The Report committee for Sukyung Park
Certifies that this is the approved version of the following report:

A Comparison of Adaptive Designs in Clinical Trials: when Multiple Treatments are tested in Multiple Stages

APPROVED BY

SUPERVISING COMMITTEE:

Peter Müller, Supervisor

James G. Scott

A Comparison of Adaptive Designs in Clinical Trials: when Multiple Treatments are tested in Multiple Stages

by

Sukyung Park, B.S.

REPORT

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN STATISTICS

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2014

Dedicated to my husband Dohyung.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor Dr. Peter Müller for his expert guidance and encouragement throughout the course of this report. The completion of this report could have not been accomplished without his suggestions and constructive advices. I would also like to thank Dr. James Scott, the reader of this report, for his patience and valuable comments. Additional thanks go to Dr. Mary Parker and Dr. Michael Daniels for teaching and helping me establish solid statistical knowledge to base this report.

My grateful thanks are also extended to my parents and friends for their support, time and help. Finally, I sincerely thank my husband Dohyung Park. Without his endless encouragement and support, I could not complete this program and the report.

A Comparison of Adaptive Designs in Clinical Trials: when Multiple Treatments are tested in Multiple Stages

Sukyung Park, M.S.Stat
The University of Texas at Austin, 2014

Supervisor: Peter Müller

In recent times, there has been an increasing interest in adaptive designs for clinical trials. As opposed to conventional designs, adaptive designs allow flexible design adaptation in the middle of a trial based on accumulated data. Although various models have been developed using both frequentist and Bayesian perspectives, relative statistical performances of adaptive designs are somewhat controversial and little is known about those of Bayesian adaptive designs. Most comparison studies also focused on single experimental treatment rather than multiple experimental treatments. In this report, both frequentist and Baysian adaptive designs were compared in terms of statistical power by a simulation study, assuming the situation when multiple experimental treatments are tested in multiple stages. The designs included in the current study are group sequential design (frequentist), adaptive design based on combination tests (frequentist), and Bayesian adaptive design (Bayesian). Based upon the results under multiple scenarios, the Bayesian adaptive design

showed the highest power, and the design based on combination tests performed better than group sequential designs when proper interim adaptation could be conducted to increase power.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Adaptive designs are statistical methodologies for clinical trials, increasingly recognized as useful tools for investigating experimental drugs, procedures, and medical devices. They allow that accumulated data can be analyzed in the middle of a trial so that appropriate interim adaptation is made during its course. Interim analyses are often recommended in clinical trials because the total process of trials can be adjusted to improve the probability of successful completion. This provides ethical and economical benefits [1]. For example, it would be desirable to move the process rapidly if an experimental drug shows overwhelming efficacy in the middle of a trial. With adaptive designs, trials can be stopped early for superiority, a treatment can be dropped for inferiority, and a samples size can be modified to increase statistical power. Conventional designs such as fixed-sample designs usually perform one final analysis without interim analyses or adaptation. Since they strictly conduct a trial only as planned, designs cannot be modified to improve efficiency or response to unexpected situations [2].

Adaptive designs incorporate multiple testing to allow interim analyses. This often results in a significant inflation of the overall type I error rate, which

can be a serious problem with respect to statistical validity. Thus, adaptive designs have been developed to provide more flexibility without compromising the overall type I error rate. One approach in adaptive designs is a group sequential design which stops a treatment at any interim analysis for either futility or superiority [3, 4, 5, 6]. Group sequential models were initially designed to allow early stopping for treatment superiority, but can be extended to incorporate treatment selection in more general settings. Another popular class of methods are adaptive designs based on combination tests, which were developed to provide additional flexibility compared to the group sequential designs [7, 8, 9, 10]. Such designs usually combine evidence from stage-wise data using a specific combination function of p-values. By introducing p-values as test statistics, a wide range of design adaptation is possible at interim analyses. A Bayesian adaptive approach also can be considered as an alternative to these designs [11, 12]. The Bayesian approach to inference is directly based on the data without concerns about type I error rate. Most design restrictions in frequentist designs, such as group sequential and combination test approaches, are related to the issue of controlling type I error rate. Therefore, more flexible designs can be constructed using Bayesian adaptive models to address various aspects of interim adaptation.

Although extensive research has been carried out on adaptive designs, debate continues about statistical performance of adaptive designs. Tsiatis and Mehta (2003) [13] first discussed systematically how group sequential designs can outperform adaptive designs based on insufficient statistics, i.e., p-values,

if more interim analyses are allowed for group sequential designs. Jennison and Turnbull (2003, 2006)[14, 15] showed by simulation that standard group sequential approaches can be more efficient than adaptive test based on combination test approaches. Kelly et al. (2005) [16] compared both designs in practical settings, indicating that the difference in power is not identified as long as asymptotic assumptions for test statistics are satisfied. For trials with multiple treatments, however, Friede and Stallard (2008) [35] reported that adaptive designs using combination tests can be better than group sequential designs in spite of the use of insufficient statistics, when it is believed a priori that all treatments are effective. Most comparisons were based on frequentist adaptive designs, and little is known about relative performances of Bayesian adaptive models in terms of statistical power.

The purpose of this report is to compare adaptive designs by simulation including both frequentist and Bayesian methods. The specific methods considered in this study are as follows: Pocock's [3] and O'Brien and Fleming's [4] for the group sequential design, Bretz et al. (2009) [1] model for the combination test approach, and a hierarchical Bayesian model based on Thall et al. (2003) [12] for the Bayesian adaptive design. The report is also aimed to investigate statistical performances of adaptive designs when multiple treatments are tested in multiple stages. Clinical trials become more complex than before and often involve two or more experimental treatments. However, there have been only a few discussions about testing multiple treatments using adaptive designs. The present report has been organized in the following way. The

chapter two first gives a brief description of frequentist adaptive methods in clinical trials, and the chapter three introduces a model for the Bayesian adaptive design. Both chapter two and three also include implementations when multiple treatments are tested in multiple stages. The chapter four summarizes scenarios of the comparison study and reports major findings. The last section concludes by discussing limitation and extensions.

Chapter 2

Frequentist Adaptive Designs

This chapter introduces two frequentist adaptive designs in clinical trials, which are used for the comparison study in this report. The first section illustrates a group sequential design, and the second section describes an adaptive design based on combination tests. For multiplicity issue of multiple hypotheses, the last section discusses the closure principle and its applications.

2.1 Group Sequential Design

Classical group sequential designs monitor superiority of an experimental treatment at each interim analysis. In a clinical trial, inferences are often made by statistical hypothesis tests, and group sequential designs examine whether or not there is evidence against a null hypothesis in accumulating data [3, 4]. In contrast to the fixed-sample designs allowing one final analysis from fully pre-planned process, group sequential designs can be considered more flexible in that a trial can be stopped at any interim analysis in response to what interim data demonstrate. However, repeating hypothesis test can increase the probability of falsely rejecting the true null hypothesis, which is also known as the type I error rate. Group sequential designs usually use

different critical regions to control overall type I error rate associated with multiple interim analyses. This section briefly describes two classical group sequential designs: Pocock's design [3] and O'Brien and Fleming's design [4]. More details about group sequential designs can be found in Yin (2012)[2] and Jennison and Turnbull (2000)[18].

2.1.1 Pocock's Design

The Pocock's Design(1977)[3] is a group sequential design which assumes equal stopping boundaries across all stages. Although the design was proposed originally for a two-sided test, it can be easily extended to a one-sided test. Suppose that we are interested in testing superiority between two treatments when a trial is scheduled up to K stages in advance. We assume that treatment 1 is the experimental treatment, while treatment 0 is the standard treatment. At each stage, outcomes are independent normal random variables such that:

$$Y_{1ij} \sim N(\mu_1, \sigma^2), \quad Y_{0ij} \sim N(\mu_0, \sigma^2), \ i = 1, 2, .., n, \ j = 1, 2, .., K \tag{2.1}$$

where Y_{1ij} and Y_{0ij} denote i-th observation of treatment 1 and 0 at stage j, respectively. Treatment 1 and 0 have the true mean μ_1 and μ_0. For simplicity, we assume that the variance σ^2 and the sample size n are known and equivalent across all treatments at all stages. The only unknown quantities are the true treatment means, while others should be determined in advance before starting a trial.

According to the notation defined above, the one-sided test for the treatment difference refers to testing the hypotheses

$$H_0 : \mu_1 - \mu_0 \leq 0, \quad H_1 : \mu_1 - \mu_0 > 0 \tag{2.2}$$

Now let $N_k = nk$ denote the cumulative sample size up to stage k for each treatment. Conditioned on the null hypothesis, the standardized test statistic Z_k for the k-th interim analysis based on the all previous observations is given by:

$$Z_k = \frac{1}{\sqrt{2N_k\sigma^2}} \sum_{j=1}^{k} \sum_{i=1}^{n} (Y_{1ij} - Y_{0ij}) = \frac{\bar{Y_1} - \bar{Y_0}}{\sqrt{2\sigma^2/N_k}} \tag{2.3}$$

where Z_k is tested at each stage against some critical value c_k. In the case of rejecting the null hypothesis at interim stages ($Z_k \geq c_k$), the trial is early stopped for superiority. Otherwise ($Z_k < c_k$), the trial proceeds to stage k+1 until the stage reaches the maximum number of stages, K.

c_k is assumed to be constant across all stages as Pocock(1977)[3] proposed, but should be larger than the single stage significant level α due to the possible inflation of the overall type I error rate. For one-sided testing, Pocock's constant $c_k = c_{po},\ k = 1, 2, .., K$ can be computed numerically according to the following definition:

$$\alpha = P(Z_k \geq c_{po} \text{ at any k in a sequential order } | H_0) \tag{2.4}$$

This can be also obtained by solving K equations with respect to the stage wise α level, α_k [2]. For example, when maximum two stages are scheduled,

c_{po} can be computed numerically by considering the following two equations.

$$\alpha_1 = P(Z_1 \geq c_{po}|H_0), \tag{2.5}$$

$$\alpha_2 = \alpha = P(Z_1 < c_{po}, Z_2 \geq c_{po}|H_0) + \alpha_1 \tag{2.6}$$

2.1.2 O'Brien and Fleming's Design

O'Brien and Fleming's Design has the exactly same procedure with Pocock's, except the critical values are not constant across K stages [4]. They proposed the critical value at stage k, c_k, such that

$$c_k = c_{of}\sqrt{K/k} \tag{2.7}$$

where c_{of} is a constant and K is the maximum number of stages. This specification is based on the idea that the rejection criterion should be more stringent at the earlier interim analyses due to the limited number of observations. As the trial proceeds, this criterion is relaxed by increasing $\sqrt{K/k}$ so that the null hypothesis is more likely rejected than earlier testing stages. For one-sided testing, O'Brien and Fleming's constant c_{of} can be computed numerically according to the following definition.

$$\alpha = P(Z_k \geq c_{of}\sqrt{K/k} \text{ at any k in a sequential order } |H_0) \tag{2.8}$$

Note that this can be also obtained by solving K equations iteratively with respect to the individual significance level of each stage k.

In general, it is expected that O'Brien and Fleming's design has more statistical power than Pocock's design because critical values decrease as trial

proceeds. In most cases, the values at the final stage in O'Brien and Fleming's are smaller than those in Pocock's, whereas they are greater than those in Pocock's at earlier stages. As such, it is more likely for the null hypothesis to be rejected in O'Brien and Fleming's design, indicating that the overall statistical power of O'Brien and Fleming's is higher than the overall statistical power of Pocock's.

2.2 Adaptive Design based on Combination Tests

Although group sequential designs are more flexible than fixed-sample designs, there are still some aspects required to specify in advance, e.g., sample sizes at each stage. The adaptive designs based on combination tests have been developed to overcome this limitation and provide more flexibility at interim stages [7, 8, 9, 10]. The key idea of the combination tests is combining stage-wise p-values according to the conditional invariance principle [19, 10, 20]. By introducing p-values as test statistics, many design features such as sample sizes can be modified at interim stages without inflation of the overall type I error rate [22]. Thus, although p-values are not sufficient statistics, adaptive designs can increase statistical power at interim stages if appropriate design modifications are applied. This section describes the conditional invariance principle and the adaptive design of Bretz et al. (2009) [1] based on the combination test approach. More detailed information can be found in Bretz et al. (2009) and the references therein.

2.2.1 Conditional Invariance Principle

The conditional invariance principle controls the type I error rate when interim results or modifications are unknown in advance. According to the conditional invariance principle, an α level test can be constructed without knowing adaptation rules, if distributions of test statistics are known and conditionally invariant with respect to the interim adaptation [20]. For example, combination test approaches usually exploit stage-wise p-values, which are always uniformly distributed under the pre-specified null hypothesis. Since this distributional aspect does not change after interim adaptation, it can be said that distributions of p-values are conditionally invariant with respect to the interim data and the mid-term design adaptation. Knowing that distributions of test statistics are (conditionally) independent of interim analyses, we can now construct an α level test in terms of p-values. The more detailed example will be following, but we also refer to Liu et al. (2002) [21] for more rigorous discussion about the conditional invariance principle.

2.2.2 Combination Test Approach

A combination test approach is a statistical method which combines stage-wise p-values through a pre-specified combination function. It allows early stopping for futility or superiority at interim stages. For simplicity, here we illustrate an adaptive design for a single null hypothesis, which is tested in two stages [1].

Suppose that there are two treatments, where one is experimental and

the other is standard. Similar to group sequential designs, we are interested in testing one-side null hypothesis H_0 for the treatment difference. Let p_1 and p_2 be the p-values from stage 1 and 2 data, respectively. At the interim stage, a combination test approach examines whether a trial can be early stopped or proceed to the second stage by comparing p-values to α_0 and α_1. α_0 and α_1 are early stopping boundaries such that $0 < \alpha_1 \leq \alpha_0 \leq 1$. If $p_1 > \alpha_0$, then H_0 is early accepted and the trial stops for futility of the experimental treatment. If $p_1 \leq \alpha_1$, then H_0 is early rejected and the trial stops for superiority. Otherwise, $\alpha_1 < p_1 \leq \alpha_0$ and the trial continues to the second stage to combine stage-wise p-values through the pre-specified combination function $C(p_1, p_2)$. This combination function defines the critical region as $C(p_1, p_2) \leq c$, where the corresponding critical value c is determined by solving the following equation:

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{[C(x,y)\leq c]} dydx = \alpha \tag{2.9}$$

which maintains the overall type I error rate at α regardless of the design adaptation after interim analyses.

Although p-values can be combined in many different ways, there are two prominent examples of combination functions in general. One is Fisher's product criterion $C(p_1, p_2) = p_1 p_2$ [10], and the other is the weighted inverse normal combination function [9]. The weighted inverse normal combination function has the following form:

$$C(p_1, p_2) = 1 - \Phi[w_1 \Phi^{-1}(1 - p_1) + w_2 \Phi^{-1}(1 - p_2)], \tag{2.10}$$

where w_1 and w_2 denote pre-specified weights such that $w_1^2 + w_2^2 = 1$, and Φ denotes the cumulative distribution function of the standard normal distribution. The weights of this function are often defined as $w_1^2 = n_1/(n_1+n_2), w_2^2 = n_2/(n_1+n_2)$, where n_1 and n_2 indicate the pre-planned sample size for the stage1 and 2, respectively. The weighted inverse normal combination function with this weight definition is equivalent to a classical group sequential when a single null hypothesis is considered without any interim adaptation [24].

The choice of α_0 for the futility stopping is sometimes important and there can be two different stopping rules for futility: Binding versus non-binding rules. Binding futility rules choose $\alpha_0 < 1$, allowing a futility stop at the interim analysis whenever $p_1 > \alpha_0$. Alternatively by fixing $\alpha_0 = 1$, non-binding rules provide more flexibility in interim stopping criteria, but they often result in less statistical power than binding rules.

2.3 Multiplicity and the Closure Principle

Clinical designs in practice involve the test of multiple hypotheses. When several experimental treatments need to be compared to one single standard treatment, multiple null hypotheses are constructed and tested simultaneously. Multiplicity, also known as the multiple testing problem, occurs in this situation because simply performing an α-level test for each hypothesis leads to inflation of the probability of rejecting at least one true null hypothesis, i.e., the family-wise error rate. For instance, suppose that we test two

hypotheses simultaneously with $\alpha = 0.05$. Then the probability of at least one null hypothesis being rejected is $1 - (0.95 \times 0.95) = 0.0975$, a much larger value than that of one single null hypothesis. However, we want this rate to be within α when multiple treatments are tested in multiple stages.

The closure principle is a general methodology to handle this issue, which strongly controls the family-wise error rate at a pre-specified α level [25]. It considers all possible intersection hypotheses and reject an elementary (original) hypothesis only when all relevant intersections are rejected at α level. More precisely, the test procedure is conducted as follows:

1. Let K be the number of hypotheses, and H_k denotes the kth hypothesis for every $k = 1, \ldots, K$.

2. Construct all intersection hypotheses. That is, define $H_I = \cap_{k \in I} H_k$ for every subset $I \subset \{1, 2, \ldots, K\}$.

3. Perform an α-test for each intersection hypothesis.

4. Reject H_k if all the intersection hypotheses H_I where $k \in I$ are rejected.

2.3.1 An Example for Group Sequential Design

Group sequential designs can be extended to incorporate multiple treatments in many ways (e.g., [26, 27]), and here is one example suggested for the closure principle. Note that the following design is summarized as a step-by-step algorithm in the section 4.1.2. Suppose we compare two experimental

treatments to one single standard in stage K using Pocock's group sequential design. The experimental treatments are named treatment 1 and 2, while the standard is treatment 0. Based on the case of the previous section, there are two elementary null hypotheses, $H_{0j} = \mu_j - \mu_0,\ j = 1, 2$ and three intersection hypotheses, $\{H_{01}, H_{02}, H_{01} \bigcap H_{02}\}$. At each interim stage, all intersection hypotheses are individually tested based on the Pocock's critical value. Note that any proper multiplicity adjustment such as Bonferroni correction [28] can be applied for intersection hypotheses involving more than one elementary hypothesis. When all intersection hypotheses that includes H_{0j} are rejected at this interim analysis, we reject the elementary null hypothesis H_{0j} and stop the trial early for the superiority of the treatment j. Otherwise, the trial continues to the next stage unless it reaches the maximum number of stage K. The multiple treatment case for O'Brien and Fleming's design follows exactly the same procedure except using different critical values.

2.3.2 An Example for Combination Test Approach

The closure principle also provides an α level test for adaptive designs for combination tests with multiple null hypotheses [1, 23]. The following procedure also can be found as a step-by-step algorithm in the section 4.1.3. Suppose two experimental treatments are compared to one single standard in two stages. There are two elementary null hypotheses, $H_{0j} = \mu_j - \mu_0,\ j = 1, 2$ and three intersection hypotheses, $\{H_{01}, H_{02}, H_{01} \bigcap H_{02}\}$. At the interim stage, all intersection hypotheses are tested individually using p-values based on the

data from the first stage. That is, all intersection hypothesis are examined whether their p-values are greater than stopping boundary for futility α_0 or less than the boundary for superiority α_1. Any proper multiplicity adjustment can be applied for p-values of intersection hypotheses involving more than one elementary hypothesis. If all intersection hypotheses that includes H_{0j} demonstrates smaller p-values than α_1 at this interim analysis, we early reject the elementary null hypothesis H_{0j} and stop the trial for superiority of the treatment j. If any intersection that includes H_{0j} has greater p-value than α_0, we early accept the elementary null hypothesis H_{0j} and drop the treatment j for futility. Otherwise, the trial continues to stage 2 with treatment j and rejects H_{0j} if combined stage-wise p-values for all relevant intersections are smaller than the critical value c. At stage 2, intersections involving both dropped and remaining treatments are replaced by intersections of remaining treatments. For example, we assume that experimental treatment 1 is dropped at interim analysis. Then, the 2nd stage p-value for the intersection hypothesis involving H_{01} and H_{02} is replaced by the 2nd stage p-value for H_{02}. H_{02} will be rejected at stage 2 if $C(p_{[1,H_{02}]}, p_{[2,H_{02}]}) < c$ and $C(p_{[1,H_{01} \bigcap H_{02}]}, p_{[2,H_{02}]}) < c$, where $p_{[i,H_{02}]}$ is the p-value at stage i for H_{02} and $p_{[i,H_{01} \bigcap H_{02}]}$ is the p-value at stage i for $H_{01} \bigcap H_{02}$.

Chapter 3

Bayesian Adaptive Design

Traditional frequentist designs such as group sequential designs and combination test approaches have been developed to provide more flexibility in clinical trials by introducing proper interim data analyses. In order to control the overall type I error rate, however, these designs still depend in part on pre-specified design components, e.g., the maximum number of stages, and statistical inferences are only meaningful when these critical design features are maintained as planned. In this regard, Bayesian methods can offer more flexibility, as they can make various posterior inferences based on the data without pre-determined design constraints or asymptotic assumptions of test statistics [11, 12]. Bayesian methods directly derive a posterior distribution of parameters given the data and thus do not require calculation of type I error in advance for mid-trial adaptation. While conceptually a Bayesian design does not require the evaluation of type I error, in practice, any real implementation would include an extensive discussion of type I error under a variety of realistic scenarios [11]. This chapter describes the concept of Bayesian methods and an Bayesian adaptive design for the comparison study, which is summarized as a step-by-step algorithm in the section 4.1.4. More extensive reviews about Bayesian adaptive designs can be found in Berry et al. (2010) [11].

3.1 A Bayesian Model for Adaptive Design

Bayesian models often incorporate additional information through prior distributions of parameters and produce a posterior distribution so that inferences are made based on the distribution of parameters given the data [29]. Priors can be constructed from various sources such as insights from experts and results from previous analyses. Priors also can be non-informative if there is no relevant prior knowledge. According to the Bayes theorem, the joint posterior distribution is obtained from the following equation:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{3.1}$$

where y is observed data and θ is a parameter vector of interest. $p(\theta)$ is an assumed prior distribution for the parameters.

3.1.1 A Bayesian Hierarchical Model for Multiplicity

One popular Bayesian approach in clinical trial designs is based on a hierarchical Bayesian model, which borrows strength from the related subpopulations. The related subpopulations can be different treatments on the same disease or the same treatment on the related disease types. In a hierarchical Bayesian approach, information from one treatment is shared with other treatments through presumed underlying structure, and the result of one treatment can provide information about the effect of the other related treatments [11, 12]. For the Bayesian adaptive model in this report, a hierarchical Bayesian model was constructed based on Thall et al. (2003) [12]. Thall et

al. (2003) [12] proposed a clinical trial design for a disease with multiple subtypes, allowing treatment effects to differ but correlated through a hierarchical structure. By using a Bayesian hierarchical model, multiple treatments also can be tested in multiple stages. For a rigorous discussion of a Bayesian view of multiple testing problems, see, for example, Scott and Berger (2010) [33].

Suppose that we are interested in testing multiple one-sided null hypotheses in K stages, i.e., comparing multiple experimental treatments to one standard treatment. We assume that there are J+1 treatments for the same disease, where one is the standard treatment (treatment 0) and the others are different experimental treatments (treatment 1,..,J). Similar to other designs in this report, outcomes are considered as normal random variables with treatment-specific means and the known common variance. Since they are independent conditioned on the unknown treatment means, i-th observation of j-th treatment, y_{ij}, has the following distribution:

$$y_{ij}|\mu_j \sim N(\mu_j, \sigma^2),\ i = 1, ..., n_{jk},\ j = 0, ..., J \tag{3.2}$$

where μ_j denotes the mean of j-th treatment, σ^2 denotes the known common variance, and the n_{jk} denotes the number of patients enrolled in j-th treatment up to stage k=1,..,K.

μ_j's are different across treatments but allowed to be correlated by presumed underlying structures. Therefore, this model assumes each μ_j follows the same normal prior distribution with mean γ and variance τ^2:

$$\mu_j \sim N(\gamma, \tau^2),\ j = 0, .., J \tag{3.3}$$

where γ is a normally distributed parameter with mean m and variance V, and τ^2 is a parameter whose inverse follows a Gamma distribution with the parameters a and b.

$$\mu \sim N(m, V) \tag{3.4}$$

$$1/\tau^2 \sim Ga(a, b) \tag{3.5}$$

These hyper parameters, m, V, a, b, are often determined in advance so that μ and τ^2 have non-informative vague priors under the absence of any relevant information. In this study, the values for m, V, a, b are determined as follows to reflect this lack of prior information.

$$m = 0 \tag{3.6}$$

$$V = 100 \tag{3.7}$$

$$a = 0.01 \tag{3.8}$$

$$b = 0.01 \tag{3.9}$$

3.1.2 Posterior Estimation

Bayesian models make inferences based on posterior distributions, which are usually estimated by Markov Chain Monte Carlo (MCMC). Markov Chain Monte Carlo is a popular technique to simulate posterior distributions when the model is too complicated to obtain direct estimation through traditional methods. In this study, Gibbs sampling can be effectively exploited to generate posterior samples of parameters [30]. Gibbs Sampling is one of MCMC

methods and needs only full conditional posterior distributions of parameters. Thus, it allows a convenient way to simulate the joint posterior distribution for the models with well-known full conditional posterior distributions. At stage k, the full conditional posterior distributions for the parameters in this study are calculated as follows:

$$p(\mu_j|\gamma,\tau^2,y) = N\left(\left(\frac{n_{jk}}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{\sum_{i=1}^{n_{jk}} y_{ij}}{\sigma^2}+\frac{\gamma}{\tau^2}\right),\ \left(\frac{n_{jk}}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\right) \tag{3.10}$$

$$p(\gamma|\theta,\tau^2,y) = N\left(\left(\frac{J+1}{\tau^2}+\frac{1}{V}\right)^{-1}\left(\frac{\sum_{j=0}^{J}\mu_j}{\tau^2}+\frac{m}{V}\right),\ \left(\frac{J+1}{\tau^2}+\frac{1}{V}\right)^{-1}\right) \tag{3.11}$$

$$p(1/\tau^2|\theta,\gamma,y) = Ga\left(\frac{J+1}{2}+a,\ \frac{\sum_{j=0}^{J}(\mu_j-\gamma)^2}{2}+b\right) \tag{3.12}$$

3.1.3 Decisions and Design

At each interim stage, experimental treatments are tested based on their estimated posterior distributions to determine whether they should be early stopped or proceed to the next stage. That is, i-th treatment is dropped for futility if the posterior probability $p(\mu_i-\mu_0>\mu_j-\mu_0|y)$ is smaller than α for the all other experimental treatment j:

$$p(\mu_i-\mu_j>0|y)<\alpha,\ \text{for all } j\neq i \tag{3.13}$$

where α is a fixed cut-off value such as 0.025 or 0.05.

If this posterior probability is not smaller than α at least for one experimental treatment, the treatment continues to the next stage after interim

adaptation such as sample size modification. When the trial reaches the final stage K, the null hypotheses with respect to on-going treatments are examined individually at the final analysis using Bayes Factor. Bayes Factor is a Bayesian approach for hypothesis testing, which summarizes evidence provided by data in favor of one hypothesis against the other hypothesis [31, 32]. Let the hypotheses for the i-th treatment effect be:

$$H_{0i} : \mu_i - \mu_0 \leq 0, \quad H_{1i} : \mu_i - \mu_0 > 0 \tag{3.14}$$

Then, Bayes Factor is the ratio of probabilities that we observe data given H_{1i} versus H_{0i} such that

$$\mathrm{BF}_{10} = \frac{p(y|H_{1i})}{p(y|H_{0i})} = \frac{p(H_{1i}|y)/p(H_{1i})}{p(H_{0i}|y)/p(H_{0i})} \tag{3.15}$$

where y is all previous data associated with the treatment i up to stage K. If BF_{10} is greater than 3, we conclude that there is substantial evidence in favor of H_1 against H_0 and reject the null hypothesis for the treatment i as the final decision. If BF_{10} is not greater than 3, we cannot reject the null hypothesis for the treatment i.

Chapter 4

Comparison Study for Adaptive Designs

This chapter illustrates a simulation study that compares statistical power of adaptive designs when multiple stages are conducted with multiple treatments. The study mainly targets adaptive designs introduced in the previous chapters, but one fixed-sample design is also included to see the difference with adaptive designs. The fixed-sample design in this study regards stages as separate trials and only use the final stage data for the final decision. The mid-stage data can be used to configure the next stage design. The detailed simulation scenarios for four designs are given in the first section, and the main findings from the study are discussed in the second section. The designs are presented and compared in the following order: 1) fixed-sample design; 2) group sequential design; 3) combination test approach; and 4) Bayesian adaptive design.

4.1 Simulation Scenarios

The simulation in this study includes seven scenarios for four designs, two for each frequentist design and one for the Bayesian design. The trial settings and assumptions are based on the simulation study of Bretz et al.

(2009) [1] which compares adaptive designs based on combination tests in various settings, but individual scenarios are extended to meet the unique feature of designs considered in this study. This section first illustrates the common trial settings for simulation, and then provides simulation algorithms (or scenarios) for each design.

We assume that a trial is conducted in two stages with three treatments, where two are the experimental treatment and the other is the standard (or control) treatment. The experimental treatments are named treatment 1 and 2, while the standard treatment is treatment 0. One interim analysis and one final analysis are allowed in this setting. As described in the previous chapters, observations are normally distributed with treatment-specific means $\mu_j,\ j = 0, 1, 2$ and the common known variance σ^2. Here we follow the Britez et al. (2009)'s assumption that $\sigma^2 = 6^2,\ \mu_0 = 0,\ \mu_1 = 2$, and μ_2 varying in the interval (0,3].

The trial is aimed to compare each experimental treatment to the standard treatment, and all tests in this study are based on one-sided hypotheses $H_{0j} : \mu_j - \mu_0 \leq 0,\ H_{1j} : \mu_j - \mu_0 > 0$. The family-wise error rate α for frequentist designs is controlled at significance level of 0.025. To address multiplicity issue, Bonferroni correction and the closure principle are used for frequentist designs as discussed before. The sample size per each treatment per stage is set to 72 so that a single test has a statistical power of 0.8 when the true mean difference is 2.

To compare performances of designs in terms of statistical power, suit-

able power concept needs to be implemented throughout the simulation. The general definition of statistical power is the probability of rejecting null hypothesis when the alternative is true, but this is not obvious for multiple treatment case with multiple null hypotheses. In this regard, this study computes power as the probability of rejecting at least one false null hypothesis [34]. Since this study only includes true alternative hypotheses (thus false null hypotheses), the simulation measures statistical power by computing the proportion of times at least one null hypothesis is rejected.

Overall, two adaptation rules are considered for the interim analysis in this study. One is to continue with all treatments to the second stage (adaptation I), and the other is to drop the inferior experiment treatment based on the first stage mean value (adaptation II), i.e. drop treatment j if $\hat{\mu_j} < \hat{\mu_i}$ (frequentist) or $p(\mu_j > \mu_i|y) < \alpha$ (Bayesian). In the case that one treatment is dropped at the interim analysis, the scheduled second stage sample size for the discontinued treatment is evenly assigned to the continued treatments. These interim adaptation rules can only be applied on frequentist designs when they can maintain the family-wise error rate at pre-specified level, whereas there is no restriction on the Bayesian design. For this reason, the above adaptation rules were not included in the simulation for group sequential designs. The fixed-sample design can exploit the rules because it derives the final decision only based on the final stage, and the combination test approach allows the interim adaptation by construction. Although the Bayesian design in this study does not have restrictions on the interim rules, only adaptation II was

included in the simulation due to the several practical reasons. For example, posterior distributions of true means given the first stage data would not be very different each other because of the relatively large variance of observations and the vague priors. This leads to similar results for both interim rules, so adaptation II was only implemented in this study (See appendix A).

Based on the trial settings above, seven simulation algorithms for the four designs were produced. The scenarios share the common assumptions and methodology to calculate statistical power, but follow the procedures corresponding to their own design features. For each scenario, 1000 trials were simulated for each 1000 different $\mu_2 \in (0, 3]$. The whole process of analyses was implemented using the statistical software R.

4.1.1 Algorithm for Fixed-Sample Design

In this study, a fixed-sample design is assumed to use only the second stage data to derive the final decision for treatment differences. The first stage data are used to determine which experimental treatment should be dropped in case that adaptation II is considered. In case of adaptation I, practical problems such as safety issues can be addressed based on the first stage data although there is no change in the design. For adaptation I case, the algorithm for the fixed-sample design is as follows:

1. Set $\mu_0 = 0, \mu_1 = 2, \mu_2 = 0.003, \sigma^2 = 6^2, \alpha = 0.025, n_1 = n_2 = 72$

2. Set success=0

3. Simulate observations for stage 1: Generate samples $Y_{0j} \sim N(\mu_0, \sigma^2), Y_{1j} \sim N(\mu_1, \sigma^2), Y_{2j} \sim N(\mu_2, \sigma^2),\ j = 1, ..., n_1,$ where Y_{ij} is j-th observation from treatment i.

4. Simulate observations for stage 2: Generate samples $Y_{0j} \sim N(\mu_0, \sigma^2), Y_{1j} \sim N(\mu_1, \sigma^2), Y_{2j} \sim N(\mu_2, \sigma^2),\ j = 1, ..., n_2$.

5. Test for the final decision: Based on the second-stage data, obtain the second stage p-values $p_{(2,1)},\ p_{(2,2)}$, and $p_{(2,12)}$ for three null hypotheses $H_{01} : \mu_1 - \mu_0 \leq 0,\ H_{02} : \mu_2 - \mu_0 \leq 0$, and $H_{01} \bigcap H_{02}$, respectively. The p-value for $H_{01} \bigcap H_{02}$ is calculated by $min\{1, 2min(p_{(2,1)}, p_{(2,2)})\}$ using Bonferroni correction. If either $[p_{(2,1)} < \alpha$ and $p_{(2,12)} < \alpha]$ or $[p_{(2,2)} < \alpha$ and $p_{(2,12)} < \alpha]$, increase success by one.

6. Calculate power: Repeat 2-5 steps 1000 times and calculate power as success/1000

7. Increase μ_2 by 0.003 and Repeat 2-6 steps by $\mu_2 = 3$

For adaptation II case, one experiment treatment is dropped based on the first stage mean value, and the sample size for the discontinued treatment is evenly reallocated to the continued treatments.

1. Set $\mu_0 = 0, \mu_1 = 2, \mu_2 = 0.003, \sigma^2 = 6^2, \alpha = 0.025, n_1 = n_2 = 72$

2. Set success=0

3. Simulate observations for stage 1: Generate samples $Y_{0j} \sim N(\mu_0, \sigma^2), Y_{1j} \sim N(\mu_1, \sigma^2), Y_{2j} \sim N(\mu_2, \sigma^2),\ j = 1, ..., n_1,$ where Y_{ij} is j-th observation from treatment i.

4. Drop the inferior treatment: Obtain the mean value of each experimental treatment and drop the treatment which has the smaller value.

5. Simulate observations for stage 2: Let $Y*_j$ be j-th observation from the continued experimental treatment and $\mu*$ be the corresponding true mean. Generate samples $Y_{0j} \sim N(\mu_0, \sigma^2), Y*_j \sim N(\mu*, \sigma^2),\ j = 1, ..., n*,$ where $n* = n_2 + n_1/2$.

6. Test for the final decision: Based on the second-stage data, obtain the second stage p-value $p_{(2,*)}$ for the null hypothesis $H_{0*} : \mu * - \mu_0 \leq 0$. If $p_{(2,*)} < \alpha$, increase success by one.

7. Calculate power: Repeat 2-6 steps 1000 times and calculate power as success/1000

8. Increase μ_2 by 0.003 and Repeat 2-7 steps by $\mu_2 = 3$

4.1.2 Algorithm for Group Sequential Design

The group sequential designs included in this study are two classic designs: 1) Pocock's design and 2) O'Brien and Fleming's Design. These group sequential designs do not allow interim adaptation other than early stopping for superiority. Thus, adaptation rule I and II for this study were

not considered here, and instead the algorithms follow their own rules, i.e., stopping an experimenal treatment at interim if test statistics are greater than pre-specified critical values. The critical values for two designs were obtained from the simulation based on three equal-mean treatments with the significance level 0.025. The calculated Pocock's critical value is 2.178, and O'Brien and Fleming's is 1.977. The following is an algorithm for the group sequential design when Pocock's design is used.

1. Set $\mu_0 = 0, \mu_1 = 2, \mu_2 = 0.003, \sigma^2 = 6^2, n_1 = n_2 = 72, c_{PO} = 2.178$

2. Set success=0

3. Simulate observations for stage 1: Generate samples $Y_{0j} \sim N(\mu_0, \sigma^2), Y_{1j} \sim N(\mu_1, \sigma^2), Y_{2j} \sim N(\mu_2, \sigma^2),\ j = 1, ..., n_1,$ where Y_{ij} is j-th observation from treatment i.

4. Test for early stopping at interim analysis: Based on the first stage data, obtain z-statistics $Z_1 = \frac{\bar{Y_1} - \bar{Y_0}}{\sqrt{2\sigma^2/n_1}}$ and $Z_2 = \frac{\bar{Y_2} - \bar{Y_0}}{\sqrt{2\sigma^2/n_1}}$ for the null hypotheses $H_{01} : \mu_1 - \mu_0 \leq 0$ and $H_{02} : \mu_2 - \mu_0 \leq 0$, respectively. The hypothesis test for $H_{01} \bigcap H_{02}$ can be conducted by $max(Z_1, Z_2) \geq c_{PO*}$ using Bonferroni correction, where c_{PO*} is a constant such that $p(Z > c_{PO*}) = p(Z > c_{PO})/2,\ Z \sim N(0, 1)$. If either $[Z_1 \geq c_{PO}$ and $max(Z_1, Z_2) \geq c_{PO*}]$ or $[Z_2 \geq c_{PO}$ and $max(Z_1, Z_2) \geq c_{PO*}]$, increase success by one and jump to step 7 (stop a trial). Otherwise, continue to step 5.

5. Simulate observations for stage 2: Generate samples $Y_{0j} \sim N(\mu_0, \sigma^2), Y_{1j} \sim N(\mu_1, \sigma^2), Y_{2j} \sim N(\mu_2, \sigma^2),\ j = 1, ..., n_2$.

6. Test for the final decision: Based on the first and second stage data, obtain z-statistics Z_1 and Z_2 for the null hypotheses $H_{01} : \mu_1 - \mu_0 \leq 0$ and $H_{02} : \mu_2 - \mu_0 \leq 0$, respectively. If either $[Z_1 \geq c_{PO}$ and $max(Z_1, Z_2) \geq c_{PO*}]$ or $[Z_2 \geq c_{PO}$ and $max(Z_1, Z_2) \geq c_{PO*}]$, increase success by one.

7. Calculate power: Repeat 2-6 steps 1000 times and calculate power as success/1000

8. Increase μ_2 by 0.003 and Repeat 2-7 steps by $\mu_2 = 3$

The O'Brien and Fleming's design has the exactly same procedure except different critical values for hypothesis testings. Let $c_{OB} = 1.977$. Instead of the critical value of Pocock's, $c_{OB}\sqrt{2}$ is used at the interim analysis, and c_{OB} is used at the final analysis. Critical values to test $H_{01} \bigcap H_{02}$ also should be computed differently for interim and final analyses based on $c_{OB}\sqrt{2}$ and c_{OB}.

4.1.3 Algorithm for Combination Test Approach

Adaptive designs based on combination tests allow early stopping at interim based on the pre-determined α_0 and α_1 values. However, the choice of these values is usually application-specific and depends on a case-by-case basis. For this reason, a non-binding rule was considered here by setting $\alpha_0 = 1$ and

$\alpha_1 = 0$. Non-binding rules can still provide flexible interim adaptation, but have more conservative power. The critical values for two adaptation rules were obtained from the simulation using three equal-mean treatments, which results in 0.0406 and 0.0300 for adaptation I and adaptation II, respectively. The stage wise p-values were combined using the weighted inverse normal combination function with weights $w_1 = \sqrt{n_1/(n_1+n_2)}, w_2 = \sqrt{n_2/(n_1+n_2)}$. For adaptation I case, the algorithm for the combination approach is:

1. Set $\mu_0 = 0, \mu_1 = 2, \mu_2 = 0.003, \sigma^2 = 6^2, c = 0.0406, n_1 = n_2 = 72$

2. Set success=0

3. Simulate observations for stage 1: Generate samples $Y_{0j} \sim N(\mu_0, \sigma^2), Y_{1j} \sim N(\mu_1, \sigma^2), Y_{2j} \sim N(\mu_2, \sigma^2),\ j = 1, ..., n_1,$ where Y_{ij} is j-th observation from treatment i.

4. Calculate the first stage p-values: Based on the first stage data, obtain the first stage p-values $p_{(1,1)},\ p_{(1,2)}$, and $p_{(1,12)}$ for three null hypotheses $H_{01} : \mu_1 - \mu_0 \leq 0,\ H_{02} : \mu_2 - \mu_0 \leq 0$, and $H_{01} \bigcap H_{02}$, respectively. The p-value for $H_{01} \bigcap H_{02}$ is calculated by $min\{1, 2min(p_{(1,1)}, p_{(1,2)})\}$ using Bonferroni correction.

5. Simulate observations for stage 2: Generate samples $Y_{0j} \sim N(\mu_0, \sigma^2), Y_{1j} \sim N(\mu_1, \sigma^2), Y_{2j} \sim N(\mu_2, \sigma^2),\ j = 1, ..., n_2$.

6. Calculate the second stage p-values: Based on the second-stage data, obtain the second stage p-values $p_{(2,1)},\ p_{(2,2)}$, and $p_{(2,12)}$ for three null

hypotheses $H_{01} : \mu_1 - \mu_0 \leq 0,\ H_{02} : \mu_2 - \mu_0 \leq 0$, and $H_{01} \bigcap H_{02}$, respectively. The p-value for $H_{01} \bigcap H_{02}$ is calculated by $min\{1, 2min(p_{(2,1)}, p_{(2,2)})\}$ using Bonferroni correction.

7. Test for the final decision: Let C(p,q) denotes the weighted inverse normal combination function with weights $w_1 = \sqrt{n_1/(n_1+n_2)}, w_2 = \sqrt{n_2/(n_1+n_2)}$. If either $[C(p_{(1,1)}, p_{(2,1)}) < c$ and $C(p_{(1,12)}, p_{(2,12)}) < c]$ or $[C(p_{(1,2)}, p_{(2,2)}) < c$ and $C(p_{(1,12)}, p_{(2,12)}) < c]$, increase success by one.

8. Calculate power: Repeat 2-7 steps 1000 times and calculate power as success/1000

9. Increase μ_2 by 0.003 and Repeat 2-8 steps by $\mu_2 = 3$

For adaptation II case, one experiment treatment is dropped at the interim based on the first stage mean value, and the sample size for the discontinued treatment is evenly reallocated to the continued treatments.

1. Set $\mu_0 = 0, \mu_1 = 2, \mu_2 = 0.003, \sigma^2 = 6^2, c = 0.0300, n_1 = n_2 = 72$

2. Set success=0

3. Simulate observations for stage 1: Generate samples $Y_{0j} \sim N(\mu_0, \sigma^2), Y_{1j} \sim N(\mu_1, \sigma^2), Y_{2j} \sim N(\mu_2, \sigma^2),\ j = 1, ..., n_1,$ where Y_{ij} is j-th observation from treatment i.

4. Calculate the first stage p-values: Based on the first stage data, obtain the first stage p-values $p_{(1,1)},\ p_{(1,2)}$, and $p_{(1,12)}$ for three null hypotheses

$H_{01} : \mu_1 - \mu_0 \leq 0,\ H_{02} : \mu_2 - \mu_0 \leq 0$, and $H_{01} \bigcap H_{02}$, respectively. The p-value for $H_{01} \bigcap H_{02}$ is calculated by $min\{1, 2min(p_{(1,1)}, p_{(1,2)})\}$ using Bonferroni correction.

5. Drop the inferior treatment: Obtain the mean value of each experimental treatment and drop the treatment which has the smaller value.

6. Simulate observations for stage 2: Let $Y*_j$ be the j-th observation from the continued experimental treatment and $\mu*$ be the corresponding true mean. Generate samples $Y_{0j} \sim N(\mu_0, \sigma^2), Y*_j \sim N(\mu*, \sigma^2),\ j = 1, ..., n*$, where $n* = n_2 + n_1/2$.

7. Calculate the second stage p-value: Based on the second stage data, obtain the second stage p-value $p_{(2,*)}$ for $H_{0*} : \mu * - \mu_0 \leq 0$.

8. Test for the final decision: Let C(p,q) denotes the weighted inverse normal combination function with weights $w_1 = \sqrt{n_1/(n_1 + n*)}, w_2 = \sqrt{n*/(n_1 + n*)}$. If $C(p_{(1,*)}, p_{(2,*)}) < c$ and $C(p_{(1,12)}, p_{(2,12)}) < c$, increase success by one.

9. Calculate power: Repeat 2-8 steps 1000 times and calculate power as success/1000

10. Increase μ_2 by 0.003 and Repeat 2-9 steps by $\mu_2 = 3$

4.1.4 Algorithm for Bayesian Adaptive Design

The Bayesian adaptive design proposed in the previous section provides multiple treatment testing based on the posterior distributions of parameters. Gibbs sampling method was exploited to generate posterior samples from the proposed Bayesian model, with one chain run with 500 iterations. The number of MCMC iterations is relatively small since each chain converges very quickly from the simulated data (See appendix B). As discussed earlier, it was expected that two adaptation cases show little difference in terms of statistical power. Thus, only adaptation II was implemented in this simulation, dropping one inferior experimental treatment based on the distributional difference $p(\mu_2 - \mu_1 > 0|y) < \alpha$ or $p(\mu_1 - \mu_2 > 0|y) < \alpha$.

1. Set $\mu_0 = 0, \mu_1 = 2, \mu_2 = 0.003, \sigma^2 = 6^2, \alpha = 0.025, n_1 = n_2 = 72$

2. Set success=0

3. Simulate observations for stage 1: Generate samples $Y_{0j} \sim N(\mu_0, \sigma^2), Y_{1j} \sim N(\mu_1, \sigma^2), Y_{2j} \sim N(\mu_2, \sigma^2),\ j = 1, ..., n_1,$ where Y_{ij} is j-th observation from treatment i.

4. Drop if there is an inferior treatment: Estimate posterior distributions given the first stage data by Gibbs sampling. Drop treatment 1 if $p(\mu_1 - \mu_2 > 0|y) < \alpha$, or drop treatment 2 if $p(\mu_2 - \mu_1 > 0|y) < \alpha$. Otherwise, proceed with all treatments.

5. Simulate observations for stage 2: If one of treatment is dropped, generate samples $Y_{0j} \sim N(\mu_0, \sigma^2), Y*_j \sim N(\mu*, \sigma^2),\ j = 1, ..., n*$, where $Y*_j$ is the j-th observation from the continued experimental treatment, $\mu*$ is the corresponding true mean, and $n* = n_2 + n_1/2$. Otherwise, generate samples $Y_{0j} \sim N(\mu_0, \sigma^2), Y_{1j} \sim N(\mu_1, \sigma^2), Y_{2j} \sim N(\mu_2, \sigma^2),\ j = 1, ..., n_2$.

6. Derive the final decision: Based on the first and second stage data, estimate posterior distributions by Gibbs sampling. If one of treatment was dropped, test $H_{0*} : \mu * - \mu_0 \leq 0$ vs. $H_{1*} : \mu * - \mu_0 > 0$ using Bayes factor. Increase success by one when the Bayes factor is greater than 3 in favor of the alternative hypothesis. If all treatments were continued to stage 2, test $H_{0i} : \mu_i - \mu_0 \leq 0$ vs. $H_{1i} : \mu_i - \mu_0 > 0$ for treatment i=1,2 individually using Bayes factor. Increase success by one if one of Bayes factors is greater than 3 in favor of the corresponding alternative hypothesis.

7. Calculate power: Repeat 2-6 steps 1000 times and calculate power as success/1000

8. Increase μ_2 by 0.003 and Repeat 2-7 steps by $\mu_2 = 3$

4.2 Results

This section presents results of the simulation study described in the previous section. Since statistical power increased in order of fixed-sample de-

signs, group sequential designs, combination test approaches, and the Bayesian adaptive design, pairwise comparisons were made between fixed-sample versus group sequential designs, group sequential versus combination test approaches, and combination test approaches versus the Bayesian adaptive design.

4.2.1 Fixed-Sample vs. Group Sequential Designs

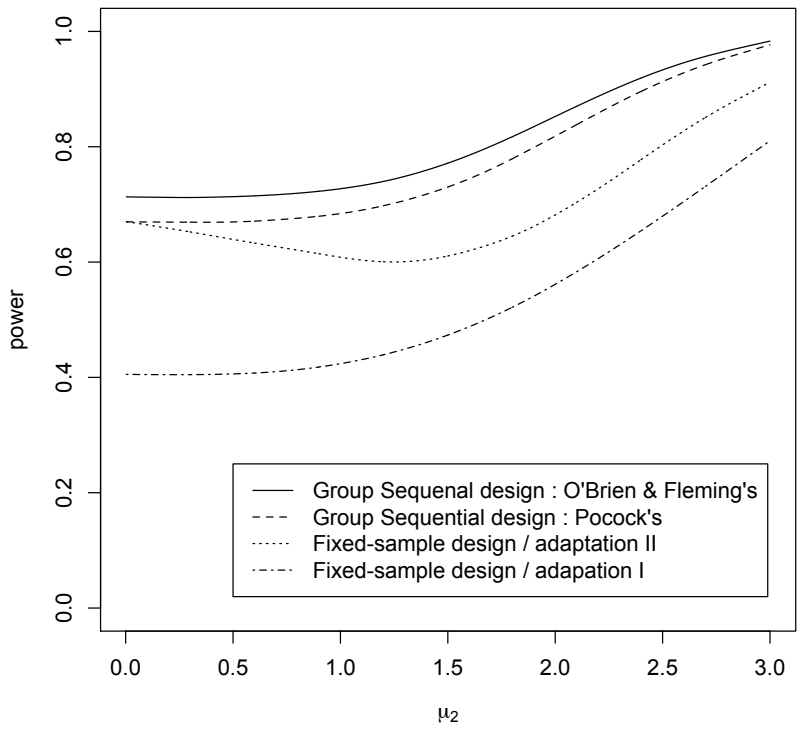

Figure 4.1: Statistical power of Fixed-sample vs. Group sequential designs

The Figure 4.1 shows the power of fixed-sample and group sequential designs. As discussed earlier, two group sequential designs, Pocock's and

O'Brien and Fleming's, cannot incorporate interim adaptation other than early stopping for superiority. For fixed-sample design, two adaptation cases were considered, where one is the case in which all treatments proceed to the second stage (adaptation I), and the other is dropping the inferior experimental treatment based on the mean value of the first stage (adaptation II). The power trend with respect to the selected values of μ_2 is also given as numbers at the end of this section (Table 4.1).

Overall, the graph demonstrates that group sequential designs have greater power than the fixed-sample design regardless of types of adaptation. Different amount of information may play an important role in this gap, since larger information leads to more power in hypothesis testing. Fixed-sample design derives the final decision only based on the second stage data, whereas group sequential designs exploit all stage data unless they stop a trial early at the interim. The difference between fixed-sample and group sequential designs decreases as the value of second treatment mean, μ_2, increases.

For the fixed-sample design, the power of adaptation II is greater than that of adaptation I due to the increased sample-sizes from the dropped treatment. Note that the power of adaptation II is not monotonous in μ_2. This can be explained by the behavior of treatment selection as follows [1, 35]. When μ_2 is much smaller than $\mu_1 = 2$, there is little chance for the treatment 2 to be selected for the stage 2. As μ_2 increases toward μ_1, however, the probability of selecting treatment 2 also increases although μ_2 is still less than μ_1. A trial may loses statistical power until it reaches the point where both treatments

equally effective, because the inferior treatment can be tested for the final decision.

For the group sequential designs, O'Brien and Fleming's design shows greater power than Pocock's design. This is not surprising because these two designs are exactly the same except the critical values used in hypothesis testings. O'Brien and Fleming's design has the smaller second stage critical value (and thus larger critical region) compared to Pocock's: 1.977 (O'Brien and Fleming's) and 2.178 (Pocock's) in this study. As a result, it is more likely that the power of O'Brien and Fleming's is higher than that of Pocock's.

4.2.2 Group Sequential vs. Combination Test Approaches

In Figure 4.2, group sequential and combination test approaches were compared in terms of statistical power. The two group sequential designs, Pocock's and O'Brien and Fleming's, allow only early stopping for superiority at the interim analysis, while the combination test approach allows two adaptation cases: continue to the second stage with all treatments (adaptation I) or only with the superior treatment based on the first stage mean value (adaptation II). In this comparison, both group sequential and combination test approaches incorporate all stage data for the final decision, but use different test statistics: z-statistics (group sequential design) and p-values (combination test approach). Designs based on p-values can be conservative than those based on z-statistics, because p-values are not sufficient statistics and thus can lose information contained in data [15]. Note that z-statistics are from

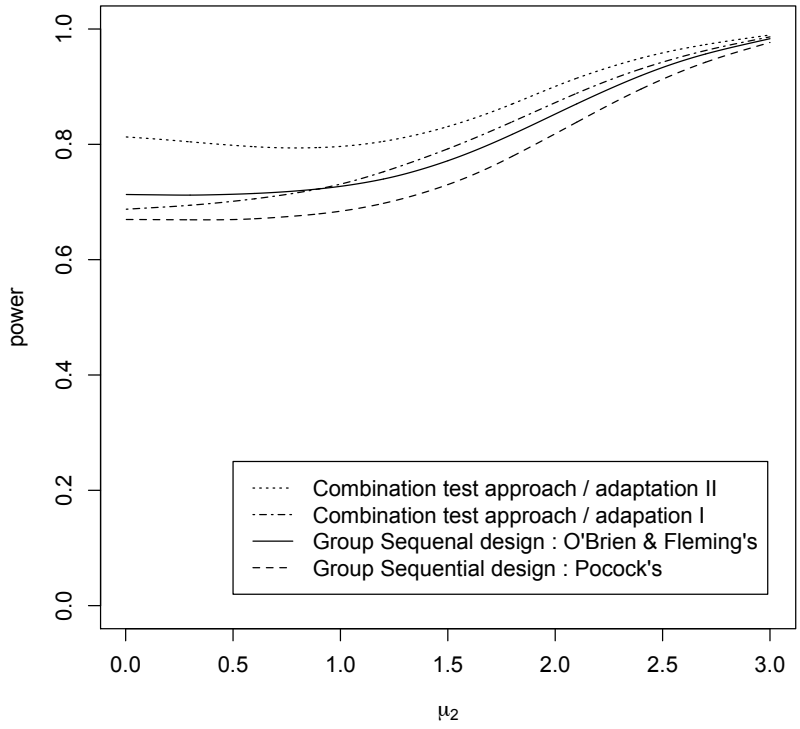

Figure 4.2: Statistical power of Group sequential vs. Combination test approaches

mean value of the data and preserve all the information in data as a sufficient statistic.

From the Figure 4.2, it can be seen that the combination test approach performs better than group sequential designs. This is consistent with the results from previous researches (i.e. [35]). The combination test approach with adaptation II has the largest power among four cases in this comparison, indicating that reallocated sample sizes compensated possible loss of information

caused by p-values. This shows that the conservative nature of test statistics in the combination test approach can be overcome by the advantage of flexible interim adaptation. The Pocock's design has the lowest power among the four cases, and the other cases are located between Pocock's and the combination test approach with adaptation II. The power trend with respect to the selected values of μ_2 is also given as numbers at the end of this section (Table 4.1).

For the combination test approach with adaptation I and the O'Brien and Fleming's design, it is hard to say that one is better than the other. O'Brien and Fleming's design shows slightly higher power than the other for small values of μ_2, but shows lower power for moderate and large values of μ_2. This can be partly explained by the difference in test statistics, since combination test approaches based on p-values and the closure principle do not lead to sufficient test statistics when multiple treatments are considered [35]. The simulation result in this study indicates that the use of insufficient statistics can lower the power, while the effect is alleviated by the large value of the true mean (μ_2 in this study).

4.2.3 Combination Test Approach vs. Bayesian Adaptive Design

For the final comparison, the power of the Bayesian adaptive design was plotted with the power of combination test approach (Figure 4.3). The combination test approach assumes two adaptation rules, I and II, and the Bayesian adaptive design only considers adaptation rule II because of the indifference of the results. Although the adaptation rule II is implemented for

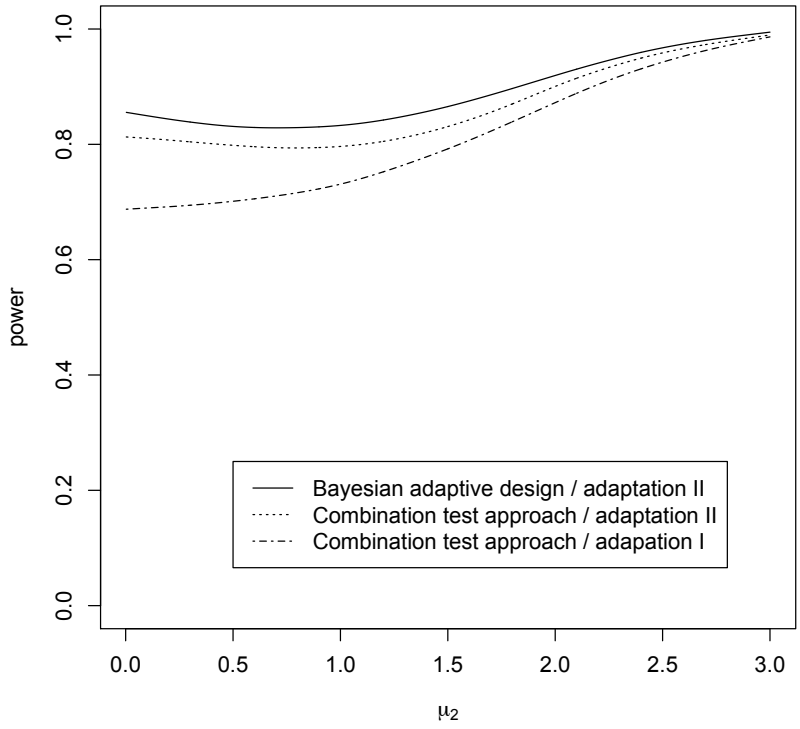

Figure 4.3: Statistical power of Combination test approaches vs. Bayesian adaptive design

both combinatio n test and Bayesian adaptive designs, dropping criterion at the interim is different each other. The combination test approach drops the inferior treatment based on the mean value of the first stage data, while the Bayesian model make this decision by comparing posterior distributions of parameters conditioned on the first stage data. For the final decision, both combination test approaches and the Bayesian adaptive design exploit all stage data but use different statistics: p-values for the combination test approach

and posterior distributions for the Bayesian model. In addition, unlike the Bayesian model, the combination test approach uses Bonferroni correction and the closure principle to deal with the multiplicity issue.

In Figure 4.3, the Bayesian model shows larger power than the combination test approaches regardless of types of adaptation. The difference is substantial for small values of μ_2 and decreases as μ_2 increases. Note that the Bayesian model is hardly dropping a treatment at the interim analysis, and thus, there are much fewer chances to increase second stage sample size than the combination test approach with the same adaptation. This indicates that Bayesian model has strong points in terms of power other than interim adaptation or sample size reallocation. There can be several possible explanations. First, the Bayesian model derives the final decision based on the posterior distributions of parameters. These distributions are conditioned on the full data, so there is usually no loss of information compared to insufficient statistics such as p-values. Second, the Bayesian model does not need to incorporate methods for type I error rate such as Bonferroni correction and the closure principle. Especially for the Bonferroni correction, it is known that the type I error rate is controlled in a very conservative sense. This means that the actual type I error rate is mostly smaller than the significance level, resulting in less power than expected. Overall, the Bayesian model performs better than all the other designs included in this study. The power trend with respect to the selected values of μ_2 is also given as numbers at the end of this section (Table 4.1).

Table 4.1: Statistical power with respect to the different treatment means μ_2

μ_2	0.300	0.600	0.900	1.200	1.500	1.800	2.100	2.400	2.700	3.000
Fixed-sample with adaptation I	0.406	0.391	0.399	0.408	0.446	0.509	0.587	0.647	0.730	0.805
Fixed-sample with adaptation II	0.642	0.647	0.639	0.627	0.642	0.657	0.691	0.790	0.878	0.914
Pocock's Design	0.696	0.656	0.663	0.686	0.735	0.775	0.867	0.903	0.942	0.970
O'Brien and Fleming's Design	0.700	0.731	0.749	0.722	0.783	0.824	0.865	0.912	0.956	0.976
Combination test with adaptation I	0.678	0.724	0.707	0.747	0.812	0.842	0.890	0.938	0.958	0.984
Combination test with adaptation II	0.779	0.800	0.811	0.795	0.837	0.862	0.921	0.953	0.969	0.985
Bayesian adaptive with adaptation II	0.841	0.832	0.837	0.841	0.858	0.868	0.925	0.952	0.982	0.990

Chapter 5

Discussion

The present report provided a comparative investigation of both frequentist and Bayesian adaptive designs when more than one experimental treatments are tested in multiple stages. Researches on the adaptive designs have been mostly restricted to limited comparison of frequentist methods, and there is little evidence for the relative performance of Bayesian models or for the situation when multiple treatments are involved in a trial. To address this issue, various adaptive designs were included in this study. They are mainly three designs: group sequential design, combination test approach, and Bayesian adaptive design. More precisely, Pocock's [3] and O'Brien and Fleming's designs [4] for the group sequential design, Bretz et al. (2009) model [1] for the combination test approach, and a hierarchical Bayesian model based on Thall et al. (2003) [12] for the Bayesian adaptive design were considered. The comparison was made through a simulation study under various scenarios, focusing on statistical power of the different designs. The simulation was based on a two-stage clinical trial with one interim analysis and assumed two different interim adaptation rules for the all designs except group sequential designs. A fixed-sample design was also included in simulation as a non-adaptive baseline design.

According to the simulation results, it was shown that the Bayesian adaptive design outperforms than all other designs in the setting of this study. The power of the combination test approach was consistently higher than O'Brien and Fleming's design when there is an early-dropped treatment which allows sample size reallocation, but not always higher when there is no dropped treatment. O'Brien and Fleming's design was better than Pocock's design as expected. It was also emphasized that all adaptive designs demonstrated higher statistical power compared to the non-adaptive fixed-sample design regardless of adaptation rules considered in this study. Overall, statistical power increased in order of fixed-sample design, group sequential design, combination test approach, and Bayesian adaptive design. The difference decreased as the true mean of the second experimental treatment increased.

The findings of the current study are consistent with Bretz et al. (2009) [1] in that combination test approaches show higher power than fixed-sample designs regardless of the interim adaptation rules. The results of these two designs from both studies seem exactly the same since the assumptions used in this study are mostly based on those of Bretz et al. (2009). The findings of the present study also confirm the findings of Tsiatis and Mehta (2003) [13] and Jennison and Turnbull(2003,2006) [14, 15] that there can be reduction of power due to the use of insufficient statistics. In the case of multiple experimental treatments, however, the reduction was small and appeared only when not all experimental treatments are sufficiently different from the standard. This result is consistent with the insight of Kelly et al (2005) [16] but was not

shown in the study of Friede and Stallard (2008) [35]. The current study also investigated the impact of interim adaptation regarding sample size reallocation, which was not considered in Friede and Stallard (2008). The superior power of the combination test approach indicates that the use of insufficient statistics can be compensated by appropriate interim adaptation. For the Bayesian adaptive design, the present study provided new evidence of relative performance in terms of statistical power. The Bayesian adaptive model demonstrated higher power with great flexibility, compared to the various frequentist adaptive designs included in this study. This result has important implication for the future design development and may help us to find better adaptive designs in clinical trials.

The findings in this report are subject to at least the following limitations. First, the study focused on only a few adaptive designs in clinical trials. There are other group sequential designs which incorporate various interim situations and other combination test approaches with different combination functions [36, 37]. The Bayesian adaptive model can also be extended to take into account various design considerations. Since it is expected that different design assumptions affect statistical power differently, the results in this study cannot be extrapolated to all adaptive designs. Second, the simulation study is based on the specific assumptions and scenarios. The design considerations are mainly influenced by the primary study objectives in practice, and the setting of this study is not always satisfied. For example, the power concept measuring the probability of rejecting at least one false null hypothesis may

not be appropriate in some settings. Power can be defined as the probability of rejecting all false null hypotheses or adjusted in many different ways to fulfill the specific goals of trials [34]. Finally, the Bayesian model in this study assumed only non-informative priors. Informative priors improve the quality of analyses by delivering additional information. In most cases, clinical trials are likely to have previous researches, trials or standard regimens. Thus, if there exist qualified prior information, the Bayesian model can result in more practical and insightful conclusions.

Appendices

Appendix A

Two Adaptation Cases in the Bayesian Model

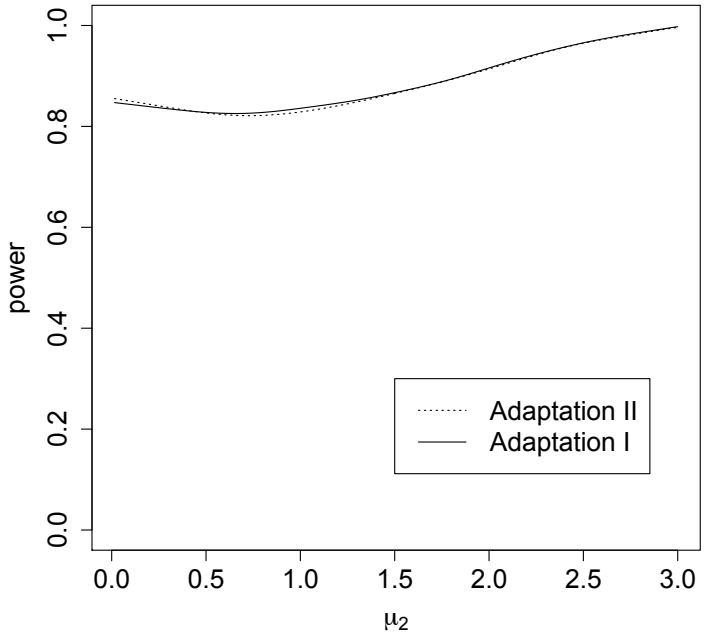

Figure A.1: Statistical power of the Bayesian model: two adaptation cases

A short simulation result of the Bayesian model comparing two adaptation cases (Figure A.1). Each case was implemented using 200 trials with 200 different μ_2 as an example. The figure demonstrates that two adaptation cases are not very different in terms of power in the setting of this study.

Appendix B

Convergence in the Bayesian Model

Posterior distributions of the Bayesian model were estimated using Gibbs sampling with 500 iterations, and samples mostly showed sufficient convergence to the target distributions. As an example, the following illustrates convergence of a posterior sample conditioned on first stage data in a trial. μ_2 was chosen to be 3.

First, Geweke diagnostic [38] was calculated as one of the convergence diagnostics. Geweke diagnostic is based on a test for equality of means of the first 10% and the last 50% of a Markov chain. This asymptotically has normal distribution under the equality and uses a standard Z-score as a test statistic. From a posterior sample drawn by Gibbs sampling, Geweke diagnostic showed 0.4529 for μ_0, -0.2952 for μ_1, 0.4215 for μ_2, -1.136 for γ, and -0.8639 for τ^2. Since no values are bigger than 1.96, it can be concluded that this chain is converged to the target distribution. (1.96 is corresponding to 0.05 significance level for a two-sided test)

Next, convergence was visually inspected through trace plots of parameters (Figure B.1). According to the plots, there is no evidence of severe lack of convergence in this sample.

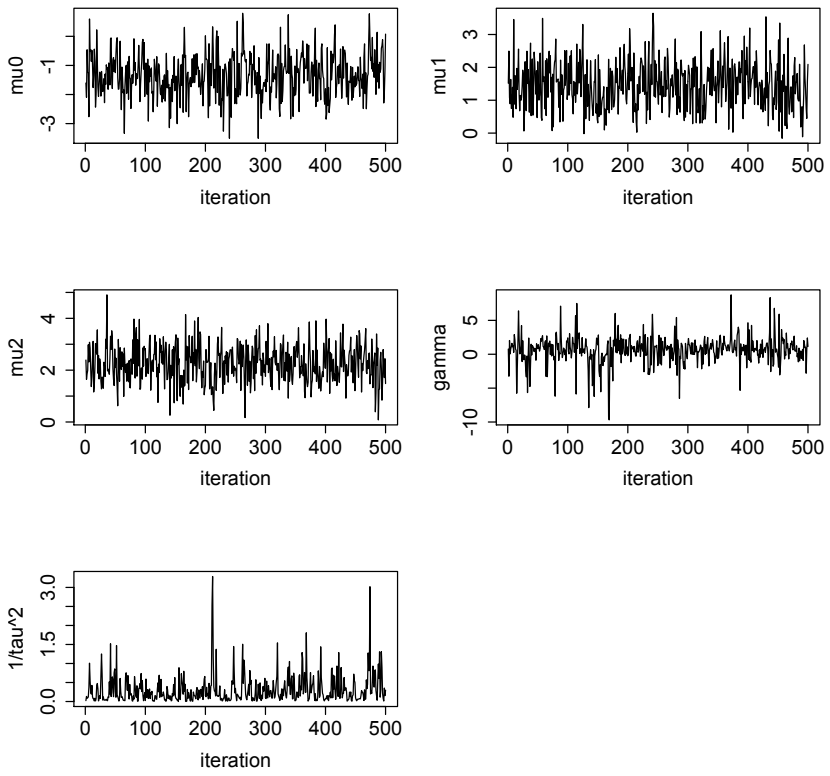

Figure B.1: Trace plots of parameters for the Bayesian model

Bibliography

[1] Bretz, F., Koenig, F., Brannath, W., Glimm, E., and Posch, M. (2009) Tutorial in biostatistics adaptive designs for confirmatory clinical trials, Statistics in Medicine; 28:1181-1217

[2] Yin, G., (2012) Clinical trial design: Bayesian and Frequentist Adaptive Methods, Hoboken; Wiley

[3] Pocock, S. J. (1977) Group sequential methods in the design and analysis of clinical trials, Biometrika; 64(2):191-199

[4] O'Brien, P. C. and Fleming, T.R. (1979) A multiple testing procedure for clinical trials, Biometrics; 35(3):549-556

[5] Whitehead, J., Whitehead, A., Todd, S., Bolland, K., and Sooriyarachchi, M. R. (2001) Mid-trial design reviews for sequential clinical trials, Statistics in Medicine; 20:165-176

[6] Stallard, N. and Friede, T (2008) Flexible group-sequential designs for clinical trials with treatment selection, Statistics in Medicine; 27:6209-6227

[7] Müller, H.-H. and Schäfer, H. (2001) Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches, Biometrics; 57:886-891

[8] Cui, L., Hung, HMJ., and Wang, SJ. (1999) Modification of sample size in group sequential clinical rials, Biometrics; 55:853-857

[9] Lehmacher, W. and Wassmer, G. (1999) Adaptive sample size calculations in group sequential trials, Biometrics; 55:1286-1290

[10] Bauer, P. and Köhne, K. (1994) Evaluation of experiments with adaptive interim analyses, Biometrics; 50:1029-1041

[11] Berry, S. M., Carlin, B. P., Lee, J. J, and Müller, Peter (2010) Bayesian Adaptive Methods for Clinical Trials; Chapman and Hall/CRC press

[12] Thall, P. F., Wathen, J. K., Bekele, B. N., Champlin, R. E., Baker, L. H., and Benjamin, R. S. (2003) Hierarchical Bayesian approaches to phase II trials in diseases with multiple subtypes, Statistics in Medicine; 22:763-780

[13] Tsiatis, A. A. and Mehta, C. (2003) On the inefficiency of the adaptive design for monitoring clinical trials, Biometrika; 90:367-378

[14] Jennison, C. and Turnbull, B. W. (2003) Mid-course sample size modification in clinical trials based on the observed treatment effect, Statistics in Medicine; 22:971-993

[15] Jennison, C. and Turnbull, B. W. (2006) Adaptive and nonadaptive group sequential tests, Biometrika; 93(1):1-21

[16] Kelly, P. J., Sooriyarachchi, M. R., Stallard, N., and Todd, S. (2005) A practical comparison of group-sequential and adaptive designs, Journal of Biopharmaceutical Statistics; 15, 719-738

[17] Friede, T. and Stallard, N. (2008) A comparison of methods for adaptive treatment selection, Biometrical Journal; 50(5):767-781

[18] Jennison, C. and Turnbull, B. W. (2000) Group sequential methods with applications to clinical trials, Boca Raton, FL; Chapman and Hall/CRC press

[19] Bauer, P. (1989) Multistage testing with adaptive designs (with Discussion), Biometrie und Informatik in Medizin und Biologie; 20:130-148

[20] Brannath, W., Koenig, F., and Bauer, P. (2007) Multiplicity and flexibility in clinical trials, Pharmaceutical Statistics; 6:205-216

[21] Liu, Q., Proschan, M. A., and Pledger, G. W. (2002) A unified theory of two-stage adaptive designs, Journal of the American Statistical Association; 97:1034-1041

[22] Vandemeulebroecke, M. (2008) Group Sequential and Adaptive Designs A Review of Basic Concepts and points of Discussion, Biometrical Journal; 50(4):541-557

[23] Kieser, M., Bauer, P., and Lehmacher, W. (1999) Inference on multiple endpoints in clinical trials with adaptive interim analyses, Biometrical Journal; 41:261-277

[24] Wassmer, G. and Vandemeulebroecke, M. (2006) A brief review on software developments for group sequential and adaptive designs, Biometrical Journal; 48:732-737

[25] Marcus, R., Peritz, E., and Gabriel, K. R. (1976) On closed testing procedure with special reference to ordered analysis of variance, Biometrika; 63:655-660

[26] Hellmich, M. (2001) Monitoring clinical trials with multiple arms, Biometrics; 57:892-898

[27] Tang, D. and Geller, N. L. (1999) Closed testing procedures for group sequential clinical trial with multiple endpoints, Biometrics; 55:1188-1192

[28] Shaffer, J. P. (1995) Multiple hypothesis testing, Annual Review of Psychology; 46:561-584

[29] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003) Bayesian data analysis; Chapman and Hall/CRC press

[30] Geman, S. and Geman, D., (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, Pattern Analysis and Machine Intelligence; 6:721-741

[31] Jeffreys, H. (1961) The theory of probability, Oxford, UK; Oxford University Press

[32] Kass, R. E. and Raftery, A. E. (1995) Bayes Factors, Journal of the American Statistical Association; 90:773-795

[33] Scott, J. G. and Berger, J. O. (2010) Bayes and Empirical-Bayes multiplicity adjustment in the variable-selection problem, The Annals of Statistics; 38(5):2587-2619

[34] Senn, S. and Bretz, F. (2007) Power and sample size when multiple endpoints are considered, Pharmaceutical Statistics; 6:161-170

[35] Friede, T. and Stallard, N. (2008) A comparison of methods for adaptive treatment selection, Biometrical Journal; 50(5):767-781

[36] Stallard, N. and Todd, S. (2003) Sequential designs for phase III clinical trials incorporating treatment selection, Statistics in Medicine; 22:689-703

[37] Bauer, P. and Kieser, M. (1999) Combining different phases in the development of meical treatments within a single trial, Statistics in Medicine; 18:1833-1848

[38] Geweke, J. (1992) Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments (with discussion). In Bayesian Statistics 4 (ed JM Bernardo et al.), Oxford, UK; Oxford University Press

Vita

Sukyung Park was born in Busan, South Korea in 1984. She received the Bachelor of Science degree in Industrial Engineering from Korea Advanced Institute of Science and Technology in 2007. She was accepted and started graduate studies in 2012.

Permanent address: skpark12@utexas.edu

This report was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.