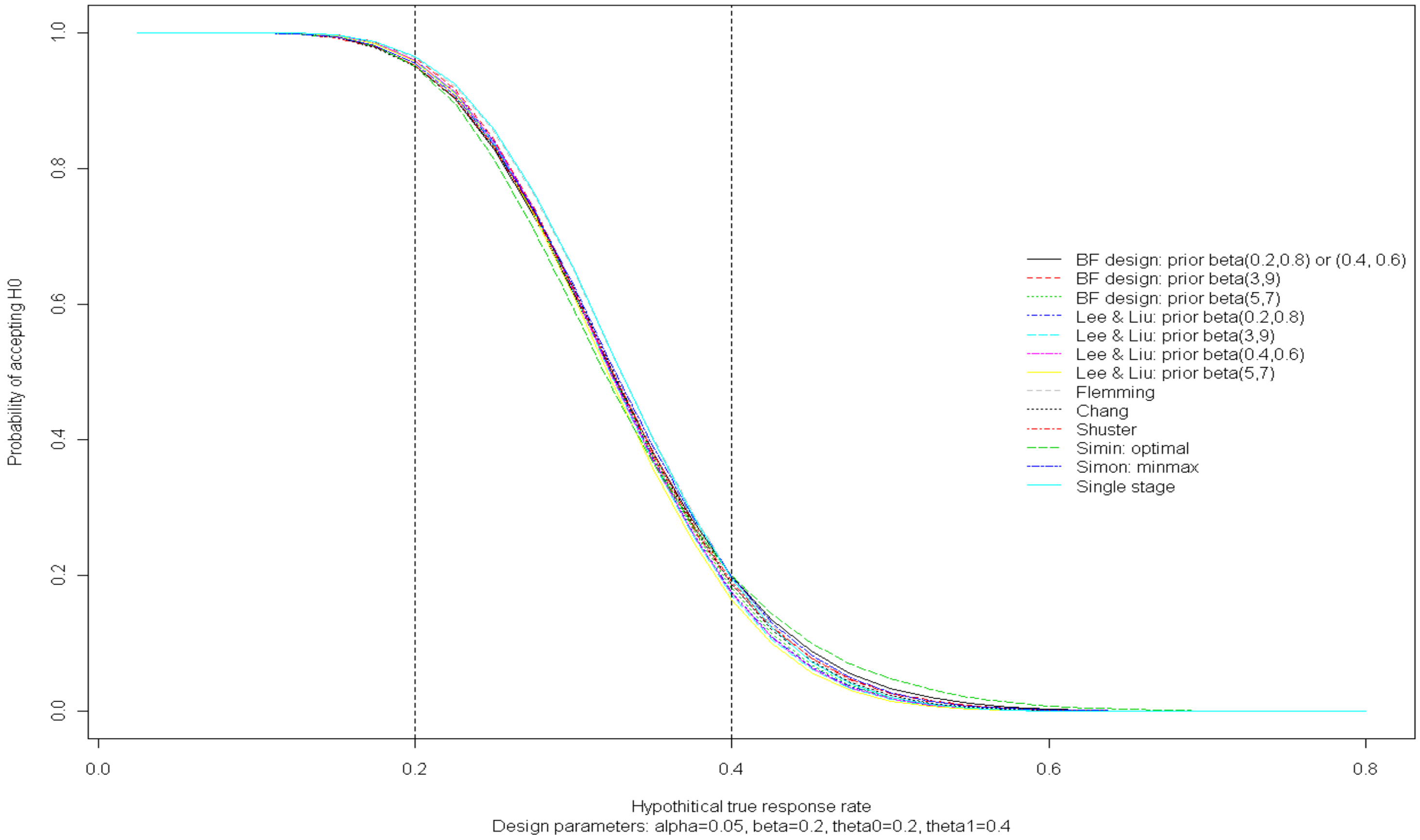


Figure 3-3 Probability of early termination (PET)

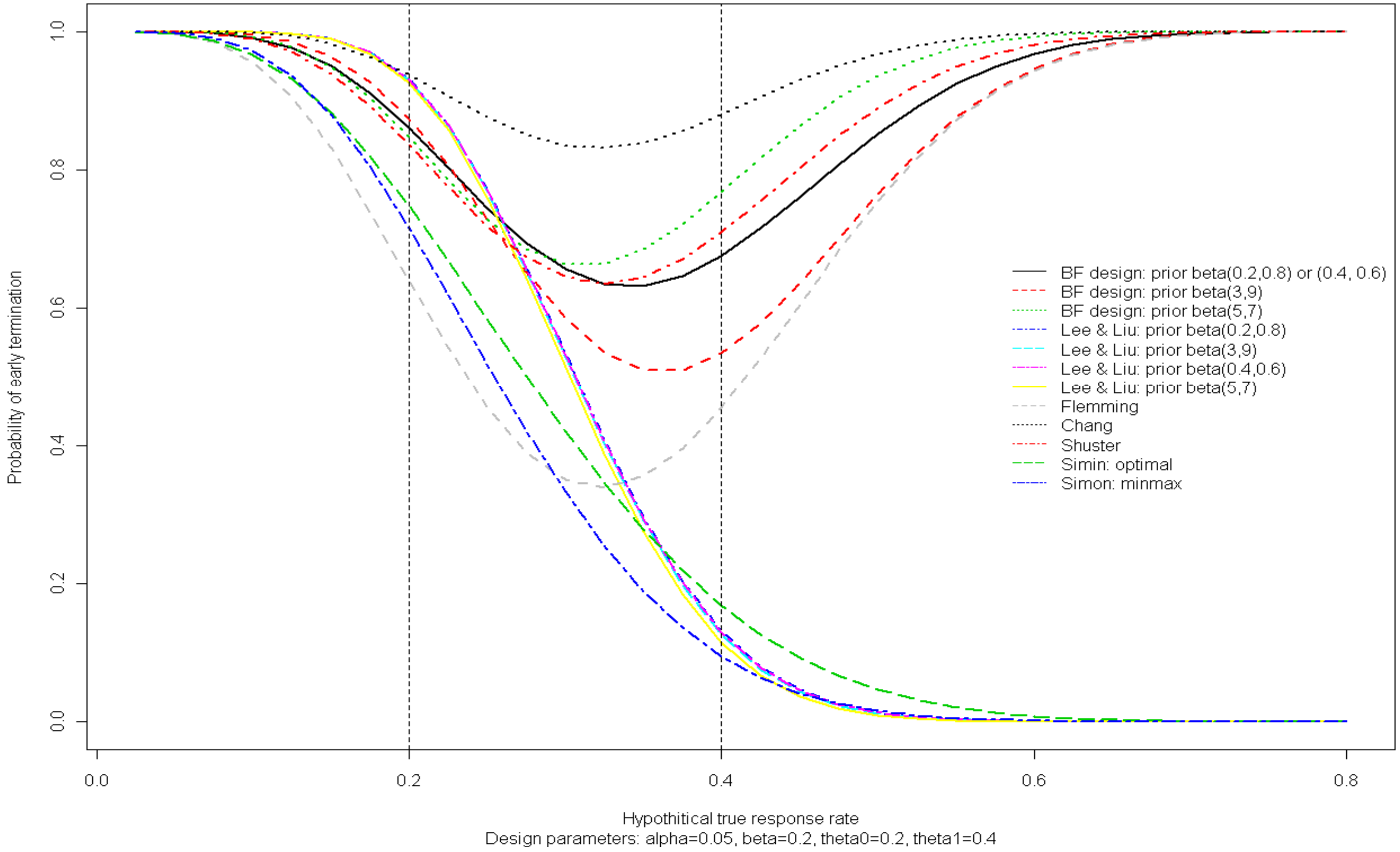


Figure 3-4 Probability of expected sample size

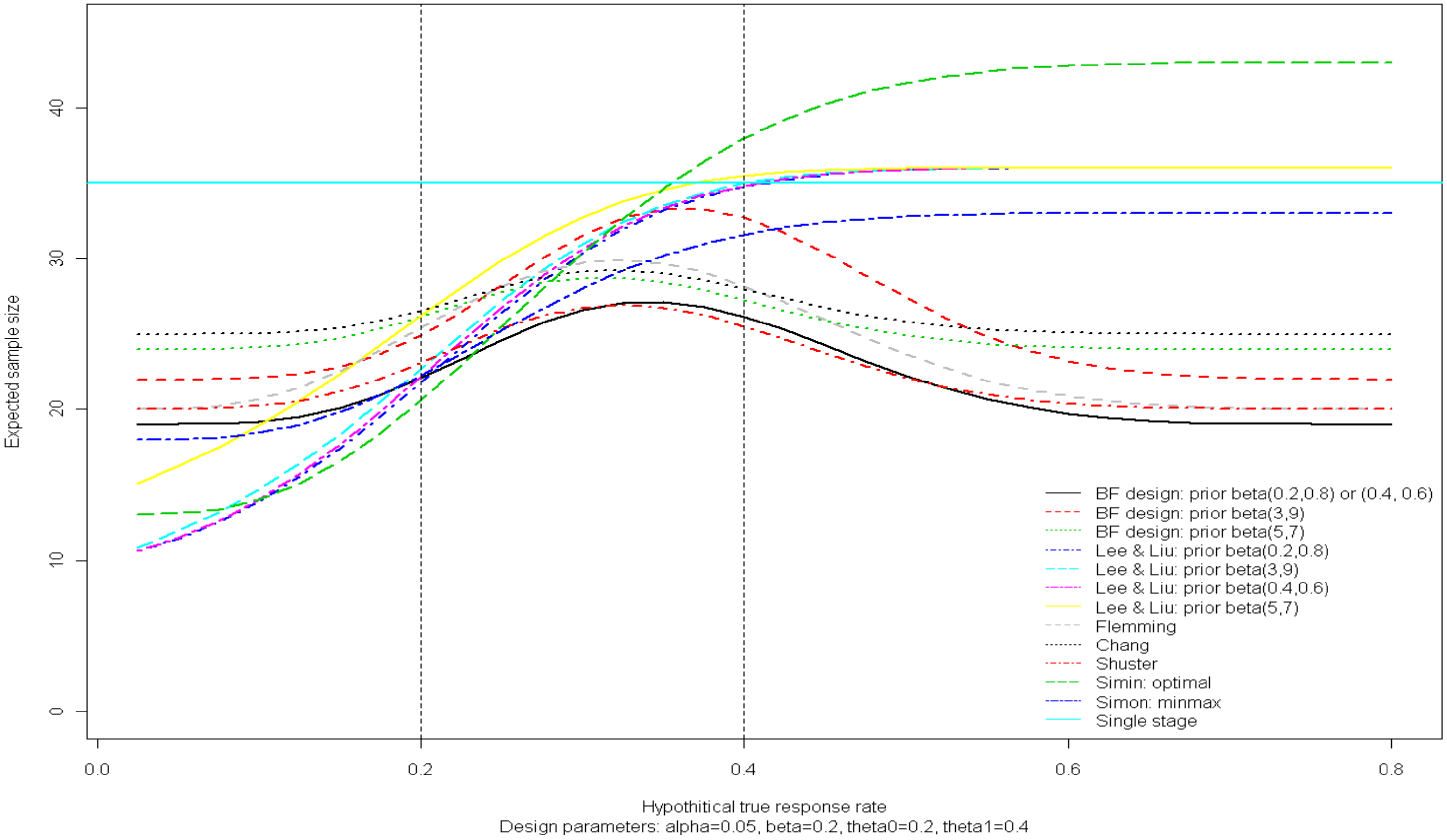
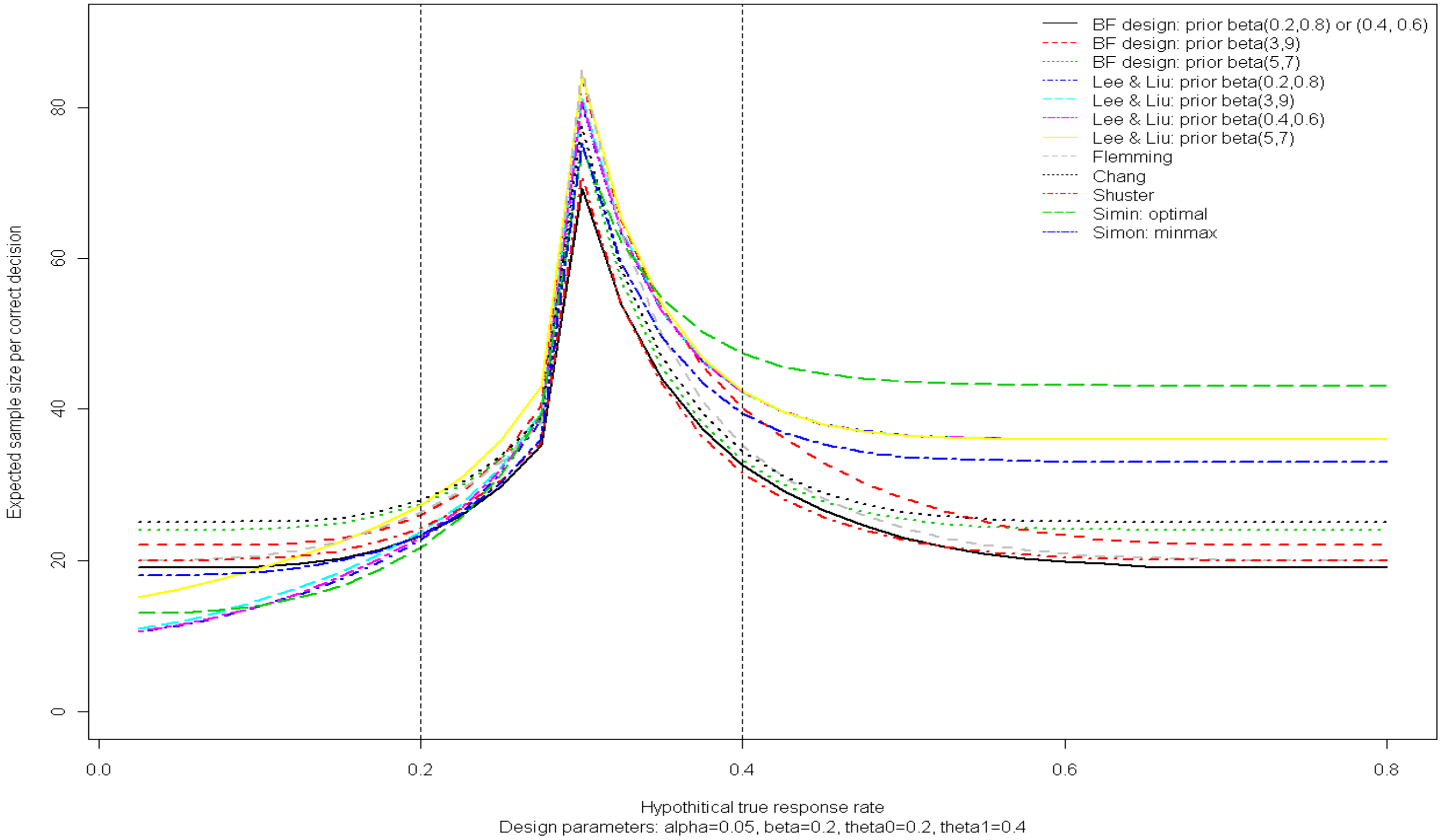


Figure 3-5 Expected sample size per correct decision



4 A Bayes factor-based two-stage design using an iMOM prior for single-arm phase II clinical trials

In Chapter 3, we presented our Bayesian-frequentist two-stage design for single-arm phase II clinical trials. The hypotheses tested in that design are $\theta \leq \theta_0$, vs. H_1: $\theta \geq \theta_1$, which are based on the frequentist setting. Under the Bayesian framework, the alternative hypothesis is not a simple negation of the null hypothesis. Instead, the Bayesian hypothesis testing requires parametric sampling density for data and prior density on model parameters. However, it is always challenging to have a prior density on model parameters. One could use a vague prior, but others could argue with an objective prior (e.g. optimistic or skeptical prior). Therefore, mis-specification of model parameter is a concern. Johnson and Rossell (2009, 2010) pointed out that conventionally Bayesian tests use local alternative priors, which assign positive probability to the regions of the parameter space that are consistent with the null hypothesis. Therefore, these tests provide exponential accumulation of evidence in favor of the true alternative hypothesis, but only sub-linear accumulation of evidence in favor of true null hypothesis. However, inverse moment (iMOM) priors provide approximately linear convergence for the logarithm of the Bayes factor under both null and alternative hypothesis. By using an iMOM prior, Johnson and Cook (2009) developed a Bayes factor-based continuous monitoring approach for single-arm phase II clinical trials. They argued that by using the Bayes factor, mis-specification of prior density from the alternative model in a single-arm phase II clinical trial setting can only decrease the expected weight of evidence in favor of the alternative model, hence the more severely the alternative model deviates from the true parameter model, the more penalty the Bayes factor-based hypothesis testing would pay, no matter whether the prior density model is optimistic or skeptical as long as it is mis-specified.

Taking advantages of the Bayes factor, in this chapter we discuss our Bayes factor-based two-stage single-arm phase II clinical trial design.

4.1 Bayes factor and weight of evidence

Suppose θ is the parameter of the response rate of the study treatment, and $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1 = \Theta - \Theta_0$ are uninteresting response rate and minimum response rate of the study treatment under null (H_0) and alternative hypothesis (H_1) respectively. The marginal distribution function $m_k(x)$ of the data X under the hypothesis H_k (k = 0, 1) can be defined as in (4.1), where $P_k(x|\theta)$ is the likelihood function and $\pi_k(\theta)$ is the prior distribution for θ under the hypothesis H_k.

$$m_k(x) = \int_{\theta \in \Theta_k} P(x \mid \theta, H_k) \pi(\theta \mid H_k) d\theta$$

$$= \int_{\theta \in \Theta_k} P_k(x \mid \theta) \pi_k(\theta) d\theta, \quad \text{k} = 0, 1 \qquad (4.1)$$

The Bayes factor BF_1 from the alternative hypothesis H_1 against the null hypothesis H_0 is defined as follows, which is the ratio of the marginal densities under the hypotheses H_1 and H_0.

$$BF_1 = \frac{m_1(x)}{m_0(x)} \qquad (4.2a)$$

Suppose that $\pi_t(\theta)$ is the true distribution of θ, and the prior density $\pi_1(\theta)$ is incorrect. This means that θ is from the true distribution $\pi_t(\theta)$ rather than the distribution $\pi_1(\theta)$. Define $m_t(x)$ similarly using $\pi_t(\theta)$ as in (4.1), then the Bayes factor BF_t from the true distribution of θ against the null hypothesis H_0 is:

$$BF_t = \frac{m_t(x)}{m_0(x)} \tag{4.2b}$$

Logarithm of Bayes factor is called weight of evidence. The expected weights of evidence EWOE$_t$ and EWOE$_1$ from the true distribution $\pi_t(\theta)$ and the incorrect prior $\pi_1(\theta)$ against the null hypothesis H$_0$ are as follows.

$$EWOE_t = \int_X m_t(x)\log(BF_t)dx = \int_X m_t(x)\log(\frac{m_t(x)}{m_0(x)})dx \tag{4.3a}$$

$$EWOE_1 = \int_X m_t(x)\log(BF_1)dx = \int_X m_t(x)\log(\frac{m_1(x)}{m_0(x)})dx \tag{4.3b}$$

As demonstrated by Johnson and Cook (2009), based on Gibbs' inequality (see Appendix A.6 for the details), we can have the following

$$EWOE_t - EWOE_1 = \int_X m_t(x)\log[\frac{m_t(x)}{m_0(x)}]dx - \int_X m_t(x)\log[\frac{m_1(x)}{m_0(x)}]dx$$

$$= \int_X m_t(x)\log[\frac{m_t(x)}{m_1(x)}]dx \geq 0 \tag{4.4a}$$

Namely,

$$EWOE_t \geq EWOE_1 \tag{4.4b}$$

The above inequality implies that the expected weight of evidence against the null hypothesis H$_0$ from the true density $\pi_t(\theta)$ is always greater than that from an incorrect density $\pi_1(\theta)$. Moreover, $EWOE_t = EWOE_1$ only if $\pi_t(\theta) = \pi_1(\theta)$. Table 4-1 shows the classifications of weight of evidence (Kass and Raftery, 1995). A negative weight of evidence indicates the evidence in favor of H$_0$.

Table 4-1 Classifications of weight of evidence (Kass and Raftery, 1995)

Evidence	Weight of evidence	
	Against H_0	In favor of H_0
Not worth more than a bare mention	0 to 1	-1 to 0
Positive	1 to 3	-3 to -1
Strong	3 to 5	-5 to -3
Very strong	>5	< -5

To demonstrate the property with respect to the mis-specification of priors for an alternative model as described in (4.4a) and (4.4b), let's consider the number of responses X following a binomial distribution with the parameter θ and sample size n, and the prior distributions $\pi_0(\theta) = I(\theta = \theta_0)$, $\pi_1(\theta) = I(\theta = \theta_1)$ and $\pi_t(\theta) = I(\theta = \theta_t)$, where I(θ) is an indicator function. For example, $\pi_0(\theta) = 1$ if $\theta = \theta_0$, otherwise $\pi_0(\theta) = 0$. Suppose $\theta_0 = 0.2$, $\theta_1 = 0.4$, $\theta_t = 0.3$ and sample size n =100, the expected weights of evidence $EWOE_t$ and $EWOE_1$ are:

$$EWOE_t = \int_X m_t(\theta)\log(BF_t)dx$$

$$= \sum_{x=0}^{n} \binom{n}{x} \theta_t^{\,x}(1-\theta_t)^{n-x} \log\left(\left(\frac{\theta_t}{\theta_0}\right)^x \left(\frac{1-\theta_t}{1-\theta_0}\right)^{n-x}\right) = 2.82$$

$$EWOE_1 = \int_X m_t(\theta)\log(BF_1)dx$$

$$= \sum_{x=0}^{n} \binom{n}{x} \theta_t^{\,x}(1-\theta_t)^{n-x} \log\left(\left(\frac{\theta_1}{\theta_0}\right)^x \left(\frac{1-\theta_1}{1-\theta_0}\right)^{n-x}\right) = 0.66$$

Therefore, as theoretically described in (4.4a) and (4.4b), the expected weight of evidence against the null hypothesis H_0 can not be increased by assigning an overly optimistic prior for the parameter of interest, which is the treatment effect under the framework of single-arm clinical trials.

4.2 iMOM prior

4.2.1 iMOM prior

One of the main advantages of Bayesian clinical trial design is to consider external information or an expert opinion as a prior for a trial design. However, sometimes it is hard to obtain a reasonable prior. For binary response, historically, conjugate beta priors are widely used. However, such priors as local alternative probability models (see definition in (4.6) in this section), as Johnson and Cook (2009) pointed out, assign positive probability to the regions of the parameter space that are consistent with the null hypothesis.

For Bayesians, under the null and alternative hypotheses, the parameter θ follows a prior distribution (Johnson and Rossell, 2008).

$H_0: \theta \sim P(\theta|H_0)$, *(4.5a)*

where $P(\theta|H_0)>0$ for any $\theta \in \Theta_0$ and $P(\theta|H_0)=0$ for any $\theta \in \Theta - \Theta_0$

$H_1: \theta \sim P(\theta|H_1)$, *(4.5b)*

where $P(\theta|H_1)>0$ for any $\theta \in \Theta - \Theta_0$

For some ε>0 and ζ>0, if the prior $P(\theta|H_1)$ satisfies the form (4.6) as defined below, then this prior is a local alternative prior density. As indicated in (4.6), this prior assigns positive densities to the regions of the parameter space that are consistent with the null hypothesis.

$P(\theta|H_1) > \varepsilon$ for all $\theta \in \Theta$ such that $\inf_{\theta_0 \in \Theta_0} |\theta - \theta_0| < \zeta$ *(4.6)*

However, for every ε >0, if there exists ζ>0 such that the prior P(θ|H$_1$) satisfies (4.7), then this prior is a non-local alternative prior density. Based on (4.7), this prior has the property of assigning negligible densities to the regions of the parameter space that are consistent with the null hypothesis.

$$P(\theta|H_1) < \varepsilon \text{ for all } \theta \in \Theta \text{ such that } \inf_{\theta_0 \in \Theta_0} |\theta - \theta_0| < \zeta \qquad (4.7)$$

Johnson and Rossell (2008) defined an inverse moment (iMOM) prior density. By adding the normalization factor Q, we define the normalized iMOM prior density in (4.8). This iMOM prior is a non-local alternative prior density, which assigns no mass to the values of the parameter θ that are consistent with the null hypothesis.

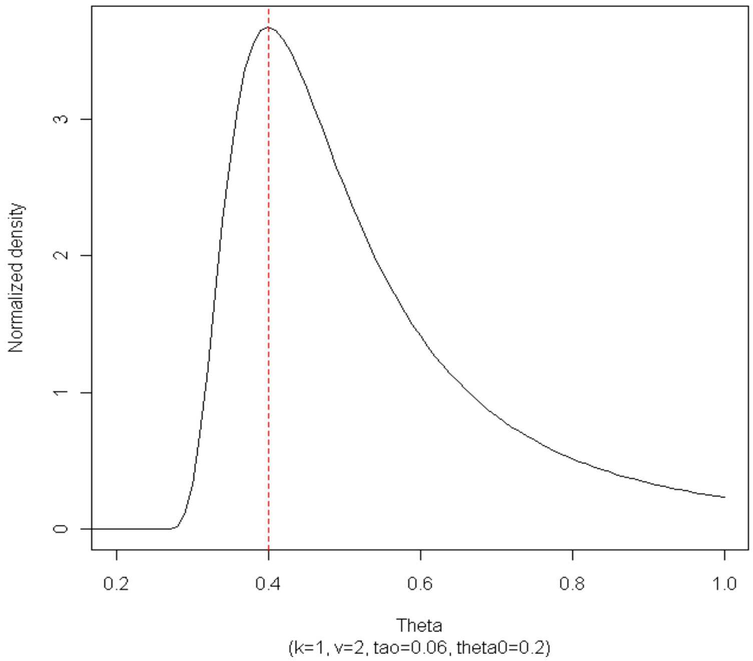
$$P_{iMOM}(\theta,\theta_0,k,v,\tau) = \begin{cases} \dfrac{1}{Q}\dfrac{k\tau^v}{\Gamma(v/2k)}\left[(\theta-\theta_0)^2\right]^{-\frac{v+1}{2}} \exp\left\{-\left[\dfrac{(\theta-\theta_0)^2}{\tau}\right]^{-k}\right\}, & \theta \in (\theta_0, 1.0) \\ 0, & elsewhere \end{cases} \qquad (4.8)$$

where k, v, and τ > 0. The normalization factor Q is calculated as

$$Q = \int_{\theta_0}^{1} \frac{k\tau^v}{\Gamma(v/2k)}\left[(\theta-\theta_0)^2\right]^{-\frac{v+1}{2}} \exp\left\{-\left[\frac{(\theta-\theta_0)^2}{\tau}\right]^{-k}\right\} d\theta$$

As θ approaches θ$_0$, the iMOM density is getting closer to 0. This is a property of non-local alternative prior density. The tail of an iMOM density is similar to the tail of a student’s t-distribution with v degrees of freedom (Johnson and Rossell, 2008). As an example, Figure 4-1 below shows an iMOM prior when k=1, v=2, τ = 0.06 and θ$_0$ =0.2.

Figure 4-1 iMOM prior density when k=1, v=2, τ = 0.06 and θ_0 =0.2



4.2.2 Mis-specification of iMOM prior

In this section, we use the iMOM prior to demonstrate the Bayes factor's property of mis-specification of priors as described in (4.4a) and (4.4b). Suppose that the number of responses X follows binomial(θ, n). Let's consider the following hypotheses.

H_0: $\theta \sim \pi_0(\theta) = I(\theta = \theta_0)$, $\theta_0 = 0.2$

H_1: $\theta \sim \pi_1(\theta) = P_{iMOM}(\theta, \theta_0, k, v, \tau)$, k = 1, v = 2 and τ = 0.06

Since the mass is condensed at θ_0 under the null hypothesis H_0, the Bayes factor BF_1 against the null hypothesis H_0 can be defined as follows.

$$BF_1 = \frac{\int_{\theta_0}^{1} \binom{n}{x} \theta^x (1-\theta)^{n-x} P_{iMOM}(\theta, \theta_0, k, v, \tau) d\theta}{\binom{n}{x} \theta_0^{\ x} (1-\theta_0)^{n-x}}$$

Let's consider the following three hypothetical true distributions for θ:

(1) $\pi_{t1}(\theta) = \pi_0(\theta) = I(\theta = \theta_0)$

(2) $\pi_{t2}(\theta) = \frac{2}{3}\pi_0(\theta) + \frac{1}{3}\pi_1(\theta)$

(3) $\pi_{t3}(\theta) = \frac{1}{3}\pi_0(\theta) + \frac{2}{3}\pi_1(\theta)$

The expected weight of evidence $EWOE_{1i}$ from the alternative hypothesis H_1 against the null hypothesis H_0 with respect to the hypothetical true distribution $\pi_{ti}(\theta)$ (i = 1, 2, 3) can be calculated as:

$$EWOE_{1i} = \int_X m_{ti}(x)\log(BF_1)dx$$

$$= \sum_{x=0}^{n}\left\{\left(\int_{\theta\in\pi_{ti}(\theta)}\binom{n}{x}\theta^x(1-\theta)^{n-x}\pi_{ti}(\theta)d\theta\right)\log\left(\frac{\int_{\theta_0}^{1}\binom{n}{x}\theta^x(1-\theta)^{n-x}P_{iMOM}(\theta,\theta_0,k,v,\tau)d\theta}{\binom{n}{x}\theta_0^{\ x}(1-\theta_0)^{n-x}}\right)\right\},$$

i = 1, 2, 3

The expected weight of evidence $EWOE_{ti}$ from the hypothetical true density $\pi_{ti}(\theta)$ (i = 1, 2, 3) against the null hypothesis H_0 is:

$$EWOE_{ti} = \int_X m_{ti}(x)\log(BF_{ti})dx$$

$$= \sum_{x=0}^{n}\left\{\left(\int_{\theta\in\pi_{ti}(\theta)}\binom{n}{x}\theta^x(1-\theta)^{n-x}\pi_{ti}(\theta)d\theta\right)\log\left(\frac{\int_{\theta\in\pi_{ti}(\theta)}\binom{n}{x}\theta^x(1-\theta)^{n-x}\pi_{ti}(\theta)d\theta}{\binom{n}{x}\theta_0^{\ x}(1-\theta_0)^{n-x}}\right)\right\},$$

i = 1, 2, 3

Suppose $\theta_0 = 0.2$ and $\theta_1 = 0.4$ (the mode of the iMOM prior). As presented in Table 4-2, the expected weights of evidence from H_1 against H_0 are -3.53, 1.37 and 6.28 when sample size n is 40 and the hypothetical true prior distributions for θ are $\pi_{t1}(\theta) = \pi_0(\theta) = I(\theta = \theta_0)$, $\pi_{t2}(\theta)= \frac{2}{3}\pi_0(\theta)+\frac{1}{3}\pi_1(\theta)$ and $\pi_{t3}(\theta)= \frac{1}{3}\pi_0(\theta)+\frac{2}{3}\pi_1(\theta)$ respectively. The corresponding expected weights of evidence are -8.26, 9.98 and 28.23 when sample size is 160. In summary of Table 4-2, we can see the following trends:

- The expected weight of evidence from H_1 against H_0 increases as hypothetical true prior for θ approaches $\pi_1(\theta)$.
- The expected weight of evidence from H_1 against H_0 is lower than that from the true prior distribution of θ.
- When the true prior distribution $\pi_t(\theta)$ is same as the distribution $\pi_0(\theta)$ under H_0, the expected weights of evidence are negative. These negative expected weights of evidence are actually in favor of the null hypothesis H_0. As sample size increases, the expected weight of evidence in favor of H_0 increases.
- When the true prior distribution $\pi_t(\theta)$ is closer to $\pi_1(\theta)$ under the alternative hypothesis H_1, as the sample size increases, the expected weight of evidence from H_1 against H_0 increases.
- As sample size increases, the expected weight of evidence from the true prior distribution for θ against H_0 increases unless the true prior distribution is $\pi_0(\theta) = I(\theta = \theta_0)$, which is the density for θ under H_0. In the later case, the expected weight of evidence is equal to 0.

Table 4-2 Expected weight of evidence against H_0

Sample size (n)	$\pi_{t1}(\theta) = I(\theta = \theta_0)$		$\pi_{t2}(\theta)= \frac{2}{3}\pi_0(\theta)+\frac{1}{3}\pi_1(\theta)$		$\pi_{t3}(\theta)= \frac{1}{3}\pi_0(\theta)+\frac{2}{3}\pi_1(\theta)$	
	$EWOE_{11}$	$EWOE_{t1}$	$EWOE_{12}$	$EWOE_{t2}$	$EWOE_{13}$	$EWOE_{t3}$
40	-3.53	0	1.37	3.23	6.28	6.97
80	-5.45	0	3.99	7.05	13.42	14.68
120	-6.97	0	6.91	10.95	20.78	22.50
160	-8.26	0	9.98	14.88	28.23	30.37

4.2.3 Property of iMOM prior

By using an iMOM prior as a non-local alternative prior density, it is much efficient to accumulate evidence in favor of the null hypothesis if the null hypothesis is true. To demonstrate this property, let's consider the following hypotheses. Again, suppose that the number of responses X follows binomial (θ, n).

H_0: $\theta \sim \pi_0(\theta) = I(\theta = \theta_0)$, $\theta_0 = 0.2$

H_{1a}: $\theta \sim P_{iMOM}(\theta,\theta_0,k,v,\tau)$, $k = 1$, $v = 2$ and $\tau = 0.06$

H_{1b}: $\theta \sim \text{Beta}(\theta, 5, 7)$, $0 < \theta \leq 1$

Under the alternative hypothesis H_{1a}, the parameter θ follows the iMOM distribution $P_{iMOM}(\theta,\theta_0,k,v,\tau)$. Whereas, under the alternative hypothesis H_{1b}, the parameter θ follows the beta distribution Beta(θ, 5, 7) that is a local alternative prior density. Beta(θ, 5, 7) is a strong optimistic prior used in Section 3.8. Both the Beta(θ, 5, 7) and $P_{iMOM}(\theta,\theta_0,k,v,\tau)$ where $k = 1$, $v = 2$ and $\tau = 0.06$ have the same mode of 0.4.

The expected weight of evidence $EWOE_a$ from H_{1a} against H_0 when H_0 is true is same as $EWOE_{11}$ as described in the previous section. $EWOE_b$ in favor of the alternative hypotheses H_{1b} when H_0 is true can be calculated as follows.

$$EWOE_b = \sum_{x=0}^{n} \binom{n}{x} \theta_0{}^x (1-\theta_0)^{n-x} \log\left(\frac{B(5+x, 7+n-x)}{B(5,7)\theta_0{}^x (1-\theta_0)^{n-x}} \right)$$

where B(a,b) is a beta function with parameters a and b.

Table 4-3 shows the expected weights of evidence $EWOE_a$, and $EWOE_b$ in favor of the alternative hypotheses H_{1a} and H_{1b} when the null hypothesis H_0 is true. These results are also graphically displayed in Figure 4-2 with the classifications of the weight of evidence per Kass and Raftery (1995). As shown in Table 4-3 and Figure 4-2, all the expected weights of evidence are negative, this means that these evidence are actually in favor of the null hypothesis H_0. As the sample size increases, both $EWOE_a$ and $EWOE_b$ in favor of H_0 increase, however, $EWOE_a$ in favor of H_0 increases much greater as the sample size increases compared to $EWOE_{b.}$ When the sample size is between 40 and approximately 70, the iMOM alternative hypothesis provides strong evidence in favor of the null hypothesis H_0 when H_0 is true, and provides very strong support to H_0 when the sample size is approximately greater then 70. Whereas, even with the sample size = 160, the local alternative hypothesis beta(θ,5,7) is unable to provide strong evidence to support H_0 when H_0 is true.

Table 4-3 Expected weight of evidence when H_0 is true

Sample size (n)	Alternative hypothesis	
	H_{1a}: iMOM	H_{1b}: beta(θ,5,7)
40	-3.53	-1.35
80	-5.45	-1.71
120	-6.97	-1.91
160	-8.26	-2.06

Figure 4-2 Expected weight of evidence when H_0 is true

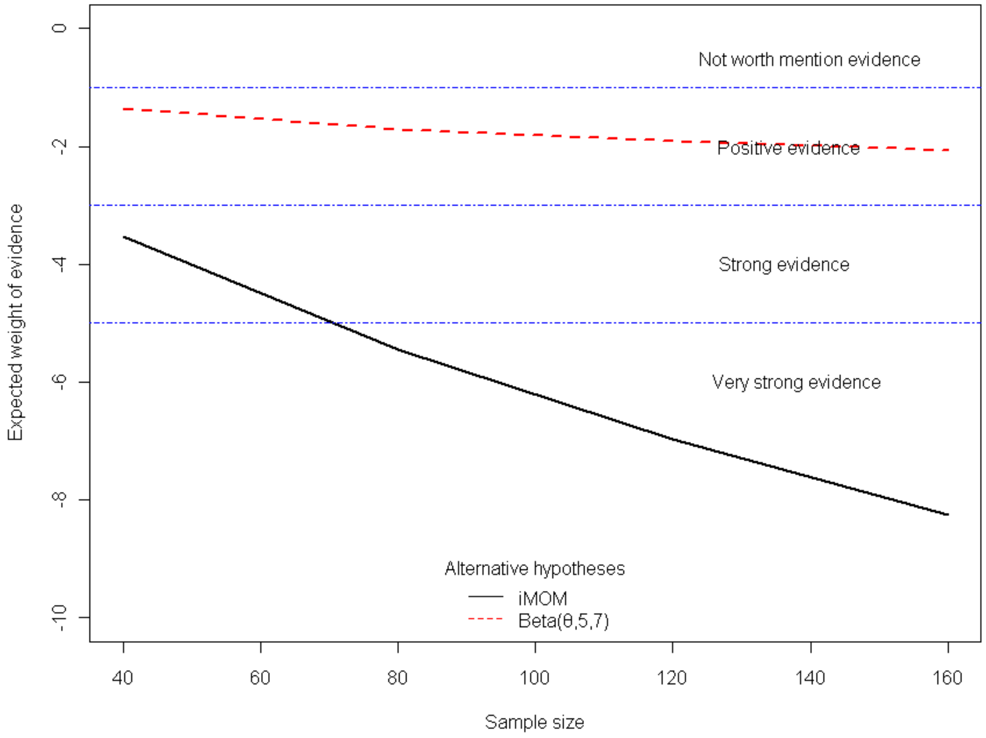
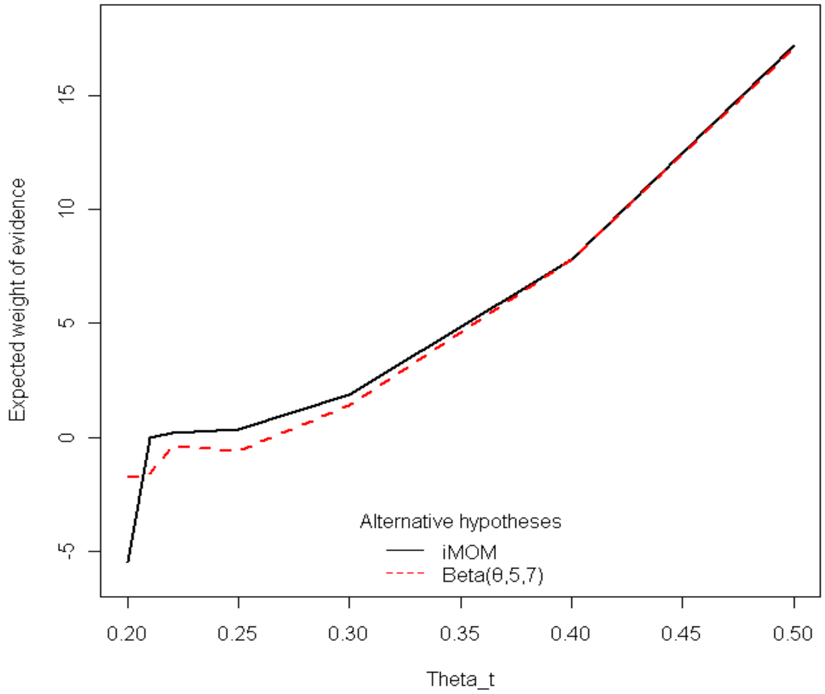


Figure 4-3 shows the expected weights of evidence when the true prior distribution for θ is $\pi_t(\theta) = I(\theta = \theta_t)$ with sample size n = 80. For this particular example, the expected weight of evidence from iMOM is equal to 0 when $\theta_t = 0.21$. From 0.2 to 0.21, a little increase in θ_t results in much larger increase in the expected weight of evidence from iMOM against H_0, whereas, only little increase in the expected weight of evidence from Beta(θ, 5, 7) is obtained. After $\theta_t = 0.21$, both iMOM and Beta(θ, 5, 7) provide the similar expected weight of evidence.

Figure 4-3 Expected weight of evidence when θ follows $\pi_t(\theta) = I(\theta = \theta_t)$ and n = 80



4.3 Two-stage design with Bayes factor and iMOM prior

For iMOM prior, as Johnson and Cook (2009) suggested, k=1 and v=2 are good in general. Set the prior mode of iMOM density at θ_1, then we can have (4.9) as below by maximizing the non-local alternative prior density as defined in (4.8).

$$\theta_1 = \theta_0 + \sqrt{\tau}\left(\frac{2k}{\nu+1}\right)^{1/2k} \tag{4.9}$$

τ can be resolved from (4.9) as follows.

$$\tau = \left(\frac{\nu+1}{2k}\right)^{k}(\theta_1 - \theta_0)^2 \tag{4.10}$$

To construct a single-arm two-stage phase II clinical trial design, following Johnson and Cook (2009), we define the null and alternative hypotheses as follows.

$$H_0: \theta \sim \pi_0(\theta) = I(\theta = \theta_0) \qquad (4.11a)$$

$$H_1: \theta \sim P_{iMOM}(\theta, \theta_0, k, v, \tau) \qquad (4.11b)$$

Let X be the data to be observed. Based on the Bayes rule, $P(H_k| X)$ can be expressed as follows.

$$P(H_k \mid X) = \frac{P(X \mid H_k)P(H_k)}{P(X \mid H_0)P(H_0) + P(X \mid H_1)P(H_1)},\ k = 0, 1 \qquad (4.12)$$

Therefore,

$$\frac{P(H_1 \mid X)}{P(H_0 \mid X)} = \frac{P(X \mid H_1)}{P(X \mid H_0)} \frac{P(H_1)}{P(H_0)} = \frac{m_1(x)}{m_0(x)} \frac{P(H_1)}{P(H_0)} = BF_1 \frac{P(H_1)}{P(H_0)} \qquad (4.13a)$$

Hence, the (4.13a) can be written as

$$\text{Posterior odds} = \text{Bayes factor} \times \text{prior odds} \qquad (4.13b)$$

Assume $P(H_1) = P(H_0) = 0.5$, then the posterior odds are equal to the Bayes factor. Let X_1 be the data observed from the first stage, and X be the data from the whole study. The decision rules for the two-stage design are constructed as follows.

- Stop the trial at the first stage for inferiority if the posterior probability $P(H_0|X_1) > P_{inf}$
- Stop the trial at the first stage for superiority if $P(H_1|X_1) > P_{sup1}$
- Claim superiority at the end of the trial if $P(H_1|X) > P_{sup2}$

where P_{inf}, P_{sup1} and P_{sup2} are threshold posterior probabilities. In practice, these threshold values need to be pre-specified. For example, $P_{inf} = 0.8$, $P_{sup1} = 0.9$, and $P_{sup2} \in (0.6, 0.8)$.

Besides the Bayesian setting, we control frequentist Type I and Type II error rates at the nominal level, for example, control Type I error rate at $\alpha = 0.05$ and Type II error rate at $\beta = 0.2$. The frequentist setting is same as what given in Section 3.2.

4.4 Examples

Similar to the examples used to demonstrate our Bayesian-frequentist two-stage design as discussed in Section 3.8, we use the same design parameters to show our Bayes factor-based two-stage design: θ_0=0.2, θ_1=0.4 (the mode of iMOM prior as in (4.8)) and the frequentist Type I error rate (α)=0.05, and Type II error rate(β)=0.2. We choose the threshold posterior probabilities: $P_{inf} = 0.8$, $P_{sup1} = 0.9$, and $P_{sup2} \in (0.6, 0.8)$. The parameters for the iMOM prior are k=1 and v=2.

In the examples presented below, the design bolded in black is the minimax design that minimizes total sample size, and the one bolded in blue is the optimal design that minimizes the expected sample size under H_0.

a) Examples

When $P_{sup2} = 0.6$

```
      a1 r1 n1  r  n PET(θ0) EN(θ0) PET(θ1) EN(θ1) Type1 Type2
 [1,]  3  8 16 11 33   0.605   22.7   0.349   27.1 0.049 0.187
 [2,]  5 10 23 12 35   0.704   26.6   0.498   29.0 0.035 0.198
 [3,]  2  6 12 12 36   0.578   22.1   0.418   26.0 0.050 0.193
 [4,]  3  7 14 12 37   0.710   20.7   0.432   27.1 0.047 0.198
 [5,]  5 10 22 13 38   0.739   26.2   0.448   30.8 0.029 0.199
 [6,]  2  7 13 13 39   0.509   25.8   0.287   31.5 0.037 0.178
 [7,]  2  7 13 13 40   0.509   26.3   0.287   32.3 0.043 0.155
 [8,]  3  8 16 14 41   0.605   25.9   0.349   32.3 0.026 0.197
 [9,]  3  7 15 14 42   0.666   24.0   0.481   29.0 0.038 0.184
[10,]  3  7 14 14 43   0.710   22.4   0.432   30.5 0.036 0.194
[11,]  3  7 15 14 44   0.666   24.7   0.481   30.1 0.048 0.152
[12,]  3  7 15 15 45   0.666   25.0   0.481   30.6 0.035 0.180
[13,]  3  7 15 15 46   0.666   25.3   0.481   31.1 0.038 0.164
[14,]  3  8 16 15 47   0.605   28.2   0.349   36.2 0.036 0.134
[15,]  3  8 16 16 48   0.605   28.6   0.349   36.8 0.024 0.165
```

```
[16,]  3  8 16 16 49  0.605  29.0  0.349  37.5 0.027 0.147
[17,]  4  8 17 16 50  0.769  24.6  0.485  34.0 0.030 0.175
```

Notation:

- PET(θ_0), PET(θ_1): Frequentist PET (probability of early termination) under null and alternative hypothesis.
- EN(θ_0), EN(θ_1): Frequentist expected sample size under null and alternative hypothesis.
- Type1, Type2: Frequentist Type I and Type II error rate.

When $P_{sup2} = 0.7$

```
      a1 r1 n1  r  n PET(θ0) EN(θ0)PET(θ1)EN(θ1) Type1 Type2
 [1,]  5 10 23 12 35  0.704  26.6  0.498  29.0 0.035 0.198
 [2,]  5 10 22 13 38  0.739  26.2  0.448  30.8 0.029 0.199
 [3,]  3  8 16 14 41  0.605  25.9  0.349  32.3 0.026 0.197
 [4,]  3  7 15 14 42  0.666  24.0  0.481  29.0 0.038 0.184
 [5,]  7 12 30 15 43  0.770  33.0  0.612  35.0 0.020 0.194
 [6,]  3  7 15 15 44  0.666  24.7  0.481  30.1 0.031 0.198
 [7,]  3  7 15 15 45  0.666  25.0  0.481  30.6 0.035 0.180
 [8,]  5 10 23 16 46  0.704  29.8  0.498  34.6 0.019 0.194
 [9,]  3  8 16 16 47  0.605  28.2  0.349  36.2 0.021 0.185
[10,]  3  8 16 16 48  0.605  28.6  0.349  36.8 0.024 0.165
[11,]  3  8 16 16 49  0.605  29.0  0.349  37.5 0.027 0.147
[12,]  4  8 18 17 50  0.733  26.6  0.531  33.0 0.025 0.187
```

When $P_{sup2} = 0.8$

```
      a1 r1 n1  r  n PET(θ0) EN(θ0)PET(θ1)EN(θ1) Type1 Type2
 [1,]  7 12 30 15 43  0.770  33.0  0.612  35.0 0.020 0.194
 [2,]  5 10 23 16 46  0.704  29.8  0.498  34.6 0.019 0.194
 [3,]  3  8 16 16 47  0.605  28.2  0.349  36.2 0.021 0.185
 [4,]  7 12 30 17 48  0.770  34.1  0.612  37.0 0.015 0.199
 [5,]  5 10 22 17 49  0.739  29.1  0.448  36.9 0.015 0.199
 [6,]  4  8 18 17 50  0.733  26.6  0.531  33.0 0.025 0.187
 [7,]  7 12 29 18 51  0.797  33.5  0.567  38.5 0.012 0.200
 [8,]  4  8 18 18 52  0.733  27.1  0.531  34.0 0.022 0.199
 [9,]  4  8 18 18 53  0.733  27.4  0.531  34.4 0.024 0.183
[10,]  4  8 18 18 54  0.733  27.6  0.531  34.9 0.026 0.168
[11,]  4  8 18 19 55  0.733  27.9  0.531  35.4 0.021 0.194
[12,]  4  9 19 19 56  0.680  30.8  0.402  41.1 0.014 0.170
[13,]  4  9 19 19 57  0.680  31.2  0.402  41.7 0.016 0.154
[14,]  4  9 19 20 58  0.680  31.5  0.402  42.3 0.012 0.183
[15,]  5  9 20 20 59  0.814  27.2  0.530  38.3 0.015 0.198
[16,]  5  9 20 20 60  0.814  27.4  0.530  38.8 0.016 0.186
```

b) Summary of characteristics of the Bayes factor-based two-stage designs using an iMOM prior

Total sample size

The characteristics of the Bayes factor-based two-stage design are summarized in Table 4-4. For optimal designs, the total sample sizes are 37, 42 and 50 corresponding to the criterion of claiming superiority when $P(H_1|X)>0.6$, >0.7 and >0.8 respectively. For minimax designs, the respective total sample sizes are 33, 35 and 43. Apparently, for both optimal and minmax designs, the more stringent criterion is used to claim superiority (e.g $P(H_1|X) > 0.8$), the larger total sample size is required.

Probability of early termination and expected sample size

For minimax designs, probabilities of early termination under the null hypothesis are 0.605, 0.704 and 0.770, and expected sample sizes are 22.7, 26.6 and 33.0 corresponding to the criterion of claiming superiority at the end of the trial by $P(H_1|X)>0.6$, >0.7 and >0.8. Therefore, as the more stringent criterion of claiming superiority is applied, although the higher probability of early termination is anticipated, the larger expected sample size is required. This is mainly due to the fact that the sample size for the first stage n_1 increases when the more stringent criterion of claiming superiority is applied. In addition, a larger total sample size is required by a more stringent criterion of claiming superiority.

The above findings also apply to the probability of early termination and the expected sample size under the alternative hypothesis, as well as for the optimal designs under the alternative hypothesis.

As expected, the minimax designs require smaller total sample size and larger expected sample size under the null hypothesis compared to the optimal designs with the same criterion of claiming superiority at the end trial. For example, the total sample sizes for the minimax designs are 33, 35 and 43, whereas the total sample sizes of 37, 42 and 50 are for the optimal designs corresponding to the criteria of $P(H_1|X)>0.6$, >0.7 and >0.8 respectively.

Table 4-4 Characteristics of Bayes factor-based two-stage designs using an iMOM prior

Bayes factor design with iMOM prior	a1	r1	n1	r	n	PET(θ_0)	EN(θ_0)	PET(θ_1)	EN(θ_1)
Optimal($P(H_1\|X)>0.6$ end of the trial)	3	7	14	12	37	0.710	20.7	0.432	27.1
Optimal($P(H_1\|X)>0.7$ end of the trial)	3	7	15	14	42	0.666	24.0	0.481	29.0
Optimal($P(H_1\|X)>0.8$ end of the trial)	4	8	18	17	50	0.733	26.6	0.531	33.0
Minimax($P(H_1\|X)>0.6$ end of the trial)	3	8	16	11	33	0.605	22.7	0.349	27.1
Minimax($P(H_1\|X)>0.7$ end of the trial)	5	10	23	12	35	0.704	26.6	0.498	29.0
Minimax($P(H_1\|X)>0.8$ end of the trial)	7	12	30	15	43	0.770	33.0	0.612	35.0

Summary

In summary of Bayes factor-based two-stage single-arm phase II clinical trial designs with an iMOM prior, the more stringent criterion of claiming superiority is used for the design, the larger total sample size is required. In addition, for the minimax design, although a higher probability of early termination is expected as a more stringent criterion of claiming superiority is applied, a larger expected sample size is anticipated since a larger sample size for the first stage is required.

4.5 Comparisons with typical frequentist and Bayesian single-arm phase II clinical trial designs

In this Section, we compare our Bayes factor-based two-stage design with typical frequentist and Bayesian two-stage designs aforementioned in Section 3.9.

Again, in order to have fair comparisons, we evaluate these designs under the same design parameters: θ_0 = 0.2, θ_1 = 0.4 (for the Bayes factor-based designs, θ_1 is the mode of an iMOM prior), Type I error α = 0.05 and Type II error rate β = 0.2. Similar to Section 3.9, we compare these designs with respect to probability of rejecting H_0, probability of accepting H_0, probability of early termination (PET), expected sample size, and expected sample size per correct decision. These operating characteristics are plotted in Figure 4-2 though Figure 4-6. The properties of the other Bayesian and frequentist designs have been provided in Section 3.9, therefore, this section only describes Bayes factor-based two-stage designs.

Probabilities of rejecting and accepting H_0

The probabilities of rejecting and accepting H_0 are shown in Figure 4-4 and Figure 4-5. Since all of these designs are performed by controlling error rates at the same levels of 0.05 for Type I error rate and 0.2 for Type II error rate, their probabilities of rejecting and accepting H_0 are very similar.

Probability of early termination

As shown in Figure 4-6, When the criterion of $P(H_1|X)>0.8$ is used, the probabilities of early termination for our iMOM minimax and optimal design are lower than that from Shuster's design, but higher than Fleming's design. When the criterion of $P(H_1|X)>0.6$ is used, the probabilities of early termination for our iMOM minimax and optimal design are slightly lower than Fleming's design. In addition, when θ is around θ_0, the probabilities of early termination for the Lee & Liu's design are higher than Bayes factor-based two-stage design with an iMOM prior.

Expected sample size

As described in Section 4.4, the more stringent criterion of claiming superiority is used, the larger expected sample size is required by Bayes factor-based design. As shown in Figure 4-7, when the criterion $P(H_1|X)>0.8$ is used, the expected sample sizes for iMOM designs (optimal and minimax designs) around $\theta \in (\theta_0, \theta_1)$ are larger than any other two-stage designs with acceptance and rejection boundaries. When the criterion $P(H_1|X)>0.6$ is used, there is no much difference in the expected sample size among iMOM designs (optimal and minimax designs), Shuster's, Fleming's or Chang's design.

Expected sample size per correct decision

The expected sample size per correct decision is presented in Figure 4-8. When the criterion $P(H_1|X)>0.8$ is used, the expected sample sizes per correct decision for iMOM designs (optimal and minimax designs) at $\theta \in (\frac{\theta_0 + \theta_1}{2}, \theta_1)$ are larger than that for other designs. Whereas, when the criteria $P(H_1|X)>0.6$ is used, the expected sample size per correct decision is lower compared to other designs.

Summary

In summary, for the Bayes factor-based two-stage design using an iMOM prior, the criterion based on posterior probability $P(H_1|X)$ at the end of the study plays an important role. In general, the criterion with a higher $P(H_1|X)$ is used, the larger samples size (including expected sample size, expected sample size per correct decision and total sample size) is required, and higher probability of early termination is expected.

Figure 4-4 Probability of rejecting H_0

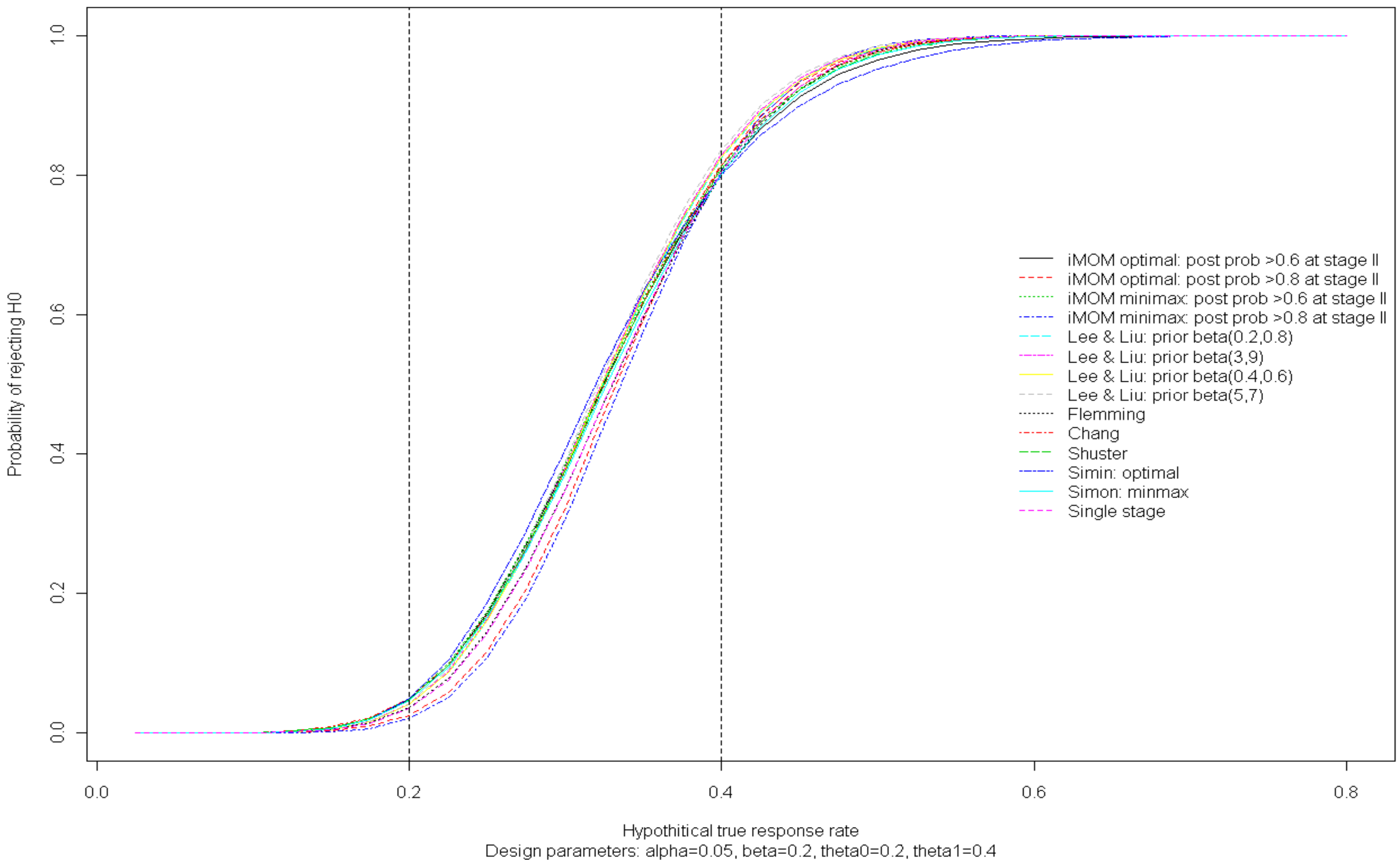


Figure 4-5 Probability of accepting H_0

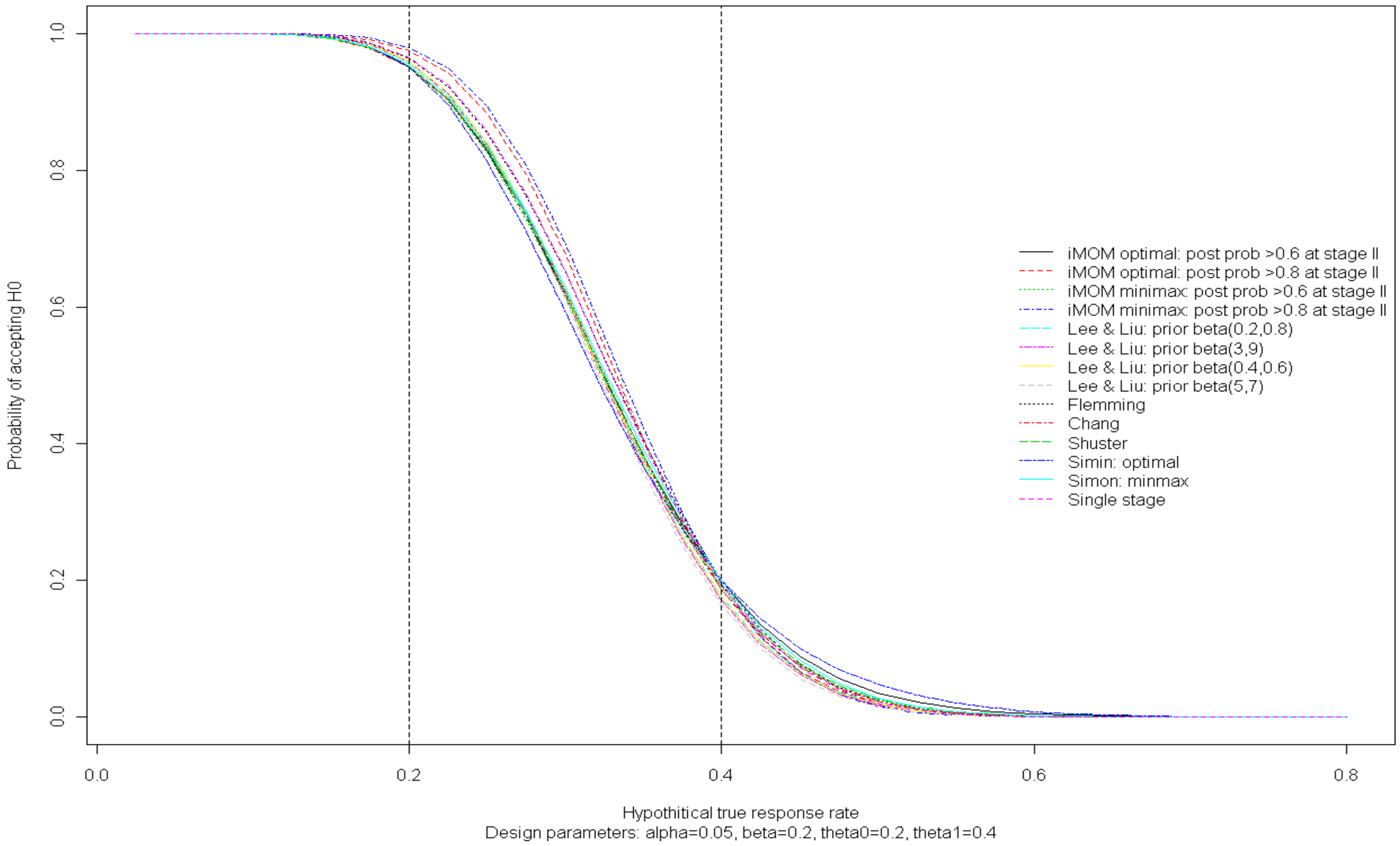


Figure 4-6 Probability of early termination (PET)

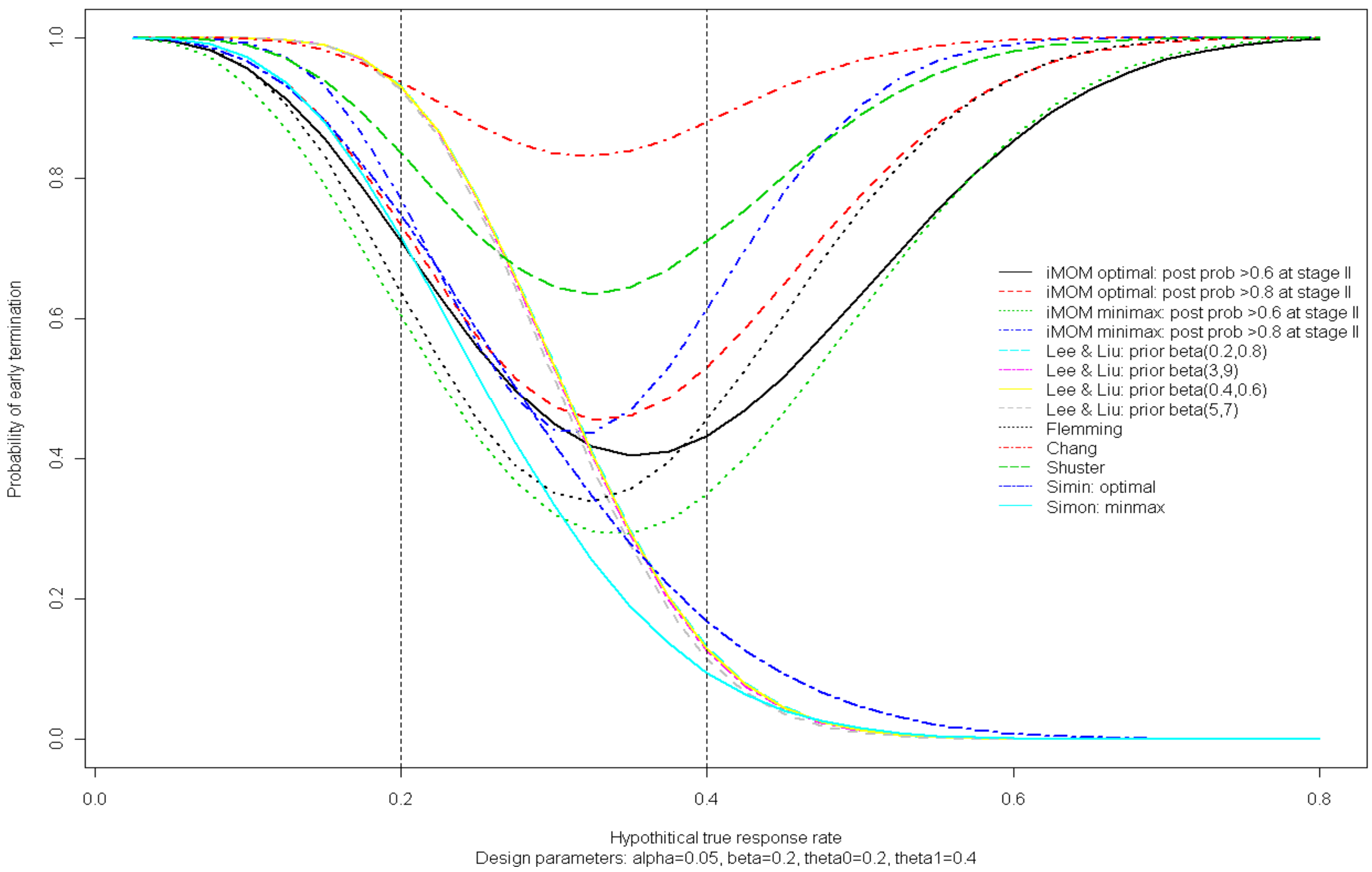


Figure 4-7 Probability of expected sample size

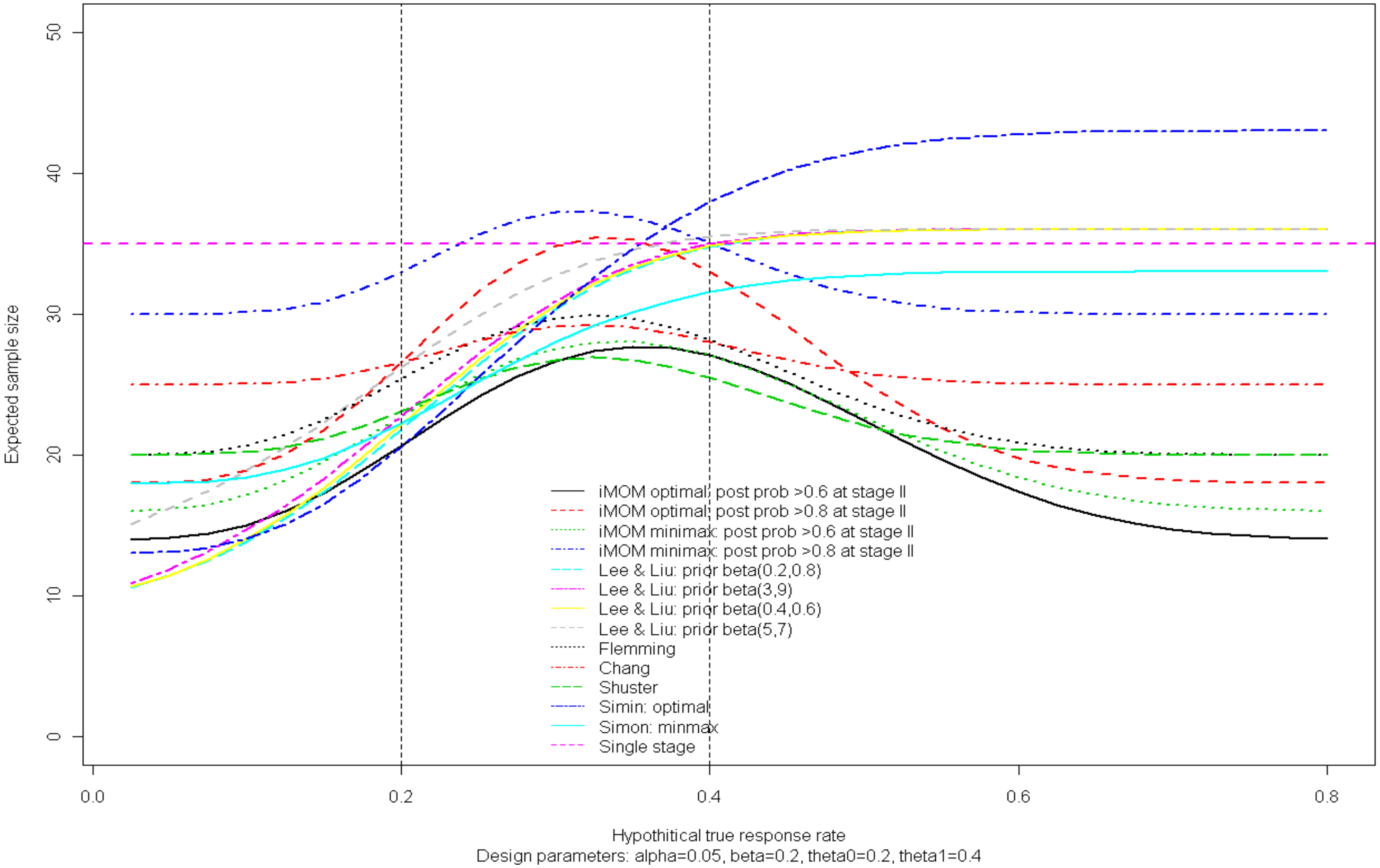
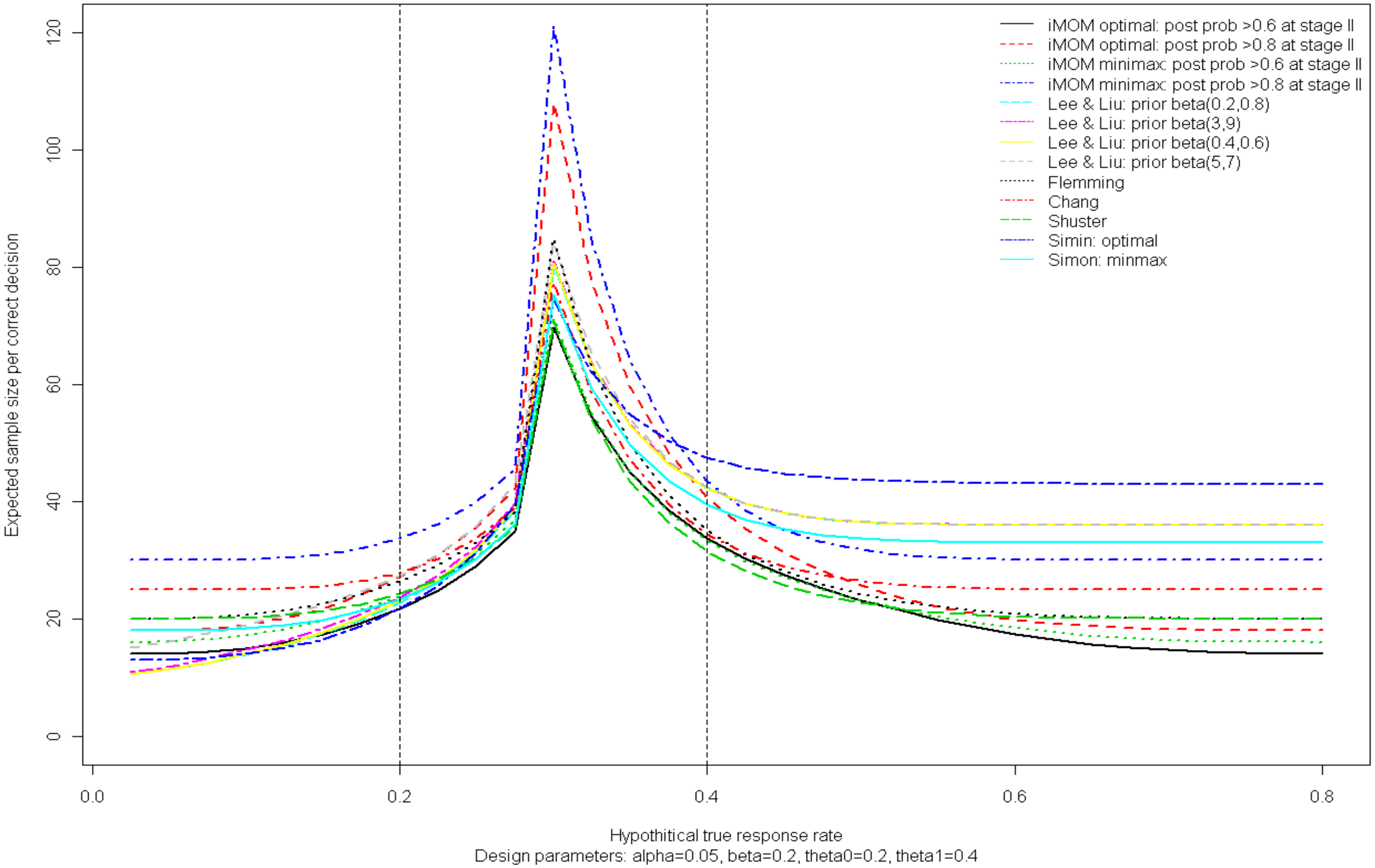


Figure 4-8 Expected sample size per correct decision



5 Literature review for phase II/III clinical trial design

Under the conventional drug development framework, phase II and Phase III clinical trials are conducted separately in terms of clinical trial operation and statistical inference. In a phase II trial, several doses of the new therapy are compared to an active control or placebo to identify the “best” one or two doses for further investigation. Upon success of phase II trial, a phase III trial as a stand-alone confirmatory trial is conducted with the goal of seeking marketing approval from health authorities.

5.1 Phase II/III clinical trial design

A phase II/III trial design is a program that addresses within a single trial the objectives that are normally achieved through separate trials in phase IIb and phase III (Gallo, 2006). This uninterrupted adaptive design has advantages of combining conventional phase II and phase III operationally and statistical inferentially into a single study (e.g. Bretz et al, 2006; Gallo, 2006; Jennison and Turnbull, 2006), particularly (1) accelerate drug development process by reducing “white space” between the two clinical trial phases; (2) gain statistical efficiency by using first stage data on the patients treated with the new therapy with the dose selected for the second stage, thus reduce the sample size needed for the second stage. (3) get long term safety data earlier since the patients in stage I are followed longer as compared to conventional phase II study.

The idea of combining phase II and phase III trials was proposed as early as in 1988 by Thall, Simon and Ellenberg for a two-stage design with binary outcome. Inoue, Thall and Berry (2002) discussed a seamless phase II/III trial using sequential Bayesian design in an oncology study. By use of mixture model-based predictive probabilities of concluding superiority of the new therapy to the active control, they repeatedly assessed whether to stop the trial early, continue or shift the phase II into phase III. Bretz et al (2006) comprehensively discussed general concepts with respect to confirmatory seamless phase II/III clinical trials, and subsequently Schmidli et al (2006) provided extensive applications and practical considerations for such seamless phase II/III clinical trial designs. Shih (2006), Jennison and Turnbull (2006), and Gould (2006) commented Bretz et al (2006)'s and Shimidli et al (2006)'s papers. In the comments on dose selection, Shih (2006) emphasized that rather than saying that a "winner" dose is selected, it should really be saying that "loser" doses are dropped; the futility condition needs to be set clearly to direct selection process, not for the sake of controlling the overall Type I error rate, but for the interpretability of the study.

Gallo et al (2006) provided executive summary of a PhRMA group on adaptive trial designs, and addressed logistics, operational, procedural and statistical challenges associated with adaptive designs in three areas: dose finding, phase II/III trial designs, and sample size reestimation. Maca et al (2006) outlined the feasibilities to conduct a phase II/III clinical trial. The most important feasibility consideration is the amount of time needed to follow up a patient to reach the study endpoint, on which the selection decision is based. If the time needed to reach this endpoint is short relative to the total enrollment time of the study, then the enrollment could be continued without interruption during this "transition" period. With respect to shortening the drug development time by the use of a phase II/III clinical trial, it is very important to consider whether this trial setup would achieve the drug development objectives within the reduced time frame. Another point to consider is drug supply (or drug packaging), which is challenging since the number of treatment groups would change during the study. A phase II/III clinical trial is more suitable if the drug regimens are not costly and not complicated.

5.2 Dose selection

Stallard and Todd (2003) and Todd and Stallard (2005) proposed an unconditional approach to select a single winner dose under normally distributed study endpoint. Shun, Lan and Soo (2008) discussed a normal approximation method to this unconditional approach. Sampson and Sill (2005) proposed a conditional approach to select a single winner dose. Bretz, Schmidli, Koenig, Racine and Maurer (2006) elaborated the closed testing procedure for confirmatory seamless phase II/III clinical trial designs. Under the closed testing principle, there is no need to pre-specify dose selection rules in order to control the multiple level type I error rate. Recently, Kimani et al (2009) proposed a dose-selection procedure with logistic dose-response relationship for seamless phase II/III clinical trials. Their approach incorporated both efficacy and safety. The choice of the doses to be continued to stage II is made by comparing the predictive power of the potential sets of the doses, which might continue.

In some clinical trials, a short-term study endpoint (or early endpoint) is considered for dose selection at the interim analysis. Liu and Pledger (2005) proposed an adaptive seamless strategy to combine phase II and phase III under a two-stage design framework with the consideration of patient long term follow-up for a clinical endpoint (or long-term endpoint). In their design, two interim analyses are planned for the first stage. The first interim analysis is used to determine a “go or no-go” decision, dose selection and sample size adjustment based on the early endpoint; the second interim analysis estimates the dose-response curve using the clinical endpoint. The second stage starts when the last patient in the first stage is randomized. For the second stage, trend statistics are adaptively chosen based on the estimated dose-response curve in the clinical endpoint of the first-stage patients. At the end of the trial, pairwise statistics for the first stage and adaptive trend statistics for the second stage of the clinical endpoint are combined to establish dose-response and to identify the lowest effective dose.

Todd and Stallard (2005) proposed a phase II/III clinical trial design using group sequential design, which incorporates treatment selection based upon a short-term endpoint and final analyses on a long-term primary study endpoint. Through an example, they demonstrated that their approach may reduce the total number of patients required for the trial.

5.3 Multiple study endpoints and multiple hypothesis testing

In general, clinical trials are designed with a primary study endpoint, a main secondary endpoint and several other secondary endpoints. The primary study endpoint is most clinically relevant in terms of characterizing treatment effect; hence, the study design and statistical analysis plan are mainly driven by this endpoint. Although the secondary endpoints are intended to enhance characterization of clinical benefits of the study treatment, they can not stand alone to demonstrate the treatment effect. To evaluate several study endpoints, frequently multiple null hypotheses are tested in a hierarchical priori order. For a clinical trial design with a fixed sample size, when the statistical test with regard to the primary endpoint achieves statistical significance, the secondary endpoints are tested at the same nominal significance level α as the primary endpoint. The hypothesis testing process continues in a pre-specified hierarchical order until the statistical significance is not achieved. This testing strategy ensures a strong control of overall Type I error rate (Hung, Wang, O'Neill, 2007).

However, in some disease areas, the primary endpoint that completely characterizes the disease and best captures treatment effect under clinical investigation is not well-established in the clinical community. Hence, changing the primary endpoint seems intuitively acceptable to some researchers (Hung, et al, 2006). For example, there is no consensus on the single most important primary endpoint in the management of primary biliary cirrhosis (PBC), which is a chronic, cholestatic disease of unknown etiology involving inflammation and subsequent obliteration of the interlobular bile ducts in the liver. Using multiple endpoints is still common in clinical trials in these diseases (Sankoh, et al, 2003). Another example of disease area is heart failure. The primary study endpoint is a composite of death, hospitalization due to worsening heart failure, myocardial infarction, etc. In this example, the primary composite endpoint may not be a good measure for the treatment effect.

In some clinical indications, the distinction between the primary and the main secondary endpoint is not clear. If the result from the primary endpoint is not statistical significant, but the findings based on the main secondary endpoint are very positive, then the question arises on whether the trial could be interpreted with this main secondary endpoint. Alosh and Huque (2010) proposed a consistency-adjusted alpha-adaptive strategy for sequential testing. Under their approach, although a larger portion of alpha is allocated to the first endpoint as the designated primary study endpoint, the alpha allocated to the main secondary endpoint is adaptive to the findings from the first endpoint if a consistency criterion is met. Even if the first endpoint does not achieve statistical significance at the pre-allocated significance level, it still has a chance to be considered significant when the first consistency criterion is satisfied and the second endpoint achieves statistical significance at the adaptive significance level. If only the second consistency criterion is met, then only the second endpoint has a chance to be used to claim trial success. However, if the first endpoint fails the hypothesis testing and no consistency criteria are met, no hypothesis testing is permitted for the second endpoint.

Frequently, multiple doses of the study drug and multiple endpoints are simultaneously considered in a clinical trial. The commonly used multiplicity adjustment approaches are Bonferroni, Hochberg and Hommel procedures. However, these conventional approaches do not consider the dose-response relationship on each endpoint. For this two-dimensional multiplicity problem, by taking into account the dose order, Quan et al (2005) proposed six procedures to control family-wise Type I error rate in the strong sense: Bonferroni-closed, Hochberg-Bonferroni, modified Hochberg-Bonferroni, improved Bonferroni-closed, improved Hochberg-Bonferroni and improved modified Hochberg-Bonferroni procedure. Through numerical examples, they showed that these newly proposed procedures in general have higher power than the commonly used procedures.

5.4 Type I error rate control

Adaptive clinical trial designs have paid much attention to improving the efficiency of current drug development processes. For example, the Pharmaceutical Research and Manufacturers of America (PhRMA) initiated a working group to facilitate a wider usage and regulatory acceptance of these designs (Gallo et al, 2006). However, it is well recognized that bias is introduced because of the opportunity to choose the successful result from among the multiplicity of options. There are two principle issues raised by adaptive design methods (FDA, 2010):

- Whether the adaptation process has led to design, analysis, or conduct flaw that have introduced bias that increases the chance of a false conclusion that the treatment is effective (a Type I error)

- Whether the adaptation process has led to positive study results that are difficult to interpret irrespective of having control of Type I error.

Controlling Type I error rate has been widely addressed by many researchers. For example, p-value combination based on Fisher's product method (Bauer and Köhne, 1994); group sequential test procedure with weighted test (Cui et al., 1999); conditional power approach (Shih et all, 2004); p-value combination based on inverse normal combination function (Lehmacher and Wassmer, 1999); adaptive group sequential design (Müller and Schäfer, 2001); and conditional error function (Proschan and Hunsberger, 1995).

In particular for a phase II/III clinical trial design, Bretz et al (2006) and Bretz (2009) discussed Type I error control in the strong sense following the closed testing procedure (Marcus, 1976) and conditional invariance principle (Brannath et al, 2002 and 2007).

5.5 Application of phase II/III clinical trial designs

Since logistics, operational, procedural and statistical challenges associated with adaptive designs are addressed (e.g. Gallo, 2006), there are some phase II/III clinical trials designed recently. A typical example is the adaptive seamless trial of integrating indacaterol dose selection in the respiratory field (Barnes et al, 2010). This is a conventional two-stage phase II/III trial. At the learning phase II stage, the total number of 805 patients were randomized into four indacaterol doses, one placebo and two active control groups in a 1:1:1:1:1:1:1 ratio. The primary objective of stage I was to determine the risk-benefit of the four doses of indacaterol in order to select two doses to be carried forward into the 2nd stage as confirmatory phase. The dose selection criteria were pre-set based on two co-primary endpoints: FEV_1 and $FEV_1AUC_{1\text{-}4h}$. At stage II, the additional patients were equally randomized into the two selected indacaterol doses, placebo and an active control groups, which resulted in a total of 1683 patients in the four treatment groups continued to the final stage. Since there were four indacterol doses studied in this trial, the 1-sided significance level of alpha = 0.025/4=.006 were used for the hypothesis testing with respect to the primary endpoint - 24 h post dose (trough) FEV_1 in patients with COPD (Chronic Obstructive Pulmonary Disease) following 12 weeks of treatment.

In the next sections, we will discuss our varying-stage adaptive phase II/III design. Similar to the trial described above, the dose selection in our design is not necessarily based on statistically significant difference between treatment arms at an interim analysis, instead, more flexibility of dose selection is granted to decision makers. However, the dose selection rules have to be pre-specified in the protocol. Different from the indacaterol trial, our design requires hypothesis testings at the interim analyses in order to determine which trial decision path to be taken. Another major difference is that p-value combination test is used to perform final statistical analysis in our design.

6 A varying-stage adaptive phase II/III clinical trial design

6.1 Introduction

Suppose a phase II/III clinical trial is planned. Although this is a stagewise adaptive study, based on the results from the first interim analysis, the number of future stages varies depending on whether an intermediate stage is needed in order to obtain adequate information for decision making, such as “go or no-go” and/or dose selection. In practice, in addition to dose selection, multiple study endpoints could be considered in a phase II study. Due to limited efficacy and safety data on the new therapy available at the time of phase II/III trial planning, it could be challenging to decide which endpoint should be regarded as primary among several potential endpoints. Therefore, the first stage of the study should be designed as a learning stage.

Figure 6.1 shows the flow chart of our varying-stage adaptive phase II/III clinical trial design. Following the initial learning stage, the first interim analysis will be performed. The goal of this interim analysis is to determine the primary study endpoint and choose the optimal dose arm(s) of the study treatment for further confirmatory investigation.

Consider a clinical trial that is initially planned with two study endpoints, up to three stages, K dose arms of the study treatment and one control arm. Let θ_{ik} be the parameter of interest with respect to the i^{th} endpoint for the k^{th} dose arm (k=0 for the control arm), p_{ijk} be the p-value of the hypothesis testing for the difference between the k^{th} dose arm (k=1, 2, … K) and the control arm with respect to the i^{th} endpoint ($i = 1, 2$) at the j^{th} stage ($j = 1, 2, 3$), and p_{ij} be the p-value of the global null hypothesis test across all dose arm(s) at the j^{th} stage with respect to the i^{th} endpoint (see Section

6.2 for statistical hypotheses). The p-values p_{ij} is based on the data from the j^{th} stage only. Let $\alpha_i^{(j)}$ be the threshold probability for the i^{th} endpoint at the j^{th} interim analysis, and $\alpha_F^{(j)}$ be futility stop level for the j^{th} interim analysis.

Initial learning stage (phase II)

The initial stage (phase II) is considered as a learning stage. Following this stage, the first interim analysis is performed. The goal of this interim analysis is to decide the primary study endpoint, drop inefficacious/harmful dose arm(s), and adjust sample size for the next stage. As shown from the flow chart in Figure 6.1, if $p_{11} < \alpha_1^{(1)}$, the endpoint 1 is kept as the primary study endpoint as initially planned. Following this, inefficacious/harmful dose arm(s) will be dropped and sample size adjustment for the final stage will be performed.

If $p_{11} \geq \alpha_1^{(1)}$ and $p_{21} < \alpha_2^{(1)}$, then the endpoint 2 will be considered as the primary study endpoint. With respect to the new primary study endpoint (endpoint 2), inefficacious/harmful dose arm(s) will be dropped, and sample size adjustment will be performed for the next stage.

Otherwise, if $p_{11} \geq \alpha_1^{(1)}$ and $p_{21} \geq \alpha_2^{(1)}$, the primary endpoint can not be decided based on the current interim data. If the p-value $p_{11} \geq \alpha_F^{(1)}$, then the trial can be terminated for futility. Otherwise, two further study stages need to be planned: one is intermediate stage; another one is confirmatory stage. The intermediate stage is considered as an extension of phase II, from which more data will be obtained, so that more informative decisions can be made particularly regarding whether the trial can be advanced to the final confirmatory stage.

Intermediate stage (extended phase II)

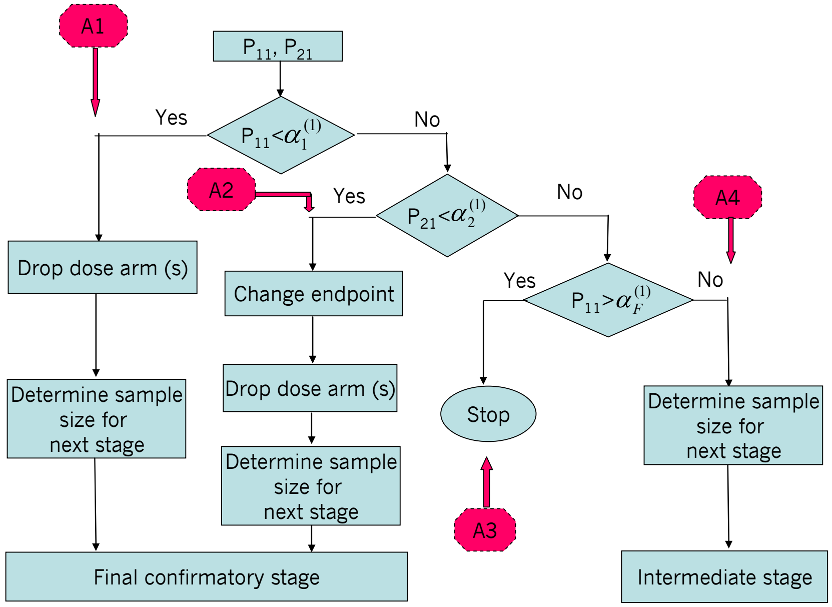
The intermediate stage is considered as an extended Phase II. The second interim analysis is conducted following the intermediate stage. For this interim analysis, a combination test is performed to incorporate data obtained from the initial stage and the intermediate stage. The combined p-values are based on a combination function $C(p_{i1}, p_{i2})$, where p_{i1} and p_{i2} are p-values from the two disjoint stages – initial stage and intermediate stage respectively for the i^th^ endpoint. There are mainly two approaches available to combine p-values from different stages: one approach is Fisher's product combination method; another approach is based on inverse normal combination function (Lehmacher and Wassmer, 1999). In this thesis, Fisher's product combination method is used to combine p-values. Hence $C(p_{i1,} p_{i2}) = p_{i1}\, p_{i2}$, i = 1, 2 for study endpoint.

If combined overall p-value $C(p_{11}, p_{12}) < \alpha_1^{(2)}$, or $C(p_{11}, p_{12}) \geq \alpha_1^{(2)}$ and $C(p_{21}, p_{22}) < \alpha_2^{(2)}$, then the similar design flow as the first interim analysis can be followed to change primary study endpoint, drop inefficacious/harmful dose arm(s), and perform sample adjustment for the final confirmatory stage.

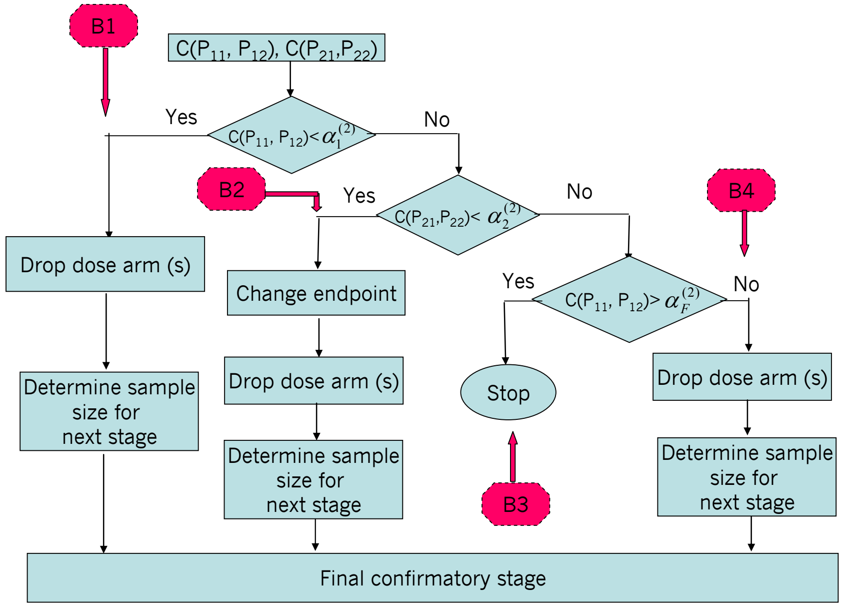
If the combined p-value $\alpha_1^{(2)} \leq C(p_{11}, p_{12}) < \alpha_F^{(2)}$ and $C(p_{21}, p_{22}) \geq \alpha_2^{(2)}$, the trial will be continued to the final stage with the endpoint 1 as the primary study endpoint. Similar to the scenarios described above, dose selection and sample size adjustment will be carried out for the final stage.

Otherwise if $C(p_{11}, p_{12}) \geq \alpha_F^{(2)}$ and $C(p_{21}, p_{22}) \geq \alpha_2^{(2)}$, the trial will be stopped for futility.

Figure 6-1 **Flow chart of varying-stage adaptive phase II/III clinical trial design**



(a) Interim I (initial learning stage)


B1
$C(P_{11}, P_{12}), C(P_{21},P_{22})$
Yes
$C(P_{11}, P_{12})<\alpha_1^{(2)}$
No
B2
Yes
$C(P_{21},P_{22})< \alpha_2^{(2)}$
No
B4
Yes
$C(P_{11}, P_{12})> \alpha_F^{(2)}$
No
Drop dose arm (s)
Change endpoint
Determine sample size for next stage
Drop dose arm (s)
Stop
Drop dose arm (s)
Determine sample size for next stage
B3
Determine sample size for next stage
Final confirmatory stage

(b) Interim II (intermediate stage)

Final confirmatory stage (phase III)

The final stage is considered as a confirmatory phase III stage. Similar to the second interim analysis, the final analysis incorporates data from previous stage(s) via a combination test. The statistical significance will be demonstrated by comparing the combined p-value $C(p_{i1}, p_{i2})$ for two-stage setting or $C(p_{i1}, p_{i2}, p_{i3})$ for three-stage setting against a critical value with respect to the primary study endpoint. Determination of the critical values will be discussed in Section 6.4.

6.2 Statistical hypotheses

Let D = {1, 2, … , K} be the index set of dose arms of the study treatment. The null hypothesis for the comparison between the kth dose arm of the study treatment and the control arm is.

H_{0ik}: $\theta_{ik} = \theta_{i0}$, vs $H_{1ik} : \theta_{ik} > \theta_{i0}$, where, i=1, 2 for endpoint; $k \in D$ (6.1)

This hypothesis is called elementary null hypothesis. The global null hypothesis with respect to the ith endpoint is

$$]H_{0i} : \bigcap_{k\in D} H_{0ik} = H_{0i1} \bigcap H_{0i2} \bigcap \ldots \bigcap H_{0iK} \text{, where i=1,2 for endpoint} \quad (6.2)$$

The global null hypothesis for both study endpoints is

$$H_0 : H_{01} \bigcap H_{02} \quad (6.3)$$

6.3 Conditional distribution of combined p-value

6.3.1 No primary study endpoint change

a) Two-stage setting

If $p_{11} < \alpha_1^{(1)}$, the trial will be designed as a two-stage study with the endpoint 1 as the primary study endpoint. Given this condition, the cumulative distribution function of p_{11} under the global null hypothesis (6.2) is as follows.

$$F(p_{11} = x \mid p_{11} < \alpha_1^{(1)}) = \Pr(p_{11} < x \mid p_{11} < \alpha_1^{(1)}) = \frac{\Pr(p_{11} < x, p_{11} < \alpha_1^{(1)})}{\Pr(p_{11} < \alpha_1^{(1)})}$$

$$= \frac{\Pr(p_{11} < x)}{\Pr(p_{11} < \alpha_1^{(1)})} = \frac{x}{\alpha_1^{(1)}}, \text{ where } 0 < \text{x} < \alpha_1^{(1)} \qquad (6.4)$$

Therefore, given the two-stage setting with $p_{11} < \alpha_1^{(1)}$ at the interim I, p_{11} under the global null hypothesis (6.2) follows the truncated uniform distribution - Uniform(0, $\alpha_1^{(1)}$).

b) Three-stage setting with $p_{11}p_{12} < \alpha_1^{(2)}$ at the interim II

If $\alpha_1^1 \leq p_{11} < \alpha_F^{(1)}$ and $p_{21} \geq \alpha_2^{(1)}$ at the interim I, an intermediate stage will be planned. At the interim II, if $p_{11}p_{12} \leq \alpha_1^{(2)}$, the study will be designed as a three-stage trial with the endpoint 1 as the primary study endpoint. In this Section, we prove that given this condition, $p_{11}p_{12}$ under the global null hypothesis (6.2) follows the truncated uniform distribution Uniform(0, $\alpha_1^{(2)}$). We assume p-values from different stages are independent.

$$F(p_{11}p_{12} = x \mid \alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)}, p_{21} \geq \alpha_2^{(1)}, p_{11}p_{12} < \alpha_1^{(2)})$$

$$= F(p_{11}p_{12} = x \mid \alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)}, p_{11}p_{12} < \alpha_1^{(2)})$$

$$= \frac{\Pr(p_{11}p_{12} < x, \alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)}, p_{11}p_{12} < \alpha_1^{(2)})}{\Pr(\alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)}, p_{11}p_{12} < \alpha_1^{(2)})}$$

$$= \frac{\Pr(p_{11}p_{12} < x, \alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)})}{\Pr(\alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)}, p_{11}p_{12} < \alpha_1^{(2)})}, \text{ x} < \alpha_1^{(2)} \qquad (6.5)$$

In practice, $\alpha_1^{(2)} < \alpha_1^{(1)}$, hence, $0 < p_{11}p_{12} < \alpha_1^{(2)} < \alpha_1^{(1)}$. The numerator in (6.5) can be derived as:

$$\Pr(p_{11}p_{12} < x, \alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)}) = \Pr(p_{12} < \frac{x}{p_{11}}, \alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)})$$

$$= \int_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} \Pr(p_{12} < \frac{x}{p_{11}} \mid p_{11}) dp_{11}$$

$$= \int_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} \frac{x}{p_{11}} dp_{11} = x \ln(p_{11}) \Big|_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} = x \ln(\frac{\alpha_F^{(1)}}{\alpha_1^{(1)}}), \; 0 \leq \text{x} < \alpha_1^{(2)} \qquad (6.6)$$

Similarly, the denominator in (6.5) can be derived as:

$$\Pr(\alpha_1^{(1)} \le p_{11} < \alpha_F^{(1)}, p_{11}p_{12} < \alpha_1^{(2)}) = \Pr(p_{12} < \frac{\alpha_1^{(2)}}{p_{11}}, \alpha_1^{(1)} \le p_{11} < \alpha_F^{(1)})$$

$$= \int_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} \Pr(p_{12} < \frac{\alpha_1^{(2)}}{p_{11}} \mid p_{11}) dp_{11}$$

$$= \int_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} \frac{\alpha_1^{(2)}}{p_{11}} dp_{11} = \alpha_1^{(2)} \ln(p_{11}) \,|_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} = \alpha_1^{(2)} \ln(\frac{\alpha_F^{(1)}}{\alpha_1^{1}}) \qquad (6.7)$$

Therefore, given the three-stage setting ($p_{11}p_{12} < \alpha_1^{(2)}$) with the endpoint 1 as the primary study endpoint, the cumulative distribution function of $p_{11}p_{12}$ under the global null hypothesis (6.2) is the ratio of (6.6) to (6.7):

$$\Pr(p_{11}p_{12} < x \mid \alpha_1^{(1)} \le p_{11} < \alpha_F^{(1)}, p_{11}p_{12} < \alpha_1^{(2)}) = \frac{x}{\alpha_1^{(2)}} \qquad (6.8)$$

where $0 < x < \alpha_1^{(2)} < \alpha_1^{(1)}$.

which is the truncated uniform distribution - Uniform(0, $\alpha_1^{(2)}$).

c) Three-stage setting with $\alpha_1^{(2)} \le p_{11}p_{12} < \alpha_F^{(2)}$ and $p_{21}p_{22} \ge \alpha_2^{(2)}$ at interim II

At the interim II, if $\alpha_1^{(2)} \le p_{11}p_{12} < \alpha_F^{(2)}$ and $p_{21}p_{22} \ge \alpha_2^{(2)}$, then the study will be continued to the final confirmatory stage with the endpoint 1 as the primary study endpoint. Given this condition, the cumulative distribution function for $p_{11}p_{12}$ is:

$$F(p_{11}p_{12} = x \mid \alpha_1^{(1)} \le p_{11} < \alpha_F^{(1)}, p_{21} \ge \alpha_2^{(1)}, \alpha_1^{(2)} \le p_{11}p_{12} < \alpha_F^{(2)}, p_{21}p_{22} \ge \alpha_2^{(2)})$$

$$= F(p_{11}p_{12} = x \mid \alpha_1^{(1)} \le p_{11} < \alpha_F^{(1)}, \alpha_1^{(2)} \le p_{11}p_{12} < \alpha_F^{(2)})$$

$$= \Pr(p_{11}p_{12} \le x \mid \alpha_1^{(1)} \le p_{11} < \alpha_F^{(1)}, \alpha_1^{(2)} \le p_{11}p_{12} < \alpha_F^{(2)})$$

$$= \frac{\Pr(p_{11}p_{12} \le x, \alpha_1^{(1)} \le p_{11} < \alpha_F^{(1)}, \alpha_1^{(2)} \le p_{11}p_{12} < \alpha_F^{(2)})}{\Pr(\alpha_1^{(1)} \le p_{11} < \alpha_F^{(1)}, \alpha_1^{(2)} \le p_{11}p_{12} < \alpha_F^{(2)})}$$

$$= \frac{\Pr(\alpha_1^{2} \le p_{11}p_{12} < x, \alpha_1^{(1)} \le p_{11} < \alpha_F^{(1)})}{\Pr(\alpha_1^{(1)} \le p_{11} < \alpha_F^{(1)}, \alpha_1^{(2)} \le p_{11}p_{12} < \alpha_F^{(2)})}, \ x < \alpha_F^{(2)} \qquad (6.9)$$

The numerator in (6.9) can be derived as:

$$\Pr(\alpha_1^{(2)} \le p_{11}p_{12} \le x, \alpha_1^{(1)} \le p_{11} < \alpha_F^{(1)}) = \Pr(\frac{\alpha_1^2}{p_{11}} \le p_{12} \le \frac{x}{p_{11}}, \alpha_1^{(1)} \le p_{11} < \alpha_F^{(1)})$$

$$= \int_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} \Pr(\frac{\alpha_1^{(2)}}{p_{11}} \le p_{12} \le \frac{x}{p_{11}} \mid p_{11}) dp_{11} = \int_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} \frac{x-\alpha_1^{(2)}}{p_{11}} dp_{11}$$

$$= (x-\alpha_1^{(2)})\ln(p_{11})\,|_{\alpha_1^1}^{\alpha_F^{(1)}} = (x-\alpha_1^{(2)})\ln(\frac{\alpha_F^{(1)}}{\alpha_1^{(1)}}),\ \mathrm{x} < \alpha_F^{(2)} \tag{6.10}$$

Similarly, the denominator in (6.9) can be derived as:

$$\Pr(\alpha_1^{(1)} \le p_{11} < \alpha_F^{(1)}, \alpha_1^{(2)} \le p_{11}p_{12} < \alpha_F^{(2)}) = \Pr(\frac{\alpha_1^{(2)}}{p_{11}} \le p_{12} < \frac{\alpha_F^{(2)}}{p_{11}}, \alpha_1^{(1)} \le p_{11} < \alpha_F^{(1)})$$

$$= \int_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} \Pr(\frac{\alpha_1^{(2)}}{p_{11}} \le p_{12} < \frac{\alpha_F^{(2)}}{p_{11}} \mid p_{11}) dp_{11} = \int_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} \frac{\alpha_F^{(2)}-\alpha_1^{(2)}}{p_{11}} dp_{11}$$

$$= (\alpha_F^{(2)}-\alpha_1^{(2)})\ln(p_{11})\,|_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} = (\alpha_F^{(2)}-\alpha_1^{(2)})\ln(\frac{\alpha_F^{(1)}}{\alpha_1^{(1)}}) \tag{6.11}$$

Therefore, given the three-stage setting ($\alpha_1^{(2)} \le$ p$_{11}$p$_{12}$ $< \alpha_F^{(2)}$ and p$_{11}$p$_{12}$ $\ge \alpha_2^{(2)}$) with the endpoint 1 as the primary study endpoint, the cumulative distribution function of p$_{11}$p$_{12}$ under the global null hypothesis (6.2) is the ratio of (6.10) to (6.11):

$$F(p_{11}p_{12} = x \mid \alpha_1^{(1)} \le p_{11} < \alpha_F^{(1)}, p_{21} \ge \alpha_2^{(1)}, \alpha_1^{(2)} \le p_{11}p_{12} < \alpha_F^{(2)}, p_{21}p_{22} \ge \alpha_2^{(2)})$$

$$= \Pr(p_{11}p_{12} \le x \mid \alpha_1^{(1)} \le p_{21} < \alpha_F^{(1)}, \alpha_1^{(2)} \le p_{11}p_{12} < \alpha_F^{(2)})$$

$$= \frac{x-\alpha_1^{(2)}}{\alpha_F^{(2)}-\alpha_1^{(2)}} \tag{6.12}$$

where $\alpha_1^{(2)} \le \mathrm{x} < \alpha_F^{(2)}$.

Therefore, the conditional distribution function of $\frac{p_{11}p_{12}-\alpha_1^{(2)}}{\alpha_F^{(2)}-\alpha_1^{(2)}}$ is Uniform($\alpha_1^{(2)}, \alpha_F^{(2)}$).

6.3.2 With primary study endpoint change

a) With primary endpoint change at the interim I

If $p_{11} \geq \alpha_1^{(1)}$ and $p_{21} < \alpha_2^{(1)}$, the endpoint 2 is considered as the primary study endpoint. Given this condition, the cumulative distribution function of p_{21} under the global null hypothesis (6.2) is as follows.

$$F(p_{21} = x \mid p_{21} < \alpha_2^{(1)}) = \Pr(p_{21} < x \mid p_{21} < \alpha_2^{(1)}) = \frac{\Pr(p_{21} < x, p_{21} < \alpha_2^{(1)})}{\Pr(p_{21} < \alpha_2^{(1)})}$$

$$= \frac{\Pr(p_{21} < x)}{\Pr(p_{21} < \alpha_2^{(1)})} = \frac{x}{\alpha_2^{(1)}}, \text{ where } 0 < x < \alpha_2^{(1)} \qquad (6.13)$$

which is the truncated uniform distribution - Uniform(0, $\alpha_2^{(1)}$).

b) With primary study endpoint change at the interim II

If $\alpha_1^{(1)} \leq p_{11} < \alpha_F^1$ and $p_{21} \geq \alpha_2^{(1)}$ at the interim I, an intermediate stage is planned. At the interim II, if $p_{11}p_{12} \geq \alpha_1^{(2)}$ and $p_{21}p_{22} < \alpha_2^{(2)}$, the endpoint 2 is considered as the primary study endpoint. Given this condition, the cumulative distribution function of $p_{21}p_{22}$ under the global null hypothesis (6.2) is.

$$F(p_{21}p_{22} = x \mid \alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)}, p_{21} \geq \alpha_2^{(1)}, p_{11}p_{12} \geq \alpha_1^{(2)}, p_{21}p_{22} < \alpha_2^{(2)})$$

$$= F(p_{21}p_{22} = x \mid p_{21} \geq \alpha_2^{(1)}, p_{21}p_{22} < \alpha_2^{(2)})$$

$$= \frac{\Pr(p_{21}p_{22} \leq x, p_{21} \geq \alpha_2^{(1)}, p_{21}p_{22} < \alpha_2^{(2)})}{\Pr(p_{21} \geq \alpha_2^{(1)}, p_{21}p_{22} < \alpha_2^{(2)})}$$

$$= \frac{\Pr(p_{21}p_{22} < x, p_{21} \geq \alpha_2^{(1)})}{\Pr(p_{21} \geq \alpha_2^{(1)}, p_{21}p_{22} < \alpha_2^{(2)})}, \ x < \alpha_2^{(2)} \qquad (6.14)$$

In practice, $\alpha_2^{(2)} < \alpha_2^{(1)}$, hence, $0 < p_{21}p_{22} < \alpha_2^{(2)} < \alpha_2^{(1)}$. The numerator in (6.14) can be derived as:

$$\Pr(p_{21}p_{22} \leq x, p_{21} \geq \alpha_2^{(1)}) = \Pr(p_{22} < \frac{x}{p_{21}}, p_{21} \geq \alpha_2^{(1)}) = \int_{\alpha_2^{(1)}}^{1} \Pr(p_{22} < \frac{x}{p_{21}} \mid p_{21}) dp_{21}$$

$$= \int_{\alpha_2^{(1)}}^{1} \frac{x}{p_{21}} dp_{21} = x \ln(p_{21}) |_{\alpha_2^{(1)}}^{1} = -x \ln(\alpha_2^{(1)}) , \text{ x} < \alpha_2^{(2)} \quad (6.15)$$

Similarly, the denominator in (6.14) can be derived as:

$$\Pr(p_{21} \geq \alpha_2^{(1)}, p_{21}p_{22} < \alpha_2^{(2)}) = \Pr(p_{22} < \frac{\alpha_2^{(2)}}{p_{21}}, p_{21} \geq \alpha_2^{(1)}) = \int_{\alpha_2^{(1)}}^{1} \Pr(p_{22} < \frac{\alpha_2^{(2)}}{p_{21}} | p_{21}) dp_{21}$$

$$= \int_{\alpha_2^{(1)}}^{1} \frac{\alpha_2^{(2)}}{p_{21}} dp_{21} = \alpha_2^{(2)} \ln(p_{21}) |_{\alpha_2^{(1)}}^{1} = -\alpha_2^{(2)} \ln(\alpha_2^{(1)}) \quad (6.16)$$

Therefore, given primary study endpoint change at the interim II, the cumulative distribution function of $p_{21}p_{22}$ under the global null hypothesis (6.2) of is the ratio of (6.15) to (6.16).

$$\Pr(p_{21}p_{22} < x | p_{21} \geq \alpha_2^{(1)}, p_{21}p_{22} < \alpha_2^{(2)}) = \frac{x}{\alpha_2^{(2)}} \quad (6.17)$$

where $0 < \text{x} < \alpha_2^{(2)} < \alpha_2^{(1)}$.

which is the truncated uniform distribution - Uniform(0, $\alpha_2^{(2)}$).

6.4 Final analysis on the primary study endpoint under H_{0i}

In this section, we discuss the final analysis on the primary study endpoint under the global null hypothesis H_{0i}. The analyses of the comparison between the selected dose and the control at multiple level α is presented in Section 7.2.

6.4.1 Two-stage setting

(a) No primary study endpoint change at the interim I

If there is no primary study endpoint change at the interim I, the combined p-value for the final analysis is defined based on the study endpoint 1 as:

$$C(p_{11}, p_{12}) = p_{11}p_{12} \quad (6.18)$$

The global null hypothesis H_{01} as specified in (6.2) is rejected if the combined p-value = $p_{11}p_{12} < \alpha_1^{(1)} c_{\alpha*}$. The critical probability $c_{\alpha*}$ is defined as follows corresponding to the significance level α^*:

$$c_{\alpha*} = \exp(-0.5\,\chi^2_{\alpha^*,4}) \tag{6.19}$$

The critical value for the combine p-value is $\alpha_1^{(1)} c_{\alpha*}$. This is because $p_{11}/\alpha_1^{(1)}$ follows Uniform(0,1) under the null hypothesis H_{01} given the two-stage setting with the endpoint 1 as the primary study endpoint as determined from the interim I.

(b) With primary study endpoint change at interim I

If primary study endpoint is changed at the interim I, the combined p-value for final analysis is defined based on the endpoint 2.

$$C(p_{21}, p_{22}) = p_{21}p_{22} \tag{6.20}$$

The global null hypothesis H_{02} as specified in (6.2) is rejected if the combined p-value = $p_{21}p_{22} < \alpha_2^{(1)} c_{\alpha*}$. The critical value of the combine p-value is determined as $\alpha_1^{(1)} c_{\alpha*}$ since conditional distribution of $p_{21}/\alpha_2^{(1)}$ follows Uniform(0,1) under H_{02}.

6.4.2 Three stage setting

Similar to two-stage setting, depending on whether the primary endpoint is changed at the interim II or not, final analysis can be carried out as follows.

(a) No primary study endpoint change at the interim II - $p_{11}p_{12} < \alpha_1^{(2)}$

If the primary study endpoint is not changed at the interim II ($p_{11}p_{12} < \alpha_1^{(2)}$), the combined p-value is defined as.

$$C(p_{11}, p_{12}, p_{13}) = p_{11}p_2p_{13} \tag{6.21}$$

The global null hypothesis H_{01} is rejected if the combined p-value = $p_{11}p_{12}p_{13} < \alpha_1^{(2)} c_{\alpha*}$. The critical value for the combine p-value is $\alpha_1^{(2)} c_{\alpha*}$. This is because $p_{11}p_{12}/\alpha_1^{(2)}$ follows Uniform(0,1) under null hypothesis H_{01} given the three-stage setting with the endpoint 1 as the primary study endpoint ($p_{11}p_{12} < \alpha_1^{(2)}$).

(b) No primary study endpoint changed at the interim II - $\alpha_1^{(2)} \le p_{11}p_{12} < \alpha_F^{(2)}$ and $p_{21}p_{22} \ge \alpha_2^{(2)}$

If $\alpha_1^{(2)} \le p_{11}p_{12} < \alpha_F^{(2)}$ and $p_{21}p_{22} \ge \alpha_2^{(2)}$, the primary study endpoint is not changed at the interim II. The combined p-value from the three stages is defined as

$$\text{P-value} = C(p_{11}, p_{12}, p_{13}) = \frac{p_{11}p_{12} - \alpha_1^{(2)}}{\alpha_F^{(2)} - \alpha_1^{(2)}} p_{13} \tag{6.22}$$

Given $\alpha_1^{(2)} \le p_{11}p_{12} < \alpha_F^{(2)}$ and $p_{21}p_{22} \ge \alpha_2^{(2)}$, conditional distribution of $\frac{p_{11}p_{12} - \alpha_1^{(2)}}{\alpha_F^{(2)} - \alpha_1^{(2)}}$ is Uniform(0,1). Therefore, the global null hypothesis H_{01} is rejected if the combined p-value = $\frac{p_{11}p_{12} - \alpha_1^{(2)}}{\alpha_F^{(2)} - \alpha_1^{(2)}} p_{13} < c_{\alpha^*}$.

c) With primary study endpoint change at the interim II

If the primary study endpoint is changed at the interim II ($p_{11}p_{12} \ge \alpha_1^{(2)}$ and $p_{21}p_{22} < \alpha_2^{(2)}$), the combined p-value is defined based on study endpoint 2 as:

$$C(p_{21}, p_{22}, p_{23}) = p_{21}p_{22}p_{23} \tag{6.23}$$

The global null hypothesis H_{02} is rejected if the combined p-value = $p_{21}p_{22}p_{23} < \alpha_2^{(2)} c_{\alpha^*}$. The critical value for the combine p-value is $\alpha_2^{(2)} c_{\alpha^*}$. This is because $p_{21}p_{22}/\alpha_2^{(2)}$ follows Uniform(0,1) under H_{02} given the primary study endpoint change at the interim II.

6.4.3 Summary

To simplify notation, we use “CP-value” for the combined p-value for the final analysis, which is

- $p_{11}p_{12}$, two-stage setting, no endpoint change at the interim I
- $p_{21}p_{22}$, two-stage setting, endpoint change at the interim I

- $p_{11}p_{12}p_{13}$, three-stage setting, no endpoint change at the interim II ($p_{11}p_{12} < \alpha_1^{(2)}$)
- $\frac{p_{11}p_{12} - \alpha_1^{(2)}}{\alpha_F^{(2)} - \alpha_1^{(2)}} p_{13}$, three-stage setting, no endpoint change at the interim II ($\alpha_1^{(2)} \le p_{11}p_{12} < \alpha_F^{(2)}$ and $p_{11}p_{12} \ge \alpha_2^{(2)}$)
- $p_{21}p_{22}p_{23}$, three-stage setting, endpoint change at the interim II

We use ac_{α^*} denote critical value for CP-value, which is corresponding to the significance level of α^*, where a is equal to

- $\alpha_1^{(1)}$, two-stage setting, no endpoint change at the interim I
- $\alpha_2^{(1)}$, two-stage setting, endpoint change at the interim I
- $\alpha_1^{(2)}$, three-stage setting, no endpoint change at the interim II ($p_{11}p_{12} < \alpha_1^{(2)}$)
- 1, three-stage setting, no endpoint change at the interim II ($\alpha_1^{(2)} \le p_{11}p_{12} < \alpha_F^{(2)}$ and $p_{11}p_{12} \ge \alpha_2^{(2)}$)
- $\alpha_2^{(2)}$, three-stage setting, endpoint change at the interim II

To perform the final analysis, the global null hypothesis H_0 as defined in (6.3) is rejected at α level if CP-value < ac_{α^*}. The proof of overall Type I error control is provided in the next section.

6.5 Proof of overall Type I error rate control

The proof provided here only considers Type I error control for hypothesis testing on overall comparison across all the dose arms over the control arm with respect to both study endpoints.

6.5.1 Trial decision paths

Trial decision paths are shown in Figure 6-2. Each area in Figure 6-2 is also indicated in Figure 6-1 with the same indicator as A_m (m=1, 2, 3, 4) or B_n (n=1, 2, 3, 4).

6.5.2 Overall Type I error rate control

Let RH_0 denote "reject H_0 | H_0 is true", and ER for Type I error rate, then ER = $\Pr(RH_0)$. The global null hypothesis H_0 is defined in (6.3), which considers both study endpoints. The Type I error rate is calculated based on the trial decision paths as shown in Figure 6-2. We use the same notations of A_m and B_n (m, n = 1, 2, 3, 4) for decision elements, and $ER_{A_mB_n}$ for the Type I error rate for the decision A_mB_n. We assume that the p-values from different stages are independent

The overall Type I error rate is

$$ER = \Pr(RH_0) = \sum_{m=1}^{4} \Pr(RH_0 |A_m)P(A_m)$$

$$= \Pr(RH_0|A_1) \Pr(A_1) + \Pr(RH_0|A_2) \Pr(A_2) + \Pr(RH_0|A_3) \Pr(A_3) + \Pr(RH_0|A_4) \Pr(A_4) \qquad (6.24)$$

Since $\Pr(RH_0|A_4) = \Pr(RH_0|B_1, A_4) \Pr(B_1 |A_4) + \Pr(RH_0|B_2, A_4) \Pr(B_2 |A_4) + \Pr(RH_0|B_3, A_4) \Pr(B_3 |A_4) + \Pr(RH_0|B_4, A_4) \Pr(B_4 |A_4)$

The total Type I error rate is

$$\begin{aligned} ER = \Pr(RH_0) &= \Pr(RH_0|A_1) \Pr(A_1) + \Pr(RH_0|A_2) \Pr(A_2) + \Pr(RH_0|A_3) \Pr(A_3) \\ &\quad + \Pr(A_4)[\Pr(RH_0|B_1, A_4) \Pr(B_1|A_4) + \Pr(RH_0|B_2, A_4) \Pr(B_2|A_4) \\ &\quad + \Pr(RH_0|B_3, A_4) \Pr(B_3|A_4) + \Pr(RH_0|B_4, A_4) \Pr(B_4|A_4)] \\ &= \Pr(RH_0|A_1) \Pr(A_1) + \Pr(RH_0|A_2) \Pr(A_2) + \Pr(RH_0|A_3) \Pr(A_3) \\ &\quad + \Pr(RH_0|B_1, A_4) \Pr(B_1, A_4) + \Pr(RH_0|B_2, A_4) \Pr(B_2, A_4) \\ &\quad + \Pr(RH_0|B_3, A_4) \Pr(B_3, A_4) + \Pr(RH_0|B_4, A_4) \Pr(B_4, A_4) \end{aligned}$$

(6.25)

Since the final analysis is performed at the significance level of α^*,

$\Pr(RH_0|A_1) = \Pr(p_{11}p_{12} < \alpha_1^{(1)} c_{\alpha^*} | A_1, H_{01}) = \alpha^*$

$\Pr(RH_0|A_2) = \Pr(p_{21}p_{22} < \alpha_2^{(1)} c_{\alpha^*} | A_2, H_{02}) = \alpha^*$

$\Pr(RH_0|A_3) = 0$

$\Pr(RH_0|B_1, A_4) = \Pr(p_{11}p_{12}\, p_{13} < \alpha_1^{(2)} c_{\alpha^*} | B_1, A_4, H_{01}) = \alpha^*$

$\Pr(RH_0|B_2, A_4) = \Pr(p_{21}p_{22}\, p_{23} < \alpha_2^{(2)} c_{\alpha^*} | B_2, A_4, H_{02}) = \alpha^*$

$\Pr(RH_0|B_3, A_4) = 0$

$$\Pr(RH_0|B_4, A_4) = \Pr\left(\frac{p_{11}p_{12} - \alpha_1^{(2)}}{\alpha_F^{(2)} - \alpha_1^{(2)}} p_{13} < c_{\alpha^*} \mid B_4, A_4, H_{01}\right) = \alpha^*$$

Therefore, Type I error rate for each decision path can be calculated as follows

$ER_{A_1} = \Pr(RH_0|A_1) \Pr(A_1)$

$= \alpha^* \Pr(A_1) = \alpha^* \Pr(p_{11} < \alpha_1^{(1)})$

$= \alpha^* \alpha_1^{(1)}$ (6.26a)

$ER_{A_2} = \Pr(RH_0|A_2) \Pr(A_2)$

$= \alpha^* \Pr(A_2) = \alpha^* \Pr(p_{11} \geq \alpha_1^{(1)}, p_{21} < \alpha_2^{(1)})$

$= \alpha^* \alpha_2^{(1)} (1 - \alpha_1^{(1)})$ (6.26b)

$ER_{A_3} = \Pr(RH_0|A_3) \Pr(A_3)$

$= 0$ (6.26c)

$ER_{A_4B_1} = \Pr(RH_0|B_1, A_4) \Pr(B_1, A_4)$

$= \alpha^* \Pr(p_{11}p_{12} < \alpha_1^{(2)}, \alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)}, p_{21} \geq \alpha_2^{(1)})$

$= \alpha^* \Pr(p_{11}p_{12} < \alpha_1^{(2)}, \alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)}) \Pr(p_{21} \geq \alpha_2^{(1)})$

$$= \alpha^{*}\ (1\text{-}\alpha_2^{(1)}) \int_{\alpha_1^1}^{\alpha_F^1} \Pr(p_{12} < \frac{\alpha_1^{(2)}}{p_{11}} \mid p_{11}) dp_{11}$$

$$= \alpha^{*}\ (1\text{-}\alpha_2^{(1)}) \int_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} \frac{\alpha_1^{(2)}}{p_{11}} dp_{11}$$

$$= \alpha^{*} \alpha_1^{(2)} (1 - \alpha_2^{(1)}) \ln(\frac{\alpha_F^{(1)}}{\alpha_1^{(1)}}) \qquad (6.26d)$$

$ER_{A_4B_2}$ = Pr(RH$_0$|B$_2$, A$_4$)Pr(B$_2$, A$_4$)

$$= \alpha^{*}\ \Pr(p_{21}p_{22} < \alpha_2^{(2)},\ p_{21} \geq \alpha_2^{(1)},\ \alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)},\ p_{11}p_{12} \geq \alpha_1^{(2)})$$

$$= \alpha^{*}\ \Pr(p_{21}p_{22} < \alpha_2^{(2)},\ p_{21} \geq \alpha_2^{(1)})\ \Pr(\alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)},\ p_{11}p_{12} \geq \alpha_1^{(2)})$$

$$= \alpha^{*} \int_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} \Pr(p_{12} \geq \frac{\alpha_1^{(2)}}{p_{11}} \mid p_{11}) dp_{11} \cdot \int_{\alpha_2^{(1)}}^{1} \Pr(p_{22} < \frac{\alpha_2^{(2)}}{p_{21}} \mid p_{21}) dp_{21}$$

$$= \alpha^{*} \int_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} (1 - \frac{\alpha_1^{(2)}}{p_{11}}) dp_{11} \cdot \int_{\alpha_2^{(1)}}^{1} \frac{\alpha_2^{(2)}}{p_{21}} dp_{21}$$

$$= -\alpha^{*} \alpha_2^{(2)} [\alpha_F^{(1)} - \alpha_1^{(1)} - \alpha_1^{(2)} \ln(\frac{\alpha_F^{(1)}}{\alpha_1^{(1)}})] \ln(\alpha_2^{(1)}) \qquad (6.26e)$$

$ER_{A_4B_3}$ = Pr(RH$_0$|B$_3$, A$_4$)Pr(B$_3$, A$_4$)

$$= 0 \qquad (6.26f)$$

$ER_{A_4B_4}$ = Pr(RH$_0$|B$_4$, A$_4$)Pr(B$_4$, A$_4$)

$$= \alpha^{*}\ \Pr(\alpha_1^{(2)} \leq p_{11}p_{12} < \alpha_F^{(2)},\ p_{21}p_{22} \geq \alpha_2^{(2)},\ \alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)},\ p_{21} \geq \alpha_2^{(1)})$$

$$= \alpha^{*}\ \Pr(\alpha_1^{(2)} \leq p_{11}p_{12} < \alpha_F^{(2)},\ \alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)}) \Pr(p_{21}p_{22} \geq \alpha_2^{(2)},\ p_{21} \geq \alpha_2^{(1)})$$

$$= \alpha^{*} \int_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} \Pr(\frac{\alpha_1^{(2)}}{p_{11}} \leq p_{12} < \frac{\alpha_F^{(2)}}{p_{11}} \mid p_{11}) dp_{11} \cdot \int_{\alpha_2^{(1)}}^{1} \Pr(p_{22} \geq \frac{\alpha_2^{(2)}}{p_{21}} \mid p_{21}) dp_{21}$$

$$= \alpha^{*} (\alpha_F^{(2)} - \alpha_1^{(2)}) \ln(\frac{\alpha_F^{(1)}}{\alpha_1^{(1)}}) [1\text{-}\alpha_2^{(1)} + \alpha_2^{(2)} \ln(\alpha_2^{(1)})] \qquad (6.26g)$$

The Type I error rate for the two-stage setting is

$$\begin{aligned} \mathrm{ER}_{\mathrm{II}} &= \sum_{m=1}^{3} ER_{A_m} \\ &= ER_{A_1} + ER_{A_2} \\ &= \alpha^{*}\,[\alpha_1^{(1)} + \alpha_2^{(1)}\,(1 - \alpha_1^{(1)})] \end{aligned} \tag{6.27a}$$

The Type I error rate for the three-stage setting is

$$\begin{aligned} \mathrm{ER}_{\mathrm{III}} &= \sum_{n=1}^{4} ER_{A_4 B_n} \\ &= ER_{A_4 B_1} + ER_{A_4 B_2} + ER_{A_4 B_4} \\ &= \alpha^{*}\,\{\alpha_1^{(2)}(1 - \alpha_2^{(1)})\ln(\frac{\alpha_F^{(1)}}{\alpha_1^{(1)}}) \\ &\quad - \alpha_2^{(2)}[\alpha_F^{(1)} - \alpha_1^{(1)} - \alpha_1^{(2)}\ln(\frac{\alpha_F^{(1)}}{\alpha_1^{(1)}})]\ln(\alpha_2^{(1)}) \\ &\quad + (\alpha_F^{(2)} - \alpha_1^{(2)})\ln(\frac{\alpha_F^{(1)}}{\alpha_1^{(1)}})\;[1 - \alpha_2^{(1)} + \alpha_2^{(2)}\ln(\alpha_2^{(1)})]\} \end{aligned} \tag{6.27b}$$

Therefore, the overall Type I error rate is

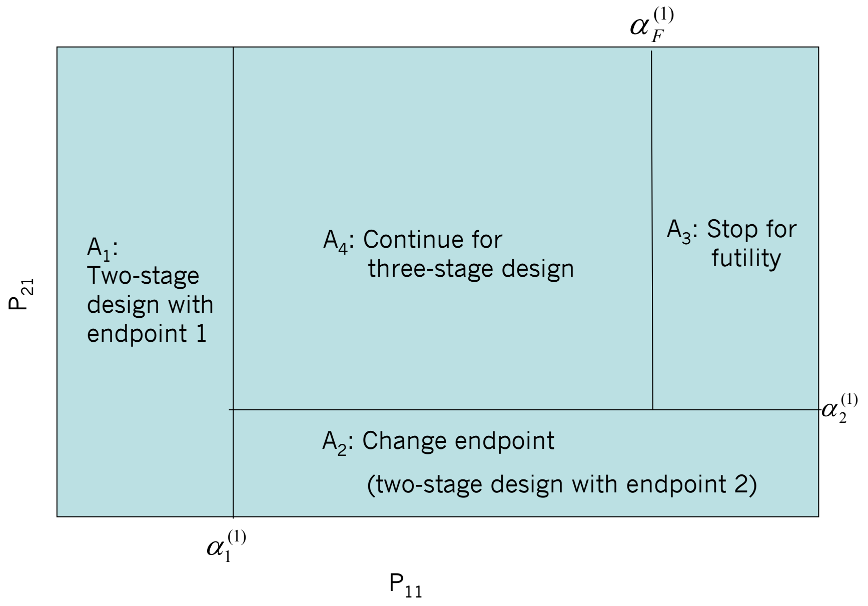
$$\begin{aligned} \mathrm{ER} &= \mathrm{ER}_{\mathrm{II}} + \mathrm{ER}_{\mathrm{III}} \\ &= \alpha^{*}\,\{\alpha_1^{(1)} + (1 - \alpha_1^{(1)})\alpha_2^{(1)} + \alpha_1^{(2)}(1 - \alpha_2^{(1)})\ln(\frac{\alpha_F^{(1)}}{\alpha_1^{(1)}}) \\ &\quad - \alpha_2^{(2)}[\alpha_F^{(1)} - \alpha_1^{(1)} - \alpha_1^{(2)}\ln(\frac{\alpha_F^{(1)}}{\alpha_1^{(1)}})]\ln(\alpha_2^{(1)}) \\ &\quad + (\alpha_F^{(2)} - \alpha_1^{(2)})\ln(\frac{\alpha_F^{(1)}}{\alpha_1^{(1)}})\;[1 - \alpha_2^{(1)} + \alpha_2^{(2)}\ln(\alpha_2^{(1)})]\} \\ &\le \alpha \end{aligned} \tag{6.28}$$

where α is nominal significance level, under which the overall Type I error is controlled. In practice for a clinical trail design, the threshold values $\alpha_1^{(1)}$, $\alpha_2^{(1)}$, $\alpha_1^{(2)}$,

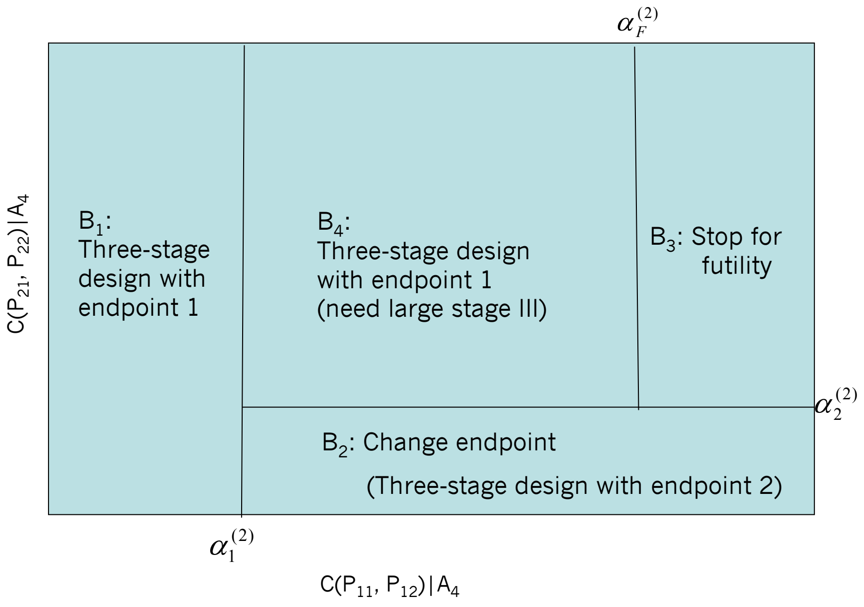
$\alpha_2^{(2)}$, $\alpha_F^{(1)}$ and $\alpha_F^{(2)}$, and allowable type I error rate α are given, the significance level α^* for the final analysis can be resolved from (6.28).

The above overall Type I error rate control (6.28) is carried out for the primary endpoint. However, the another endpoint (the second endpoint) is also a pre-determined endpoint that could be potentially kept or switched as the primary endpoint based on the interim data. Therefore, in practice, it is of interest to test whether this endpoint achieves statistical significance. Further, this endpoint could be a co-primary endpoint in some therapeutic area. In our future research, we plan to develop a statistical method to test the second endpoint under Type I error rate control.

Figure 6-2 Diagram of trial decision paths


$\alpha_F^{(1)}$
A1: Two-stage design with endpoint 1
A4: Continue for three-stage design
A3: Stop for futility
$\alpha_2^{(1)}$
A2: Change endpoint (two-stage design with endpoint 2)
$\alpha_1^{(1)}$
P21
P11

(a) Interim I (initial learning stage)


$\alpha_F^{(2)}$
B1: Three-stage design with endpoint 1
B4: Three-stage design with endpoint 1 (need large stage III)
B3: Stop for futility
$\alpha_2^{(2)}$
B2: Change endpoint (Three-stage design with endpoint 2)
$\alpha_1^{(2)}$
C(P21, P22)|A4
C(P11, P12)|A4

(b) Interim II (intermediate stage)

7 A varying-stage adaptive phase II/III clinical trial design: Dose selection and multiple comparisons on the primary study endpoint

7.1 Dose selection

At the time of phase II/III clinical trial planning, since limited efficacy and safety information on the study treatment is available, several doses are included in the initial learning phase or intermediate stage to cover wide range of dosing. The goal of the interim analyses is to select doses, and determine sample size to ensure adequate statistical power for the final confirmatory stage.

With respect to dose selection, as described in Section 5.2, many researchers have proposed different approaches. For our varying-stage adaptive phase II/III clinical trial design, we use closed testing procedure to perform hypothesis testings. This approach has been used by many researchers, such as Bauer and Kieser (1999), Bretz, Schmidli, Koenig, Racine and Maurer (2006), and Koenig, Brannath, Bretz, and Posch (2008). The details of hypothesis testing under closed testing procedure at multiple level α for the final analysis is provided in Section 7.2. Since closed testing procedure is used, the dose selection rules do not impact Type I error control. Therefore, the specific dose selection methods are not covered in this thesis. To apply our design, one may choose dose selection criteria based on their own situation. Again, Type I error rate is protected by using closed testing procedure here.

For illustration purpose, in Section 7.2.2 and Chapter 9, two dose arms are selected for the final stage; and in our simulation as presented in Section 9.2, the dose arm with the smallest p-value or the smallest combined p-value is chosen for the final confirmatory stage.

7.2 Final analysis of multiple comparisons on the primary study endpoint

7.2.1 Family-wise error rate under multiple comparisons

The issue of multiplicity rises in practice under multiple comparisons for multiple endpoints or multiple treatment arms. In these cases, Family-wise error (FWE) rate has to be considered for overall Type I error rate control. FWE is defined as the probability that we reject one or more of true null hypotheses (see Section 6.2 for null hypotheses) in a set of comparisons. FWE is controlled in a 'weak' sense if FWE rate from a multiple testing procedure is less than or equal to a specified Type I error rate (α) when all null hypotheses are true simultaneously, while FWE is controlled in a 'strong' sense if the FWE rate is maintained less than or equal to a specified α level for any subset of true null hypotheses, regardless of which and how many of the individual null hypotheses are true. The former FWE control is referred as global level α control, and the later FWE control is referred as multiple level α control (Bauer, 19991).

We have addressed the global level α control in Section 6.5. In the next two sections, we discuss closed testing procedure and multiple level α control.

7.2.2 Closed testing procedure

Marcus et al (1976) introduced the closed testing procedure. Let W be a set of null hypotheses, and w_i and w_j be arbitrary elements of W. The set W is closed under intersection if $w_i \bigcap w_j$ is also an element of W. Following the closed testing principle, an elementary null hypothesis is rejected if it is rejected at significance level α and all other intersection null hypotheses containing this elementary null hypothesis are rejected at the same significance level α. This closed testing procedure preserves FWE rate in the strong sense.

Apparently, the global null hypothesis H_0 defined in (6.3) is closed under intersection since all the intersection null hypotheses are nested within the global null hypothesis H_0. Hence the closed testing procedure can be performed in a stepdown process. That is, the global null hypothesis H_0 is tested first, if it is rejected, then test the intersection null hypotheses one level below, and so on. This process is conducted until the level of the elementary null hypothesis. All these hypotheses should be tested at the same significance level.

7.2.3 Multiple level α control

Multiple level α control is carried out under framework of closed testing procedure, so that FWE rate is controlled in the strong sense. As described in Section 6.4, the global null hypothesis H_0 is rejected if CP-value < $ac_{\alpha*}$. In this Section, we discuss final analysis to reject elementary null hypothesis H_{0is} (i=1, 2 for endpoint, s for a selected dose arm) as defined in (6.1) under closed testing procedure.

For phase II/III clinical trials, based on interim results, usually one or two doses of the study drug are chosen for the final confirmatory stage. Let S be the index set of the doses chosen for the final confirmatory phase III. To reject the null hypothesis H_{0is}, $s \in S$, all the intersection null hypotheses $H_{0i,S_1}: \bigcap_{k \in S_1} H_{0ik}$ containing H_{0is}, $s \in S_1 \subseteq S$ have to be rejected. For example, suppose there are four doses included in the initial stage, D = {1, 2, 3, 4}. Based on the interim results, the dose arm 2 and 4 are chosen for the final confirmatory phase III stage, S = {2, 4}. To reject H_{0i2}, both the intersection null hypothesis $H_{0i,\{2,4\}}$ and the elementary null hypothesis H_{0i2} need to be rejected under closed testing procedure.

However, the intersection null hypothesis testing described above is only based on the set S for the dose arms chosen for the final stage. To take full scale of the closed testing procedure, the intersection null hypotheses should not be limited to the dose arms in the final stage. Rather, all the intersection null hypotheses including dose arms dropped from an interim analysis should be considered as long as these intersection null hypotheses contain H_{0is}, such that $H_{0i,D_1}: \bigcap_{k \in D_1} H_{0ik}$, $s \in D_1 \subseteq D = \{1,2,..., K\}$. For the previous example, to reject H_{0i2}, all the null hypotheses $H_{0i,\{1,2,3,4\}}$, $H_{0i,\{1,2,3\}}$, $H_{0i,\{1,2,4\}}$, $H_{0i,\{2,3,4\}}$, $H_{0i,\{1,2\}}$, $H_{0i,\{2,3\}}$, $H_{0i,\{2,4\}}$, and H_{0i2} have to be rejected at the same significance level.

To perform hypothesis testing for the intersection null hypothesis $H_{0i,D_1}: \bigcap_{k \in D_1} H_{0ik}$, the p-values from the final stage and previous stage(s) are combined. For example, the combined p-value – CP-value = $p_{i1}{}^{(D_1)} p_{i2}{}^{(S_1)}$ $s \in S_1 \subseteq D_1 \subseteq D = \{1,2,..., K\}$ under two-stage setting. To simplify notation, we use $CP-value^{(D_1,S_1)}$ to denote the combined p-value for the final analysis to test an intersection null hypothesis, which containing H_{0is} and is constructed based on the set S_1 and D_1, such that $s \in S_1 \subseteq D_1 \subseteq D = \{1,2,..., K\}$. Specifically, $CP-value^{(D_1,S_1)}$ is defined as:

- $p_{11}{}^{(D_1)} p_{12}{}^{(S_1)}$, two-stage setting, no primary endpoint change at interim I
- $p_{21}{}^{(D_1)} p_{22}{}^{(S_1)}$, two-stage setting, primary endpoint change at interim I
- $p_{11}{}^{(D_1)} p_{12}{}^{(D_1)} p_{13}{}^{(S_1)}$, three-stage setting, no primary endpoint change at interim II ($p_{11}p_{12} < \alpha_1^{(2)}$)
- $\dfrac{p_{11}{}^{(D_1)} p_{12}{}^{(D_1)} - \alpha_1^{(2)}}{\alpha_F^{(2)} - \alpha_1^{(2)}} p_{13}{}^{(S_1)}$, three-stage setting, no endpoint change at interim II ($\alpha_1^{(2)} \leq p_{11}p_{12} < \alpha_F^{(2)}$ and $p_{11}p_{12} \geq \alpha_2^{(2)}$)
- $p_{21}{}^{(D_1)} p_{22}{}^{(D_1)} p_{23}{}^{(S_1)}$, three-stage setting, primary endpoint change at interim II

where i =1, 2 for a study endpoint.

The null hypothesis H_{0is} is rejected if the $CP-value^{(D_1,S_1)}$ <$ac_{\alpha*}$ for any D_1 and S_1 such that $s \in S_1 \subseteq D_1 \subseteq D = \{1,2,...,K\}$. One should note that the critical value $ac_{\alpha*}$ for the combined p-value under this closed testing procedure is same as those described in Section 6.4 for global null hypothesis (H_0) testing. Bauer and Kieser (1999) applied the same approach regarding the use of critical value for the combined p-values. Also one should note that some researchers consider the intersection null hypotheses based on the set D_1 since S_1 is a reduced set from D_1 or S_1 is a subset of D_1 with dropped dose arms excluded.

The above hypothesis testing strategy follows the closed testing procedure. Hence the Type I error rate is controlled in the strong sense.

Go back to the previous example, to reject the null hypothesis H_{0i2} (i=1,2 for endpoint) under the closed testing procedure, all the intersection null hypotheses containing H_{0i2} and the elementary null hypothesis H_{0i2} have to be rejected. Hence, the following combined p-values have to be calculated and compared to their corresponding critical values: CP-value$^{(\{1,2,3,4\},\ \{2,4\})}$, CP-value$^{(\{2,3,4\},\ \{2,4\})}$, CP-value$^{(\{1,2,4\},\ \{2,4\}}$, CP-value$^{(\{2,4\},\ \{2,4\})}$, CP-value$^{(\{1,2,3\},\ \{2\})}$, CP-value$^{(\{2,3\},\ \{2\})}$, CP-value$^{(\{1,2\},\ \{2\})}$ and CP-value$^{(\{2\},\ \{2\})}$.

7.2.4 Stagewise p-value adjustment for intersection null hypothesis testing

There are statistical methods available to test intersection null hypotheses. For multivariate normally distributed study endpoints, Hotelling's T^2 test can be used. For normal and other types of response variables, Bonferroni, Sidak and Simes test can be used for stagewise p-value adjustment for intersection null hypothesis testings. Suppose there are m comparisons; Bonferroni-adjusted stagewise p-value is min(1,

$m*\min(p_k))$, $k=1, 2, \ldots, m$, where m is the total number of comparisons; Sidak-adjusted p-value is $1-[1-\min(p_k)]^m$; Simes-adjusted p-value is $m\, p_{(k)}/k$, where $p_{(k)}$ are ordered p-values.

It is well-known that Bonferroni-type p-value adjustments are conservative. Our varying stage adaptive phase II/III clinical trial design is more complex by considering varying number of stages, dropping inefficacious/harmful doses and choosing a more sensitive study endpoint as the primary study endpoint. To fully utilize pre-specified α (allowable Type I error rate) and achieve better statistical power, we use statistical models (e.g. ANOVA, ANCOVA, or mixed effect models) to obtain stagewise p-values for intersection null hypothesis testings.

8 A varying-stage adaptive phase II/III clinical trial design: Statistical power and sample size determination

8.1 Distribution of p-value under alternative hypothesis

Statistical power calculation is performed under the alternative hypothesis with effect size τ. For our proposed design, the test statistic is a combined p-value. Therefore, to define statistical power, we need to know the distribution of p-value under alternative hypothesis.

The distribution of p-value under alternative hypothesis has been studied by many researchers, particularly by Hung, O'Neil, Bauer and Kohne (1997). The density function of a p-value p under the alternative hypothesis for a single treatment arm trial is

$$g(p) = \phi(Z_{1-p} - \sqrt{m}\tau) / \phi(Z_{1-p}), \quad 0<p<1 \qquad (8.1)$$

where ϕ is the density function of standard normal distribution, τ is effect size calculated as treatment effect δ divided by square root of variance σ ($\tau = \delta / \sigma$). Z_{1-p} is the $(1\text{-}p)^{th}$ percentile of the standard normal distribution.

For two sample scenario with equal number of patients in each treatment arm, $\sigma = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}$, where σ_1^2 and σ_2^2 are variance for both treatment groups.

m = n/2, where n is sample size per treatment group.

8.2 Statistical power

For our varying stage phase II/III clinical trial design, the statistical power is calculated based on the trial decision paths as shown in Figure 6-2. We use the same notations of A_m and B_n (m, n = 1, 2, 3, 4) for decision elements, $PW_{A_mB_n}$ for the statistical power for the decision A_mB_n. Again, we assume that the p-values from different stages are independent.

- Decision path A_1: Two-stage setting with the study endpoint 1 as the primary study endpoint

Under decision the path A_1, for which $p_{11}<\alpha_1^{(1)}$, the statistical power PW_{A_1} is calculated as

$$PW_{A_1} = \Pr(p_{11}p_{12} < \alpha_1^{(1)} c_{\alpha*} | A_1, \tau_1)\Pr(A_1)$$

$$= \Pr(p_{11}p_{12} < \alpha_1^{(1)} c_{\alpha*}, p_{11}<\alpha_1^{(1)} | \tau_1)$$

$$= \int_0^{\alpha_1^{(1)}} \left[\int_0^{\alpha_1^{(1)} c_{\alpha*} / p_{11}} g(p_{12})dp_{12} \right] g(p_{11})dp_{11} \tag{8.2a}$$

To claim treatment success, $p_{11}p_{12} < \alpha_1^{(1)} c_{\alpha*}$ has to be satisfied. However, based on the interim I data, if $p_{11}<\alpha_1^{(1)} c_{\alpha*}$ is obtained, then the 2nd stage is deemed not necessary, since no matter how much p_{12} would be, $p_{11}p_{12} < \alpha_1^{(1)} c_{\alpha*}$ will be met anyway. In this case, the trial could be terminated early due to efficacy. Hence, (8.2a) can be partitioned as follows with respect to $\alpha_1^{(1)} c_{\alpha*} \le p_{11} < \alpha_1^{(1)}$ and $p_{11} < \alpha_1^{(1)} c_{\alpha*}$.

$$PW_{A_1} = \int_{\alpha_1^{(1)} c_{\alpha*}}^{\alpha_1^{(1)}} \left[\int_0^{\alpha_1^{(1)} c_{\alpha*} / p_{11}} g(p_{12})dp_{12} \right] g(p_{11})dp_{11}$$

$$+ \int_0^{\alpha_1^{(1)} c_{\alpha*}} g(p_{11})dp_{11} \tag{8.2b}$$

- Decision path A_2: Two-stage setting with the study endpoint 2 as the primary study endpoint

$$PW_{A_2} = \Pr(p_{21}p_{22} < \alpha_2^{(1)} c_{\alpha*} | A_2, \tau_1, \tau_2)\Pr(A_2)$$

$$= \Pr(p_{21}p_{22} < \alpha_2^{(1)} c_{\alpha*}, p_{21} < \alpha_2^{(1)}, p_{11} \ge \alpha_1^{(1)} | \tau_1, \tau_2)$$

$= \Pr(p_{21}p_{22} < \alpha_2^{(1)} c_{\alpha*}, p_{21} < \alpha_2^{(1)} \mid \tau_2) \Pr(p_{11} \geq \alpha_1^{(1)} \mid \tau_1)$

Since if $p_{21} < \alpha_2^{(1)} c_{\alpha*}$ and $p_{11} \geq \alpha_1^{(1)}$, the 2nd stage is deemed not necessary, PW_{A_2} can be expressed as

$$PW_{A_2} = \{\int_{\alpha_2^{(1)} c_{\alpha*}}^{\alpha_2^{(1)}} \left[\int_0^{\alpha_2^{(1)} c_{\alpha*} / p_{21}} g(p_{22}) dp_{22} \right] g(p_{21}) dp_{21} + \int_0^{\alpha_2^{(1)} c_{\alpha*}} g(p_{21}) dp_{21} \}$$

$$\cdot \int_{\alpha_1^{(1)}}^{1} g(p_{11}) dp_{11} \tag{8.3}$$

- Decision path A_3: early stop due to futility

Under the decision path A3, since the trial is stopped due to futility, the statistical power is equal to zero.

$$PW_{A_3} = 0 \tag{8.4}$$

- Decision path A_4B_1: Three-stage setting with the endpoint 1 as the primary endpoint ($p_{11}p_{12} < \alpha_1^{(2)}$)

$$PW_{A_4B_1} = \Pr(p_{11}p_{12}p_{13} < \alpha_1^{(2)} c_{\alpha*} \mid A_4 \& B_1, \tau_1, \tau_2) \Pr(A_4 \& B_1)$$

$$= \Pr(p_{11}p_{12}p_{13} < \alpha_1^{(2)} c_{\alpha*}, \ p_{11}p_{12} < \alpha_1^{(2)}, \ \alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)} \mid \tau_1) \Pr(p_{21} \geq \alpha_2^{(1)} \mid \tau_2)$$

$$\tag{8.5a}$$

Similar to the previous arguments, $p_{11}p_{12} < \alpha_1^{(2)} c_{\alpha*}$ leads to trial success without the need for the final stage. Therefore, the statistical power for the decision path A_4B_1 is

$$PW_{A_4B_1} = \{\int_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} \left\{ \int_{\alpha_1^{(2)} c_{\alpha*} / p_{11}}^{\alpha_1^{(2)} / p_{11}} \left[\int_0^{\alpha_1^{(2)} c_{\alpha*} / p_{11} p_{12}} g(p_{13}) dp_{13} \right] g(p_{12}) dp_{12} \right\} g(p_{11}) dp_{11}$$

$$+ \int_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} \left[\int_0^{\alpha_1^{(2)} c_{\alpha*} / p_{11}} g(p_{12}) dp_{12} \right] g(p_{11}) dp_{11} \} \cdot \int_{\alpha_2^{(1)}}^{1} g(p_{21}) dp_{21} \tag{8.5b}$$

- Decision path A_4B_2: Three-stage setting with the endpoint 2 as the primary endpoint

$$PW_{A_4B_2} = \Pr(p_{21}p_{22}p_{23} < \alpha_2^{(2)} c_{\alpha^*} \mid A_4\&B_2, \tau_1, \tau_2)\Pr(A_4\&B_2)$$

$$= \Pr(p_{21}p_{22}p_{23} < \alpha_2^{(2)} c_{\alpha^*}, p_{21}p_{22} < \alpha_2^{(2)}\ p_{21} \geq \alpha_2^{(1)} \mid \tau_1)$$

$$\cdot \Pr(p_{11}p_{12} \geq \alpha_1^{(2)},\ \alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)} \mid \tau_2)$$

$$= \int_{\alpha_2^{(1)}}^{1} \left\{ \int_0^{\alpha_2^{(2)}/p_{21}} \left[\int_0^{\alpha_2^{(2)} c_{\alpha^*}/p_{21}p_{22}} g(p_{23})dp_{23} \right] g(p_{22})dp_{22} \right\} g(p_{21})dp_{21}$$

$$\cdot \int_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} \left[\int_{\alpha_1^{(2)}/p_{11}}^{1} g(p_{12})dp_{12} \right] g(p_{11})dp_{11} \qquad (8.6a)$$

Similarly, partition (8.6a) with respect to $\alpha_2^{(2)} c_{\alpha^*} \leq p_{21}p_{22} < \alpha_2^{(2)}$ and $p_{21}p_{22} < \alpha_2^{(2)} c_{\alpha^*}$, we have

$$PW_{A_4B_2} = \{ \int_{\alpha_2^{(1)}}^{1} \left\{ \int_{\alpha_2^{(2)} c_{\alpha^*}/p_{21}}^{\alpha_2^{(2)}/p_{21}} \left[\int_0^{\alpha_2^{(2)} c_{\alpha^*}/p_{21}p_{22}} g(p_{23})dp_{23} \right] g(p_{22})dp_{22} \right\} g(p_{21})dp_{21}$$

$$+ \int_{\alpha_2^{(1)}}^{1} \left\{ \int_0^{\alpha_2^{(2)} c_{\alpha^*}/p_{21}} g(p_{22})dp_{22} \right\} g(p_{21})dp_{21} \}$$

$$\cdot \int_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} \left[\int_{\alpha_1^{(2)}/p_{11}}^{1} g(p_{12})dp_{12} \right] g(p_{11})dp_{11} \qquad (8.6b)$$

- Decision path A_4B_3: stop due to futility at interim II

Since the trial is stopped due to futility under the decision A_4B_3, the statistical power is equal to zero.

$$PW_{A_4B_3} = 0 \qquad (8.7)$$

- Decision path A_4B_4:Three-stage setting with the study endpoint 1 ($\alpha_1^{(2)} \leq p_{11}p_{12} < \alpha_F^{(2)}$ and $p_{21}p_{22} \geq \alpha_2^{(2)}$) as the primary study endpoint

$$PW_{A_4B_4} = \Pr(\frac{p_{11}p_{12} - \alpha_1^{(2)}}{\alpha_F^{(2)} - \alpha_1^{(2)}} p_{13} < c_{\alpha^*} \mid A_4\&B_4, \tau_1, \tau_2)\Pr(A_4\&B_4)$$

$$= \Pr\left(\frac{p_{11}p_{12} - \alpha_1^{(2)}}{\alpha_F^{(2)} - \alpha_1^{(2)}} p_{13} < c_{\alpha^*},\ \alpha_1^{(2)} \le p_{11}p_{12} < \alpha_F^{(2)},\ \alpha_1^{(1)} \le p_{11} < \alpha_F^{(1)} \mid \tau_1\right)$$

$$\cdot \Pr(p_{21}p_{22} \ge \alpha_2^{(2)},\ p_{21} \ge \alpha_2^{(1)} \mid \tau_2)$$

$$= \int_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} \left\{ \int_{\alpha_1^{(2)}/p_{11}}^{\alpha_F^{(2)}/p_{11}} \left[\int_0^{c_{\alpha^*}(\alpha_F^{(2)} - \alpha_1^{(2)})/(p_{11}p_{12} - \alpha_1^{(2)})} g(p_{13})dp_{13} \right] g(p_{12})dp_{12} \right\} g(p_{11})dp_{11}$$

$$\cdot \int_{\alpha_2^{(1)}}^{1} \left[\int_{\alpha_2^{(2)}/p_{21}}^{1} g(p_{22})dp_{22} \right] g(p_{21})dp_{21} \qquad (8.8a)$$

To satisfy $\frac{p_{11}p_{12} - \alpha_1^{(2)}}{\alpha_F^{(2)} - \alpha_1^{(2)}} p_{13} < c_{\alpha^*}$, we need to have $p_{13} < \frac{c_{\alpha^*}(\alpha_F^{(2)} - \alpha_1^{(2)})}{p_{11}p_{12} - \alpha_1^{(2)}}$.

However, if $\alpha_1^{(2)} \le p_{11}p_{12} < c_{\alpha^*}(\alpha_F^{(2)} - \alpha_1^{(2)}) + \alpha_1^{(2)}$, $\frac{p_{11}p_{12} - \alpha_1^{(2)}}{\alpha_F^{(2)} - \alpha_1^{(2)}} p_{13} < c_{\alpha^*}$ satisfies no matter how much p_{13} would be. Therefore, (8.8a) can be partitioned with respect to $c_{\alpha^*}(\alpha_F^{(2)} - \alpha_1^{(2)}) + \alpha_1^{(2)} \le p_{11}p_{12} < \alpha_F^{(2)}$ and $p_{11}p_{12} < c_{\alpha^*}(\alpha_F^{(2)} - \alpha_1^{(2)}) + \alpha_1^{(2)}$.

$$PW_{A_4B_4} = \left\{ \int_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} \left\{ \int_{[c_{\alpha^*}(\alpha_F^{(2)} - \alpha_1^{(2)}) + \alpha_1^{(2)}]/p_{11}}^{\alpha_F^{(2)}/p_{11}} \left[\int_0^{c_{\alpha^*}(\alpha_F^{(2)} - \alpha_1^{(2)})/(p_{11}p_{12} - \alpha_1^{(2)})} g(p_{13})dp_{13} \right] g(p_{12})dp_{12} \right\} g(p_{11})dp_{11} \right.$$

$$\left. + \int_{\alpha_1^{(1)}}^{\alpha_F^{(1)}} \left\{ \int_{\alpha_1^{(2)}/p_{11}}^{[c_{\alpha^*}(\alpha_F^{(2)} - \alpha_1^{(2)}) + \alpha_1^{(2)}]/p_{11}} g(p_{12})dp_{12} \right\} g(p_{11})dp_{11} \right\}$$

$$\cdot \int_{\alpha_2^{(1)}}^{1} \left[\int_{\alpha_2^{(2)}/p_{21}}^{1} g(p_{22})dp_{22} \right] g(p_{21})dp_{21} \qquad (8.8b)$$

Overall statistical power is

$$\text{Power} = \sum_{m=1}^{3} PW_{A_m} + \sum_{n=1}^{4} PW_{A_4B_n}$$

$$= PW_{A_1} + PW_{A_2} + PW_{A_4B_1} + PW_{A_4B_2} + PW_{A_4B_4} \qquad (8.9)$$

8.3 Threshold probabilities and sample size determination

8.3.1 $\alpha_1^{(1)}$ and $\alpha_F^{(2)}$

Under the decision path A_4B_3, the conditions $p_{11}\, p_{12} \geq \alpha_F^{(2)}$ and $\alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)}$ have to be met. Hence $p_{12} \geq \alpha_F^{(2)} / p_{11} \geq \alpha_F^{(2)} / \alpha_1^{(1)}$. In order to have the p-value $p_{12} \leq 1$, the following condition has to be met.

$$\alpha_F^{(2)} \leq \alpha_1^{(1)} \tag{8.10}$$

8.3.2 $\alpha_1^{(1)}$, $\alpha_F^{(1)}$ and sample size n_1

We assume the two study endpoints follow normal distributions. As described previously, based on the first interim analysis, the study could be designed as a two-stage trial or a three-stage trial. In our proposed design, we determine the sample size n_1 for the initial stage to detect the treatment effect at significance level α_1^* with statistical power $1-\gamma_1$.

$$n_1 = \frac{2(Z_{1-\alpha_1^*} + Z_{\gamma_1})^2}{\tau^2} \tag{8.11}$$

We recommend $\alpha_1^{(1)} = 0.05 \sim 0.1$, and $\alpha_F^{(1)} = 0.2 \sim 0.35$. We noticed that, in general, if $\alpha_1^* = \alpha_1^{(1)}$ is used, the sample size n_1 per (8.11) is relatively larger, whereas if $\alpha_1^* = \alpha_F^{(1)}$ is used, the smaller sample size n_1 is obtained. Although one may use a α_1^* level he/she prefers, $\alpha_1^* = (\alpha_1^{(1)} + \alpha_F^{(1)})/2$ gives us reasonable sample size n_1. On the other hand, one may use a feasible sample size n_1 for the first stage based on their financial resource and drug development timeframe.

In practice, γ_1 and τ are given for a trial design. The formula (8.11) provides sample size per treatment group for the initial stage. One may use different effect size τ for each dose arm and get a different sample size for each treatment group. Also one may use an average effect size, and plug it into (8.11) to get a sample size per treatment group. In this thesis, the latter approach is used for simplicity, but without loss of generality.

8.3.3 Sample size $n_{2(3)}$ under three-stage setting

Under the three-stage setting, the sample size for the intermediate stage $n_{2(3)}$ can be determined based on a updated effect size τ^* (e.g average effect size observed from the first stage) at significance level α_2^* with power $1-\gamma_{2(3)}$.

$$n_{2(3)} = \frac{2(Z_{1-\alpha_2^*} + Z_{\gamma_{2(3)}})^2}{(\tau^*)^2} \tag{8.12}$$

In practice, the same stagewise significance level can be used for the intermediate stage and the initial stage, therefore, α_2^* can be set as follows although this is not required and one may use other reasonable significance levels.

$$\alpha_2^* = \exp(-0.5\,\chi^2_{\alpha_1*,4})/p_{11}$$

In practice, $\gamma_{2(3)}$ is given (e.g $\gamma_{2(3)} = 80\%$).

8.3.4 Sample size determination for the final confirmatory stage under two-stage setting

Following the concept of Shih et al (2004), for our varying stage adaptive phase II/III clinical trial design, we propose the use of conditional power to determine sample size for the final confirmatory stage. In this section, we introduce sample size determination for the final confirmatory stage under two-stage setting - $n_{2(2)}$. Sample size determination under three-stage setting will be discussed in the next section.

As described in Section 6.1, if $p_{11}<\alpha_1^{(1)}$, or $p_{11}>\alpha_1^{(1)}$ and $p_{12} < \alpha_2^{(1)}$, the trial is designed under two-stage setting with the 2nd stage as the confirmatory stage. Given the p-value p_{i1} from the first interim analysis and the updated effect size τ^* (based on internal and/or external information), the conditional power is calculated as.

$$\Pr(p_{i1}p_{i2} < ac_{\alpha^*} \mid p_{i1}, \tau^*) = \Pr(p_{i2} < ac_{\alpha^*}/p_{i1} \mid p_{i1}, \tau^*)$$

$$= \int_0^{ac_{\alpha^*}/p_{i1}} g(p_{i2} \mid p_{i1}, \tau^*) dp_{i2}$$

$$> 1\text{-}\gamma_{2(2)} \qquad (8.13)$$

where i =1, 2 for endpoint; 1- $\gamma_{2(2)}$ is the pre-specified conditional power; a = $\alpha_1^{(1)}$ if the primary study endpoint is not changed; a = $\alpha_2^{(1)}$ if the primary study endpoint is changed to the endpoint 2.

Since the density function g for the p-value is a function of sample size $n_{2(2)}$, by solving the inequality (8.13), the sample size $n_{2(2)}$ can be determined.

8.3.5 Sample size determination for the final confirmatory stage under three-stage setting

Under three stage setting, when $p_{11}p_{12}<\alpha_1^{(2)}$, or $p_{11}p_{12}>\alpha_1^{(2)}$ and $p_{21}p_{22}<\alpha_2^{(2)}$, the conditional power is calculated as follows given the combined p-values $p_{i1}p_{i2}$ from the 2nd interim analysis and updated effect size τ*.

$$\Pr(p_{i1}p_{i2}p_{i3} < ac_{\alpha^*} \mid p_{i1}p_{i2}, \tau^*) = \Pr(p_{i3} < ac_{\alpha^*}/p_{i1}p_{i2}, \tau^*)$$

$$= \int_0^{ac_{\alpha^*}/p_{i1}p_{i2}} g(p_{i3} \mid p_{i1}p_{i2}, \tau^*) dp_{i3}$$

$$> 1\text{-}\gamma_3 \qquad (8.14)$$

where i = 1, 2 for endpoint, and γ_3 is pre-specified conditional power.

If $\alpha_1^{(2)} \le p_{11}p_{12} < \alpha_F^{(2)}$ and $p_{21}p_{22} \ge \alpha_2^{(2)}$, the significance of trial results is evaluated by whether the combined p-value $\frac{p_{11}p_{12}-\alpha_1^{(2)}}{\alpha_F^{(2)}-\alpha_1^{(2)}} p_{13}$ exceeds the critical value c_{α^*}. Therefore, the conditional power is

$$\Pr\left(\frac{p_{11}p_{12}-\alpha_1^{(2)}}{\alpha_F^{(2)}-\alpha_1^{(2)}} p_{13} < c_{\alpha^*} \mid p_{11}p_{12}, \tau^*\right) = \Pr(p_{13} < c_{\alpha^*}(\alpha_F^{(2)}-\alpha_1^{(2)})/(p_{11}p_{12}-\alpha_1^{(2)}), \tau^*)$$

$$= \int_0^{c_{\alpha^*}(\alpha_F^{(2)}-\alpha_1^{(2)})/(p_{11}p_{12}-\alpha_1^{(2)})} g(p_{13} \mid p_{11}p_{12}, \tau^*) dp_{13}$$

$$> 1-\gamma_3 \qquad (8.15)$$

The conditional power as presented in (8.14) or (8.15) is a function of sample size n_3 given the combine p-value and updated effect size, therefore, the sample size n_3 for the final confirmatory stage under three-stage setting can be obtained by solving the inequality (8.14) or (8.15).

9 A varying-stage adaptive phase II/III clinical trial design: simulations and a special case of the proposed design

9.1 Illustration

To demonstrate our proposed design, in this section, we illustrate the design with a hypothetical example. Suppose an adaptive phase II/III clinical trial is planned, for which two potential endpoints and four doses of the study drug and a control are considered. Hence, D = {1,2,3,4} and K = 4. Assume the two endpoints follow normal distributions, and anticipated effect size τ = 1/3 by average over the four dose arms. Effect size here is defined as anticipated treatment difference between a dose arm and the control arm divided by the common standard deviation.

Suppose the trial is designed with overall Type I error rate $\alpha = 0.05$ (two-sided) and power = $1 - \beta = 80\%$. We choose $\alpha_1^{(1)} = \alpha_2^{(1)} = 0.15$, and $\alpha_F^{(1)} = 0.35$. To have the thresholds probabilities at the 2nd stage same to the 1st stage, we choose $\alpha_1^{(2)} = \alpha_2^{(2)} = \exp(-0.5\,\chi^2_{0.15,4}) = 0.034$, and $\alpha_F^{(2)} = \exp(-0.5\,\chi^2_{0.35,4}) = 0.109$. Plug in these parameters into (6.28), we get $\alpha^* = 0.1378$ and $c_{\alpha^*} = \exp(-0.5\,\chi^2_{0.1378,4}) = 0.0307$.

Per (8.11), the sample size for the initial stage is 42 per treatment group, which is determined to detect effect size τ = 1/3 at significance level of $\alpha_1^* = (\alpha_1^{(1)} + \alpha_F^{(1)})/2 = 0.25$ with statistical power $1 - \gamma_1 = 80\%$. Suppose analyses based on ANOVA models are performed with respect to the study endpoints. From the first interim analysis, $p_{11} = 0.1 < \alpha_1^{(1)} = 0.15$ is obtained for the global null hypothesis $H_{01:} \bigcap_{k \in D} H_{01k} = H_{011} \cap H_{012} \cap H_{013} \cap H_{014}$. Hence the trial is designed as a two-stage study. In comparing efficacy and safety profiles, suppose the dose arm 2 and 4 are chosen for the final stage (S = {2. 4}).

Suppose the updated effect size τ^* is same as initial effect size τ. From (8.13), we get the sample size for the final confirmatory stage (2nd stage) is 115 per treatment arm, which satisfies conditional power = 80%.

With respect to the primary study endpoint – endpoint 1, suppose the p-values from both stages are obtained as summarized in the following table under the corresponding null hypotheses.

Stage	Null hypothesis	P-value
I	$H_{01,\{1,2,3,4\}}$	0.100
	$H_{01,\{1,2,3\}}$	0.141
	$H_{01,\{1,2,4\}}$	0.091
	$H_{01i,\{2,3,4\}}$	0.096
	$H_{01,\{1,2\}}$	0.130
	$H_{01,\{2,3\}}$	0.135
	$H_{01,\{2,4\}}$	0.070
	H_{012}	0.085
II	$H_{01,\{2,4\}}$	0.031
	H_{012}	0.025

The critical value for combined p-value = $\alpha_1^{(1)} c_{\alpha*}$ = 0.15*0.0307 = 0.0046. The Combined p-values are as follows:

CP-value$^{(\{1,2,3,4\},\ \{2,4\})}$ = 0.100*0.031= 0.0031< $\alpha_1^{(1)} c_{\alpha*}$

CP-value$^{(\{1,2,4\},\ \{2,4\})}$ = 0.091*0.031= 0.0028< $\alpha_1^{(1)} c_{\alpha*}$

CP-value$^{(\{2,3,4\},\ \{2,4\})}$ = 0.096*0.031 = 0.0030 < $\alpha_1^{(1)} c_{\alpha*}$

CP-value$^{(\{2,4\},\ \{2,4\})}$ = 0.070*0.031 = 0.0022 < $\alpha_1^{(1)} c_{\alpha*}$

CP-value$^{(\{1,2,3\},\ \{2\})}$ = 0.141*0.025 = 0.0035> $\alpha_1^{(1)} c_{\alpha*}$

CP-value$^{(\{1,2\},\{2\})}$ = 0.130*0.025 = 0.0033< $\alpha_1^{(1)}$ c$_{\alpha*}$

CP-value$^{(\{2,3\},\{2\})}$ = 0.135*0.025 = 0.0034< $\alpha_1^{(1)}$ c$_{\alpha*}$

CP-value$^{(\{2\},\{2\})}$ = 0.085*0.025 = 0.0021< $\alpha_1^{(1)}$ c$_{\alpha*}$

Apparently, the CP-values for all the intersection null hypothesis containing H_{012} as well as for the elementary null hypothesis H_{012} are less than the critical value $\alpha_1^{(1)}$ c$_{\alpha*}$. Hence, we can reject H_{012} at significance level of two-sided α = 0.05 in the strong sense, and claim dose arm 2 is superior compared to the control with respect to the primary endpoint – endpoint 1, which is chosen based on interim I analysis.

The similar analysis can be carried out for the dose arm 4.

9.2 Simulations

The objective of this simulation is to demonstrate the properties of the proposed design. Suppose for an adaptive phase II/III clinical trial, two potential endpoints and three doses of the study drug and a control are considered. Hence, D = {1, 2, 3} and K = 3. Assume both endpoints follow normal distributions. The following are assumed distributions for each treatment arm and study endpoint. The average effect size $\tau_1 = 0.35$ for both endpoint 1 and endpoint 2.

Study Endpoint	Distribution			
	Control	Dose 1	Dose 2	Dose 3
1	$N(40, 20^2)$	$N(45, 20^2)$	$N(47, 20^2)$	$N(49, 20^2)$
2	$N(10, 15^2)$	$N(14, 15^2)$	$N(15, 15^2)$	$N(17, 15^2)$

We choose $\alpha_1^{(1)} = \alpha_2^{(1)} = 0.1$, and $\alpha_F^{(1)} = 0.3$. To have the threshold probabilities at the 2nd stage same to the 1st stage although this setup is not necessary, we choose $\alpha_1^{(2)} = \alpha_2^{(2)} = \exp(-0.5\,\chi^2_{0.1,4}) = 0.0205$, and $\alpha_F^{(2)} = \exp(-0.5\,\chi^2_{0.3,4}) = 0.0872$. We control the family wise Type I error rate α = 0.05 (two-sided). By plugging in these parameters into (5.28), we obtain $\alpha^* = 0.1778$ and c$_{\alpha*}$ = $\exp(-0.5\,\chi^2_{0.1778,4}) = 0.0428$.

Based on (8.11), the sample size for the initial stage is 47 per treatment group, which is determined to detect effect size τ = 0.35 at significance level of $\alpha_1^* = (\alpha_1^{(1)} + \alpha_F^{(1)})/2 = 0.2$ with statistical power $1 - \gamma_1 = 80\%$. To determine sample size for next stage based on interim results, we use 80% conditional power.

Same as illustration as presented in the previous section, to control Type I error rate in the strong sense, the closed testing procedure (Marcus, 1976) is used in the simulations.

9.2.1 Dunnett test

For simplicity, we choose the dose arm with the smallest p-value from stage I or the smallest combined p-value from the first two stages as the dose arm together with the control for the final stage under two-stage or three-stage setting respectively. To compute the p-values, Dunnett test (Dunnett, 1955) under one-way ANOVA analysis is used for many-to-one comparisons (each treatment vs control). Dunnett's test holds the family wise error rate to a level not exceeding the pre-specified allowable Type I error rate α. Let $\overline{X_k}$ (k =1, 2, …, K) and $\overline{X_0}$ be the mean values of the primary endpoint, and n_k and n_0 be the number of subjects for the treatment k and the control group. Define t_k representing two-sample t-statistic as follows.

$$t_k = \frac{\overline{X_k} - \overline{X_0}}{s\sqrt{\frac{1}{n_k} + \frac{1}{n_0}}} \tag{9.1}$$

where s is the square root of the ANOVA mean square error (MSE). The treatment k is significantly different from the control (two-sided test) if the following condition is satisfied.

$$| t_k | \geq d(\alpha, K, v, \rho_1, \rho_2, ..., \rho_K) \tag{9.2}$$

where $d(\alpha, K, v, \rho_1, \rho_2, ..., \rho_K)$ is the critical value of the "many-to-one t statistic" (Miller 1981; Krishnaiah and Armitage 1966) for the treatment group k to be compared to a control, with $v = \sum_{i=1}^{K+1} n_i - (K+1)$ degrees of freedom and correlations $\rho_1, \rho_2, ..., \rho_K$, where $\rho_i = \sqrt{n_i / (n_0 + n_i)}$. The correlation terms arise because each of the treatment means is being compared to the same control.

The critical value $d(\alpha, K, v, \rho_1, \rho_2, ..., \rho_K)$ is the solution of the following equation with respect to the variable q (Hsu 1996).

$$\int_0^\infty \left\{ \int_{-\infty}^\infty \phi(y) \prod_{i=1}^K \left[\Phi(\frac{\rho_i y + |q| s}{\sqrt{1-\rho_i^2}}) - \Phi(\frac{\rho_i y - |q| s}{\sqrt{1-\rho_i^2}}) \right] dy \right\} \gamma(s) ds = 1 - \alpha \quad (9.3)$$

where ϕ and Φ are probability density function and cumulative distribution function of standard normal distribution. $\gamma(s)$ is defined as

$$\gamma(s) = \frac{v^{v/2}}{\Gamma(v/2) 2^{v/2-1}} s^{v-1} e^{-\frac{vs^2}{2}}$$

The following formula (9.4) defines the Dunnett adjusted p-value (two-sided) under the global null hypothesis H_0. The adjusted p-values for intersection null hypotheses can be calculated in a similar way. For the final stage, since only one dose arm is chosen together with the control arm in the simulations to be presented in the next section, the Dunnett test is reduced to a t-test.

$$P(\max(|t_1|, |t_2|, ..., |t_K|) > d(\alpha, K, v, \rho_1, \rho_2, ..., \rho_K)) \quad (9.4)$$

9.2.2 Simulation on overall Type I error rate

As defined in (6.1), the elementary null hypothesis for k^th^ treatment vs the control is H_{0ik}: $\theta_{ik} = \theta_{i0}$ (k=1, 2, 3 for treatment group, i = 1, 2 for endpoint). As defined in (6.2), the global null hypotheses with respect to the endpoint 1 and endpoint 2 are $H_{01}: H_{011} \bigcap H_{012} \bigcap H_{013}$ and $H_{02}: H_{021} \bigcap H_{022} \bigcap H_{023}$. The global null hypothesis for both study endpoints is $H_0: H_{01} \bigcap H_{02}$.

Table 9-1a Type I error rates (per simulation vs per theory)

Decision path	Per simulation	Per theory
A_1	0.0169	0.0178
A_2	0.0166	0.0160
A_4B_1	0.0036	0.0036
A_4B_2	0.0017	0.0015
A_4B_4	0.0107	0.0111
Total	**0.0495**	**0.05**

For each decision path A_mB_n (m, n = 1, 2, 3, 4), the Type I error rate under the global null hypothesis H_0 can be calculated theoretically per (6.26) as derived in Section 6.5.2. Table 9-1a presents Type I error rates per simulation vs per theoretical calculation for each decision path. The simulated overall Type I error rate is 0.0495 based on 100,000 replicates of adaptive phase II/III clinical trials. This simulated Type I error rate is very close to the theoretically calculated rate of 0.05. For each decision path, the difference in Type I error rate between the simulation result and theoretical result is less than 0.001.

Further simulations are performed with the consideration of the correlation between the two endpoints within a stage. Table 9-1b summarizes the simulated Type I error rates based on 20,000 replicates. The simulated overall (total) Type I error rates are 0.05, 0.0483, 0.0472 and 0.0398 when the correlation coefficient between the two endpoints is assumed 0, 0.25, 0.5 and 0.75 respectively. Apparently, as the two endpoints are more correlated, the overall Type I error rate becomes smaller. Therefore, it is conservative to assume the independence of the two endpoints when to calculate Type I error rate.

Table 9-1b Simulated Type I error rates

Decision path	Correlation coefficient			
	$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$
A_1	0.0173	0.0192	0.0192	0.0175
A_2	0.0156	0.0132	0.0130	0.0080
A_4B_1	0.0035	0.0032	0.0034	0.0031
A_4B_2	0.0020	0.0014	0.0013	0.0006
A_4B_4	0.0116	0.0114	0.0102	0.0106
Total	**0.0500**	**0.0483**	**0.0472**	**0.0398**

9.2.3 Simulation on statistical power

Simulation on statistical power under the alternative hypothesis for the endpoint 1

Under the alternative hypothesis H_{A1}: θ_{11}, θ_{12} or θ_{13} differ from θ_{10} for the endpoint 1, we simulate 50,000 adaptive Phase II/III trials. Table 9-2 presents the simulated statistical powers vs theoretical results per (8.2a) through (8.9) as derived in Section 8.2. The simulated overall statistical power is 70.99%, which is similar to the power of 72.40% per theoretical calculation. As expected, the statistical powers concentrated on the decision paths A_1, A_4B_1 and A_4B_4 since the simulation is performed under the alternative hypothesis with respect to the endpoint 1 and null hypothesis for the endpoint 2. The powers for the decision path A_2 and A_4B_2 are actually Type I error rates under the null hypothesis H_{02} with respect to the endpoint 2. The total Type I error rate from the decision path A_2 and A_4B_2 is $0.0064 + 0.0007 = 0.0071$, which is much less than $\alpha = 0.05$ (two-sided). The mean sample sizes are 38 and 39 for the final stage under the decision path A_1 and A_2; 33 for the intermediate stage; and 45, 31 and 61 for the final stage under the decision path A_4B_1, A_4B_2 and A_4B_4 respectively.

Table 9-2 **Statistical power under the alternative hypothesis for the endpoint 1**

Decision path	Per simulation	Per theory
A_1	0.5719	0.5714
A_2	0.0064	0.0060
A_4B_1	0.0822	0.0924
A_4B_2	0.0007	0.0003
A_4B_4	0.0487	0.0539
Total	0.7099	0.7240

Simulation on statistical power under the alternative hypothesis for the endpoint 2

Similar to the simulation given in the previous section, 50,000 adaptive Phase II/III trials are simulated under alternative hypothesis H_{A2}: θ_{21}, θ_{22} or θ_{23} differ from θ_{20} for the endpoint 2. Table 9-3 summarizes the simulated statistical powers vs theoretical results. The simulated overall statistical power is 58.17%, which is close to the power 56.21% per theoretical calculation. The mean sample sizes are 38 and 37 for the final stage under the decision path A_1 and A_2; 37 for the intermediate stage; and 38, 43 and 43 for the final stage under the decision path A_4B_1, A_4B_2 and A_4B_4 respectively.

Table 9-3 **Statistical power under the alternative hypothesis for endpoint 2**

Decision path	Per simulation	Per theory
A_1	0.0177	0.0178
A_2	0.5338	0.5117
A_4B_1	0.0014	0.0014
A_4B_2	0.0264	0.0292
A_4B_4	0.0025	0.0020
Total	0.5817	0.5621

Simulation on statistical power under the alternative hypotheses for both endpoints

Table 9-4a shows the simulated statistical powers vs theoretical results. The simulations are carried out with 50,000 adaptive phase II/III trials under both alternative hypotheses H_{A1}: θ_{11}, θ_{12} or θ_{13} differ from θ_{10} for the endpoint 1 and H_{A2}: θ_{21}, θ_{22} or θ_{23} differ from θ_{20} for the endpoint 2. The simulated overall statistical power is 83.57%, which is similar to the power 82.67% per theoretical calculation. The mean sample sizes are 38 and 37 for the final stage under decision path A_1 and A_2; 33 for the intermediate stage; and 45, 45 and 61 for the final stage under the decision path A_4B_1, A_4B_2 and A_4B_4 respectively. Since the statistical powers mainly come from the decision paths A_1 and A_2, more weights are put on the two-stage setting in this varying-stage adaptive Phase II/III design.

To consider the correlation of the two endpoints within a stage, we performed further simulations on statistical power. Table 9-4b presents the simulation results based on 15,000 replicates.

Table 9-4a Statistical powers under alternative the hypotheses for both endpoints

Decision path	Per simulation	Per theory
A_1	0.5721	0.5714
A_2	0.2094	0.1927
A_4B_1	0.0306	0.0388
A_4B_2	0.0126	0.0136
A_4B_4	0.0111	0.0102
Total	0.8357	0.8267

The simulated overall (total) statistical powers are 83.18%, 80.58%, 78.18% and 75.78% when the correlation coefficient between the two endpoints is assumed 0, 0.25, 0.5 and 0.75 respectively. From this simulation, as the two endpoints are more correlated, the simulated overall statistical power becomes smaller.

Table 9-4b Simulated statistical powers under alternative the hypotheses for both endpoints

Decision path	Correlation coefficient			
	$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$	$\rho = 0.75$
A_1	0.5693	0.5677	0.5647	0.5696
A_2	0.2091	0.1782	0.1472	0.1076
A_4B_1	0.0310	0.0363	0.0411	0.0487
A_4B_2	0.0114	0.0105	0.0113	0.0103
A_4B_4	0.0110	0.0131	0.0175	0.0217
Total	**0.8318**	**0.8058**	**0.7818**	**0.7578**

9.3 A special case of the varying-stage adaptive phase II/III clinical trial design

For the varying-stage adaptive phase II/III clinical trial design presented in the previous sections, the probability of having the three-stage setting is much lower than that for the two-stage setting. For the example presented in the previous simulations with $\alpha_1^{(1)} = \alpha_2^{(1)} = 0.1$, $\alpha_F^{(1)} = 0.3$, $\alpha_1^{(2)} = \alpha_2^{(2)} = 0.0205$, and $\alpha_F^{(2)} = 0.0872$, the probabilities of the decision paths under the global null hypothesis are

$$\Pr(A_1) = \Pr(p_{11} < \alpha_1^{(1)}) = \alpha_1^{(1)} = 0.1$$

$$\Pr(A_2) = \Pr(p_{11} \geq \alpha_1^{(1)}, p_{21} < \alpha_2^{(1)}) = \alpha_2^{(1)}(1-\alpha_1^{(1)}) = 0.09$$

$$\Pr(A_3) = \Pr(p_{11} > \alpha_F^{(1)}, p_{21} \geq \alpha_2^{(1)}) = (1-\alpha_F^{(1)})(1-\alpha_2^{(1)}) = 0.63$$

$$\Pr(A_4B_1) = \Pr(p_{11}p_{12} < \alpha_1^{(2)},\ \alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)}, p_{21} \geq \alpha_2^{(1)})$$

$$= \alpha_1^{(2)}(1-\alpha_2^{(1)})\ln\left(\frac{\alpha_F^{(1)}}{\alpha_1^{(1)}}\right) = 0.020$$

$$\Pr(A_4B_2) = \Pr(p_{21}p_{22} < \alpha_2^{(2)}, p_{21} \geq \alpha_2^{(1)},\ \alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)}, p_{11}p_{12} \geq \alpha_1^{(2)})$$

$$= -\alpha_2^{(2)}[\alpha_F^{(1)} - \alpha_1^{(1)} - \alpha_1^{(2)} \ln(\frac{\alpha_F^{(1)}}{\alpha_1^{(1)}})]\ln(\alpha_2^{(1)}) = 0.008$$

$$\Pr(A_4B_3) = \Pr(p_{11}p_{12} > \alpha_F^{(2)},\ p_{21}p_{22} \geq \alpha_2^{(2)},\ \alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)},\ p_{21} \geq \alpha_2^{(1)})$$

$$= [(\alpha_F^{(1)} - \alpha_1^{(1)}) - \alpha_F^{(2)} \ln(\frac{\alpha_F^{(1)}}{\alpha_1^{(1)}})]\,[1 - \alpha_2^{(1)} + \alpha_2^{(2)} \ln(\alpha_2^{(1)})] = 0.089$$

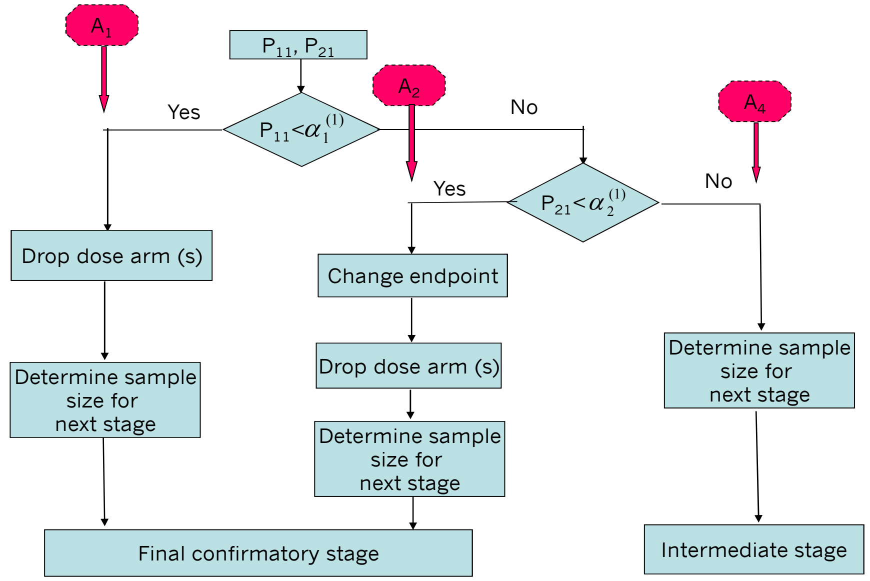
$$\Pr(A_4B_4) = \Pr(\alpha_1^{(2)} \leq p_{11}p_{12} < \alpha_F^{(2)},\ p_{21}p_{22} \geq \alpha_2^{(2)},\ \alpha_1^{(1)} \leq p_{11} < \alpha_F^{(1)},\ p_{21} \geq \alpha_2^{(1)})$$

$$= (\alpha_F^{(2)} - \alpha_1^{(2)})\ln(\frac{\alpha_F^{(1)}}{\alpha_1^{(1)}})\,[1 - \alpha_2^{(1)} + \alpha_2^{(2)} \ln(\alpha_2^{(1)})] = 0.062$$

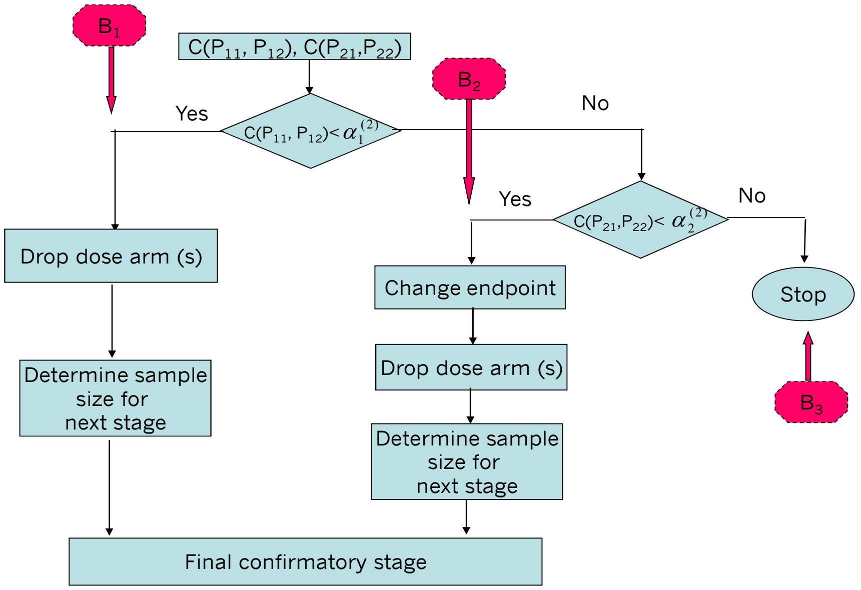
Therefore, the probability of having the three-stage setting = $\Pr(A_4B_1)$ + $\Pr(A_4B_2)$ + $\Pr(A_4B_4)$ = 0.020 + 0.008 + 0.062 = 0.09, which is much lower than the probability for the two-stage setting = $\Pr(A_1) + \Pr(A_2)$ = 0.1 + 0.09 = 0.19. However, both of these two probabilities are relatively low compared to the probability for the futility stopping at interims, which is $\Pr(A_3) + \Pr(A_4B_3)$ = 0.63 + 0.089 = 0.719.

In clinical practice, the sample size for the initial stage is usually small; hence, the information obtained from the initial stage may not be adequate for decision making. In this case, an intermediate stage will be carried out to get more data. To fully utilize the advantage of this intermediate stage, by dropping the decision paths A_3 and B_4 as shown in Figure 9-1, a special case of the varying-stage adaptive phase II/III clinical trial design is discussed in this section. In this special case, the threshold value $\alpha_F^{(1)}$ is set to 1 (no stopping for futility at the interim I) and $\alpha_F^{(2)}$ is set to 0 (stopping for futility if neither endpoint is promising per data cumulated to the interim II).

Figure 9-1 Flow chart of the special case of the posposed design


A_1
P_{11}, P_{21}
A_2
Yes
$P_{11} < \alpha_1^{(1)}$
No
A_4
Yes
$P_{21} < \alpha_2^{(1)}$
No
Drop dose arm (s)
Change endpoint
Determine sample size for next stage
Drop dose arm (s)
Determine sample size for next stage
Determine sample size for next stage
Final confirmatory stage
Intermediate stage

(a) Interim I (initial stage)


B_1
$C(P_{11}, P_{12}), C(P_{21}, P_{22})$
B_2
Yes
$C(P_{11}, P_{12}) < \alpha_1^{(2)}$
No
Yes
$C(P_{21}, P_{22}) < \alpha_2^{(2)}$
No
Drop dose arm (s)
Change endpoint
Stop
Drop dose arm (s)
Determine sample size for next stage
B_3
Determine sample size for next stage
Final confirmatory stage

(b) Interim II (intermediate stage)

9.3.1 Alpha allocation

For this special case of the proposed design, the cumulative distribution function of p_{11}, p_{21}, $p_{11}p_{12}$ and $p_{21}p_{22}$ under the global null hypothesis (6.3) are same as what derived in (6.4), (6.13), (6.8) and (6.17) respectively. That is, conditionally, $p_{11} \sim$ Uniform(0, $\alpha_1^{(1)}$) and $p_{21} \sim$ Uniform(0, $\alpha_2^{(1)}$), $p_{11}p_{12} \sim$ Uniform(0, $\alpha_1^{(2)}$) and $p_{21}p_{22} \sim$ Uniform(0, $\alpha_2^{(2)}$) under the global null hypothesis (6.3).

Let λ be the percent of alpha allocated for the two-stage setting. The Type I error rate for the two-stage setting is

$$ER_{II} = \alpha^* \, [\alpha_1^{(1)} + (1-\alpha_1^{(1)})\alpha_2^{(1)}] = \lambda\alpha \qquad (9.5)$$

The Type I error rate for the three-stage setting is

$$ER_{III} = \alpha^* \{ -\alpha_1^{(2)}(1-\alpha_2^{(1)})\ln(\alpha_1^{(1)}) - \alpha_2^{(2)}\ln(\alpha_2^{(1)})[1-\alpha_1^{(1)} + \alpha_1^{(2)}\ln(\alpha_1^{(1)})] \}$$

$$= (1-\lambda)\alpha \qquad (9.6)$$

Without loss of generality for clinical practice, let $\alpha_1^{(1)} = \alpha_2^{(1)}$ and $\alpha_1^{(2)} = \alpha_2^{(2)}$. Given λ and $\alpha_1^{(1)}$, the significance level α^* can be obtained from (9.5). Apply α^* to (9.6), $\alpha_1^{(2)}$ (or $\alpha_2^{(2)}$) can be solved.

As shown in Appendix A.7, under null hypothesis of no treatment difference, the logarithm of the combined p-value $p_{i1}p_{i2}$ (i = 1, 2 for endpoint) multiplied by -2 follow the Chi-square distribution with four degrees of freedom, namely $-2\ln(p_{i1}p_{i2}) \sim \chi_4^2$. Based on the threshold $\alpha_1^{(2)}$ or $\alpha_2^{(2)}$ for the combined p-values $p_{11}p_{12}$ and $p_{21}p_{22}$, we define the probability α_2 be the threshold value as follows.

$$\alpha_1^{(2)} = \alpha_2^{(2)} = \exp(-0.5\,\chi^2_{\alpha_2,4}) \qquad (9.7)$$

$$\alpha_2 = 1 - F_{\chi^2}(-2\ln(\alpha_1^{(2)}),4) \qquad (9.8)$$

where $F_{\chi^2}(x,k)$ is the cumulative distribution function of the Chi-square distribution with k degrees of freedom, and x ~ χ_k^2. As an example, when $\alpha = 0.05$ (two-sided), $\lambda = 60\%$ and $\alpha_1^{(1)} = \alpha_2^{(1)} = 0.07$, $\alpha_2 = 0.0932$ is obtained from (9.5) through (9.8), which enables 60% and 40% of alpha to be allocated to the two-stage and three-stage setting respectively.

Table 9-5, 9-6 and 9-7 present the threshold α_2 and the critical values $\alpha_1^{(1)} c_{\alpha*}$ and $\alpha_1^{(2)} c_{\alpha*}$ for various scenarios of λ and $\alpha_1^{(1)}$ when $\alpha = 0.05$ (two-sided), $\alpha_1^{(1)} = \alpha_2^{(1)}$ and $\alpha_1^{(2)} = \alpha_2^{(2)}$. Given the threshold value $\alpha_1^{(1)}$ or $\alpha_2^{(1)}$ ($\alpha_1^{(1)} = \alpha_2^{(1)}$), the lower value of α_2, the higher critical value of $\alpha_1^{(1)} c_{\alpha*}$ and lower critical value of $\alpha_1^{(2)} c_{\alpha*}$ are required as the percent (λ) of alpha allocated in the two-stage setting increases. This means that, given $\alpha_1^{(1)}$ or $\alpha_2^{(1)}$, to allocate more alpha for the two-stage setting, the less stringent requirement on the critical value $\alpha_1^{(1)} c_{\alpha*}$ is needed to reject the null hypothesis H_{0i} (i = 1, 2 for endpoint) for the two-stage setting. However, as an expense, the more stringent requirement on $\alpha_1^{(2)} c_{\alpha*}$ has to be put in place regarding rejecting H_{0i} for the three-stage setting. On the other hand, given λ, the higher value of α_2 is allowed as $\alpha_1^{(1)}$ or $\alpha_2^{(1)}$ increases. This implies that, in order to have the same alpha allocation, the requirement on the threshold α_2 (to have three-stage setting) relaxes as the threshold value $\alpha_1^{(1)}$ or $\alpha_2^{(1)}$ (to have two-stage setting) increases. However, the change in $\alpha_1^{(1)} c_{\alpha*}$ is very minimal, but the requirement on $\alpha_1^{(2)} c_{\alpha*}$ is less stringent as $\alpha_1^{(1)}$ increases.

Table 9-5 **Threshold value α_2 when $\alpha = 0.05$ (two-sided), $\alpha_1^{(1)} = \alpha_2^{(1)}$ and $\alpha_1^{(2)} = \alpha_2^{(2)}$**

$\alpha_1^{(1)}$	λ					
	50%	55%	60%	65%	70%	75%
0.05	0.0888	0.0752	0.0634	0.0531	0.0438	0.0355
0.06	0.1089	0.0780	0.0924	0.0654	0.0541	0.0438
0.07	0.1298	0.1102	0.0932	0.0781	0.0647	0.0525
0.08	0.1513	0.1286	0.1088	0.0913	0.0757	0.0615
0.09	0.1736	0.1476	0.1250	0.1050	0.0871	0.0708
0.10	0.1965	0.1672	0.1417	0.1191	0.0989	0.0805

Table 9-6 **Critical value $\alpha_1^{(1)} c_{\alpha^*}$ when $\alpha = 0.05$ (two-sided), $\alpha_1^{(1)} = \alpha_2^{(1)}$ and $\alpha_1^{(2)} = \alpha_2^{(2)}$**

$\alpha_1^{(1)}$	λ					
	50%	55%	60%	65%	70%	75%
0.05	0.0128	0.0141	0.0154	0.0167	0.0179	0.0192
0.06	0.0129	0.0142	0.0155	0.0168	0.0180	0.0193
0.07	0.0130	0.0142	0.0155	0.0168	0.0181	0.0194
0.08	0.0130	0.0143	0.0156	0.0169	0.0182	0.0195
0.09	0.0131	0.0144	0.0157	0.0170	0.0183	0.0196
0.10	0.0132	0.0145	0.0158	0.0171	0.0184	0.0197

Table 9-7 **Critical value $\alpha_1^{(2)} c_{\alpha^*}$ when $\alpha = 0.05$ (two-sided), $\alpha_1^{(1)} = \alpha_2^{(1)}$ and $\alpha_1^{(2)} = \alpha_2^{(2)}$**

$\alpha_1^{(1)}$	λ					
	50%	55%	60%	65%	70%	75%
0.05	.0045	.0040	.0036	.0031	.0027	.0022
0.06	.0049	.0044	.0039	.0034	.0029	.0024
0.07	.0053	.0047	.0042	.0036	.0031	.0026
0.08	.0056	.0050	.0044	.0039	.0033	.0027
0.09	.0060	.0054	.0047	.0041	.0035	.0029
0.10	.0064	.0057	.0050	.0044	.0037	.0031

9.3.2 Relation of statistical power and sample size to λ and $\alpha_1^{(1)}$

In this section, we use simulations to demonstrate the relationship of statistical power and sample size to the parameters λ and $\alpha_1^{(1)}$. For each scenario of λ and $\alpha_1^{(1)}$ as displayed in Table 9-5, we simulate 15,000 trials with the same two endpoints assumed in Section 9.2. The 15,000 simulations imply that the simulated overall power has standard error of 0.0039 or less. For all the simulations, the conditional powers are set to 80% and allowable Type I error rate α = 0.05 (two-sided). Following Section 8.3.2, the sample size for the initial stage is determined as follow.

$$n_1 = \frac{2(Z_{1-\alpha_1^*} + Z_{\gamma_1})^2}{\tau^2}$$

where $\alpha_1^* = (\alpha_1^{(1)} + \alpha_F^{(1)})/2 = (\alpha_1^{(1)} + 1.0)/2$, and τ is the initial treatment effect size. Again in the simulation $\gamma_1 = 80\%$ is used. To avoid the sample size $n_{2(3)}$ for the intermediate stage too large, the smaller of the following sizes is chosen as the sample size for the intermediate stage in the simulation.

- Calculated based on (8.12) as described in Section 8.3.3

  $$n_{2(3)} = \frac{2(Z_{1-\alpha_2^*} + Z_{\gamma_{2(3)}})^2}{(\tau^*)^2}$$

  Where $\alpha_2^* = \exp(-0.5\,\chi^2_{\alpha_2,4})/p_{11}$, τ* is the updated treatment effect size, and the power $\gamma_{2(3)}$ is set to 80%.

- $0.65{*}n - n_1$, where n is the sample size for a single-stage design, and n_1 is the sample size determined previously for the initial stage.

In this section, we simulate for the following parameters.

- Statistical power for each decision path and total power
  - Probability for each decision path ($\Pr ob_{D_i}$, D_i for i^th decision path)

- Sample size for each decision path (n_{D_i}) and average sample size for the selected dose arm and the control
- Sample size for each stage and average total sample size

The average sample size (EN) for the selected dose arm and the control is calculated as follows.

$$\mathrm{EN} = \sum_{i=1}^{5} n_{D_i} \Pr ob_{D_i}$$

$$= \sum_{j=1}^{2} n_{A_j} \Pr ob_{A_j} + \sum_{k=1}^{3} n_{A_4B_k} \Pr ob_{A_4B_k}$$

Let n_1 denote the sample size for the first stage, n_{21} and n_{22} for the 2nd stage under two-stage setting for the endpoint 1 and 2 as the primary endpoint respectively, $n_{2(3)}$ for the 2nd stage under the three-stage setting, n_{31} and n_{32} for the 3rd stage for the endpoint 1 and 2 as the primary endpoint Respectively. The sample sizes for each decision path are as follows.

$$n_{A_1} = n_1 + n_{21}$$

$$n_{A_2} = n_1 + n_{22}$$

$$n_{A_4B_1} = n_1 + n_{2(3)} + n_{31}$$

$$n_{A_4B_2} = n_1 + n_{2(3)} + n_{32}$$

$$n_{A_4B_3} = n_1 + n_{2(3)}$$

For the example presented in this simulation with three doses of the study treatment and one control group in the initial stage, and one selected dose arm and the control are continued to the final confirmatory stage, the average total sample size is calculated as

$$\mathrm{EN_T} = 4{*}n_1 + 2{*}n_{21}{*} \Pr ob_{A_1} + 2{*}n_{22}{*} \Pr ob_{A_2}$$

$$+ 4*n_{2(3)}*(\Pr ob_{A_3} + \Pr ob_{A_4B_1} + \Pr ob_{A_4B_2})$$

$$+ 2*n_{31}* \Pr ob_{A_4B_1} + 2*n_{32}* \Pr ob_{A_4B_2}$$

The simulations are performed under the alternative hypothesis (hypotheses): (1) for the endpoint 1; (2) for the endpoint 2; (3) for both endpoints. Table 9-8a, 9-8b and 9-8c present the simulation results under alternative hypothesis for the endpoint 1. From these simulation results, we can see the following trends.

- Given λ, as $\alpha_1^{(1)}$ increases, the power for the decision path A_1, total power, sample sizes (n_{21}, n_{22}, n_{31}, n_{31}, EN and EN_T) increase; the probability of the decision path A_1 increases and the probability of the decision path A_4B_3 decreases; however, there is no much change in the power for the decision path A_4B_1 and the probability of the decision path A_4B_1.
- Given $\alpha_1^{(1)}$, as λ increases, the total power, sample sizes (n_{21}, n_{22}, n_{31}, n_{31}, EN and EN_T) decrease; the probability of the decision path A_4B_3 increase; however, there is no much change in the power and probability of the decision path A_1.
- The total power of the decision paths A_2 and A_4B_2 is less than 0.015. As matter of fact, this is Type I error rate under the null hypothesis H_{02}.

The simulation results under alternative hypothesis for the endpoint 2 are provided in Table 9-9a, 9-9b and 9-9c. The trends of the simulation results are very similar to those under alternative hypothesis for the endpoint 1, but for the parameters related to the endpoint 2.

Table 9-10a, 9-10b and 9-10c summarize the simulation results under the alternative hypothesis for both endpoints. From these simulation results, we can see the following trends.

- Given λ, as $\alpha_1^{(1)}$ increases, the power for the decision path A_1, total power, sample sizes (n_{21}, n_{22}, n_{31}, n_{31}, EN and EN_T) and the probability of the decision path A_1 increase; the probability of the decision path A_4B_1 decrease; however, there is no much change in the power and probability for other decision paths.
- Given $\alpha_1^{(1)}$, as λ increases, the sample sizes (n_{21}, n_{22}, n_{31}, n_{31}, EN and EN_T) decrease; however, there is no much change in power and probability for any decision paths.

In summary, the total power and sample size increase as $\alpha_1^{(1)}$ increases given λ; and sample size decreases as λ increases given $\alpha_1^{(1)}$. Hence, the parameters $\alpha_1^{(1)}$ and λ play important roles in clinical trial designs. These parameters can be determined based on the feasibility of sample size, statistical power, and anticipated alpha allocation. For example, under alternative hypotheses for both endpoints, when λ = 50% and $\alpha_1^{(1)}$=0.10, the average total sample size is 303.5 (n_{21}= 66, n_{22}= 65, n_{31}= 52, n_{31}= 48 and EN=106.8) and power = 0.9020; whereas, when λ = 70% and $\alpha_1^{(1)}$=0.05, the average total sample size is 256.9 (n_{21}= 29, n_{22}= 29, n_{31}= 25, n_{31}= 24 and EN=76.9) and power = 0.8404.

Table 9-8a Simulated power under the alternative hypothesis for the endpoint 1

λ	$\alpha_1^{(1)}$	Decision path				Total power
		A_1	A_2	A_4B_1	A_4B_2	
50%	0.05	0.3802	0.0065	0.3198	0.0035	0.7100
	0.06	0.4129	0.0069	0.3075	0.0023	0.7296
	0.07	0.4291	0.0057	0.3065	0.0013	0.7426
	0.08	0.4441	0.0061	0.3051	0.0024	0.7577
	0.09	0.4714	0.006	0.2972	0.0024	0.7770
	0.10	0.4941	0.0053	0.2890	0.0013	0.7897
60%	0.05	0.3815	0.0083	0.2792	0.0029	0.6719
	0.06	0.4080	0.0078	0.2929	0.0027	0.7114
	0.07	0.4202	0.0076	0.2798	0.0024	0.7100
	0.08	0.4539	0.0084	0.2711	0.0024	0.7358
	0.09	0.4688	0.0076	0.2635	0.0013	0.7412
	0.10	0.4711	0.0066	0.2722	0.0024	0.7523
70%	0.05	0.3873	0.0101	0.2417	0.0022	0.6413
	0.06	0.4164	0.0096	0.2348	0.0028	0.6636
	0.07	0.4289	0.0092	0.2457	0.0013	0.6851
	0.08	0.4522	0.0100	0.2404	0.0023	0.7049
	0.09	0.4676	0.0097	0.2291	0.0022	0.7086
	0.10	0.4711	0.0085	0.2380	0.0018	0.7194

Table 9-8b **Simulated sample size for each stage under the alternative hypothesis for the endpoint 1**

λ	$\alpha_1^{(1)}$	n_1	n_{21}	n_{22}	$n_{2(3)}$	n_{31}	n_{32}	EN_T
50%	0.05	36	42	65	48	32	63	312.8
	0.06	36	47	73	48	39	75	317.7
	0.07	35	53	80	49	42	82	322.4
	0.08	35	58	88	49	45	84	328.9
	0.09	35	63	92	49	47	89	332.3
	0.10	34	66	100	50	51	97	333.4
60%	0.05	36	35	55	48	28	55	300.4
	0.06	36	44	67	48	36	67	310.0
	0.07	35	46	74	49	37	71	311.8
	0.08	35	50	76	49	41	79	312.7
	0.09	35	54	83	49	45	89	317.3
	0.10	34	60	90	50	48	83	323.6
70%	0.05	36	29	46	48	26	47	290.6
	0.06	36	35	56	48	30	53	293.8
	0.07	35	40	63	49	34	63	298.2
	0.08	35	45	71	49	39	63	302.7
	0.09	35	49	75	49	41	71	305.0
	0.10	34	52	77	50	45	76	308.9

Table 9-8c Simulated sample size for each decision path under the alternative hypothesis for the endpoint 1

λ	$\alpha_1^{(1)}$	Decision path										EN
		A_1		A_2		A_4B_1		A_4B_2		A_4B_3		
		Prob	n	Prob	n	Prob	n	Prob	n	Prob	n	
50%	0.05	0.4169	78	0.0281	101	0.3478	116	0.0106	147	0.1966	84	93.8
	0.06	0.4539	83	0.0337	109	0.3340	123	0.0110	159	0.1674	84	98.2
	0.07	0.4723	88	0.0362	115	0.3348	126	0.0129	166	0.1438	84	102.1
	0.08	0.4844	93	0.0416	123	0.3345	129	0.0141	168	0.1254	84	106.2
	0.09	0.5142	98	0.0431	127	0.3216	131	0.0145	173	0.1066	84	109.5
	0.10	0.5422	100	0.0452	134	0.3146	135	0.0113	181	0.0867	84	112.1
60%	0.05	0.4195	71	0.0295	91	0.3021	112	0.0094	139	0.2395	84	87.7
	0.06	0.4487	80	0.0338	103	0.3187	120	0.0108	151	0.1880	84	95.0
	0.07	0.4628	81	0.0373	108	0.3042	121	0.0102	155	0.1855	84	95.5
	0.08	0.4949	85	0.0397	113	0.2959	125	0.0106	163	0.1589	84	98.6
	0.09	0.5147	89	0.0428	118	0.2896	129	0.0104	173	0.1425	84	102.0
	0.10	0.5233	94	0.0472	124	0.2972	132	0.0111	167	0.1212	84	106.3
70%	0.05	0.4236	65	0.0292	82	0.2621	110	0.0070	131	0.2781	84	83.0
	0.06	0.4578	71	0.0314	92	0.2553	114	0.0076	137	0.2479	84	86.4
	0.07	0.4681	75	0.0364	98	0.2680	118	0.0066	147	0.2209	84	89.8
	0.08	0.4952	80	0.0399	106	0.2615	123	0.0079	147	0.1955	84	93.6
	0.09	0.5159	84	0.0429	110	0.2497	125	0.0076	155	0.1839	84	95.9
	0.10	0.5184	86	0.0461	111	0.2598	125	0.0090	160	0.1667	84	97.6

Table 9-9a Simulated power under the alternative hypothesis for the endpoint 2

λ	$\alpha_1^{(1)}$	Decision path				Total power
		A_1	A_2	A_4B_1	A_4B_2	
50%	0.05	0.0125	0.3764	0.0080	0.3072	0.7041
	0.06	0.0141	0.4003	0.0070	0.2873	0.7087
	0.07	0.0136	0.4147	0.0061	0.2904	0.7248
	0.08	0.0126	0.4313	0.0078	0.2802	0.7319
	0.09	0.0123	0.4417	0.0081	0.2643	0.7264
	0.10	0.0127	0.4433	0.0083	0.2667	0.7310
60%	0.05	0.0155	0.3763	0.0068	0.2782	0.6768
	0.06	0.0140	0.4010	0.0054	0.2845	0.7049
	0.07	0.0138	0.4123	0.0039	0.2703	0.7003
	0.08	0.0165	0.4312	0.0054	0.2581	0.7112
	0.09	0.0157	0.4439	0.0046	0.2471	0.7113
	0.10	0.0169	0.4509	0.0039	0.2394	0.7111
70%	0.05	0.0186	0.3777	0.0036	0.2452	0.6451
	0.06	0.0164	0.4073	0.0047	0.2314	0.6598
	0.07	0.0188	0.4238	0.004	0.2309	0.6775
	0.08	0.0170	0.4267	0.0040	0.2224	0.6701
	0.09	0.0182	0.4460	0.049	0.2185	0.6876
	0.10	0.0180	0.4435	0.0042	0.2207	0.6864

Table 9-9b Simulated sample size for each stage under the alternative hypothesis for the endpoint 2

λ	$\alpha_1^{(1)}$	n_1	n_{21}	n_{22}	$n_{2(3)}$	n_{31}	n_{32}	EN_T
50%	0.05	36	65	41	48	65	31	311.7
	0.06	36	72	47	48	73	38	319.2
	0.07	35	80	51	49	83	41	322.8
	0.08	35	88	56	49	88	45	330.0
	0.09	35	92	61	49	93	47	335.3
	0.10	34	98	64	50	95	48	337.9
60%	0.05	36	53	34	48	56	28	300.3
	0.06	36	67	42	48	71	34	309.9
	0.07	35	71	46	49	74	37	310.8
	0.08	35	78	49	49	77	41	313.9
	0.09	35	84	53	49	85	45	319.2
	0.10	34	87	58	50	92	45	321.8
70%	0.05	36	47	28	48	48	25	290.1
	0.06	36	56	34	48	55	31	293.5
	0.07	35	63	39	49	64	35	296.1
	0.08	35	70	44	49	72	38	302.0
	0.09	35	75	47	49	73	40	304.4
	0.10	34	80	52	50	82	43	309.6

Table 9-9c Simulated sample size for each decision path under the alternative hypothesis for the endpoint 2

λ	$\alpha_1^{(1)}$	Decision path										EN
		A_1		A_2		A_4B_1		A_4B_2		A_4B_3		
		Prob	n	Prob	n	Prob	n	Prob	n	Prob	n	
50%	0.05	0.0529	111	0.4115	77	0.0288	149	0.3307	115	0.1761	84	94.6
	0.06	0.0617	108	0.4367	83	0.0372	157	0.3094	122	0.1550	84	99.6
	0.07	0.0694	115	0.4512	86	0.0376	164	0.3107	125	0.1311	84	102.8
	0.08	0.0793	123	0.4629	91	0.0426	172	0.2995	129	0.1157	84	107.6
	0.09	0.0882	127	0.4807	96	0.0491	177	0.2848	131	0.0972	84	111.5
	0.10	0.0978	132	0.4798	98	0.0494	179	0.2858	132	0.0872	84	113.8
60%	0.05	0.0479	89	0.4101	70	0.0222	140	0.3006	112	0.2192	84	88.2
	0.06	0.0568	103	0.4362	78	0.0258	155	0.3043	118	0.1769	84	94.6
	0.07	0.0687	106	0.4472	81	0.0235	158	0.2909	121	0.1697	84	96.7
	0.08	0.0814	113	0.4673	84	0.0268	161	0.2783	125	0.1462	84	99.8
	0.09	0.0886	119	0.4846	88	0.0320	169	0.2656	129	0.1292	84	103.7
	0.10	0.1048	121	0.4870	92	0.0330	176	0.2601	129	0.1151	84	106.5
70%	0.05	0.0480	83	0.4084	64	0.0117	132	0.2640	109	0.2679	84	82.9
	0.06	0.0584	92	0.4423	70	0.0140	139	0.2500	115	0.2353	84	86.8
	0.07	0.0713	98	0.4606	74	0.0159	148	0.2484	119	0.2038	84	90.1
	0.08	0.0808	105	0.4658	79	0.0179	156	0.2396	122	0.1959	84	93.8
	0.09	0.0918	110	0.4815	82	0.0201	157	0.2351	124	0.1715	84	96.3
	0.10	0.1046	114	0.4840	86	0.0228	166	0.2379	127	0.1507	84	100.2

Table 9-10a Simulated power under the alternative hypotheses for both endpoints

λ	$\alpha_1^{(1)}$	Decision path				Total power
		A_1	A_2	A_4B_1	A_4B_2	
50%	0.05	0.3826	0.2334	0.1843	0.0709	0.8712
	0.06	0.4048	0.2338	0.1753	0.0643	0.8782
	0.07	0.4231	0.2368	0.1743	0.0547	0.8889
	0.08	0.4312	0.2352	0.1691	0.0551	0.8906
	0.09	0.4614	0.2389	0.1526	0.0436	0.8965
	0.10	0.4768	0.2353	0.1534	0.0365	0.9020
60%	0.05	0.3779	0.2322	0.1715	0.0769	0.8585
	0.06	0.4061	0.2380	0.1672	0.0627	0.8740
	0.07	0.4238	0.2380	0.1526	0.0671	0.8815
	0.08	0.4574	0.2305	0.1471	0.0516	0.8866
	0.09	0.4672	0.2440	0.1364	0.0457	0.8933
	0.10	0.4684	0.2415	0.1367	0.0469	0.8935
70%	0.05	0.3851	0.2352	0.1436	0.0765	0.8404
	0.06	0.4120	0.2298	0.1350	0.0736	0.8504
	0.07	0.4255	0.2385	0.1355	0.0692	0.8692
	0.08	0.4448	0.2313	0.1306	0.0630	0.8697
	0.09	0.4659	0.2393	0.1243	0.0554	0.8849
	0.10	0.4779	0.2385	0.1221	0.0484	0.8869

Table 9-10b Simulated sample size for each stage under the alternative hypotheses for both endpoints

λ	$\alpha_1^{(1)}$	n_1	n_{21}	n_{22}	$n_{2(3)}$	n_{31}	n_{32}	EN_T
50%	0.05	36	42	41	48	33	31	279.9
	0.06	36	47	46	48	37	38	285.8
	0.07	35	52	52	49	43	41	290.4
	0.08	35	58	57	49	47	44	299.3
	0.09	35	62	61	49	48	48	300.5
	0.10	34	66	65	50	52	48	303.5
60%	0.05	36	35	35	48	28	28	260.1
	0.06	36	43	42	48	35	35	263.2
	0.07	35	46	45	49	38	38	268.2
	0.08	35	51	49	49	42	38	272.0
	0.09	35	54	53	49	44	45	273.9
	0.10	34	58	58	50	48	44	280.0
70%	0.05	36	29	29	48	25	24	256.9
	0.06	36	34	34	48	30	29	262.0
	0.07	35	40	39	49	35	34	266.7
	0.08	35	45	45	49	37	37	273.1
	0.09	35	47	48	49	43	40	274.7
	0.10	34	54	51	50	43	43	278.4

Table 9-10c Simulated sample size for each decision path under the alternative hypotheses for both endpoints

λ	$\alpha_1^{(1)}$	Decision path										EN
		A_1		A_2		A_4B_1		A_4B_2		A_4B_3		
		Prob	n	Prob	n	Prob	n	Prob	n	Prob	n	
50%	0.05	0.4226	78	0.2577	77	0.1993	117	0.0762	115	0.0442	84	88.6
	0.06	0.4483	83	0.2565	82	0.1912	121	0.0684	122	0.0356	84	92.7
	0.07	0.4692	87	0.2561	87	0.1890	127	0.0598	125	0.0259	84	96.8
	0.08	0.4786	93	0.2555	92	0.1843	131	0.0589	128	0.0227	84	101.6
	0.09	0.5113	97	0.2576	96	0.1658	132	0.0463	132	0.0190	84	103.9
	0.10	0.5289	100	0.2520	99	0.1667	136	0.0397	132	0.0127	84	106.8
60%	0.05	0.4141	71	0.2514	71	0.1854	112	0.0829	112	0.0662	84	82.9
	0.06	0.4477	79	0.2591	78	0.1808	119	0.0686	119	0.0438	84	88.9
	0.07	0.4645	81	0.2578	80	0.1661	122	0.0720	112	0.0396	84	89.9
	0.08	0.5036	86	0.2504	84	0.1600	126	0.0555	124	0.0305	84	93.9
	0.09	0.5165	89	0.2610	88	0.1484	128	0.0493	129	0.0248	84	96.4
	0.10	0.5163	92	0.2620	92	0.1490	132	0.0512	128	0.0215	84	99.6
70%	0.05	0.4216	65	0.2556	65	0.1540	109	0.0835	108	0.0853	84	76.9
	0.06	0.4541	70	0.2509	70	0.1462	114	0.0801	113	0.0687	84	80.8
	0.07	0.4658	75	0.2593	74	0.1474	119	0.0738	118	0.0537	84	84.9
	0.08	0.4884	80	0.2525	80	0.1441	121	0.0672	121	0.0478	84	88.9
	0.09	0.5080	82	0.2586	83	0.1343	127	0.0600	124	0.0391	84	90.9
	0.10	0.5247	88	0.2565	85	0.1313	127	0.0523	127	0.0352	84	94.3

10 A varying-stage adaptive phase II/III clinical trial design: Practical issues and trial implementation

10.1 What need to be specified in the clinical study protocol

For our varying-stage adaptive phase II/III design, the followings need to be specified in the clinical study protocol.

- Trial decision paths as shown with the flow charts in Figure 6-1 or Figure 9-3.
- Threshold values: $\alpha_1^{(1)}$, $\alpha_2^{(1)}$, $\alpha_1^{(2)}$, $\alpha_2^{(2)}$, $\alpha_F^{(1)}$, $\alpha_F^{(2)}$ and λ. Usually, the same threshold values can be applied to both study endpoints ($\alpha_1^{(1)} = \alpha_2^{(1)}$ and $\alpha_1^{(2)} = \alpha_2^{(2)}$). In order to have the thresholds at the same level for both initial stage and intermediate stage, the threshold values for the intermediate stage can be specified as $\alpha_1^{(2)} = \alpha_2^{(2)} = \exp(-0.5\,\chi^2_{\alpha_1^{(1)},4})$ and $\alpha_F^{(2)} = \exp(-0.5\,\chi^2_{\alpha_F^{(1)},4})$. In addition, to ensure plausibility of thresholds, as described in Section 8.3.1, the requirement of $\alpha_F^{(2)} \le \alpha_1^{(1)}$ has to be met.

  When $\alpha_F^{(1)} = 1$ and $\alpha_F^{(2)} = 0$, the varying-stage adaptive two-stage phase II/III clinical trial design is reduced to the special case as presented in Section 9.3.
- Statistical method to perform hypothesis testing. In our simulations, we assume that the two study endpoints follow normal distributions; therefore, Dunnett test under one-way ANOVA is used. However, the principle of statistical analysis to perform hypothesis testing for our design is adaptive combination test, which is not limited to normally distributed study endpoints. Even for normally distributed study endpoints, one may use

Dunnett test under general linear model (Hsu, 1996) or mixed effect model.

- Type I error rate. Our design is more complex, we propose to use closed testing procedure (Marcus, et al, 1976) to preserve Type I error rate in the strong sense.
- Conditional statistical power for sample size determination. In this thesis, we propose to use conditional power to determine sample size for the final stage. Based on our experience, ≥ 80% conditional power is sufficient to ensure adequate overall statistical power.
- Overall statistical power and simulation. Due to trial design complexity, no derivation of overall statistical power can be obtained theoretically. We recommend performing simulations under various scenarios to characterize the design and to ensure sufficient statistical power for the trial with feasible sample size.

10.2 What does not need to be specified in the clinical study protocol

There is no need to specify dose selection rules in the study protocol. The decision of dose selection can be made using trial interim data and/or information external to the trial, thus, more flexibility is granted to decision makers on dose selection. We propose to use closed testing procedure (Marcus, et al, 1976) to perform the final analyses. This procedure preserves Type I error rate in the strong sense, obviating the need to specify the specific dose selection rules in the study protocol.

10.3 Blinding issues

In our design, the interim decisions need to be made to adaptively switch the primary study endpoint, to select a dose (doses), and to determine sample size for the next stage. Although these decisions are based on unblended interim analysis results with stagewise p-value or combined p-value, in general, in order to preserve the integrity of the trial, the clinical trial team should be blinded to treatment allocation related data during the trial execution. The interim analyses should be carried out by an independent team including independent statisticians and independent clinicians who do not involve in daily activities of the trial.

The Data Monitoring Committee (DMC) plays important roles in reviewing interim data and making interim decision. Therefore, the scope of interim analyses and adaptation strategies should be clearly specified in the DMC charter. Our design is more complex, the sponsor should ensure the transparent and efficient communication with the DMC. In addition, the appropriate documentation of adaptive process should be put in place with restricted access only to the independent team during trial execution.

10.4 Unblinded sample size determination

During past decades, many researchers studied sample size re-estimation. If the sample size re-estimation is based on nuisance parameters such as common variance for a normally distributed primary study endpoint or pooled event rate for a binary response variable, the Type I error rate will not be materially inflated (e.g. Gould & Shih, 1992; Shih & Gould, 1995; Shih & Zhao, 1997; Shih & Long, 1998).

Regarding sample size re-estimation using unblended interim results, Chen, DeMets and Lan (2004) pointed out that increasing the sample size when conditional power under current trend is great than 0.5 will only decrease the Type I error rate conditional on the data observed. In our design, as aforementioned, we recommend 80% or higher conditional power to determine sample size for the next stage, therefore, sample size determination in our design using updated treatment effect τ^* from the data cumulated to the interim analysis should not inflate Type I error rate.

10.5 Clinical utility index

Recently, some researchers (e.g. Poland et al, 2009) proposed to use clinical utility index (CUI) to support drug development decision. The CUI is defined as.

$$CUI = \sum_{i=1}^{m} w_i U_i(x_i) \qquad (10.1)$$

where m is total number of drug attributes, w_i is the importance weight to the i^{th} attribute, and $U_i(x_i)$ is a utility function for that drug attribute. The attributes typically limited to the drug efficacy and safety profiles, and those related to market value, development cost or time are omitted from the CUI. Therefore, CUI can be used to evaluate the study drug's benefit-risk (efficacy-safety) ratio; hence, it can be a useful tool to facilitate trial interim decisions. Further, some researchers (e.g Skrivanek, 2008) propose to use CUI to perform adaptive randomization, in which patients are randomized to the dose arm with the best CUI at the time being. It is worthwhile to investigate in the future regarding how CUI can be incorporated in our design.

Appendix

A.1 Relation of the posterior probability $p(\theta|s)$ at stage II and data (s_1, n_1) obtained from stage I

Sambucini (2008) showed the relation of the posterior probability $p(\theta|s)$ at stage II and data (s_1, n_1) obtained from stage I for a two-stage setting with an acceptance boundary at the interim I. In this appendix, we extend this work to the two-stage setting with both acceptance and rejection boundaries at the interim I.

Proposition A1: the data (s_1, n_1) obtained from stage I do not affect the posterior probability of $p(\theta|s)$ at stage II.

Proof:

a) Stage I

Prior: $\theta \sim beta\,(a, b)$ (A.1.1)

Hypothetical data: $S_1 \mid \theta \sim binomial(n_1, \theta)$ (A.1.2)

Posterior: $p(\theta|\, a_1<S_1<r_1) \propto p(\theta)p(a_1 < S_1 < r_1|\,\theta)$

$$\propto \theta^{a-1}(1-\theta)^{b-1}[Bin(r_1-1, n_1, \theta)-Bin(a_1, n_1, \theta)] \qquad (A.1.3)$$

b) Stage II

Likelihood: $p(s|a_1 < S_1 < r_1, \theta) = \dfrac{p(S=s, a_1<S_1<r_1 \mid \theta)}{p(a_1<S_1<r_1 \mid \theta)}$

$$= \frac{\sum_{s_1=a_1+1}^{r_1-1} bin(s_1, n_1, \theta) bin(s-s_1, n_2, \theta)}{Bin(r_1-1, n_1, \theta) - Bin(a_1, n_1, \theta)}$$

$$= \frac{\sum_{s_1=a_1+1}^{r_1-1} \binom{n_1}{s_1} \theta^{s_1}(1-\theta)^{n_1-s_1} \binom{n_2}{s-s_1} \theta^{s-s_1}(1-\theta)^{n_2-s+s_1}}{Bin(r_1-1, n_1, \theta) - Bin(a_1, n_1, \theta)}$$

$$= \frac{\theta^{s}(1-\theta)^{n-s} \sum_{s_1=a_1+1}^{r_1-1} \binom{n_1}{s_1} \cdot \binom{n_2}{s-s_1}}{Bin(r_1-1, n_1, \theta) - Bin(a_1, n_1, \theta)} \qquad (A.1.4)$$

posterior probability : $p(\theta|s, a_1 < S_1 < r_1) \propto p(\theta| a_1<S_1<r_1)\, p(S|a_1 < S_1 < r_1, \theta)$

$$\propto \theta^{a-1}(1-\theta)^{b-1}\theta^{s}(1-\theta)^{n-s}\sum_{s_1=a_1+1}^{r_1-1}\binom{n_1}{s_1}\cdot\binom{n_2}{s-s_1}$$

$$= \theta^{a+s-1}(1-\theta)^{b+n-s-1}\sum_{s_1=a_1+1}^{r_1-1}\binom{n_1}{s_1}\cdot\binom{n_2}{s-s_1}$$

$$\propto Beta(a+s, b+n-s) \qquad (A.1.5)$$

Therefore, $p(\theta|s, a_1 < S_1 < r_1)= p(\theta|s)$. This indicates that the data (s_1, n_1) obtained from stage I do not affect the posterior probability of $p(\theta|s)$ at stage II, and explains why the predictive probability is constructed as in (3.8).

A.2 Prior predictive probability

$$p(s \mid a_1 < S_1 < r_1) = \int_0^1 p(S = s \mid a_1 < S_1 < r_1, \theta)\, p(\theta \mid a_1 < S_1 < r_1)\, d\theta$$

$$\propto \sum_{s_1=a_1+1}^{r_1-1}\binom{n_1}{s_1}\cdot\binom{n_2}{s-s_1}\int_0^1 \theta^{a+s-1}(1-\theta)^{b+n-s-1}\, d\theta$$

$$= \frac{\Gamma(a+s)\Gamma(b+n-s)}{\Gamma(a+b+n)}\sum_{s_1=a_1+1}^{r_1-1}\binom{n_1}{s_1}\cdot\binom{n_2}{s-s_1} \qquad (A.2.1)$$

$$\textit{Normalization constant } C = \sum_{S=s_1}^{n_2+s_1}\left\{\frac{\Gamma(a+s)\Gamma(b+n-s)}{\Gamma(a+b+n)}\sum_{s_1=a_1+1}^{r_1-1}\binom{n_1}{s_1}\cdot\binom{n_2}{s-s_1}\right\}$$

A.3 Predictive probability of continuing the trial to stage II

$$PP(a_1<S_1<r_1|n_1, n_2) = \sum_{S_2=s_1}^{n_2+s_1} \mathrm{P}(\mathrm{S}_2 \mid \mathrm{a}_1 < \mathrm{S}_1 < \mathrm{r}_1)\cdot \mathrm{I}\{\mathrm{P}(\theta \ge \theta_0 \mid \mathrm{a}_1 < \mathrm{S}_1 < \mathrm{r}_1, \mathrm{S}_2) \ge P_{\mathrm{T}}\} \qquad (A.3.1)$$

A.4 Monotonic property of predictive probability

Proposition A2: Predictive probability defined in (3.8) is a non-decreasing function with respect to the responses s_1.

Proof: Let s_2 denote number of future responses at stage II, Whitehead at el. (2008) demonstrated that the following posterior probability is an increasing function with respect to number of responses (s_1) at stage I.

$$P(s_1 \mid n_1, s_2, n_2) = \Pr ob(\theta > \theta_0 \mid (s_1, n_1), (s_2, n_2)) \qquad (A.4.1)$$

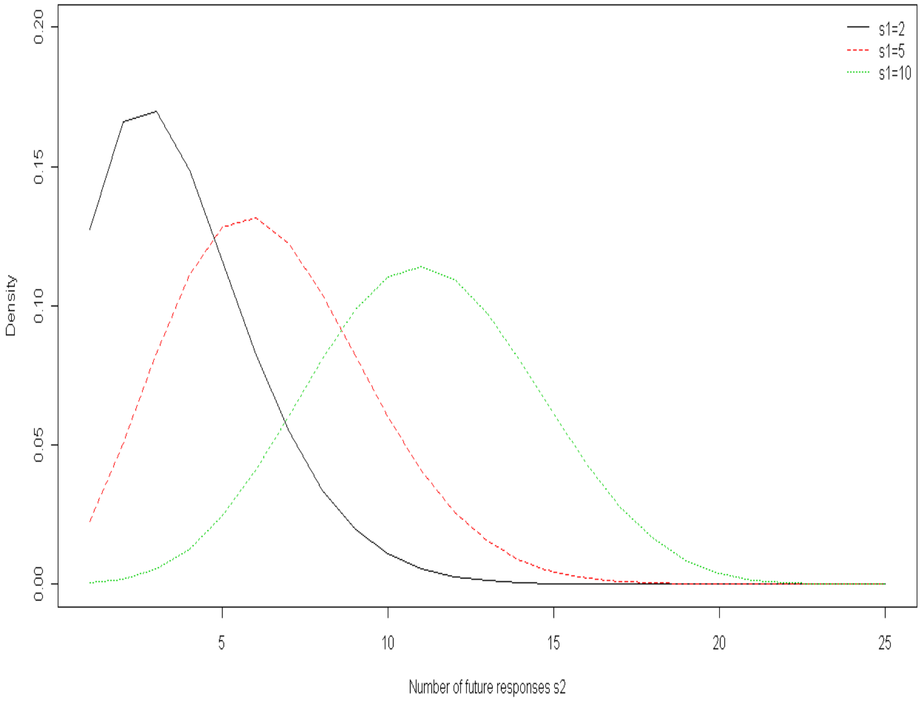
Given conjugate beta prior beta(a, b), and responses s_1 among n_1 patients at stage I, the future s_2 responses among n_2 patients at stage II follows beta-binomial distribution, the mean and variance can be expressed as follows.

$$E[S_2 \mid (s_1, n_1, n_2)] = \frac{n_2(a+s_1)}{a+b+n_1} = n_2 E[\theta \mid (s_1, n_1)] \qquad (A.4.2)$$

$$Var[S_2 \mid (s_1, n_1, n_2)] = \frac{n_2(a+s_1)(b+n_1-s_1)(a+b+n)}{(a+b+n_1)^2(a+b+n_1+1)} \qquad (A.4.3)$$

From above mean value and variance expression, also as shown in Figure A.4.1, the distribution of the future s_2 responses moves toward right side as response s_1 increases given size n_1 and n_2 for stage I and stage II respectively, therefore, sum of $P(s_2 \mid (s_1, n_1, n_2))$ is non-decreasing with respect to s_1. Hence predictive probability defined in (3.8) is a non-decreasing function, and boundaries a_1 and r_1 for stage I exist.

Figure A.4.1 **Beta-binomial distribution of s_2 given s_1, n_1=20, n_2=25, a=2 and b=4**



A.5 Some properties of beta and binomial distribution

A.5.1 Relation of beta and binomial probability calculations

The regularized (normalized) incomplete beta function is defined based on the incomplete beta function $B_\theta(u,v)$ and the complete beta function $B(u,v)$ as follows.

$$I_\theta(u,v)=\frac{\int_0^\theta t^{u-1}(1-t)^{v-1}dt}{\int_0^1 t^{u-1}(1-t)^{v-1}dt}=\frac{B_\theta(u,v)}{B(u,v)},\quad \text{u>0 and v>0} \qquad \text{(A.5.1)}$$

Using integration by parts, the regularized incomplete beta function can be expressed as (formula 6.6.4, Abramowitz and Stegun, 1965)

$$I_\theta(a,n-a+1)=\sum_{j=a}^{n}\binom{n}{j}\theta^{i}(1-\theta)^{n-j} \qquad \text{(A.5.2)}$$

Let $u = a$, $v = n - a + 1$ and $n = u + v - 1$, then the above formula can be written as

$$I_\theta(u,v) = \sum_{i=u}^{u+v-1} \binom{u+v-1}{i} \theta^i (1-\theta)^{u+v-1-i} \qquad \text{(A.5.3)}$$

Revise the term at the right side of (A.5.3) based on binomial distribution,

$$I_\theta(u,v) = \sum_{i=u}^{u+v-1} bin(\theta, i, u+v-1) = 1 - Bin(\theta, u-1, u+v-1) \qquad \text{(A.5.4)}$$

where bin and Bin are binomial density function and cumulative distribution function.

The left side of (A.5.4) is the probability of $x \le \theta$ based on $x \sim beta(u, v)$, therefore,

$$Beta(\theta, u, v) = 1 - Bin(\theta, u-1, u+v-1) \qquad \text{(A.5.5a)}$$

where beta and Beta are density function and cumulative distribution function for beta distribution. The (A.5.5a) can be written as

$$Bin(\theta, x, m) = 1 - Beta(\theta, x+1, m-x) \qquad \text{(A.5.5b)}$$

The formula (A.5.5a) shows that beta probability can be calculated based on a binomial distribution; vice verse, binomial probability can be calculated based on a beta distribution as shown in (A.5.5b).

A.5.2 A monotonic property of beta probability function

Proposition A3: *Beta(θ, u, v)* is a monotonic decreasing function as u increases given *u+v* is fixed.

Proof: The probability *Bin(u-1, u+v-1, θ | u+v)* is a monotonic increasing function as u increases, therefore, from *(A.5.5a), Beta(θ, u, v)* is a monotonic decreasing function as u increases given *u+v* is fixed.

A.5.3 Beta(θ, u, v) ≥ Beta(θ, u+k, v), k>0

Proposition A4: Beta(θ, u, v) ≥ Beta(θ, u+k, v), k>0

Proof: Abramowitz and Stegun (1965) provided the following as formula 26.5.16,

$$I_\theta(u,v) = \frac{\Gamma(u+v)}{\Gamma(u+1)\Gamma(v)} \theta^a (1-\theta)^b + I_\theta(u+1, v) \qquad \text{(A.5.6)}$$

Apparently, $Beta(\theta, u, v) \geq Beta(\theta, u+1, v)$. This was also described by Thall (1994). As consequence of such decomposition, $Beta(\theta, u, v) \geq Beta(\theta, u+1, v) \geq Beta(\theta, u+2, v) \geq \ldots\ldots$. In short expression,

$$Beta(\theta, u, v) \geq Beta(\theta, u+k, v),\ k\geq 1 \qquad \text{(A.5.7)}$$

A.5.4 Bin(*θ,* x, m)≤Bin(*θ,* x+k, m+k)

Proposition A5: $Bin(\theta, x, m) \leq Bin(\theta, x+k, m+k),\ k>0.$

Proof: Following (A.5.5a), (A.5.7) can be expressed in binomial form as

$$1\text{-}Bin(\theta, u\text{-}1, u+v\text{-}1) \geq 1\text{-}Bin(u\text{-}1+k, u+v\text{-}1+k)$$

Let x=u-1 and m=u+v-1, then we have

$$Bin(\theta, x, m) \leq Bin(\theta, x+k, m+k),\ k>0 \qquad \text{(A.5.8)}$$

A.5.5 A monotonic property of binomial distribution

To study monotonic property of binomial distribution, let's compare binomial density functions $bin(\theta,$ k, $m)$ and $bin(\theta,$ k-1, $m)$.

$$\frac{bin(\theta, k, m)}{bin(\theta, k-1, m)} = \frac{\binom{m}{k}\theta^{k}(1-\theta)^{m-k}}{\binom{m}{k-1}\theta^{k-1}(1-\theta)^{m-k+1}} = \frac{m-k+1}{k}\frac{\theta}{1-\theta} = 1 + \frac{(m+1)\theta - k}{k(1-\theta)}$$

(A.5.9)

From (A.5.9) , apparently, binomial density function $bin(k, m, \theta)$ is monotone increasing as k increases for $x < (m + 1)\theta$; and monotone decreasing as k increases for $k > (m + 1)\theta$. There exists an integer k that satisfies the following (A.5.10), which is known as the most probable (most likely) outcome of Bernoulli trials; when $(m + 1)\theta$ is an integer, there are two maximum binomial density values for $k = (m+ 1)\theta$ and $k - 1$.

$$(m + 1)\theta - 1 < k \leq (m + 1)\theta \qquad \text{(A.5.10)}$$

A.5.6 Compliment property of binomial probability

Proposition A6: compliment property of binomial probability:

$$Bin\ (\theta,\ x,\ m) = 1 - Bin(1\text{-}\ \theta,\ m\text{-}x\text{-}1,\ m) \tag{A.5.11a}$$

$$1 - Bin(\theta,\ x,\ m)\ = Bin(1\text{-}\ \theta,\ m\text{-}x\text{-}1,\ m) \tag{A.5.11b}$$

Proof: The following was given by Abramowitz and Stegun (1965, formula 6.6.3),

$$I_\theta(u,v) = 1 - I_{1-\theta}(v,u) \tag{A.5.12}$$

Let $u = x + 1,\ v = m\text{-}x$, then (A.5.12) can be written as

$$1 - I_\theta(x+1, m-v) = I_{1-\theta}(m-x, x+1) \tag{A.5.13}$$

Apply (A.5.5a) to replace incomplete beta functions in both side of (A.5.13) with binomial probabilities, and then we can obtain the formula (A.5.11a). The formula (A.5.11b) can be derived similarly.

A.5.7 Incremental property of binomial probability

Proposition A7: for 0<k<m, the following incremental property of binomial probability is true if $x < m\theta$.

$$Bin(\theta,\ x,\ m) < Bin(\theta,\ x+k\theta,\ m+k) \tag{A.5.14}$$

Proof: The proof of (A.5.14) is carried out by finding a condition which satisfies (A.5.14). Use normal approximation to binomial probability, (A.5.14) can be written as

$$\frac{x - m\theta}{\sqrt{m\theta(1-\theta)}} < \frac{x + k\theta - (m+k)\theta}{\sqrt{(m+k)\theta(1-\theta)}} \tag{A.5.15}$$

(A.5.15) can be simplified as

$$\left(\frac{1}{\sqrt{m}} - \frac{1}{\sqrt{m+k}}\right)\theta < \frac{1}{\sqrt{m}}\frac{x+k\theta}{m+k} - \frac{1}{\sqrt{m+k}}\frac{x}{m} \tag{A.5.16}$$

Multiply (A.5.16) by $\sqrt{m(m+k)}$ and further simplify,

$$(\sqrt{m+k} - \sqrt{m})\theta < \frac{m\sqrt{m+k} - \sqrt{m}(m+k)}{m(m+k)}x + \frac{k}{\sqrt{m+k}}\theta \tag{A.5.17}$$

Solve x from (A.5.17),

$$x < (\sqrt{m+k} - \sqrt{m} - \frac{k}{\sqrt{m+k}})\frac{\sqrt{m(m+k)}}{\sqrt{m} - \sqrt{m+k}}\theta$$

$$= (\frac{k}{\sqrt{m+k} - \sqrt{m}} - \sqrt{m+k})\sqrt{m}\theta \qquad \text{(A.5.18)}$$

Let 0<k<m, equivalently, $\frac{k}{m} = c$, 0<c<1, (A.5.18) can be simplifies as

$$x < (\frac{c}{\sqrt{c+1} - 1} - \sqrt{c+1})m\theta \qquad \text{(A.5.19)}$$

For any 0<c<1, $\frac{c}{\sqrt{c+1} - 1} - \sqrt{c+1}$ is approximately equal to 1 (the difference between this term and 1 is less than 10^{-10} by numerical calculation), therefore, the incremental property specified in (A.5.14) holds if number of responses *x* satisfies (A.5.20).

$$x < m\theta \qquad \textit{(A.5.20)}$$

A.6 Gibbs' inequality (Gibbs, 1902, 1960)

Let P = {p_1, p_2, ..., p_n} and Q = {q_1, q_2, ..., q_n} are two probability distributions. The following Gibbs' inequality holds.

$$\sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \geq 0$$

This was first presented by Josiah Willard Gibbs in the 19th century.

Proof: Because log(x) ≤ x- 1 for any x≥0, we have

$$p_i \log \frac{q_i}{p_i} \leq p_i(\frac{q_i}{p_i} - 1) = q_i - p_i$$

Sum up the both sides of the above equation , and then we have

$$\sum_{i=1}^{n} p_i \log \frac{q_i}{p_i} \leq \sum_{i=1}^{n} q_i - \sum_{i=1}^{n} p_i = 0 - 0 = 0$$

Hence,

$$\sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \geq 0$$

Moreover, the equality holds if and only if P=Q.

When p(x) and q(x) are continuous probability density function for $x \in R$, then

$$\int_R p(x) \log \frac{p(x)}{q(x)} \geq 0, \quad x \in R$$

A.7 Proof of unconditional distribution of a combining p-value under the null hypothesis

a) Proof based on logarithmic transformation

P-value under the null hypothesis (H_0) of no treatment difference follows Uniform(0,1), namely p ~ Uniform(0,1), where $0<p<1$.

Let x = -2ln(p), $0<x<\infty$. Then the probability density function (pdf) for x can be derived as follows by using Jacobin transfer.

$f(x) = f(p)\left|\frac{dp}{dx}\right| = \frac{1}{2}e^{-x/2}$, which is pdf of χ_2^2.

Therefore, x = -2ln(p) ~ χ_2^2, where $0<x<\infty$ and $0<p<1$.

Let p_1 and p_2 are independent p-values from stage I and stage II, such that p_1~Uniform(0,1) and p_2~Uniform(0,1) under H_0. Following Fisher's product method, p_1 and p_2 can be combined as follows.

$Y = C(p_1, p_2) = p_1p_2$.

Hence, $-2\ln(y) = -2\ln(p_1) - 2\ln(p_2) \sim \chi_4^2$

The cumulative distribution function (cdf) of y is

$F(y) = \Pr(Y \leq y) = \Pr(-2\ln(Y) \geq -2\ln(y)) = \Pr(w \geq -2\ln(y))$, where $w = -2\ln(Y) \sim \chi_4^2$, herefore, $F(y) = F_{\chi_4^2}(-2\ln(y)) = 1 - \int_0^{-\ln(y)} te^{-t}dt = y - y\ln(y)$

Hence the probability density function (pdf) of y is

$$f(y) = -\ln(y)$$

The Figure A.7.1 and A.7.2 show the cdf and pdf of y.

Figure A.7.1 Cumulative density function (cdf) of combined p-value (y)

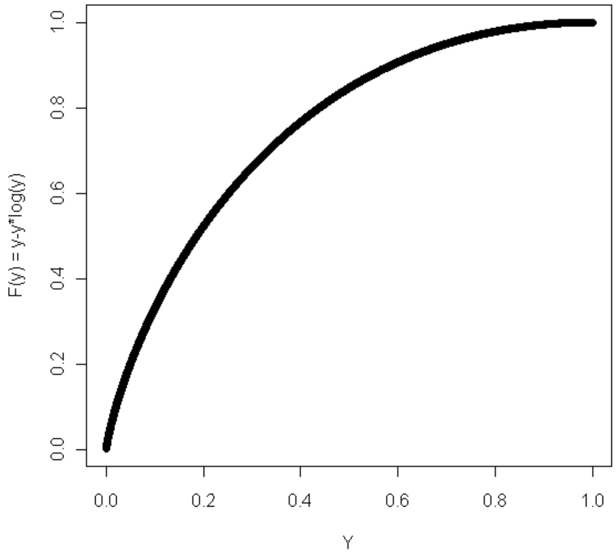
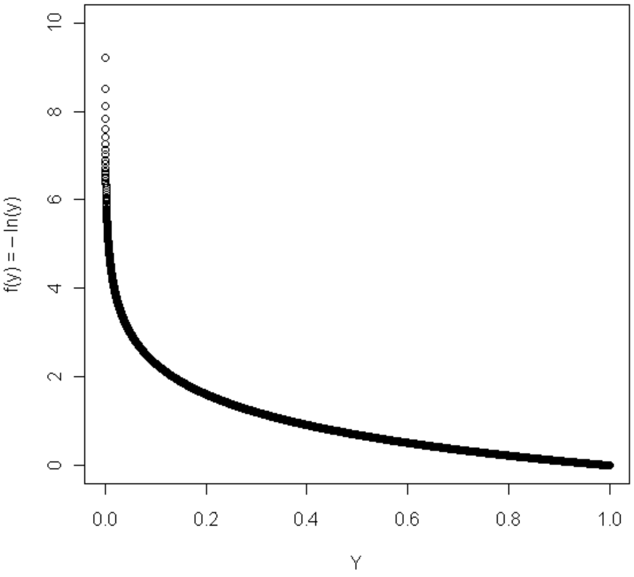


Figure A.7.2 Probability density function (pdf) of combined p-value (y)



b) Alternative proof based on Jacobian transformation:

As indicated before, p_1~Uniform(0,1) and p_2~Uniform(0,1) under H_0.

Assume p_1 and p_2 are independent, the join pdf of p_1 and p_2 is as follows

$$f(p_1, p_2) = 1, \quad 0 \le p_1 \le 1 \text{ and } 0 \le p_2 \le 1$$

Let

$$\begin{cases} Y = C(p_1, p_2) = p_1, p_2 \\ Z = p_2 \end{cases}, \quad 0 \le p_1 \le 1 \text{ and } 0 \le p_2 \le 1$$

Hence,

$$\begin{cases} p_1 = \dfrac{Y}{Z} \\ p_2 = Z \end{cases}, \quad 0 \le Y \le 1 \text{ and } Y \le Z \le 1$$

The Jacobian transformation determinant is

$$J = \begin{vmatrix} \frac{\partial p_1}{\partial y} & \frac{\partial p_1}{\partial z} \\ \frac{\partial p_2}{\partial y} & \frac{\partial p_2}{\partial z} \end{vmatrix} = \begin{vmatrix} \frac{1}{z} & -\frac{y}{z^2} \\ 0 & 1 \end{vmatrix} = \frac{1}{z}$$

The join pdf of Y and Z is

$$f(y,z) = f(p_1, p_2)|J| = \frac{1}{z}, \quad y \le z \le 1$$

The marginal pdf of Y is

$$f(y) = \int_y^1 \frac{1}{z} dz = \ln(z)\,|_y^1 = -\ln(y), \qquad 0 \le y \le 1$$

Therefore, the cdf of Y is

$$F(y) = \int_0^y -\ln(x) dx = y - y\ln(y), \qquad 0 \le y \le 1$$

c) Alternative proof based on definition of cdf:

$$F(y) = \Pr(Y \le y) = \Pr(p_1 p_2 \le y) = \Pr(p_1 \le \frac{y}{p_2}) = \int_0^y \Pr(p_1 \le \frac{y}{p_2} \mid p_2) dp_2 + \int_y^1 \Pr(p_1 \le \frac{y}{p_2} \mid p_2) dp_2$$

When $0 \le p2 \le y$, $\Pr(p_1 \le \frac{y}{p_2} \mid p_2) = 1$;

When $y \le p2 \le 1$, $\Pr(p_1 \le \frac{y}{p_2} \mid p_2) = \frac{y}{p_2}$;

Therefore,

$$F(y) = \int_0^y \Pr(p_1 \le \frac{y}{p_2} \mid p_2) dp_2 + \int_y^1 \Pr(p_1 \le \frac{y}{p_2} \mid p_2) dp_2$$

$$= \int_0^y dp_2 + \int_y^1 \frac{y}{p_2} dp_2 = p_2 \mid_0^y - y\ln(p_2) \mid_y^1 = y - y\ln(y), \quad 0 \le y \le 1$$

Hence,

$f(y) = -\ln(y), \qquad 0 \le y \le 1$

A.8 Notation and glossary

θ – Response rate of the testing treatment ($0 \le \theta \le 1$)

beta(a, b) – Beta prior with parameters a and b

PP – Predictive probability defined in (3.8)

P_L, P_U – Lower and upper threshold predictive probability

P_T – Threshold posterior provability to evaluate trial success

θ_0 – Maximum uninteresting response rate

θ_1 – Expected or target response rate of testing treatment

R – Rejection of null hypothesis H_0: $\theta \le \theta_0$

A – Acceptance of null hypothesis H_0: $\theta \le \theta_0$

α^f, β^f – Frequentist Type I, and Type II error rate

α^B, β^B – Bayesian Type I, and Type II error rate

s_1, s_2 – Number of responses at stage I, and stage II.

s – Number of responses from the whole trial ($s = s_1 + s_2$).

r_1, r – Upper boundary to declare $\theta \ge \theta_1$ at stage I, and stage II.

a_1, a – Lower boundary to reject $\theta \ge \theta_1$ at stage I, and stage II.

n_1, n_2 – Sample size for stage I, and stage II.

n – Sample size for the whole trial ($n = n_1 + n_2$).

PET_f, PET_B – Probability of early termination under frequentist, and Bayesian framework.

bin – Binomial probability density function

Bin – Cumulative binomial distribution function

B – Beta function

$I_\theta(u, v)$ – Incomplete beta function

beta – Beta probability density function

Beta – Cumulative beta distribution function

$\pi_0(\theta)$, $\pi_t(\theta)$ – Distribution for θ under H_0 and true prior distribution for θ

BF_1, BF_t - Bayes factor under H_1 and $\pi_t(\theta)$

$EWOE_1$, $EWOE_t$ – expected weight of evidence from H_1 and $\pi_t(\theta)$

$m_k(x)$, $m_t(x)$ – Marginal density of X under the hypothesis H_k ($k = 0, 1$) and $\pi_t(\theta)$

iMOM – Inverse moment prior

ER – Type I error rate

PW – Statistical power

$C(p_{ij}, p_{ij`})$ – Combination of the p-values $p_{ij}$ and $p_{ij`}$ from j^{th} and $j`^{th}$ stage for the i^{th} study endpoint

λ - Percent of alpha allocated for the two-stage setting

References

1. Abramowitz M and Stegun I. A, eds. (1965) Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. New York: Dover
2. Alosh M, Huque M. A consistency-adjusted alpha-adaptive strategy for sequential testing Stat Med. 2010; 29(15):1559–1571
3. Barnes PJ, Pocock SJ, Magnussen H, Iqbal A, Kramer B, Higgins M, Lawrence D. Integrating indacaterol dose selection in a clinical study in COPD using an adaptive seamless design. Pulm Pharmacol Therapeutics. 2010 Jun;23(3):165-71
4. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. Stat Med. 1999 Jul 30;18(14):1833-48
5. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. Biometrics. 1994 Dec;50(4):1029-41
6. Berry DA. A guide to drug discovery: Bayesian clinical trials. Nature Reviews Drug Discovery 2006; 5: 27–36.
7. Bischoff W, Miller F. A seamless phase II/III design with sample-size re-estimation. J Biopharm Stat. 2009 Jul;19(4):595-609.
8. Brannath W, Koenig F, Bauer P. Multiplicity and flexibility in clinical trials. Pharmaceutical Statistics 2007; 6:205–216.
9. Bretz F, Schmidli H, Koenig F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts (with Discussion). Biometrical Journal 2006; 48:623–634.
10. Bretz F, Koenig F, Brannath W, Glimm E, Posch M. Adaptive designs for confirmatory clinical trials. Stat Med. 2009;

11. Bretz F, Wang SJ. From Adaptive Design to Modern Protocol Design for Drug Development: Part II. Success Probabilities and Effect Estimates for Phase 3 Development Programs. Drug Information Journal 2010 : 44(03)
12. Burman CF, Sonesson C. Are flexible designs sound? Biometrics. 2006 Sep;62(3):664-9
13. Chang M, Chow SC. A hybrid Bayesian adaptive design for dose response trials. J Biopharm Stat. 2005;15(4):677-91
14. Chang, M.N.; Therneau, T.N.; Wieand, H.S.; Cha, S.S. Designs for Group Sequential Phase II Clinical Trials. Biometrics 1987, 43, 865–874.
15. Chen TT. Optimal three-stage designs for phase II cancer clinical trials. Stat Med 1997; 16: 2701–11.
16. Chen J, DeMets D, Lan G. Increasing the sample size when the unblinded interim result is promising. Stat. Med 2004; 23:1023-1038
17. Chen TT, Ng TH. Optimal flexible designs in phase II clinical trials. Stat Med. 1998;17(20):2301-12
18. Cui, L., Hung, H. M. J., Wang, S. J. Modification of sample size in group sequential clinical trials. Biometrics 1999; 55:853–857.
19. Daimon T. Predictive checking for Bayesian interim analyses in clinical trials. Contemp Clin Trials. 2008;29(5):740-50
20. Dmitrienko A, Wang MD. Bayesian predictive approach to interim monitoring in clinical trials. Stat Med 2006; 25:2178–95.
21. Fleming, T.R. One Sample Multiple Testing Procedure for Phase II Clinical Trials. Biometrics 1982, 38, 143–151.
22. Food and Drug Administration (FDA), Guidance for the use of Bayesian statistics in medical device clinical trials. 2010. Available at http://www.fda.gov/

downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071121.pdf. Accessed August 6, 2010.

23. Food and Drug Administration (FDA), Adaptive design clinical trials for drugs and biologics (draft FDA guidance). 2010. Available at http://www.fda.gov/downloads/Drugs/guidancecomplianceregulatoryinformation/guidances/ucm201790.pdf. Accessed August 6, 2010.
24. Gajewski BJ, Mayo MS. Bayesian sample size calculations in phase II clinical trials using a mixture of informative priors. Stat Med 2006; 25: 2554–66.
25. Gallo P, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, Pinheiro J. Aptive designs in clinical drug development - an Executive Summary of the PhRMA Working Group. J Biopharm Stat. 2006 May;16(3):275-83
26. Gibbs W. J, Elementary Principles in Statistical Mechanics (in particular, Theorem III, p. 131). 1902; Constable, London. Reprinted by Dover, New York, 1960
27. Gould AL. How practical are adaptive designs likely to be for confirmatory trials? Biom J. 2006 Aug;48(4):644-9
28. Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. Bayesian data analysis. 2nd Ed. 2004; London: Chapman and Hall.
29. Gould AL, Shih WJ. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. Communications in Statistics – Theory and Methods 1992; 21(10): 2833-2853
30. Green SJ, Dahlberg S. Planned versus attained design in phase II clinical trials. Stat Med 1992; 11(7):853-62
31. Grieve AP. Predictive probability in clinical trials. Biometrics 1991; 47: 323–30.

32. Hall WJ, Yakir B. Inference about a secondary process following a sequential trial. Biometrika, 2003; 90: 597 - 611
33. Heitjan DF. Bayesian interim analysis of phase II cancer clinical trials. Stat Med 1997; 16: 1791–802.
34. Herson J. Predictive probability early termination plans for phase II clinical trials. Biometrics 1979; 35:775–83.
35. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med. 1990; 9: 811-818
36. Holm S. A simple sequentially rejective multiple test procedure, Scandinavian Journal of Statistics, 1979; 6:65-70
37. Hung HM, O'Neill RT, Bauer P, Köhne K. The behavior of the P-value when the alternative hypothesis is true. Biometrics. 1997 Mar;53(1):11-22.
38. Hung HM, O'Neill R, Wang SJ, Lawrence J. A regulatory view on adaptive/flexible clinical trial design. Biometrical J. 2006;48 (4):565-573
39. Hung HM, Wang SJ, O'Neill R. Statistical considerations for testing multiple endpoints in group sequential or adaptive clinical trials. J Biopharm Stat. 2007;17(6):1201-10
40. Hsu, J. C. Multiple Comparisons: Theory and Methods, London: Chapman & Hall. 1996
41. ICH E-9 Expert Working Group. Statistical principles for clinical trials (ICH Harmonized Tripartite Guideline E-9). Stat Med 1999; 18:1905-1942
42. Inoue LY, Thall PF, Berry DA. Seamlessly expanding a randomized phase II trial to phase III. Biometrics. 2002 Dec;58(4):823-31
43. Jennison C, Turnbull B. W. Mid-course sample size modification in clinical trials based on observed treatment effect. Stat Med. 2003; 22(6): 971–993.

44. Jennison C, Turnbull B. W. Confirmatory Seamless Phase II/III Clinical Trials with Hypotheses Selection at Interim: Opportunities and Limitations. Biometrical Journal 2006; 48: 650-655
45. Jennison C, Turnbull BW. Adaptive seamless designs: selection and prospective testing of hypotheses. J Biopharm Stat. 2007;17(6):1135-61
46. Johns D, Andersen JS. Use of predictive probabilities in phase II and phase III clinical trials. J Biopharm Stat. 1999 Mar;9(1):67-79
47. Johnson VE, Cook JD. Bayesian design of single-arm phase II clinical trials with continuous monitoring. Clin Trials. 2009;6(3):217-26
48. Johnson V, Rossell D. Non-local prior densities for default Bayesian hypothesis tests. Available at http://www.bepress.com/mdandersonbiostat/paper42/. Accessed August 1, 2010.
49. Johnson V, Rossell D. On the use of non-local prior densities in Bayesian hypothesis tests. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2010; 72(2): 143-170
50. Jung SH, Lee T, Kim KM, George SL. Admissible two-stage designs for phase II cancer clinical trials. Stat Med 2004; 23: 561–9.
51. Kass, R.E. and Raftery, A.E. Bayes Factors. Journal of the American Statistical Association. 1995; 90: 773–795.
52. Kimani PK, Stallard N, Hutton JL. Dose selection in seamless phase II/III clinical trials based on efficacy and safety. Stat Med. 2009 Mar 15;28(6):917-36
53. Kelly, P.J., Stallard, N., Todd, S. An adaptive group sequential design for Phase II/III clinical trials that select a single treatment from several. J. Biopharm. Statist. 2005; 15:641-658.

54. Kieser M, Friede T. Simple procedures for blinded sample size adjustment that do not affect the type I error rate. Statistics in Medicine 2003; 22:3571–3581.
55. Koenig F, Brannath W, Bretz F, Posch M. Adaptive Dunnett tests for treatment selection. Stat Med. 2008 May 10;27(10):1612-25
56. Krishnaiah, P. R. and Armitage, J. V. "Tables for Multivariate "t Distribution, Sankhya, Series B 1966; 31 - 56.
57. Lee JJ, Liu DD, A predictive probability design for phase II cancer, Clin Trials. 2008;5(2):93-106
58. Lee SJ, Zelen M. Clinical trials and sample size considerations: another perspective (with discussion). Statistical Science 2000; 15: 95–110.
59. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. Biometrics. 1999 Dec;55(4):1286-90
60. Li G, Shih WJ, Wang Y. Two-stage adaptive design for clinical trials with survival data. J Biopharm Stat. 2005;15(4):707-18.
61. Li G, Shih WJ, Xie T, Lu J. A sample size adjustment procedure for clinical trials based on conditional power. Biostatistics. 2002 Jun;3(2):277-87
62. Lin Y, Shih WJ. Adaptive two stage designs for single-arm phase IIA cancer clinical trials, Biometrics. 2004;60(2):482-90
63. Liu Q, Pledger G. Phase 2 and 3 combination designs to accelerate drug development. Journal of the American Statistical Association. 2005; 100: 493-502
64. Maca J, Bhattacharya S, Dragalin V, Gallo P, Krams M. Adaptive seamless phase II/III designs—background, operational aspects, and examples. Drug Information Journal 2006; 40:463–473.
65. Marcus R, Peritz E, Gabriel K. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 1976; 63(3):655-660

66. Mayo MS, Gajewski BJ. Bayesian sample size calculations in phase II clinical trials using informative conjugate priors. Cont Clin Trials 2004; 25: 157–67.
67. Miller, R. G. J. Simultaneous Statistical Inference, New York: Springer-Verlag. 1981
68. Müller, H. H., Schäfer, H. Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. Biometrics 2001; 57:886–891.
69. Poland B, Hodge FL, Khan A, Clemen RT, Wagner JA, Dykstra K, and Krishna R. The clinical utility index as a practical multiattribute approach to drug development decisions. Clinical Pharmacology & Therapeutics 2009; 86(1):105-108
70. Posch M, Koenig F, Branson M, Brannath W, Dunger-Baldauf C, Bauer P. Testing and estimation in flexible group sequential designs with adaptive treatment selection. Stat Med. 2005; 24:3697–3714
71. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. Biometrics 1995;51:1315–1324
72. Proschan MA, Liu Q, Hunsberger S. Practical midcourse sample size modification in clinical trials. Control Clin Trials. 2003 Feb;24(1):4-15
73. Quan H, Luo X, Capizzi T. Multiplicity adjustment for multiple endpoints in clinical trials with multiple doses of an active treatment. Stat Med. 2005; 24:2151-2170.
74. Sambucini V. A Bayesian predictive two-stage design for phase II clinical trials, Stat Med. 2008;27(8):1199-224
75. Sampson AR, Sill MW. Drop-the-losers design: normal case. Biom J. 2005 Jun;47(3):257-68; discussion 269-81

76. Sankoh A, D’Agostino R, Huque M. Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. Stat Med 2003; 22: 3133-3150
77. Schmidli H, Bretz F, Racine A, Maurer W. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: applications and practical considerations. Biometrical Journal 2006**;** 48:635–643
78. Schultz JR, Nichol FR, Elfring GL, Weed SD: Multi-stage procedures for drug screening. Biometrics 1973; 29:293-300
79. Shao Y, Mukhi V, Goldberg JD. A hybrid Bayesian-frequentist approach to evaluate clinical trial designs for tests of superiority and non-inferiority. Stat Med. 2008 Feb 20;27(4):504-19
80. Shih WJ. Sample size re-estimation - journey for a decade. Stat Med. 2001;20(4):515-8
81. Shih WJ. Group sequential, sample size re-estimation and two-stage adaptive designs in clinical trials: a comparison. Stat Med. 2006 Mar 30;25(6):933-41
82. Shih WJ. Plan to be flexible: a commentary on adaptive designs. Biom J. 2006 Aug;48(4):656-9; discussion 660-2
83. Shih WJ, Gould AL. Re-evaluating design specifications of longitudinal clinical trials without unblinding when the key response is rate of change. Stat. Med. 1995; 14: 2239-2248
84. Shih WJ, Long J. Blinded sample size re-estimation with un-equal variance and center effect in clinical trials. Communication in Statistics – Theory and Methods 1998; 27:395-408

85. Shih WJ, Quan H, Li G. Two-stage adaptive strategy for superiority and non-inferiority hypotheses in active controlled clinical trials. Stat Med. 2004;23(18):2781-98
86. Shih WJ, Zhao P. Design for sample size re-estimation with interim data for double-blind clinical trials with binary outcomes. Stat Med. 1997; 16(17): 1913-1923
87. Shun Z, Lan KK, Soo Y. Interim treatment selection using the normal approximation approach in clinical trials. Stat Med. 2008 Feb 20;27(4):597-618
88. Shuster, J, Optimal two-stage designs for single arm phase II cancer trials, J Biopharm Stat. 2002 Feb;12(1):39-51
89. Simon R. Optimal two-stage designs for phase II clinical trials, Control Clin Trials. 1989 Mar;10(1):1-10
90. Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. Bayesian approaches to clinical trials and health-care evaluation. 2000; New York: Wiley
91. Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. Stat Med. 2003; 22: 689-703
92. Sylvester RJ. A Bayesian approach to the design of phase II clinical trials. Biometrics 1998; 44: 823–36.
93. Tan SB, Machin D. Bayesian two-stage designs for phase II clinical trials. Stat Med 2002; 21: 1991-2012.
94. Thall PF, Simon R. Practical Bayesian guidelines for phase IIB clinical trials. Biometrics 1994; 50: 337–49.
95. Thall PF, Wooten LH, Logothetis CJ, Millikan RE, Tannir NM Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. Stat Med. 2007;26(26):4687-702

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

97. Todd S, Stallard N. A new clinical trial design combining phase 2 and 3: Sequential designs with treatment selection and a change of endpoint. Drug information Journal. 2005; 39: 109-118

98. Tsiatis, A. A, Mehta, C. On the inefficiency of the adaptive design for monitoring clinical trials. Biometrika 2003; 90(2):367–378.

99. Wang SJ, Hung HMJ, O'Neill RT. Adapting the sample size planning of a phase III trial based on phase II data. Pharmaceutical Statistics 2007; 5(2):81–97

100. Wang YG, Leung DHY, Li M, Tan SB. Bayesian designs with frequentist and Bayesian error rate considerations. Statistical Methods in Medical Research 2005; 14: 445–56.

101. Wang SJ, Bretz F. From Adaptive Design to Modern Protocol Design for Drug Development: Part I. Editorial and Summary of Adaptive Designs Session at the Third FDA/DIA Statistics Forum. Drug Information Journal 2010; 44(03)

102. Whitehead J, Valdés-Márquez E, Johnson P, Graham G. Bayesian sample size for exploratory clinical trials incorporating historical data. Stat Med. 2008 Jun 15;27(13):2307-27.

103. Wu Y, Shih WJ, Moore DF. Elicitation of a beta prior for Bayesian inference in clinical trials. Biom J. 2008;50(2):212-23

104. Wu Y, Shih WJ. Approaches to handling data when a phase II trial deviates from the pre-specified Simon's two-stage design. Stat Med. 2008;27(29): 6190-6208

Vita

Gaohong Dong

2010	Ph.D. in Biostatistics, School of Public Health, University of Medicine and Dentistry of New Jersey
1999 – 2001	M.S. in Statistics, University of Cincinnati
1989 – 1992	B.E. in Civil engineering, Lanzhou Railway University, China