2014

Adaptive methodologies in multi-arm dose response and biosimilarity clinical trials

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

ADAPTIVE METHODOLOGIES IN MULTI-ARM DOSE RESPONSE AND BIOSIMILARITY CLINICAL TRIALS

by

JOSEPH MOON WAI WU

B.Soc.Sc., The University of Hong Kong, 1992
M.A., Trinity International University, 2002
M.A., Boston University, 2011

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2014

Approved by

First Reader

Mark Chang, Ph.D.
Adjunct Professor of Biostatistics

Second Reader

Gheorghe Doros, Ph.D.
Associate Professor of Biostatistics

Third Reader

Sandeep Menon, Ph.D.
Adjunct Professor of Biostatistics

Dedicated To: Miu-King Tse & Michael J.M. Holmes

Acknowledgements

I have to say this dissertation represents the confluence of the support and guidance of many individuals. First of all, I would like to express my deepest gratitude to all of my thesis committee members, especially Dr. Mark Chang, my primary advisor, for introducing me to the topic of adaptive clinical trials. Mark has not only given so much inspiration to my thesis, but also has exemplified a strong passion for statistical research. Here, I salute Mark for his great mentorship! I also want to thank Dr. Gheorghe Doros for showing me how to be a "Bayesian" and for all the books he lent me. Also, I am grateful to Dr. Sandeep Menon for giving me the opportunities to be a strong writer in the pharmaceutical industry and connecting me to the resources I needed.

As I look back at the five and a half years time at Boston University, I am greatly appreciative of the faculty and staff in the Department of Biostatistics. In particular, I want to thank Dr. Adrienne Cupples and Dr. Howard Cabral for offering me the opportunity to pursue this doctoral degree as well as for the confidence they placed in me to bring it to completion. Over the years, I was glad to be able to work with Dr. Mayetri Gupta on a computation biology project in which I have learned a lot of the proper Bayesian inference. Besides that, I want to thank Dr. Joseph Massaro and Dr. Robert Lew for connecting me to the real world of clinical trials. Through these collective experiences I have grown to be a better researcher.

In this arduous journey to completing my dissertation, I have made many great friends who have shared laughter and tears together. I shall not forget the light-hearted chats we had in the hallway or classrooms, besides talking about homework and exams. I especially want to thank Revathi Ananthakrishnan for her great friendship and encouragement, and I find her frank and congenial spirit has inspired me to always look at the bright side! I am glad for the friendships of Si-yan Xu for being a good listener, Han Chen, Bai-yun Yao, and many others for the kind words we shared.

There are also many great friends who have kept encouraging me behind the "scene",

particularly the Foleys, the Allens and the entire “village” for their support. And of course, from a thousand miles and across the ocean, my friends Edmund, Bowie, Dee, Tony, Gary, and Leo for constantly checking up on me and my sanity. Besides these, I want to say thanks to my family in the US: Al and Chris Brissette, Richard and Susan Holmes, the Chus, the Chows and those in Hong Kong: Miu-king Tse, my mother, who has always loved me and has raised me to be a hard-working person. Last, I want to thank Michael Holmes, my husband, whom I love dearly for putting up with me through all these years, and spending time with me at Starbucks for hours every single weekend while I worked on my dissertation. I bow in thanks to all of you!

ADAPTIVE METHODOLOGIES IN MULTI-ARM DOSE RESPONSE AND BIOSIMILARITY CLINICAL TRIALS

(Order No.)

JOSEPH MOON WAI WU

Boston University, Graduate School of Arts and Sciences, 2014

Major Professor: Mark Chang, Adjunct Professor of Biostatistics

ABSTRACT

As most adaptive clinical trial designs are implemented in stages, well-understood methods of sequential trial monitoring are needed. In the frequentist paradigm, examples of sequential monitoring methodologies include the p-value combination tests, conditional error, conditional power, and alpha spending approaches. Within the Bayesian framework, posterior and predictive probabilities are used as monitoring criteria, with the latter being analogous to the conditional power approach.

In a placebo or active-contolled dose response clinical trial, we are interested in achieving two objectives: selecting the best therapeutic dose and confirming this selected dose. Traditional approach uses the parallel group design with Dunnett's adjustment. Recently, some two-stage Seamless II/III designs have been proposed. The drop-the-losers design considers selecting the dose with the highest empirical mean after the first stage, while another design assumes a dose-response model to aid dose selection. These designs however do not consider prioritizing the doses and adaptively inserting new doses. We propose an adaptive staggered dose design for a normal endpoint that makes minimal assumption regarding the dose response and sequentially adds doses to the trial. An alpha spending function is applied in a novel way to monitor the doses across the trial. Through numerical and simulation studies, we confirm that optimistic alpha spending coupled with informative dose ordering jointly produce some desirable operating characteristics when compared to drop-the-losers and model-based Seamless designs. In addition, we show how the design parameters can

be flexibly varied to further improve its performance and how it can be extended to binary and survival endpoints.

In a biosimilarity trial, we are interested in establishing evidence of comparable efficacy between a follow-on biological product and a reference innovator product. So far, no standard method for biosimilarity has been endorsed by regulatory agency. We propose a Bayesian hierarchical bias model and a non-inferiority hypothesis framework to prove biosimilarity. A two-stage adaptive design using predictive probability as early stopping criterion is proposed. Through simulation study, the proposed design controls the type I error better than the frequentist approach and Bayesian power is superior when biosimilarity is plausible. Two-stage design further reduces the expected sample size.

Contents

List of Tables

List of Figures

Chapter 1

Rise of Adaptive Methods in Clinical Trials

1.1 Introduction

Society and government regulatory agencies have high expectations for the production of safe and efficacious drugs from the pharmaceutical industry. However, the success rate of new drugs remains low or may even be in decline. In the past decade, the submission rate of new drug applications in the United States has shown a downward trend while the investment cost has risen (Woodcock and Woosley, 2008). In 2004, the U.S. Food and Drug Administration (FDA) launched the *Critical Path Initiative*, a project that is intended to improve the drug and medical device development processes, the quality of evidence generated, and the outcomes of clinical use of these products. This has generated considerable discussion and debate among drug developers, academics, and patient advocacy groups. After extensive consultation with stakeholders, the FDA issued the *Critical Path Report and List* in 2006. This report enumerated several important areas of scientific improvement - development and utilization of biomarkers, modernizing clinical trials processes and methodologies, aggressive use of bioinformatics, and improvement in manufacturing technologies. In the clinical trial community, it has been interpreted as encouragement for the use of innovative adaptive design methods.

Recent achievements in the methodology of adaptive designs provide new ways of drug

development that have the potential to improve quality, speed, and efficiency of decision making. By introducing adaptivity into trial design, this approach saves resources through identifying failures early and increases efficiency through focusing precious patient resources on treatments that have higher probability of success. While clearly this is advantageous to the drug development program, this is also ethically appealing as it restricts patient exposure to ineffective treatments (Dragalin, 2006).

By definition, an *adaptive design* of a clinical trial is a design that allows adaptations or modifications to some aspects of the trial after its initiation without undermining the validity and integrity of the trial (Chow, Chang, and Pong, 2005). Maintaining validity requires ensuring consistency between different stages, minimizing operational bias, and providing correct statistical inference; maintaining integrity requires providing convincing results to the broader scientific community. An adaptive design requires the trial to be conducted in multiple stages with convenient access to the accumulated data so that prospectively planned adaptations can be readily implemented after interim data are examined. It is important to emphasize that adaptations that are not prospectively planned are not *by design* but *ad hoc* adaptations and they are neither recommended nor encouraged. According to Chow and Chang (2008), modifications can be applied to either trial or statistical procedures. Examples of trial procedures include eligibility criteria, study endpoints, or treatment duration while those of statistical procedures are randomization schedule, hypotheses, sample size, or analysis plan. The following sections briefly describe these adaptations.

1.2 Types of Adaptive Methods

1.2.1 Stopping Rule

Stopping rules are intended to protect patients from clearly unsafe or ineffective treatments or to hasten the approval of a beneficial treatment when overwhelmingly strong evidence of efficacy is observed. Many stopping rules are constructed based on boundary-crossing

methodology. The classical group sequential method is the most representative of this type of adaptation. At any stage in the trial, a test statistic is calculated and compared with the given stopping boundaries. If the boundary is crossed, the trial is stopped and the corresponding conclusion is drawn; otherwise, the trial will continue to the next stage. Armitage, McPherson, and Rowe (1969) showed that repeating tests of significance at a fixed level on accumulating data increase the probability of getting a significant result under the null hypothesis. Jennison and Turnbull (2000) offered a comprehensive treatment on group sequential methodology. For example, for a two-arm K-stage group sequential design, if Z_k is the standardized test statistic comparing the two treatments at the kth interim analysis, then given the set of efficacy stopping boundaries, $\boldsymbol{b} = (b_1, b_2, ..., b_K)'$, and futility stopping boundaries, $\boldsymbol{a} = (a_1, a_2, ..., a_K)'$, the probability of rejecting the null hypothesis when it is true at the kth interim is $\psi_k = P_{\theta=0}(a_1 < Z_1 < b_1, ..., a_{k-1} < Z_{k-1} < b_{k-1}, Z_k > b_k)$ and the probability of accepting the null hypothesis when there is a true treatment difference δ is $\xi_k = P_{\theta=\delta}(a_1 < Z_1 < b_1, ..., a_{k-1} < Z_{k-1} < b_{k-1}, Z_k < a_k)$. Since $\mathbf{Z} = (Z_1, Z_2, ..., Z_K)'$ follows multivariate normal distribution, the boundary sets can be calculated numerically by controlling the overall type I error, $\Sigma_{i=1}^{K}\psi_k \leqslant \alpha$ and type II error, $\Sigma_{i=1}^{K}\xi_k \leqslant \beta$. Very often, a flexible alpha spending function $\alpha(t)$ is used to monitor the type I error, and a beta spending function $\beta(t)$ is used to monitor the type II error, where t is the information fraction (Lan and DeMets, 1983).

1.2.2 Sampling Rule

Sample size adjustment involves the re-calculation of the sample size for subsequent stages. This can be based on an interim estimate of a nuisance parameter such as the variance, or based on an interim estimate of the treatment effect. For most practical purposes, this is performed in a two-stage design in which sample size for the second stage is re-estimated using data from the first stage. However, this can also be performed in a multiple-stage design. There are many approaches to sample size adjustment. Blinded sample size re-

estimation uses the estimate of the nuisance parameter without unmasking treatment codes (Proschan, 2005; Bauer and Kieser, 1999). This is less controversial than unmasking to obtain the pooled variance or the estimated treatment differences. In this case, the sample size for the next stage is determined by either the estimated effect size or the conditional power. Cui, Hung, and Wang (1999) proposed increasing the sample size based on interim estimate of treatment difference as well as a new group sequential test procedure by changing the weights used in the traditional repeated significance two-sample mean test. The new test can provide a substantial gain in power with the increase of sample size while keeping overall type I error at target level. For a two-stage design, conditional power is defined as the probability of rejecting the null hypothesis at the end of study conditional on the observed test statistic from the first stage. Denne (2001) shows that the conditional power CP for a two-stage sample size re-estimation design is given by

$$CP_\theta(n_2, c_2|Z_1 = z_1) = 1 - \Phi\left[\frac{c_2\sqrt{n_2} - z_1\sqrt{n_1} - \frac{(n_2-n_1)\theta}{\sqrt{2\sigma^2}}}{\sqrt{n_2 - n_1}}\right]$$

where θ is the treatment effect of interest; n_1 and n_2 are the first-stage sample size and cumulative total sample size respectively; z_1 is the observed value of the standardized test statistic Z_1 in the first stage; c_2 is the critical value for the overall standardized test statistic Z_2; and σ^2 can be replaced by the sample variance in the first stage, s_1^2. The value of n_2 can be calculated to ensure $CP = 1 - \beta$.

1.2.3 Adaptive Model-Based Dose Finding

The main goal of an early-phase dose finding study is to establish the dose response relationship, which in turn provides estimates of the maximum tolerated dose (MTD) or the minimum effective dose (MED). This is also known as a dose escalation study, since the decision to escalate the experimental dose for subsequent cohorts of patients is based on the observation of dose limiting toxicity (DLT). A commonly used method is the rule-based *3+3*

conventional design. However, a more efficient method is to assume a prior dose response model such as a logistic model and use data from each cohort of patients to continuously update this model until the target dose is estimated with specified accuracy. This design is sometimes referred to as the *Continuous Reassessment Method (CRM)*, originally proposed by O'Quigley, Pepe, and Fisher (1990). Denoting the dose levels as x_i $(i = 1, 2, ..., k)$, and Y_j as the binary toxicity response for the jth patient, a one-parameter monotonic dose response model can be $E(Y_j) = \psi(x_i, a)$. We are interested in estimating x^* such that $\psi(x^*, a) = \theta$ where θ is the target probability of toxicity. Usually using a Bayesian framework, a prior distribution is elicited for a such as $\pi(a)$, then it is updated continuously with accumulating data, $\mathbf{Y}$, to obtain the posterior distribution, $\pi(a|\mathbf{Y})$ such that x^* can be estimated accordingly. A number of improved versions of CRM such as the CRM $-$ Escalation with Overdose Control (Babb, Rigatko, and Zacks, 1998) have also been proposed.

1.2.4 Seamless Designs

A seamless design combines two trials into one single trial and achieves objectives normally achieved in separate trials. For example, a seamless Phase 2/3 design is characterized by a single trial with a learning stage followed by a confirmation stage (Jennison and Turnbull, 2007). Maca *et al.* (2006) described it as a fusion of treatment selection techniques and hypothesis testing for integration, both operationally and inferentially, of phase 2 and 3 into a single trial. This type of design may incorporate elements of several adaptation rules. There are advantages in combining two trials into one. Seamless designs can substantially reduce time by eliminating the wait between two trials, and gain efficiency with a smaller sample size than that for two trials. However, feasibility may only be restricted to a primary endpoint with relatively short follow-up time. As for inference, the *P-value Combination Test* has been proposed to combine data from both stages (Posch *et al.*, 2005). Denoting p_1 as the p-value based on data from the first stage, if $p_1 \leqslant a$, then null hypothesis is rejected early, or if $p_1 > b$, then null hypothesis is accepted due to futility. If we let p_2 be the p-value

based only on the second stage data, and $C(p_1, p_2)$ be the combination function, then the null hypothesis is rejected if $C(p_1, p_2) \leqslant c$ and therefore we have

$$a + \int_a^b \int_0^1 I_{[C(x,y)\leqslant c]}\, dy\, dx = \alpha$$

where I is an indicator function. Examples of combination functions are (1) the method of product of p-values (MPP), $C(p_1, p_2) = p_1 p_2$, which is also known as the Fisher's criterion, (Bauer and Kohne, 1994) (2) the method of sum of p-values (MSP), $C(p_1, p_2) = p_1 + p_2$ (Chang, 2007), and (3) the method of weighted inverse normal combination of p-values (MINP), $C(p_1, p_2) = 1 - \Phi[\omega_1 \Phi^{-1}(1 - p_1) - \omega_2 \Phi^{-1}(1 - p_2)]$ where $\omega_1^2 + \omega_2^2 = 1$ (Lehmacher and Wassmer, 1999). The conditional power approach can be used to evaluate and compare the operating characteristics of these methods.

1.2.5 Decision Rule

This category summarizes any additional decision rules such as changing test statistic, re-designing multiple endpoints, selecting which hypothesis are to be tested (e.g. switching from superiority to non-inferiority), changing the hierarchical order of hypotheses, or changing the study population as the trial continues. Lang, Auterith, and Bauer (2000) proposed a two-stage adaptive trend test with scores corresponding to the unknown shape of a dose response curve updated using data from the first stage. Another example of adaptive modeling is choosing the right covariate at interim stage to be included at interim stage which can increase the precision of estimating treatment effects (Wang and Hung, 2005). Hommel (2001) investigated how one can modify the hypotheses in a trial after an interim analysis. He suggested possible modifications such as reducing the set of hypotheses, changing the weights of hypotheses, changing the *a priori* order of the hypotheses, or even adding new hypotheses. Jenkins, Stone, and Jennison (2011) proposed a seamless phase 2/3 design with a subpopulation selection at interim. Many of these adaptive decision rules require a learning stage before adaptations are planned and implemented for the subsequent stages.

1.2.6 Allocation Rule

Randomization or random allocation of patients is used to maximize balance of all known and unknown, observed and unobserved covariates (prognostic factors) at baseline between treatments. Fixed allocation rule uses allocation probabilities that are determined in advance and are not changed during the trial. However, an adaptive allocation rule dynamically alters the allocation probabilities to reflect the accruing data on the trial. Response-adaptive allocation uses interim data to modify the allocation probabilities in favor of the treatment arms showing superior outcomes. The earliest application of this method in a clinical trial is the *play-the-winner rule* proposed by Zelen (1969), although this was first studied by Robbins (1952). In some situations, the allocation rule removes inferior treatment arms completely from a further randomization schedule.

In summary, although statistical methods have been developed to allow for these types of adaptations, these methods should never be used as a substitute for careful planning in the statistical design of a clinical trials. Before the trial, an adaptive design must be detailed in the protocol to avoid the introduction of bias.

1.3 FDA Perspectives

1.3.1 Major Concerns

In response to the increasing popularity of adaptive methods in clinical trials, the FDA released a draft guidance titled "Guidance for Industry - Adaptive Design Clinical Trials for Drugs and Biologics" in February 2010 to unfold the agency's current perspectives on adaptive designs and to invite comments and suggestions from stakeholders. Gallo *et al.* (2010), members of the Pharmaceutical Research and Manufacturers of America (PhRMA), have delineated their viewpoints on the document in a white paper. These authors recognized the concerns raised by FDA and proposed recommendations to promote

discussion and learning between leaders from the drug development industry and from the regulatory authority.

In Section V of the document, the FDA encourages adaptive approaches that they described as *well-understood*. These approaches include adaptive exploratory studies on patient characteristics such as enrollment rate or eligibility criteria, adaptations that use only blinded interim data, interim analyses on an outcome unrelated to efficacy, and the classical group sequential methods. However, in Section IV of the document, the FDA cautions several areas of challenges in using adaptive design in drug development. First, adaptive design has the potential to increase the chance of a false positive conclusion. It is possible, during the adaptation process, for operational bias to be introduced. This operational bias is due to subjective decision-making during the course of a trial. One well-known example is selection bias. This type of bias is difficult to quantify and renders the study result difficult to interpret. Statistical bias due to statistical procedures such as selecting the best performing treatment out of many, by random chance, that is more favorable than the true value, can also over-estimate the true effect. Without proper control, the type I error can be inflated. Second, adaptive design, by shortening the trial duration or eliminating the wait time between trials such as in seamless combining of two traditionally separate trials, can limit identifying gaps in knowledge. Lack of extended time allocated to fully explore the data between trials may also lead to inadequate recognition of safety issues. On the other hand, a seamless design may allow longer follow-up time for safety assessment using subjects enrolled in the first stage, particularly when these safety issues are rare and take longer time to appear. Third, complex adaptive designs may potentially increase the trial planning time. Fourth, trial analyses and revisions that are not prospectively planned have the potential to increase false positive rate and difficulty in interpreting the study result. Any unplanned revisions to the conduct and analysis of the trial may result in a trial that is different from the one originally planned and may not answer the originally stated objectives and questions. Some adaptive methods that are prone to the above pitfalls are listed in Sections VI and VII. These methods that are all based on unblinding at interim in order

to estimate comparative treatment effects are termed as *less well-understood* designs. Cited examples of these less well-understood designs are outcome-adaptive randomization, sample size adjustment based on interim effect size estimate, modification of patient population enrolled, and endpoint selection based on interim treatment effect estimate.

1.3.2 Recommendations

Although the document made cautious statements regarding the application of adaptive methods, the FDA encourages their use in exploratory studies which have less impact on regulatory approval decisions and where type I error is of lesser importance. The remaining sections of the document attempt to provide constructive guidelines for scientists or trialists to consider when planning an adaptive clinical trials. Some of these suggestions are:

- Thorough understanding of the statistical properties and operating characteristics of the proposed adaptive design. This can be achieved through extensive computer simulations which are intended to characterize and quantify the level of statistical uncertainty in each adaptation and its impact on type I error, power, and bias.
- Early interactions with the FDA when planning and conducting an adaptive design. Sponsors who have questions about the adaptive designs should seek the FDA feedback and review comments from FDA.
- Adequate documentation of adaptations in a protocol, statistical analysis plan (SAP), and supportive documents. The SAP for an adaptive trial is likely to be more detailed and complex than for a non-adaptive trial.
- Prospective specification of study design and analysis, particularly where unblinded interim analyses are planned. Any ad hoc or retrospective adaptations will likely risk the inflation of type I error.

In the past few years, much statistical research has been conducted by statisticians from both academia and industry to characterize different types of adaptive designs. In addi-

tion, a surge in journal submission on the topics of adaptive designs was also observed, however extensive applications of adaptive methods still remain low. Kairalla *et al.* (2012) has attributed this slow absorption of adaptive methods into common practice to the lack of infrastructure and software implementation. This is clearly an area that may require arduous future effort.

Chapter 2

Approaches to Sequential Clinical Trial Monitoring

2.1 Introduction

In Chapter 1, we presented a brief overview on the various types of adaptive methods in clinical trials. A distinctive and common feature of these adaptive designs is that a trial is monitored sequentially in stages when inspection of the interim data is carried out. The primary purpose for monitoring clinical trials is to provide an ongoing evaluation of risk-to-benefit profile that addresses the uncertainty necessary to continue. Some of the reasons that may lead to early termination of a clinical trial are (1) experimental treatment is found to be convincingly better than the control, (2) experimental treatment is found to be convincingly worse than the control, (3) side effects or toxicity are too severe to continue, or (4) study integrity has been undermined. Therefore, interim monitoring is usually carried out by an independent and impartial group called the data and safety monitoring board (DSMB) or simply data monitoring committee (DMC).

Much methodological research on adaptive methods has mainly focused on proving early efficacy. In order to control the overall type I error, many methods of hypothesis testing have been proposed to achieve the objectives of a trial while preserving this error rate under a specified α level, whether the analytical approach is frequentist or Bayesian. The following sections will be devoted to review the statistical methods, both within the frequentist and

the Bayesian paradigms, that are used to sequentially monitor a clinical trial.

2.2 Frequentist Approaches to Sequential Monitoring

As seen in Section 1.2.4, the different types of combination tests described provide additional flexibility and adaptability to allow for trial monitoring based on independent stage-wise p-values, calculated using data from sub-samples. Another approach is the conditional power monitoring approach (Lan, DeMets, and Halperin, 1984). In this approach, a trial can be terminated based on the promising conditional power calculated using the interim statistic as defined in Section 1.2.2. Lan and Wittes (1988) transformed the Z-statistic into a statistic that is independent of the sample size, but is only dependent on the information fraction, and this statistic is called the B-value, which is used in the evaluation of the interim conditional power. Proschan and Hunsberger (1995) proposed the use of an increasing conditional error function which is based on the interim data. It specifies the amount of conditional type I error for the next stage and therefore, the average conditional error is controlled under α level. If the conditional error function is the same as the combination function, it is essentially the same as the combination test.

Another prominent and flexible approach to clinical trial monitoring is the use of the alpha (or error) spending plan which was proposed by Lan and DeMets (1983). As introduced in Section 1.2.1, an alpha spending function distributes the total probability of false positive risk as a continuous function of the information time or fraction in a sequential trial. In this case, the test statistic of choice can be the normal Z-statistic, stochastic increment W-statistic, or independent p-values. The corresponding stopping boundary values for a K-stage trial can be computed based on the distribution of the chosen statistic and the alpha spending function. The Haybittle-Peto boundary, which fixes a constant p-value for the first $K-1$ stages, can be used to monitor the p-values across the stages (Haybittle, 1971). The fixed boundary methods consider a shape parameter that determines the boundary shape.

The O'Brien-Fleming-type boundary and the Pocock-type boundary are examples of the fixed boundary method (O'Brien and Fleming, 1979; Pocock, 1977). Finally, the most flexible approach is the error spending function. The error spending function is usually expressed in term of the information fraction, t. The Lan-DeMets spending function (1983) has two forms:

$$\begin{aligned} \alpha(t) &= \begin{cases} 2 - 2\Phi\left(\frac{Z_{\alpha/2}}{\sqrt{t}}\right) & \text{for one-sided test} \\ 4 - 4\Phi\left(\frac{Z_{\alpha/4}}{\sqrt{t}}\right) & \text{for two-sided test} \end{cases} \\ \alpha(t) &= \alpha\, ln\,(1 + (e-1)t) \end{aligned}$$

where α is the overall level the trial type I error is controlled at. The first form resembles the O'Brien-Fleming-type boundary while the second one resembles the Pocock-type spending. The Gamma spending function was proposed by Hwang, Shih, and DeCani (1990). The functional form is

$$\alpha(t) = \begin{cases} \alpha\frac{(1-e^{-\gamma t})}{(1-e^{-\gamma})} & \text{if } \gamma \neq 0 \\ \alpha t & \text{if } \gamma = 0. \end{cases}$$

Negative values of γ yield convex spending functions that increase in conservatism as γ decreases, while positive values of γ yield concave spending functions that increase in aggressiveness as γ increases. The choice of $\gamma = 0$ spends the error linearly. When $\gamma = -4$, it resembles the O'Brien-Fleming boundaries, while $\gamma = 1$ produces boundaries that resemble those of Pocock. Power (also known as Rho) spending function was first proposed by Kim and DeMets (1987), and was further generalized. The function form is simply represented as

$$\alpha(t) = \alpha t^{\rho}, \rho > 0.$$

When $\rho = 1$, the corresponding stopping boundaries resemble the Pocock stopping boundaries, but when $\rho = 3$, the boundaries resemble the O'Brien-Fleming boundaries. Large

values of ρ yield increasingly conservative boundaries, while $0 < \rho < 1$ gives aggressive boundaries. These methods ensure the false positive rate is controlled under the α-level.

2.3 Bayesian Approaches to Sequential Monitoring

Bayesian methods for statistical monitoring of clinical trials have a growing appeal in the biopharamceutical industry and among clinicians. The theories of Bayesian inference are also well-established, but applications to modern clinical trials are still rare. Spiegelhalter, Freedman, and Parmar (1994) argued that "the Bayesian approach allows a formal basis for using external evidence and provides a rational way for dealing with issues such as the monitoring of accumulating data and the prediction of the consequences of continuing a study." Using a Bayesian inferential framework, there are two main methods in monitoring clinical trials, (1) posterior probability, and (2) predictive probability. Thall and Simon (1994) illustrated the use of posterior probability for a binary outcome x_i $(i = 1, 2, ..., n)$ to monitor a single-armed Phase 2 clinical trial with null hypothesis $H_0 : \pi_E \leqslant \pi_C + \delta$ comparing the experimental treatment E to control treatment C. After prior distributions are elicited for the probabilities of a response (π_E, π_C), the decision rule of the trial at interim stage k is based on posterior probability such that

$$\begin{aligned} P(\pi_E > \pi_C + \delta|\boldsymbol{x}) \geqslant \theta_U &\quad : \quad \text{reject } H_0 \text{ and terminate trial to declare success} \\ P(\pi_E > \pi_C + \delta|\boldsymbol{x}) \leqslant \theta_L &\quad : \quad \text{accept } H_0 \text{ and terminate trial to declare futility} \\ \theta_L < P(\pi_E > \pi_C + \delta|\boldsymbol{x}) < \theta_U &\quad : \quad \text{continue to next stage} \end{aligned}$$

where δ is the minimum difference in the probabilities and (θ_L, θ_U) are the stopping boundaries.

Dmitrienko and Wang (2006) described another approach to interim monitoring, the posterior predictive probability or simply the predictive probability. This method is also known as the Bayesian stochastic curtailment because it is analogous to the conditional power

approach in the frequentist paradigm (Spiegelhalter, Freedman, and Blackburn, 1986). The predictive probability is the probability of a successful trial if the trial hypothetically continues to the end, conditional on the current interim data. Since the predictive samples are not yet observed, the predictive probability is defined as an average over all possible values of the predictive samples. Using the same notation in the previous example, the predictive probability P^* is defined as

$$P^* = \int I\left\{P(\pi_E > \pi_C + \delta | \boldsymbol{x}, \boldsymbol{x}^*) > \eta)\right\} P(\boldsymbol{x}^*) d\boldsymbol{x}^*$$

where $I\{.\}$ represents an indicator function, $\boldsymbol{x}^*$ as the future samples, $P(\boldsymbol{x}^*)$ as the joint probability of the future samples, and η as a high pre-specified cutoff such that $\eta \in [0.85, 0.95]$, for instance. If P^* is greater than a upper stopping boundary θ_U, then the trial can be stopped based on convincing early evidence of efficacy, otherwise the trial can continue until it reaches the planned end if interim monitoring fails to reject the null hypothesis. Based on this definition, the monitoring using predictive probability has a similar interpretation of the conditional power, given the strength of the currently observed data. It is important to emphasize that the stopping boundaries (θ_L, θ_U) need to be calibrated through simulations to achieve desirable trial performance such as control of Bayesian type I error.

In the following sections, we will discuss two types of confirmatory clinical trials: multi-arm dose response trial and biosimilarity trial. After that, we will briefly introduce how selected sequential monitoring techniques described earlier can be applied to these two types of trials to achieve their respective trial objectives.

2.4 Multi-Arm Dose-Response Clinical Trials

During the development of a new drug, a phase 2 trial with multiple treatment arms consisting of a control dose (usually an active control) and a few experimental doses is conducted

to determine the therapeutic dose of the drug for a phase 3 confirmatory clinical trial. Note that we use the term *dose* interchangeably with *treatment arm*. The objectives of a phase 2 trial are to examine (1) whether the population means for the primary endpoint increase monotonically with the doses, (2) the shape of the dose-response function, and (3) what dose is appropriate as a therapeutic dose for use in a confirmatory trial (Wakana, Yoshimura, and Hamada, 2007). Selection of one or more doses to carry into confirmatory phase 3 trials is one of the most difficult decisions that needs to be made during drug development.

When multiple doses are considered, statistical methods have been proposed to combine both the objective of selecting one or more therapeutic doses and the objective of confirming these selected doses. Many of these designs are carried out in stages. Some of these designs extend the two-arm group sequential design to multiple arms with a control. Follmann, Proschan, and Geller (1994) proposed a design to monitor multi-armed clinical trials that strongly controls the type I error rate. At any interim stage, there is a collection of pairwise test statistics and critical values. Based on this information, the decision to retain or drop individual arms or to stop the trial entirely is made. At the end of the trial, it is possible more than one arm can be selected. When only one stage is considered, this procedure is reduced to the standard Dunnett's adjustment design. A more recent multi-armed group sequential drop-the-losers design is given by Chen, DeMets, and Lan (2010). In addition to the use of non-binding futility boundary, they suggested two methods of calculating the efficacy boundaries. The first approach is called joint monitoring in which all dose-control comparisons are monitored simultaneously by using one single alpha spending function, $\alpha^*(t)$, where t is the information fraction, $0 < t < 1$. Under the global null hypothesis,

$$
\begin{aligned}
& P\left(\begin{pmatrix} Z_1(t_1) < c_1, Z_1(t_2) < c_2, ..., Z_1(t_{k-1}) < c_{k-1} \\ Z_2(t_1) < c_1, Z_2(t_2) < c_2, ..., Z_2(t_{k-1}) < c_{k-1} \\ \cdots \\ Z_J(t_1) < c_1, Z_J(t_2) < c_2, ..., Z_J(t_{k-1}) < c_{k-1} \end{pmatrix} \bigcap \bigcup_{j=1}^{J} (Z_j(t_k) \geqslant c_k)\right) \\
= \quad & \alpha^*(t_k) - \alpha^*(t_{k-1})
\end{aligned}
$$

is assumed for J treatment arms at the kth interim stage where $j = 1, 2, ..., J$ and $k = 1, 2, ..., K$. The second approach is called marginal monitoring where a marginal alpha level α_j and an alpha spending function $\alpha_j(t)$ for *each* dose-control comparison is specified. Therefore, each dose j is monitored independently by its own alpha spending function. For example, under null hypothesis, H_{j0}:

$$P\left(Z_j(t_1) < c_1, Z_j(t_2) < c_2, ..., Z_j(t_{k-1}) < c_{k-1}, Z_j(t_k) \geqslant c_k\right) = \alpha_j^*(t_k) - \alpha_j^*(t_{k-1})$$

for $k = 1, 2, ..., K$ and $j = 1, 2, ..., J$. For both approaches, numerical integration is used to calculate the stopping boundaries (i.e. critical values) since the joint standardized test statistics follow multivariate normal distribution. This design allows more than one dose to be selected at the end of the trial.

When a trial design has two stages, with the first stage selecting doses and the second stage confirming the selected doses, it is sometimes known as the Seamless Phase 2/3 design or the drop-the-losers design described in Section 1.2.4. Sampson and Sill (2005) described a drop-the-losers design for a normal endpoint. In the first stage, n_1 subjects are randomized to each of the J experimental treatments and a control. At the end of the first stage, the sample means $\overline{x}_j$ are computed. The dose j^* that shows the empirically largest mean $\overline{x}_{j^*} = \overline{x}_{(1)} = max(\overline{x}_1, \overline{x}_2, ..., \overline{x}_J)$ will be selected for continuation into the second stage and n_2 subjects are randomized to the selected treatment and control. Final inference on the selected dose is based on the data from both stages.

In order to improve the selection of the minimum effective dose (MED) to be confirmed in the second stage, some methods assume a dose response model in the first stage and select the dose based on the estimated model. An example is given by Huang, Liu, and Hsiao (2011). These authors suggested estimating a linear dose response model at the end of the first stage such as

$$E(Y_{ji}) = \beta_0 + \beta_1 d_j$$

and test if $\beta_1 > c$ where c is a specified threshold. The dose that achieves an effect compared to a control dose by at least a magnitude of δ will be selected for the second stage.

2.5 Biosimilarity Clinical Trials

The second type of clinical trial we want to consider here is a biosimilarity trial. In 2010, the passage of the Biologics Price Competition and Innovation Act (BPCI) created an abbreviated licensure pathway in section 351(k) of the Public Health Service Act (PHS). This new law allows for an expeditious approval process for a generic follow-on biological product shown to be biosimilar to a licensed reference biological product. Due to their large and complex molecular structures, biological products are fundamentally disparate from small synthetic drugs, and so are their mechanisms of action. Traditional statistical methods used to test for bioequivalence as in a generic drug development may not be the most efficient way to establish biosimilarity. The FDA has released a guidance document called "Scientific Considerations in Demonstrating Biosimilarity to a Reference Product" in 2012. This document provides the definition of biosimilarity and some philosophical guidelines for the biopharmaceutical community in developing statistical methods for proving biosimilarity. Two important principles were suggested by the document: (1) *totality of evidence* approach, and (2) *step-wise* approach. This is often interpreted by the scientific community as a series of pre-clinical and clinical studies to show the plausibility of biosimilarity. A thorough review of the literature on recently proposed statistical methodologies can be found in Chapter 5.

2.6 Direction of Thesis

In this dissertation, we want to develop an innovative trial design for a multi-arm dose-response clinical trial as described in Section 2.4. The objective of this trial is to select the maximum biological dose (MBD) based on a dose response model and to increase efficiency of

the trial. However, when some prior qualitative knowledge of the dose response relationship is given such as from prior animal dose response studies, we can design a trial that allows the explicit prioritization of doses as clinicians may not consider all doses to be equally effective. In addition, we want to design this trial such that doses are added or inserted to the trial only if previous doses do not show statistical evidence of efficacy. This feature of adaptively adding doses may reduce the expected sample size needed since not all doses are used in the trial. We explore the operating characteristics of this adaptive design using a normal endpoint in Chapter 3. Under this design, we are able to show that by pre-specifying the order of the doses or hypotheses in decreasing order of efficacy, we can employ an alpha spending function that favors earlier doses to select the efficacious doses earlier. We also show how the design parameters can be flexibly varied and that it can be extended to both binary and survival endpoints in Chapter 4. When we allow more doses to be explored per interim stage or when we reduce the number of stages per dose, we may be able to gain better statistical power or to reduce the expected sample size.

Finally, we want to look into developing a testing framework for establishing biosimilarity in a confirmatory clinical trial. Since historical trials are involved in designing a biosimilarity trial as the reference innovator biological product was approved in the past, a Bayesian framework will be a logical choice to allow the synthesis of historical and current evidence. We propose an adaptive two-stage design that uses predictive probability as a monitoring criterion. We show that it provides better control of type I error than the frequentist approach even when the constancy assumption is slightly violated and that it gives better power than frequentist approach when biosimilarity is highly plausible. This will be presented in Chapter 5.

Chapter 3

Adaptive Staggered Dose Design for a Normal Endpoint

3.1 Introduction

Clinical drug development usually follows distinct phases and in each phase one or more trials are conducted to answer a specific set of questions. For example, within phase 2, a small phase 2A study such as a proof-of-concept study is conducted to establish the dose to efficacy relationship. If efficacy is established, a phase 2B dose-ranging study is conducted to estimate the shape of the dose to efficacy relationship and to find an optimal dose that can be carried to a phase 3 confirmatory trial. However, in most practical situations, when there is still uncertainty about the efficacy and safety of this one dose, due to the use of surrogate endpoint in dose-ranging study, the investigational team may be more inclined to keep several potential doses rather than only one definitive dose. In other situations, the team may want to compare several drug schedules or regimens instead of doses. In this case, a single trial design that combines both the selection of a dose from multiple doses as in a dose-ranging trial and the confirmation of the selected dose as in a confirmatory trial is desirable.

This idea has given rise to the development of a class of adaptive trial designs that seamlessly

combines both the selection of a therapeutic dose out of multiple doses and the confirmation of the selected dose. This type of trial design is sometimes referred to as the Seamless Phase 2/3 design (Gallo *et al.*, 2006; Maca *et al.*, 2006). These designs allow trial adaptation such as selection of hypothesis, subpopulation, or treatment after the first stage and aim at controlling the experiment-wise type I error as required in a confirmatory trial. In our present context, we focus only on the seamless designs that select treatment at interim.

One of the many approaches is the two-stage ranking and selection procedure and is sometimes called the drop-the-losers design, first described by Sampson and Sill (2005). In this design, k experimental treatments and a control are administered in the first stage. At interim, the empirically best treatment is selected for continuation into the second stage, along with the control. At the end of the second stage, inference comparing the selected treatment and the control is conducted using the data from both stages. More recently, other two-stage designs were also proposed (Li *et al.*, 2009; Wang *et al.*, 2011). Depending on the ranking results at the end of the first stage, these designs allow the selection of more than one dose to be carried to the second stage. They also allow for early termination of the study if at least one dose shows statistical significance at the end of the first stage.

Another approach is the multi-stage group sequential procedure. These are designs that drop inferior treatments based on testing at interim. Follmann, Proschan, and Geller (1994) extended the two-arm group sequential design to multiple arms and provided critical values that could strongly control the type I error. Other similar designs that used the efficient score statistics developed by Whitehead (1997) also appeared in the literature. These designs allowed for sequential monitoring for any type of outcome - binary, normal, and survival (Stallard and Todd, 2003; Stallard and Friede, 2008). The major challenge in implementing multi-stage group sequential methods involves the calculation of stopping boundaries. The most recent design by Chen, DeMets, and Lan (2010) suggested two methods of calculating the efficacy boundaries. The first approach is called joint monitoring in which all dose-control comparisons are monitored simultaneously by one single alpha

spending function. The second approach is called marginal monitoring where a marginal alpha level and an alpha spending function are specified for each dose-control comparison. Computation of stopping boundaries remains intensive in this approach.

Besides ranking and hypothesis testing as methods of interim treatment selection, other innovative approaches were also found in the literature. Kimani, Stallard, and Hutton (2009) considered the selection of dose in a Seamless Phase 2/3 trial using criteria that incorporated both efficacy and safety. They proposed a Bayesian method that used prior distribution of dose response relationship to inform the selection of dose for the confirmation stage. Huang, Liu, and Hsiao (2011) considered estimating and testing for a linear dose response model to inform the selection of the best dose for the second stage. Wang and Cui (2007) applied outcome-adaptive randomization to dose selection. The allocation ratio is continuously updated based on calculation of conditional power with higher probability of assigning subjects to dose groups with higher conditional power. Friede *et al.* (2011) designed a seamless design that used early surrogate endpoint to aid treatment selection. Bretz, Pinheiro, and Branson (2005) proposed combining multiple comparisons and parametric modeling techniques in selecting both a model and a dose that gives a desired level of efficacy. Although the above adaptive approaches have made significant strides in improving statistical power and reducing the expected sample size, the derivation of the distribution of test statistics and the computation of stopping boundaries remain intensive, particularly for the rank-based selection procedure. Here we are proposing a staggered dose design that starts off with only one or a subset of the doses, and depending on interim results, remaining doses may be added to the trial if the previous doses do not show evidence of efficacy. This design is proposed for meeting the objectives of late phase 2 and confirmatory phase 3. It actively incorporates information from previous dose-ranging studies and allows for the prioritization of doses, while keeping type I error under nominal level of α.

It is important to note that, when planning a trial, sometimes the clinical team may not consider all candidate doses to be of equal importance, but would be interested in exploring

these doses one after the other starting with doses with assumed better responses and proceeding to doses with uncertain responses if the earlier doses do not show statistical evidence of efficacy. This is especially true when some prior qualitative evidence of dose response relationship exists, and investigators are willing to consider J doses which can be arranged in decreasing order of priority. We can let $j = \{1, 2, 3, ..., J\}$ be this order of priority with one being of the highest priority and J the lowest priority. We can represent the actual prioritized dose levels as $d_1, d_2, ..., d_J$. There is a need to emphasize that this order of priority does *not* necessarily imply an increasing ($d_1 < d_2 < ... < d_J$) or decreasing ($d_1 > d_2 > ... > d_J$) order of dosage. This *a priori* ordering that is based on previous knowledge offers additional flexibility for the team to explore the doses one after the other knowing that we only have a limited number of patients in a dose selection and confirmation trial. Traditional fixed dose parallel group design and some adaptive designs treat all doses to be equally important and the experiment randomizes patients equally to all of the doses. These designs may be inefficient when an informative prior model on the dose response relationship exists. The designs reviewed earlier, although making improvement in dose selection, do not consider the option of adaptively inserting new doses.

By adaptively allowing the option of adding new doses, we propose a design that can start off with fewer doses and sequentially drop inferior doses and add new doses at the same time. This new design, *adaptive staggered dose procedure*, may further reduce the expected sample size while gaining information about the efficacious doses more quickly. One major condition is that the clinical team has to provide an assumed best case of dose ordering by ranking the candidate J doses in decreasing order of priority based on their clinical judgment or previous evidence on dose response. The gain in design efficiency necessitates this assumed dose ordering because potentially better doses will be studied earlier. We define the optimal dose as the biological dose that has the highest efficacy. If the ordering is right, this design may perform better than both the current drop-the-losers design and the fixed dose parallel group design in terms of reduced expected sample size and perhaps experimental time. If the *a priori* ordering is weak, this proposed design can perform as

good as the drop-the-losers design but generally still better than the fixed dose parallel group design in most scenarios under some assumed conditions.

The following sections are organized as follows. A general version of the proposed adaptive staggered dose design will be described in Section 3.2.1. We will illustrate this design using a specific version in Section 3.2.2. In Section 3.3, we expound more on other important design considerations such as futility, type I error control, and error spending functions. The operating characteristics of this proposed design will be examined in detail in Section 3.4. Based on the simulation results, we will discuss the strengths and weaknesses of this design and recommend its suitable applications in Section 3.5

3.2 Adaptive Staggered Dose Design

3.2.1 General Design

As a motivating example, we can consider a topical ophthalmic solution containing a new histamine receptor antagonist to prevent ocular itching due to allergic conjunctivitis. A trial is conducted to select one from several safe dosages (e.g. 0.15%, 0.2%, 0.25%, and 0.3%) and to confirm the strength of the selected dosage against its placebo vehicle (0%). The primary efficacy endpoint is an Ocular Itching Score (OIS) at 7-minute post-dosing after ocular allergen challenge (CAC). For illustration purpose, we assume higher score corresponds to symptom improvement, i.e. lessened itching. Therefore a trial design that selects one or two optimal doses from J possible doses for late-phase pivotal trials will be useful. These J doses can be arranged by the investigators in the order of *decreasing* priority $j = 1, 2, ..., J$ as $\{d_1, d_2, ..., d_J\}$. It is worth-noting that $\{d_1, d_2, ..., d_J\}$ are not necessarily in increasing or decreasing order of dosage. It is generally assumed that these doses are well within the range of acceptable safety. Acceptable safety may refer to adverse events that are mild and reversible or that are not related to treatment. Although if the team decides that it is important to escalate the doses for simultaneous assessment of safety and efficacy,

a dose escalation order can be adopted, and in this case, $d_1 < d_2 < ... < d_J$. In addition, we want to compare these doses with a control dose $d_0\,(j = 0)$ and adopt a randomization ratio of $1 : R$ for control to each of the experimental doses. The primary clinical endpoint such as OIS above is assumed to be normally distributed with known common variance ($\sigma^2 = 1$) across dose groups. We can let Y_{ji} represent the endpoint in the ith subject receiving dose d_j such that

$$Y_{ji} \sim N(\mu_j, 1). \tag{3.2.1}$$

If μ is related to d, then a dose response function may exist such that $\mu_j = f(d_j)$. However, $f(d_j)$ may or may not be a monotonic function.

In this design, we consider looking at a maximum of D experimental doses ($D < J$) at each of the K *global* stages. Therefore, we do not start off with all J doses but a subset containing the first D doses, $\{d_1, d_2, ..., d_D\}$. Depending on the assessment of efficacy at interim stages, we drop doses from the D doses that show convincing futility or lack of efficacy and simultaneously add the next doses from the pre-specified dose ordering to maintain D current doses in the next interim stage. If the interim results demonstrate that at least one dose shows evidence of efficacy, the trial can be stopped early and no more doses will need to be added. Therefore, a maximum of D doses may be selected before or at the final (Kth) stage or none of the J doses shows efficacy at the final stage. If no dose is selected and confirmed at the end of the experiment, this drug development program will halt. We restrict the number of patients allocated to each experimental dose by setting the same minimum and maximum numbers of subjects allocated. For example, if we let c be the cohort size per dose and per stage, then the minimum number of subjects will be c and the maximum number of subjects will be cM, where M is the maximum number of *per-dose* stages. However, the control dose will always have c/R subjects randomized per stage. It is important to emphasize that only the control subjects that are randomized *simultaneously* with the corresponding experimental dose are compared to the experimental subjects. This

ensures the control subjects and the experimental subjects are comparable in all known and unknown prognostic factors except the doses assigned.

3.2.2 Specific Version of Design

To illustrate the decision rule of this design, we consider a specific version of the design where we study one dose at a time ($D = 1$), with a maximum of two per-dose stages ($M = 2$), and the maximum number of global stages of $K = 2J$. Setting $K = MJ = 2J$ allows the experiment to fully explore each of the doses with an equal amount of information, if added to the trial. Figure 3.1 graphically illustrates this specific adaptive staggered dose algorithm for $J = 4$ and a control to dose randomization ratio of $1 : R$ at each stage. The number of global stages K is necessarily bounded, $J \leqslant K \leqslant MJ$. The lower bound J ensures that each of the J doses will have at least one per-dose stage, while the upper bound ensures that each dose will have at most M per-dose stages. If the set of one-sided hypotheses are

$$H_{j0} : \mu_j \leqslant \mu_0, \;\; H_{ja} : \mu_j > \mu_0 \tag{3.2.2}$$

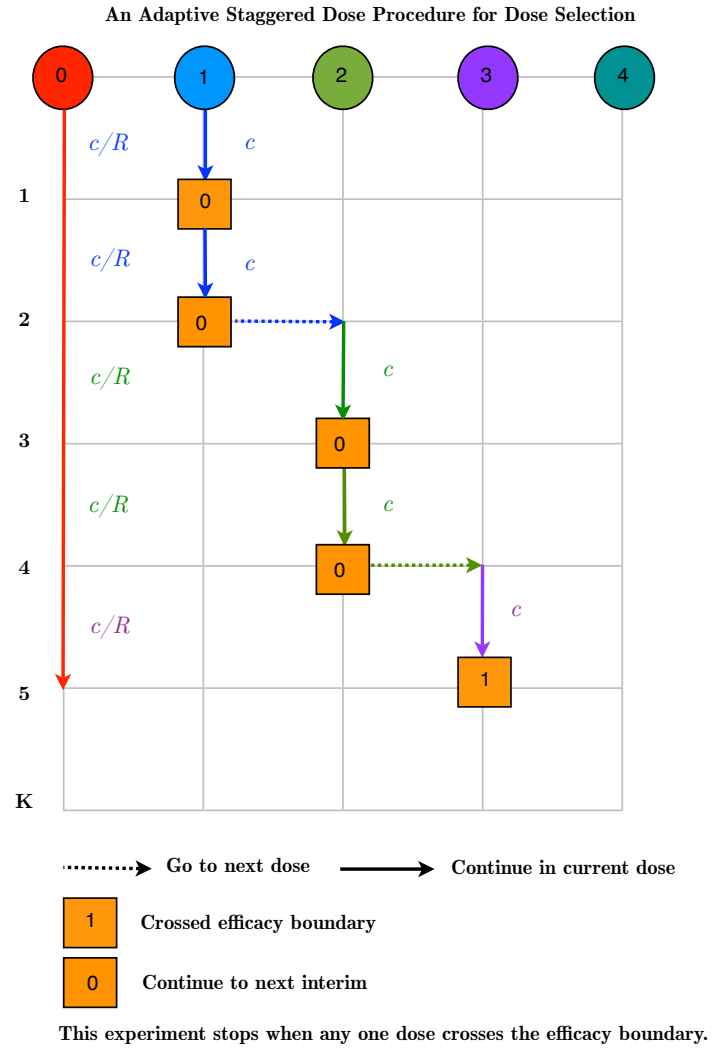
for $j = 1, 2, ..., J$, then the standardized test statistic for dose d_j, conditioned on the outcomes of previous doses, will be

$$Z_{jm} = \frac{\overline{Y}_{jm} - \overline{Y}_{0m}}{\sqrt{\frac{R+1}{cm}}} \tag{3.2.3}$$

where $\overline{Y}_{jm} = \sum_{i=1}^{cm} Y_{ji}/(cm)$ and $\overline{Y}_{0m} = \sum_{i=1}^{(cm)/R} Y_{0i}/(\frac{cm}{R})$ for $m = 1, 2$. It is important to stress that Y_{0i}'s are the responses only for control subjects that are randomized simultaneously with the experimental subjects for dose d_j. The control subjects randomized simultaneously with subjects to the previous doses, $d_1, ..., d_{j-1}$, will not be included in the calculation of this test statistic since they may not be comparable in prognostic factors. The conditional distributions of these statistics are

$$Z_{j1} \sim N\left(\frac{\mu_j - \mu_0}{\sqrt{\frac{R+1}{c}}}, 1\right), (Z_{j1}, Z_{j2})' \sim N_2\left(\begin{pmatrix} \frac{\mu_j-\mu_0}{\sqrt{\frac{R+1}{c}}} \\ \frac{\mu_j-\mu_0}{\sqrt{\frac{R+1}{2c}}} \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 1 \end{pmatrix}\right). \quad (3.2.4)$$

Figure 3.1: A graphical illustration of the proposed adaptive staggered dose procedure with $J = 4, D = 1, M = 2, R = 2$ and $K = 2J = 8$.



Under this specific design, the decision rule for dose d_j at the kth interim is described as follows.

1. If, whether $m = 1$ or 2, the test statistic for dose d_j is $Z_{jm} > b_k$ where b_k is the efficacy boundary for the kth interim analysis, then it is declared statistically significant and H_{j0} is rejected. This trial will stop and the remaining doses $\{d_{j+1}, ..., d_J\}$ will not enter the trial.

2. If, when $m = 1$, the test statistic for dose d_j at its first per-dose interim is $a_k < Z_{j1} \leqslant b_k$, where a_k ($a_k < b_k$) is the futility boundary for the kth interim analysis, then it will continue to its second per-dose stage with an additional cohort of c subjects allocated. If, when $m = 2$, the test statistic is $Z_{j2} \leqslant b_{k+1}$, where $a_{k+1} = b_{k+1}$, then this dose has reached its maximum allowable samples of $2c$ and it will be dropped due to lack of efficacy. At the same time, a new dose d_{j+1} in the next priority will be added to the experiment.

3. If, when $m = 1$, the test statistic for dose d_j is $Z_{j1} \leqslant a_k$, then it is dropped due to convincing futility and H_{j0} is accepted, and a new dose d_{j+1} in the next priority will be added to the experiment.

Therefore, at any global stage k, only one dose is being considered. This trial design allows early termination if (1) any one dose is declared efficacious, (2) all of the J doses are declared futile, or (3) no dose is declared efficacious at the end of the Kth (final) stage.

Under this staggered dose selection procedure, we can achieve greater gain in design efficiency if we have strong prior knowledge about the dose response relationship. In this adaptive staggered dose design, we do not need to assign patients to all of the doses at the beginning. At the end of the trial, if we cannot select a significantly efficacious dose, then this drug development program will stop. Since the trial can stop at any global stage for futility or efficacy, the number of stages the trial goes through before stopping is therefore a random variable and so is the sample size used. For a pre-specified K, the total number of planned subjects for this entire trial is therefore equal to $cK(\frac{1}{R} + 1)$. The operating characteristics of this proposed design will be examined in the next sections.

3.3 Other Design Considerations

3.3.1 Futility Analysis

In the specific design described in Section 3.2.2, we want to drop the interim analysis of futility at the moment and focus only on the proof of efficacy. In this case, we let $a_k = -\infty$ and hence $P(Z_{jm} \leqslant a_k) = 0$ for all j and $m = 1$. Therefore, each dose will always go through all M per-dose stages if no statistical evidence of efficacy is shown at interim stages. Investigators can still drop a dose before its Mth per-dose interim stage, if safety issues arise, without inflating the family-wise type I error, and this can provide additional flexibility. In this setting, dose d_j is considered at only two global interim stages: $k = 2j-1$, if k is odd, or $k = 2j$ if k is even (see Figure 3.1). The general form of the type I error at the kth stage will be derived in the next section.

3.3.2 Family-wise Type I Error

For this specific design without futility analysis, we denote the probability of rejecting the null hypothesis H_{j0} for dose d_j at the kth interim as $\psi_{j,k}$. Under H_{j0}, the standardized test statistics for dose d_j, using (3.2.4) for all j, will have the following distributions:

$$Z_{j1} \sim N(0,1)\,,\ (Z_{j1}, Z_{j2})' \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 1 \end{pmatrix}\right). \tag{3.3.1}$$

Since these distributions, under null hypothesis, are the same for all doses, we can simply replace Z_{j1} with Z and (Z_{j1}, Z_{j2}) with (Z_1, Z_2). As we have seen, when no futility analysis

is performed, the probability of rejecting H_{j0} given it is true can be given by

$$\psi_{j,k} = \begin{cases} P(Z > b_k) & \text{if } k = 1 \\ P(Z_1 \leqslant b_{k-1}, Z_2 > b_k) & \text{if } k = 2 \\ \left(\prod_{i=1}^{\frac{k-1}{2}} P(Z_1 \leqslant b_{2i-1}, Z_2 \leqslant b_{2i})\right) P(Z > b_k) & k = 2j-1 \\ \left(\prod_{i=1}^{\frac{k-2}{2}} P(Z_1 \leqslant b_{2i-1}, Z_2 \leqslant b_{2i})\right) P(Z_1 \leqslant b_{k-1}, Z_2 > b_k) & k = 2j. \end{cases} \tag{3.3.2}$$

This follows from the design setting that the test statistics Z_{jm}'s for any given dose d_j are calculated only using the responses from the control cohort that are randomized simultaneously with this dose but not the responses from previous doses, they are considered independent from the test statistics of previous doses. The family-wise type I error under the global null hypothesis for this trial is therefore given by

$$\psi = \sum_{i=1}^{J} (\psi_{i,2i-1} + \psi_{i,2i}) \tag{3.3.3}$$

and we are interested in keeping it under a target alpha level, $Sup\, \psi \leqslant \alpha$. The next section will discuss how the set of efficacy stopping boundaries $(b_1, b_2, ..., b_K)$ is derived.

3.3.3 Alpha Spending Functions

Classical group sequential methodology employs flexible *alpha spending functions* to monitor the trial's test statistic as multiple interim looks during a trial can cause inflation of type I error (Armitage, McPherson, and Rowe, 1969; Lan and DeMets, 1983). However, it is important to distinguish the use of alpha spending function in the traditional group sequential setting versus in this adaptive staggered dose design setting. In the traditional two-arm trial setting, alpha spending function is applied to monitor the test statistic of one dose-to-control comparison as information accrues through the stages in order to control type I error. However, in this adaptive staggered dose selection design, since earlier doses are

investigated at earlier stages and later doses at later stages, the use of one alpha spending function across all global stages monitors test statistics of the dose-to-control comparisons in a staggered fashion. Therefore, this design presents a novel application of the alpha spending function. For the kth interim, if $\alpha(t)$ represents the alpha spending function with t $(0 < t < 1)$ being the information fraction, then the doses and their corresponding hypotheses can be monitored through

$$\psi_{j,k} = \alpha(t_k) - \alpha(t_{k-1}) \tag{3.3.4}$$

where $\alpha(0) = 0, \alpha(1) = \alpha$, $t_k = k/(2J)$, and $k = 1, 2, ..., 2J$ since c, the cohort size per stage and dose, is a constant. Alternatively, one can flexibly re-define the cohort size parameter c as information fraction t_c during the first interim stage given a maximum samples allowable for each dose. For either parameterization method, the efficacy stopping boundary values can be computed via numerical method using the null distributions in (3.3.1).

Some of the common alpha spending functions used in traditional group sequential designs such as the Pocock, O'Brien & Fleming, Lan & DeMets and the Rho alpha spending functions can be used in this adaptive design. Rho alpha spending scheme, a one-parameter function, offers more flexibility by simply adjusting the parameter ρ in

$$\alpha(t) = \alpha t^{\rho}. \tag{3.3.5}$$

If $\rho < 1$, it allocates more alpha to earlier stages, hence favoring earlier doses; while for $\rho > 1$, it allows more alpha spent at later stages, and thus favoring later doses by pushing the trials to later stages. In fact, Pocock's approach is similar to the former and O'Brien & Fleming's approach to the latter. The choice of a suitable value for ρ reflects how optimistic or conservative the investigators are toward the ordering of the doses, as well as how informative the dose response relationship $\mu_j = f(d_j)$ they are willing to assume.

It is important to emphasize that once a dose is accepted or its null hypothesis is rejected, the

remaining doses cannot enter the trial nor their hypotheses be tested as this may inflat the type I error. It is because the stopping boundaries derived are based on the condition that the previous doses do not cross their respective stopping boundaries. Under this design, we can assert that by controlling the probability of a false positive conclusion under the *global null hypothesis*, that is, when all H_{j0}'s are true, we have strong control of the family-wise type I error. This assertion is proved in Section 3.6.4 of the Appendix.

Another alternative to jointly monitoring using one global alpha spending function is to allow different alpha spending functions $\alpha_j(t)$'s for different doses d_j's since we know exactly which dose is investigated at the kth stage. In this case, the stopping boundaries for dose d_j will not only depend on the alpha spending function specified for dose d_j, which is $\alpha_j(t)$, but also on the specified alpha levels of the previous doses, $(\alpha_1, \alpha_2, ..., \alpha_{j-1})$. This will be explored in Chapter 4.

3.3.4 Efficacy Stopping Boundaries

Given the general form of the type I error in (3.3.2), and a chosen alpha spending function in (3.3.4), we can numerically solve for the set of efficacy stopping boundaries, $\boldsymbol{b} = (b_1, b_2, ..., b_k, .., b_K)$ using the null distributions in (3.3.1). Numerical methods to evaluate probabilities of multivariate normal distributions have been proposed, such as the *GenzBretz method* proposed by Genz *et al.* (2011) implemented in the $\mathcal{R}$ package, *mvtnorm*.

3.3.5 Expected Stages and Sample Sizes

One of the major advantages of this adaptive staggered dose procedure is stopping the trial early if a dose showing evidence of efficacy is selected. In addition, the doses following the selected dose under the given dose ordering do not have to enter the trial, and therefore, saving patients from further allocation to potentially inferior doses. Under the proposed design, the number of stages the trial goes through before stopping for efficacy, $\mathcal{K}$, is a

random variable. The expected number of stages is given by

$$E(\mathcal{K}) = 2J - \left[\sum_{j=1}^{J-1}((2J-2j+1)\xi_{j,2j-1} + (2J-2j)\xi_{j,2j})\right] - \xi_{J,2J-1} \tag{3.3.6}$$

where $\xi_{j,k}$ is the probability of rejecting the null hypothesis H_{j0}, evaluated under the alternative hypothesis H_{ja}. In other words, $\xi_{j,k}$ is the statistical power for testing dose d_j against the control dose at the kth interim stage. The mathematical forms of $\xi_{j,k}$ and $E(\mathcal{K})$ are given in Sections 3.6.2 and 3.6.3 of the Appendix. We can see from (3.3.6) that $E(\mathcal{K}) \ll 2J$ if the stopping probabilities, $\xi_{j,k}$'s are large. The sample size $\mathcal{S}$ for this specific design is also random and its expectation is given by

$$E(\mathcal{S}) = c\left(\frac{1}{R}+1\right)E(\mathcal{K}), \tag{3.3.7}$$

since the control arm will always require an allocation of c/R subjects per stage. Table 3.1 summarizes the six design parameters of a general version of this staggered dose design.

Table 3.1: Definitions of design parameters of the adaptive staggered dose procedure

Parameter	Definition
J	Total number of experimental doses considered in the entire trial
D	Maximum number of experimental doses under study at each interim stage ($D < J$)
M	Maximum number of interim stages allowable to each experimental dose (i.e. *per-dose* stage)
c	Cohort sample size allocated to an experimental dose per interim stage
K	Total number of interim stages including final stage of the entire trial (i.e. *global* stage)
R	Randomization ratio of control dose to experimental dose is $1 : R$

3.4 Simulation Study

3.4.1 Simulation Plan

In this section, we want to investigate the operating characteristics of this proposed adaptive staggered dose procedure using extensive simulations. We consider the specific trial design illustrated in Figure 3.1 using the same design parameters. In this specific procedure, we have four experimental doses ($J = 4$), maximum of eight global stages ($K = 2J = 8$) and only one dose being compared to the control at each interim stage ($D = 1$). Each experimental dose will have a maximum of two per-dose stages ($M = 2$). A fixed randomization ratio of $1 : R = 1 : 2$ is used for the control dose to each of the experimental doses throughout the trial. In other words, if c is the cohort size for each experimental dose per stage, then the control will have a cohort of $c/2$ per stage.

For operating characteristics, we are interested in evaluating, for a given dose response model, the cohort size c required to attain statistical powers of 0.8 and 0.9 under different combinations of dose orderings and error spending schemes. This cohort size c can be evaluated numerically using the mathematical form of statistical power stated in Section 3.6.2 of the Appendix. This cohort size c will guide the research team to decide and plan for the recruitment of the total sample size for the entire trial under given pre-specified scenarios. As we have seen earlier, this design allows for early stopping when an experimental dose is selected due to statistically significant evidence of efficacy. Therefore, the number of global stages the trial goes through before stopping for efficacy is a random variable and so is the sample size used. Therefore, we want to know the expected number of global stages and expected trial sample size using the cohort size corresponding to each of the dose orderings and error spending schemes. Also, the experimental doses will be selected with different probabilities under different dose orderings and error spending schemes. We also want to characterize the variation of these dose selection probabilities, particularly the probability of selecting the dose with best efficacy.

We choose the following dose levels, $d = 0, 2, 4, 6$, and 8 for illustration. They can correspond to studying $0\%, 0.15\%, 0.2\%, 0.25\%$ and 0.3% concentrations in our previous example of topical ophthalmic drop for ocular itching. The first dose level of zero refers to the control dose, d_0, with $\mu_0 = f(d_0 = 0) = 0$ and the second to fifth are the four increasing levels of the experimental drug. We consider four dose response models - Flat, Linear, Emax, and Umbrella, for the alternative hypothesis. Table 3.2 describes these four dose response curves, $E(Y_{ji}) = \mu_j = f(d_j)$. These four curves are chosen because they are the commonly assumed dose response models in clinical trials (Bornkamp *et al.*, 2007; Antonijev *et al.*, 2010). Although not common, non-monotonic umbrella dose response has been observed in the therapeutic targeting of angiogenesis in cancer (Reynolds, 2010). Also when doses are high, downturn in response is common when cytolethality occurs (Simpson and Margolin, 1986; Combes, 1997).

Table 3.2: Four dose response models with $\mu_0 = f(d_0) = 0$

Dose Response	$\mu_j = f(d_j)$	$f(d_0, d_1, d_2, d_3, d_4) = f(0, 2, 4, 6, 8)$
Flat	$\mu_j = 0.35$	(0.000, 0.350, 0.350, 0.350, 0.350)
Linear	$\mu_j = 0.04375 d_j$	(0.000, 0.088, 0.175, 0.263, 0.350)
Emax	$\mu_j = 0.4375 \frac{d_j}{2+d_j}$	(0.000, 0.219, 0.292, 0.328, 0.350)
Umbrella	$\mu_j = 0.117 d_j - 0.0097 d_j^2$	(0.000, 0.194, 0.311, 0.350, 0.311)

We consider looking into three ordering schemes for the four experimental doses - dose escalation, informative, and uninformative orderings. In dose escalation ordering, these four experimental doses are ordered in increasing dosage such that $(d_1, d_2, d_3, d_4) = (2, 4, 6, 8)$. In this case, the trial proceeds from the lowest dose to the highest dose. If a strong dose response model based on pre-clinical or earlier clinical experience exists, we can adopt an informative ordering. In informative ordering, we arrange the doses in the presumed order of decreasing priority. For example, for the umbrella dose response model described above,

we can order the four experimental doses such that the best dose is explored earlier, then $(d_1, d_2, d_3, d_4) = (6, 4, 8, 2)$. In this case, we prefer to study $f(d_1 = 6) = 0.35$ first and $f(d_4 = 2) = 0.194$ last. In uninformative ordering, we will use all of the 24 ($J! = 4!$) permuted orderings and take the average of all the calculated cohort sizes. This is the expected cohort size required if we randomly pick an ordering out of the 24 permuted orderings with equal probabilities. This simulates the situation when there is no prior knowledge of the dose response relationship.

Next, we select eight different error spending plans. We include the Pocock, O'Brien & Fleming error spendings, Rho error spendings with parameter ρ pre-specified at $0.3, 0.5, 1, 2$, and 3, and a final error spending that fixes a constant efficacy stopping boundary b_k across all $K = 8$ stages. For Rho error spending, we choose the above values of ρ to study how a wide spectrum of error spending can affect the cohort size, the expected number of stages to stop the trial and the probabilities of dose selection. The family-wise type I error is set at one-sided $\alpha = 0.05$ for all of the selected error spending plans. The objective of this part of the simulation study is that we can compare the operating characteristics for different dose orderings and error spending plans for a given dose response model.

In the next part of the simulation study, we would like to compare this adaptive staggered dose design with three comparator designs: the traditional parallel group design using Dunnett's adjustment (Dunnett, 1955), the two-stage drop-the-losers (pick-the-winner) design, and a two-stage seamless design with dose selection based on an informative dose response in the first stage. We use the expected trial sample size evaluated under uninformative ordering for a given dose response model and a given error spending plan attaining power of 0.9 as the trial sample size to simulate the statistical power of the first two comparator designs. We denote the expected trial sample size as $E(\mathcal{S})$ as in (3.3.7). When simulating the statistical power of the last comparator design, the two-stage dose-response informed seamless design, we use the expected trial sample size evaluated under informative ordering for the proposed adaptive procedure since we assume informative prior dose response model

here. The simulated powers for the three comparator designs can be compared to the target power of 0.9 under the proposed adaptive procedure.

In the parallel group design, we use balanced allocation ratio for the control dose and the four experimental dose groups. Therefore, each arm will receive $E(\mathcal{S})/5$ subjects. If at least one dose is found to show statistical evidence of efficacy under the Dunnett's multiplicity adjustment with one-sided $\alpha = 0.05$, then the trial will be declared a success. For the drop-the-losers design, we have two stages. In the first stage, balanced allocation ratio is used for the control dose and the four experimental dose groups with each group receiving w subjects. At interim, the experimental dose, which demonstrates the largest sample mean, will continue to the second stage with the control dose, while the remaining three experimental doses will be dropped from the trial without further randomization. This second stage will use balanced allocation ratio to randomize the remaining subjects to both the selected dose and the control dose, with each receiving another w subjects. Therefore, w is equal to $E(\mathcal{S})/7$. There is no testing of hypothesis at the first interim stage, but only dose selection. However, the test statistic at the final stage comparing the selected dose and the control dose will use all $2w$ subjects allocated in both stages. The efficacy stopping boundary for the final stage, evaluated by simulation, will keep the family-wise type I error at one-sided $\alpha = 0.05$. If this selected dose shows evidence of efficacy, then this drop-the-losers trial is a success. For the two-stage dose-response informed design, similar to the drop-the-losers design, $w = E(\mathcal{S})/7$ subjects will be randomized to each dose. At the end of the first stage, the parameters of the dose response curve, which is assumed to be known, will be estimated using linear or non-linear regression analysis. The dose with the highest estimated mean response based on the dose response model estimated using the data from the first stage will be selected for the second stage. Only the selected dose and the control will continue to the second stage for further randomization. Observed responses from both stages will be combined for final analysis. However, no hypothesis testing is performed at the end of first stage as the first stage is the learning and dose selection stage. If the selected dose demonstrated evidence of efficacy over control at the end of second stage, the trial is

considered a success. Lastly, we will simulate data from normal distribution with common known unit variance ($\sigma^2 = 1$) and the number of simulated trials is 10,000 unless otherwise stated. All of the simulations are conducted in $\mathcal{R}$ software.

3.4.2 Efficacy Stopping Boundaries

Table 3.3 shows the stage-wise efficacy boundaries and their corresponding alphas spent using the eight selected error spending plans under the global null hypothesis. These boundaries values are evaluated numerically. We have seen earlier that, without futility analysis at interim stages, dose d_1 is studied at interim stages $k = 1, 2$; dose d_2 at $k = 3, 4$ and so on. The constant efficacy boundary scheme offers similar α_k spent for each of the four doses. Dose d_1 gets an alpha of 0.0127, d_2 of 0.0126, d_3 of 0.0124 and d_4 of 0.0122. In this case, we are not showing strong favoritism to any of the doses.

Table 3.3: Eight alpha spending schemes and their corresponding stage-wise efficacy stopping boundaries (b_k) and errors (α_k) for $J = 4, D = 1, M = 2, R = 2$, and $K = 2J = 8$. Family-wise type I error is controlled at one-sided $\alpha = 0.05$. No futility is adopted, $a_k = -\infty$ for all k.

Error Spending Scheme	k	1	2	3	4	5	6	7	8
1. Constant efficacy boundary	b_k	2.442	2.442	2.442	2.442	2.442	2.442	2.442	2.442
	α_k	0.0073	0.0054	0.0072	0.0054	0.0071	0.0053	0.0070	0.0052
2. Pocock-type boundary	b_k	2.337	2.291	2.451	2.399	2.534	2.480	2.599	2.544
	α_k	0.0097	0.0081	0.0070	0.0061	0.0055	0.0049	0.0045	0.0041
3. O'Brien & Fleming-type boundary	b_k	5.421	3.750	3.015	2.600	2.426	2.220	2.232	2.078
	α_k	2.96e-08	8.84e-05	0.0013	0.0042	0.0076	0.0104	0.0125	0.0139
4. Rho error spending $\rho = 0.3$	b_k	1.930	2.277	2.619	2.613	2.755	2.728	2.837	2.802
	α_k	0.0268	0.0062	0.0043	0.0034	0.0028	0.0024	0.0022	0.0020
5. Rho error spending $\rho = 0.5$	b_k	2.104	2.273	2.526	2.493	2.625	2.577	2.685	2.632
	α_k	0.0177	0.0073	0.0056	0.0047	0.0042	0.0038	0.0035	0.0032
6. Rho error spending $\rho = 1.0$	b_k	2.498	2.407	2.493	2.402	2.489	2.397	2.484	2.392
(i.e Constant error spending)	α_k	0.0062	0.0063	0.0063	0.0063	0.0063	0.0063	0.0063	0.0063
7. Rho error spending $\rho = 2.0$	b_k	3.163	2.797	2.659	2.474	2.451	2.289	2.310	2.151
	α_k	0.0008	0.0023	0.0039	0.0055	0.0070	0.0086	0.0102	0.0117
8. Rho error spending $\rho = 3.0$	b_k	3.725	3.190	2.901	2.638	2.512	2.289	2.236	2.016
	α_k	9.77e-05	0.0007	0.0019	0.0034	0.0060	0.0089	0.0124	0.0165

Under Pocock-type boundary set, we are favoring the earlier doses by spending more alphas in the earlier stages, while under O'Brien & Fleming-type boundary set, we are doing the opposite by spending more alphas in later stages. The boundary sets under Rho error spendings with $\rho = 0.3$ and 0.5 have similar properties to the Pocock-type boundary set, while those with $\rho = 2$ and 3 are similar to the O'Brien & Fleming-type boundary set. Similar to the constant boundary set, the boundary set under Rho error spending with $\rho = 1$ also has the same α_k spent for each of the four doses. Each dose gets an alpha spending of 0.0125. This is almost similar to the constant efficacy boundary scheme except for the fact that constant efficacy boundary favors the first per-dose stage $m = 1$ slightly more than the second per-dose stage $m = 2$ for each of the doses, that is, $\alpha_1 > \alpha_2$, $\alpha_3 > \alpha_4$, etc.

In order to confirm if the derived sets of stopping boundaries in Table 3.3 are able to control the family-wise type I error under one-sided $\alpha < 0.05$, we conducted simulation with 50,000 simulated trials. Table 3.4 displays the simulated type I error under the global null hypothesis. It can be seen that these boundary sets are able preserve the family-wise type I error.

Table 3.4: Simulated type I error rates for the derived stopping boundary sets under the selected eight alpha spending plans in Table 3.3

	Constant boundary	Pocock	O'Brien-Fleming	$\rho = 0.3$
Simulated type I error	0.0496	0.0482	0.04856	0.0481
	$\rho = 0.5$	$\rho = 1.0$	$\rho = 2.0$	$\rho = 3.0$
Simulated type I error	0.0501	0.0488	0.0490	0.0481

3.4.3 Cohort Sizes and Planned Trial Sample Sizes

Table 3.5 displays the cohort sizes, expected stages, expected sample sizes, and dose selection probabilities under this proposed adaptive procedure to obtain statistical power of 0.8. Table 3.6 displays the same characteristics under this proposed adaptive procedure to obtain statistical power of 0.9. The following summary and discussion focus on Table 3.6, but similar, parallel interpretation can be drawn for Table 3.5. For all dose response models except the flat model, the required cohort size is smaller when a trial uses informative ordering and error spending schemes that favor earlier stages, namely Pocock, Rho with $\rho = 0.3$ and 0.5 than a trial that uses other orderings and error spending schemes that favor later stages. Considerable saving in patient resources in terms of sample size and time can be achieved under these optimistic scenarios. When the prior knowledge of the dose-response relationship is strong, we can plan the trial by putting presumably efficacious doses first and applying an optimistic error spending function that favors earlier stages. On the other hand, if we couple an informative ordering with an error spending plan that favors later stages and doses, such as O'Brien & Fleming, Rho with $\rho = 2$ and 3, we will have to increase the cohort size to maintain the same statistical power. O'Brien & Fleming error spending performs the worst in this case due to its strong favoritism on the later stages. In this case, this type of error spending plan is not complementary to the informative dose ordering.

For these same dose response models except the flat model, the constant efficacy boundary and Rho with $\rho = 1$ have similar results with the latter showing only a small advantage. It can be seen that for these two error spending plans, informative ordering does not offer additional advantage over dose escalation ordering. It is apparent that the flat dose response model is immune to dose ordering and all error spending plans perform similarly under this model.

Table 3.5: Cohort size per stage (c), expected global stage to stop for efficacy ($E(\mathcal{K})$), expected trial sample size $\left(\frac{c(1+R)}{R}E(\mathcal{K})\right)$, and dose selection probabilities for attaining statistical power of 80%. The number of simulated trials is 10,000.

Dose		Constant efficacy boundary			Pocock-type boundary			O'Brien & Fleming boundary			Rho error spending ρ=0.3		
Response		Escalation	Informative	Uninformative	Escalation	Informative	Uninformative	Escalation	Informative	Uninformative	Escalation	Informative	Uninformative
1. Flat	cohort size	46	46	46	45	45	45	49	49	49	51	51	51
	expected stage	4.4	4.4	4.4	4.3	4.3	4.3	5.8	5.8	5.8	3.8	3.8	3.8
	expected sample size	307	307	307	289	289	289	430	430	430	295	295	295
	dose selection prob	(0.42, 0.28, 0.18, 0.12)	(0.42, 0.28, 0.18, 0.12)	(0.25, 0.25, 0.25, 0.25)	(0.48, 0.26, 0.16, 0.10)	(0.48, 0.26, 0.16, 0.10)	(0.25, 0.25, 0.25, 0.25)	(0.05, 0.34, 0.37, 0.24)	(0.05, 0.34, 0.37, 0.24)	(0.25, 0.25, 0.25, 0.25)	(0.60, 0.20, 0.12, 0.08)	(0.60, 0.20, 0.12, 0.08)	(0.25, 0.25, 0.25, 0.25)
2. Linear	cohort size	87	87	87	92	81	87	72	140	98	111	84	98
	expected stage	6.2	3.3	4.6	6.0	3.3	4.5	6.7	4.3	5.7	5.9	2.9	4.2
	expected sample size	804	433	606	835	396	583	721	901	833	985	372	610
	dose selection prob	(0.06, 0.19, 0.36, 0.39)	(0.01, 0.05, 0.17, 0.76)	(0.04, 0.11, 0.28, 0.57)	(0.09, 0.20, 0.34, 0.37)	(0.01, 0.04, 0.16, 0.79)	(0.04, 0.12, 0.29, 0.55)	(0.00, 0.11, 0.42, 0.47)	(0.04, 0.13, 0.39, 0.44)	(0.03, 0.10, 0.27, 0.60)	(0.14, 0.17, 0.32, 0.37)	(0.01, 0.03, 0.11, 0.86)	(0.04, 0.13, 0.29, 0.54)
3. Emax	cohort size	60	60	60	61	58	60	56	75	66	73	64	68
	expected stage	5.2	3.9	4.5	5.0	3.8	4.3	6.2	4.3	5.8	4.6	3.4	3.9
	expected sample size	467	355	406	457	330	386	518	593	571	505	325	396
	dose selection prob	(0.21, 0.32, 0.28, 0.19)	(0.05, 0.13, 0.27, 0.55)	(0.12, 0.23, 0.30, 0.35)	(0.27, 0.32, 0.25, 0.17)	(0.04, 0.11, 0.25, 0.60)	(0.12, 0.23, 0.30, 0.35)	((0.01, 0.26, 0.42, 0.30)	(0.12, 0.31, 0.45, 0.13)	(0.11, 0.22, 0.30, 0.36)	(0.37, 0.27, 0.21, 0.15)	(0.03, 0.08, 0.18, 0.71)	(0.13, 0.23, 0.30, 0.34)
4. Umbrella	cohort size	61	61	61	63	59	61	58	77	67	74	65	69
	expected stage	5.1	3.9	4.5	4.9	3.8	4.3	5.9	5.3	5.8	4.6	3.4	3.9
	expected sample size	468	361	414	465	335	395	517	608	582	515	329	405
	dose selection prob	(0.17, 0.38, 0.31, 0.14)	(0.04, 0.24, 0.56, 0.16)	(0.10, 0.27, 0.36, 0.27)	(0.22, 0.39, 0.28, 0.11)	(0.03, 0.23, 0.61, 0.13)	(0.10, 0.27, 0.36, 0.27)	(0.01, 0.32, 0.46, 0.21)	(0.09, 0.40, 0.13, 0.37)	(0.09, 0.27, 0.37, 0.27)	(0.32, 0.34, 0.25, 0.10)	(0.02, 0.16, 0.72, 0.10)	(0.11, 0.27, 0.35, 0.27)

Dose		Rho error spending ρ=0.5			Rho error spending ρ=1.0			Rho error spending ρ=2.0			Rho error spending ρ=3.0		
Response		Escalation	Informative	Uninformative	Escalation	Informative	Uninformative	Escalation	Informative	Uninformative	Escalation	Informative	Uninformative
1. Flat	cohort size	47	47	47	44	44	44	45	45	45	48	48	48
	expected stage	4.1	4.1	4.1	4.5	4.5	4.5	5.2	5.2	5.2	5.6	5.6	5.6
	expected sample size	289	289	289	300	300	300	350	350	350	405	405	405
	dose selection prob	(0.54, 0.23, 0.14, 0.09)	(0.54, 0.23, 0.14, 0.09)	(0.25, 0.25, 0.25, 0.25)	(0.41, 0.28, 0.19, 0.13)	(0.41, 0.28, 0.19, 0.13)	(0.25, 0.25, 0.25, 0.25)	(0.24, 0.31, 0.26, 0.19)	(0.24, 0.31, 0.26, 0.19)	(0.25, 0.25, 0.25, 0.25)	(0.14, 0.29, 0.32, 0.25)	(0.14, 0.29, 0.32, 0.25)	(0.25, 0.25, 0.25, 0.25)
2. Linear	cohort size	99	82	91	85	85	85	75	102	87	72	122	93
	expected stage	6.0	3.1	4.3	6.2	3.4	4.7	6.5	3.7	5.2	6.7	3.9	5.5
	expected sample size	890	381	589	790	429	597	732	564	672	726	718	764
	dose selection prob	(0.11, 0.19, 0.33, 0.37)	(0.01, 0.03, 0.13, 0.83)	(0.04, 0.12, 0.29, 0.54)	(0.06, 0.19, 0.36, 0.39)	(0.01, 0.05, 0.17, 0.76)	(0.04, 0.11, 0.28, 0.57)	(0.02, 0.15, 0.38, 0.45)	(0.02, 0.08, 0.23, 0.67)	(0.03, 0.10, 0.27, 0.60)	(0.01, 0.10, 0.38, 0.51)	(0.04, 0.10, 0.26, 0.61)	(0.03, 0.10, 0.26, 0.61)
3. Emax	cohort size	65	60	63	59	59	59	56	64	60	56	71	64
	expected stage	4.8	3.6	4.1	5.2	4.0	4.5	5.7	4.5	5.2	6.1	4.9	5.6
	expected sample size	472	325	387	461	351	402	483	434	463	514	527	531
	dose selection prob	(0.31, 0.29, 0.23, 0.16)	(0.03, 0.10, 0.22, 0.65)	(0.13, 0.23, 0.30, 0.34)	(0.21, 0.32, 0.28, 0.19)	(0.05, 0.13, 0.27, 0.55)	(0.12, 0.23, 0.30, 0.35)	(0.09, 0.30, 0.35, 0.26)	(0.09, 0.20, 0.33, 0.39)	(0.11, 0.22, 0.35, 0.54)	(0.04, 0.25, 0.39, 0.32)	(0.12, 0.25, 0.35, 0.27)	(0.11, 0.22, 0.30, 0.37)
4. Umbrella	cohort size	67	61	64	60	60	60	57	65	61	58	73	65
	expected stage	4.8	3.6	4.1	5.1	4.0	4.5	5.6	4.5	5.2	5.9	4.9	5.6
	expected sample size	483	329	396	462	357	410	480	441	472	516	539	542
	dose selection prob	(0.26, 0.36, 0.27, 0.11)	(0.03, 0.19, 0.66, 0.12)	(0.10, 0.27, 0.35, 0.27)	(0.17, 0.38, 0.31, 0.14)	(0.04, 0.25, 0.56, 0.16)	(0.10, 0.27, 0.36, 0.27)	(0.07, 0.36, 0.38, 0.19)	(0.07, 0.30, 0.40, 0.23)	(0.09, 0.27, 0.37, 0.27)	(0.03, 0.31, 0.43, 0.23)	(0.09, 0.32, 0.29, 0.30)	(0.09, 0.27, 0.38, 0.27)

Table 3.6: Cohort size per stage (c), expected global stage to stop for efficacy ($E(\mathcal{K})$), expected trial sample size $\left(\frac{c(1+R)}{R}E(\mathcal{K})\right)$, and dose selection probabilities for attaining statistical power of 90%. The number of simulated trials is 10,000.

Dose		Constant efficacy boundary			Pocock-type boundary			O'Brien & Fleming boundary			Rho error spending ρ=0.3		
Response		Escalation	Informative	Uninformative	Escalation	Informative	Uninformative	Escalation	Informative	Uninformative	Escalation	Informative	Uninformative
1. Flat	cohort size	60	60	60	60	60	60	65	65	65	67	67	67
	expected stage	3.7	3.7	3.7	3.5	3.5	3.5	5.2	5.2	5.2	3.1	3.1	3.1
	expected sample size	334	334	334	313	313	313	509	509	509	309	309	309
	dose selection prob	(0.49, 0.27, 0.15, 0.09)	(0.49, 0.27, 0.15, 0.09)	(0.25, 0.25, 0.25, 0.25)	(0.55, 0.25, 0.13, 0.07)	(0.55, 0.25, 0.13, 0.07)	(0.25, 0.25, 0.25, 0.25)	(0.08, 0.40, 0.35, 0.17)	(0.08, 0.40, 0.34, 0.17)	(0.25, 0.25, 0.25, 0.25)	(0.66, 0.19, 0.10, 0.06)	(0.66, 0.19, 0.10, 0.06)	(0.25, 0.25, 0.25, 0.25)
2. Linear	cohort size	114	114	114	120	108	114	97	176	128	142	111	127
	expected stage	5.7	2.5	4.0	5.6	2.5	3.8	6.3	3.5	5.1	5.5	2.2	3.5
	expected sample size	982	433	680	1,010	397	651	917	916	973	1,165	367	668
	dose selection prob	(0.07, 0.22, 0.39, 0.33)	(0.01, 0.03, 0.13, 0.83)	(0.03, 0.11, 0.29, 0.56)	(0.09, 0.23, 0.37, 0.31)	(0.01, 0.03, 0.12, 0.85)	(0.04, 0.12, 0.30, 0.54)	(0.00, 0.14, 0.46, 0.40)	(0.02, 0.08, 0.32, 0.57)	(0.02, 0.10, 0.26, 0.62)	(0.14, 0.20, 0.35, 0.31)	(0.00, 0.02, 0.08, 0.90)	(0.04, 0.13, 0.30, 0.52)
3. Emax	cohort size	79	79	79	81	77	79	75	98	87	94	84	89
	expected stage	4.5	3.2	3.8	4.3	3.0	3.6	5.7	4.5	5.2	4.0	2.6	3.2
	expected sample size	538	373	446	525	346	421	636	666	672	558	331	419
	dose selection prob	(0.25, 0.35, 0.26, 0.14)	(0.03, 0.10, 0.24, 0.63)	(0.12, 0.23, 0.30, 0.35)	(0.31, 0.34, 0.23, 0.12)	(0.02, 0.08, 0.22, 0.68)	(0.12, 0.23, 0.30, 0.34)	(0.02, 0.33, 0.43, 0.23)	(0.07, 0.25, 0.48, 0.20)	(0.11, 0.22, 0.31, 0.37)	(0.41, 0.28, 0.19, 0.11)	(0.02, 0.06, 0.16, 0.77)	(0.13, 0.23, 0.30, 0.33)
4. Umbrella	cohort size	81	81	81	82	78	81	77	100	89	96	85	90
	expected stage	4.5	3.1	3.8	4.3	3.0	3.6	5.4	4.5	5.2	4.0	2.6	3.2
	expected sample size	544	379	457	531	351	430	625	682	686	574	335	429
	dose selection prob	(0.20, 0.42, 0.28, 0.10)	(0.02, 0.22, 0.64, 0.12)	(0.09, 0.27, 0.36, 0.27)	(0.25, 0.42, 0.25, 0.08)	(0.02, 0.20, 0.69, 0.10)	(0.10, 0.27, 0.35, 0.27)	(0.01, 0.40, 0.44, 0.15)	(0.06, 0.44, 0.21, 0.30)	(0.08, 0.27, 0.38, 0.27)	(0.35, 0.36, 0.22, 0.07)	(0.01, 0.14, 0.77, 0.07)	(0.11, 0.27, 0.34, 0.27)

Dose		Rho error spending ρ=0.5			Rho error spending ρ=1.0			Rho error spending ρ=2.0			Rho error spending ρ=3.0		
Response		Escalation	Informative	Uninformative	Escalation	Informative	Uninformative	Escalation	Informative	Uninformative	Escalation	Informative	Uninformative
1. Flat	cohort size	62	62	62	59	59	59	60	60	60	63	63	63
	expected stage	3.3	3.3	3.3	3.7	3.7	3.7	4.4	4.4	4.4	4.9	4.9	4.9
	expected sample size	309	309	309	331	331	331	397	397	397	465	465	465
	dose selection prob	(0.60, 0.22, 0.12, 0.07)	(0.60, 0.22, 0.12, 0.07)	(0.25, 0.25, 0.25, 0.25)	(0.49, 0.27, 0.15, 0.09)	(0.49, 0.27, 0.15, 0.09)	(0.25, 0.25, 0.25, 0.25)	(0.32, 0.33, 0.23, 0.13)	(0.32, 0.33, 0.23, 0.13)	(0.25, 0.25, 0.25, 0.25)	(0.20, 0.33, 0.29, 0.17)	(0.20, 0.33, 0.29, 0.17)	(0.25, 0.25, 0.25, 0.25)
2. Linear	cohort size	128	109	118	111	112	112	100	131	114	96	154	121
	expected stage	5.6	2.3	3.7	5.8	2.6	4.0	6.1	2.9	4.5	6.4	3.1	4.9
	expected sample size	1,066	378	652	962	431	673	917	566	770	916	717	882
	dose selection prob	(0.12, 0.21, 0.36, 0.31)	(0.00, 0.02, 0.10, 0.88)	(0.04, 0.13, 0.30, 0.53)	(0.07, 0.22, 0.39, 0.33)	(0.01, 0.03, 0.13, 0.83)	(0.03, 0.11, 0.29, 0.56)	(0.02, 0.18, 0.42, 0.38)	(0.01, 0.05, 0.18, 0.76)	(0.03, 0.10, 0.28, 0.60)	(0.01, 0.13, 0.43, 0.43)	(0.02, 0.06, 0.21, 0.71)	(0.02, 0.09, 0.26, 0.63)
3. Emax	cohort size	86	79	83	77	78	78	74	84	79	75	92	84
	expected stage	4.2	2.8	3.4	4.6	3.2	3.8	5.2	3.7	4.4	5.6	4.1	4.9
	expected sample size	535	335	416	530	370	441	573	466	524	626	572	613
	dose selection prob	(0.35, 0.31, 0.21, 0.12)	(0.02, 0.07, 0.19, 0.72)	(0.13, 0.23, 0.30, 0.34)	(0.25, 0.35, 0.26, 0.15)	(0.03, 0.10, 0.24, 0.63)	(0.12, 0.23, 0.30, 0.35)	(0.12, 0.35, 0.34, 0.19)	(0.05, 0.15, 0.32, 0.48)	(0.11, 0.22, 0.31, 0.37)	(0.05, 0.31, 0.39, 0.24)	(0.08, 0.20, 0.36, 0.37)	(0.10, 0.22, 0.31, 0.37)
4. Umbrella	cohort size	87	81	84	79	79	79	76	85	81	77	95	85
	expected stage	4.2	2.8	3.4	4.5	3.2	3.8	5.0	3.7	4.4	5.4	4.1	4.9
	expected sample size	545	340	426	535	376	451	571	474	534	620	585	624
	dose selection prob	(0.29, 0.39, 0.24, 0.08)	(0.02, 0.17, 0.73, 0.10)	(0.10, 0.27, 0.35, 0.27)	(0.20, 0.42, 0.28, 0.10)	(0.02, 0.22, 0.64, 0.12)	(0.09, 0.27, 0.36, 0.27)	(0.09, 0.42, 0.36, 0.13)	(0.04, 0.29, 0.49, 0.18)	(0.08, 0.27, 0.37, 0.27)	(0.04, 0.38, 0.42, 0.17)	(0.06, 0.32, 0.38, 0.24)	(0.08, 0.27, 0.38, 0.27)

3.4.4 Expected Stages and Expected Trial Sample Sizes

Table 3.6 also tabulates the expected number of global stages out of $K = 8$ stages for statistical power of 0.9. For all of the dose response models except the flat model, informative dose ordering results in substantial reduction in the number of stages a trial goes through for efficacy, regardless of which error spending function is used. This reduction in stages is particularly greater, $E(\mathcal{K}) < 4$, under Pocock, Rho with $\rho = 0.3, 0.5$, and 1 error spendings. We can only achieve $E(\mathcal{K}) < 6$ if we use O'Brien & Fleming, Rho with $\rho = 2$ and 3 under informative ordering. Again, constant efficacy boundary and Rho $\rho = 1$ perform similarly. As expected, the flat model does not depend on dose ordering, but Pocock, Rho with $\rho = 0.3$ and 0.5 error spendings can offer an advantage to stop the trial early for efficacy better than the other error spending plans can.

We can also see from Table 3.6 that dose escalation ordering does not perform well in reducing the expected number of stages a trial goes through before stopping for efficacy. It is because the better doses are placed at the later stages, so trials that use dose escalation ordering and error spending plans which favor later stages will tend to proceed through the later stages. Table 3.6 combines the results of cohort sizes and expected number of stages to obtain the expected trial sample sizes under statistical power of 0.9. This expected trial sample sizes are expected to be substantially smaller than the planned trial sample sizes.

3.4.5 Probability of Selecting the Best Dose

Table 3.6 also shows the probabilities of dose selection. Under dose escalation ordering, the best dose is usually located last except for flat and umbrella dose response models. Error spending functions that favor later stages such as O'Brien & Fleming, Rho with $\rho = 2$ and 3 tend to push the trial to these later stages and therefore increase the probability of selecting the best dose. On the other hand, error spending functions that favor earlier stages are unlikely to move a trial to later stages, and thus they decrease the probabilities that these

better doses are selected. For example, under the linear model and dose escalation ordering, if Rho error spending with $\rho = 0.3$ is used, the probability that $d_4 = 8$ with $\mu_4 = 0.35$ is selected is 0.31, but this probability increases to 0.43 if $\rho = 3$ is used. Under informative ordering, the best dose is located first and when it is coupled with a complementary error spending function like Pocock, Rho with $\rho = 0.3$ or 0.5 that favor earlier dose selection, the probability that this best dose is selected is much higher. Using the same example of linear model but using informative ordering, the probability of selecting $d_4 = 8$ with $\mu_4 = 0.35$ is 0.9 for $\rho = 0.3$ or 0.88 for $\rho = 0.5$.

However, under the flat model, the later doses, although having similar mean effect, are not likely selected. If one insists on fully exploring all of the experimental doses, a Rho error spending plan with $\rho = 3$ or higher can help a trial goes through all the stages, allowing doses under the flat dose response model to have comparable probabilities of being selected.

It is important to note that the probability of selecting the best dose not only depends on the mean responses of the doses but also on the ordering of the doses and the error spending plan chosen for the trial. Generally, error spending plans that favor the later stages tend to push the trials all the way through the stages, and therefore most of the doses will be studied. This will increase the probability of selecting the later doses.

3.4.6 Comparison of Statistical Power

Table 3.7 displays the statistical powers of the parallel group fixed dose design, the drop-the-losers design, and the seamless two-stage dose response informed design. We can compare these simulated powers to the 0.9 level. This table also displays the probabilities of dose selection for the last two comparator designs. For parallel group design using traditional Dunnett's adjustment, its statistical power is generally lower than that of the adaptive staggered dose procedure except under the following three error spending plans - O'Brien & Fleming, Rho with $\rho = 2$ and 3. As we have seen earlier, these are the error spending plans that favor later stages and therefore they need more expected number of global stages for

the trial to stop for efficacy. As a result, they also tend to require larger expected sample sizes under the proposed adaptive design and hence we will lose efficiency if we use these types of error spending functions.

When compared to the drop-the-losers and the two-stage dose response informed designs, the proposed adaptive procedure can outperform in statistical power only when certain conditions are met. Figures 3.2, 3.3, 3.4, and 3.5 graphically compare the statistical powers of the designs. Their powers are plotted against the cohort size that gives the same expected trial sample size under the adaptive staggered dose design. It can be noted that the power curves for the proposed design rise more steeply as cohort size increases than those of the other designs.

Under the flat, emax or umbrella dose response models, the adaptive procedure performs better than both the parallel group and drop-the-losers designs when using an error spending that favors earlier doses such as Pocock or Rho with $\rho \leqslant 1$. This can be confirmed from Table 3.6. Under the linear dose response model, this adaptive design performs worse than all three comparator designs, except under informative ordering coupled with Pocock or Rho with $\rho \leqslant 1$, but seamless two-stage design outperforms the proposed design when dose response model is informative and cohort size is smaller. Generally, the adaptive design performs better than the parallel group and drop-the-losers designs when using informative ordering and optimistic error spending functions such as Pocock or Rho with $\rho \leqslant 1$. If one has to use the dose escalation ordering, then the cohort size must be increased substantially to achieve comparable statistical power.

As for selecting the best dose, using the example of the linear dose response model discussed earlier, the drop-the-loser design has a probability of selecting $d_4 = 8$ ($\mu_4 = 0.35$) equal to 0.72, comparing to 0.90 when we use Rho with $\rho = 0.3$ and informative dose ordering. However, the two-stage dose-response informed design always selects the best dose under the informative monotonic dose response model when the estimated dose response is significant at interim. The probability of selecting the best dose can be higher only if we use

an optimistic error spending function and when knowledge of the dose response model is strong for the proposed design. If dose safety is not of great concern, assuming a monotonically increasing dose response curve, dose de-escalation ordering is considered as the best informative ordering and can result in a higher probability of selecting the best dose.

Figure 3.2: Comparison of statistical power under flat dose response model (3,000 simulated trials)

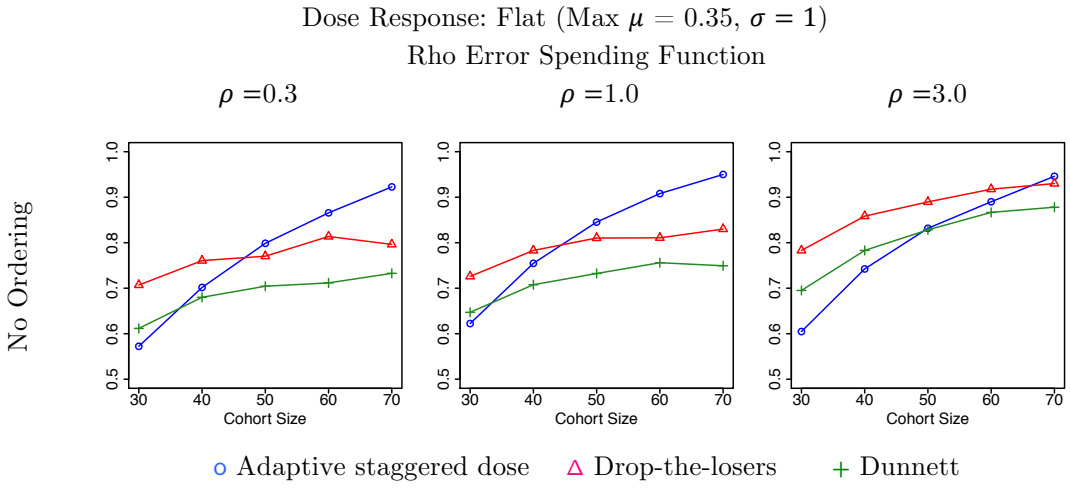


Table 3.7: Simulated powers from parallel group design with Dunnett's adjustment, drop-the-losers (pick-the-winner) design, and seamless dose response informed design. The former two designs used expected sample sizes from uninformative ordering while the last design used expected sample sizes from informative ordering corresponding to 90% power of the proposed adaptive staggered dose design. The number of simulated trials is 10,000.

Dose		Constant efficacy boundary			Pocock-type boundary			O'Brien & Fleming boundary			Rho error spending ρ=0.3		
Response		Dunnett's	Drop-the-losers	Seamless	Dunnett's	Drop-the-losers	Seamless	Dunnett's	Drop-the-losers	Seamless	Dunnett's	Drop-the-losers	Seamless
1. Flat	power	0.7509	0.8252	—	0.7267	0.8052	—	0.8910	0.9401	—	0.7218	0.8067	—
	dose selection prob	—	(0.25, 0.25,	—	—	(0.25, 0.25,	—	—	(0.25, 0.25,	—	—	(0.25, 0.25,	—
		—	0.25, 0.25)	—	—	0.25, 0.25)	—	—	0.25, 0.25)	—	—	0.25, 0.25)	—
2. Linear	power	0.8385	0.9067	0.8547	0.8143	0.8972	0.8433	0.9334	0.9749	0.9910	0.8207	0.9036	0.7970
	dose selection prob	—	(0.01, 0.05,	(0.00, 0.00,	—	(0.01, 0.05,	(0.00, 0.00,	—	(0.00, 0.03,	(0.00, 0.00,	—	(0.01, 0.05,	(0.00, 0.00,
		—	0.23, 0.72)	0.00, 1.00)	—	0.23, 0.71)	0.00, 1.00)	—	0.20, 0.76)	0.00, 1.00)	—	0.23, 0.72)	0.00, 1.00)
3. Emax	power	0.7763	0.8453	0.8233	0.7478	0.8334	0.8030	0.9021	0.9515	0.9597	0.7467	0.8301	0.7760
	dose selection prob	—	(0.07, 0.21,	(0.00, 0.00,	—	(0.07, 0.21,	(0.00, 0.00,	—	(0.06, 0.19,	(0.00, 0.00,	—	(0.07, 0.20,	(0.00, 0.00,
		—	0.32, 0.41)	0.00, 1.00)	—	0.32, 0.41)	0.00, 1.00)	—	0.33, 0.43)	0.00, 1.00)	—	0.32, 0.41)	0.00, 1.00)
4. Umbrella	power	0.7818	0.8519	0.8440	0.7466	0.8380	0.8273	0.9035	0.9535	0.971	0.7425	0.8268	0.7983
	dose selection prob	—	(0.05, 0.26,	(0.00, 0.24,	—	(0.05, 0.27,	(0.00, 0.24,	—	(0.03, 0.26,	(0.00, 0.18,	—	(0.05, 0.26,	(0.00, 0.24,
		—	0.42, 0.27)	0.40, 0.36)	—	0.42, 0.27)	0.38, 0.38)	—	0.45, 0.26)	0.51, 0.31)	—	0.42, 0.27)	0.38, 0.38)

Dose		Rho error spending ρ=0.5			Rho error spending ρ=1.0			Rho error spending ρ=2.0			Rho error spending ρ=3.0		
Response		Dunnett's	Drop-the-losers	Seamless	Dunnett's	Drop-the-losers	Seamless	Dunnett's	Drop-the-losers	Seamless	Dunnett's	Drop-the-losers	Seamless
1. Flat	power	0.8116	0.8008	—	0.8195	0.8209	—	0.8703	0.8769	—	0.9107	0.9210	—
	dose selection prob	—	(0.25, 0.25,	—	—	(0.25, 0.25,	—	—	(0.25, 0.25,	—	—	(0.25, 0.25,	—
		—	0.25, 0.25)	—	—	0.25, 0.25)	—	—	0.25, 0.25)	—	—	0.25, 0.25)	—
2. Linear	power	0.8385	0.8918	0.8247	0.8143	0.9033	0.8693	0.9334	0.9325	0.9390	0.8207	0.9587	0.971
	dose selection prob	—	(0.01, 0.05,	(0.00, 0.00,	—	(0.01, 0.05,	(0.00, 0.00,	—	(0.01, 0.04,	(0.00, 0.00,	—	(0.00, 0.04,	(0.00, 0.00,
		—	0.23, 0.71)	0.00, 1.00)	—	0.23, 0.71)	0.00, 1.00)	—	0.23, 0.73)	0.00, 1.00)	—	0.22, 0.74)	0.00, 1.00)
3. Emax	power	0.7386	0.8174	0.7867	0.7666	0.8465	0.8233	0.8239	0.8110	0.8880	0.8736	0.9280	0.9383
	dose selection prob	—	(0.07, 0.21,	(0.00, 0.00,	—	(0.07, 0.21,	(0.00, 0.00,	—	(0.06, 0.20,	(0.00, 0.00,	—	(0.06, 0.20,	(0.00, 0.00,
		—	0.32, 0.41)	0.00, 1.00)	—	0.31, 0.41)	0.00, 1.00)	—	0.32, 0.42)	0.00, 1.00)	—	0.32, 0.42)	0.00, 1.00)
4. Umbrella	power	0.7445	0.8293	0.8193	0.7618	0.8476	0.8433	0.8309	0.9001	0.9073	0.8793	0.9335	0.9390
	dose selection prob	—	(0.05, 0.27,	(0.00, 0.26,	—	(0.05, 0.27,	(0.00, 0.23,	—	(0.04, 0.26,	(0.00, 0.21,	—	(0.04, 0.26,	(0.00, 0.20,
		—	0.42, 0.26)	0.37, 0.37)	—	0.42, 0.26)	0.40, 0.37)	—	0.43, 0.27)	0.44, 0.35)	—	0.43, 0.27)	0.45, 0.35)

Figure 3.3: Comparison of statistical power under linear dose response model (3,000 simulated trials)

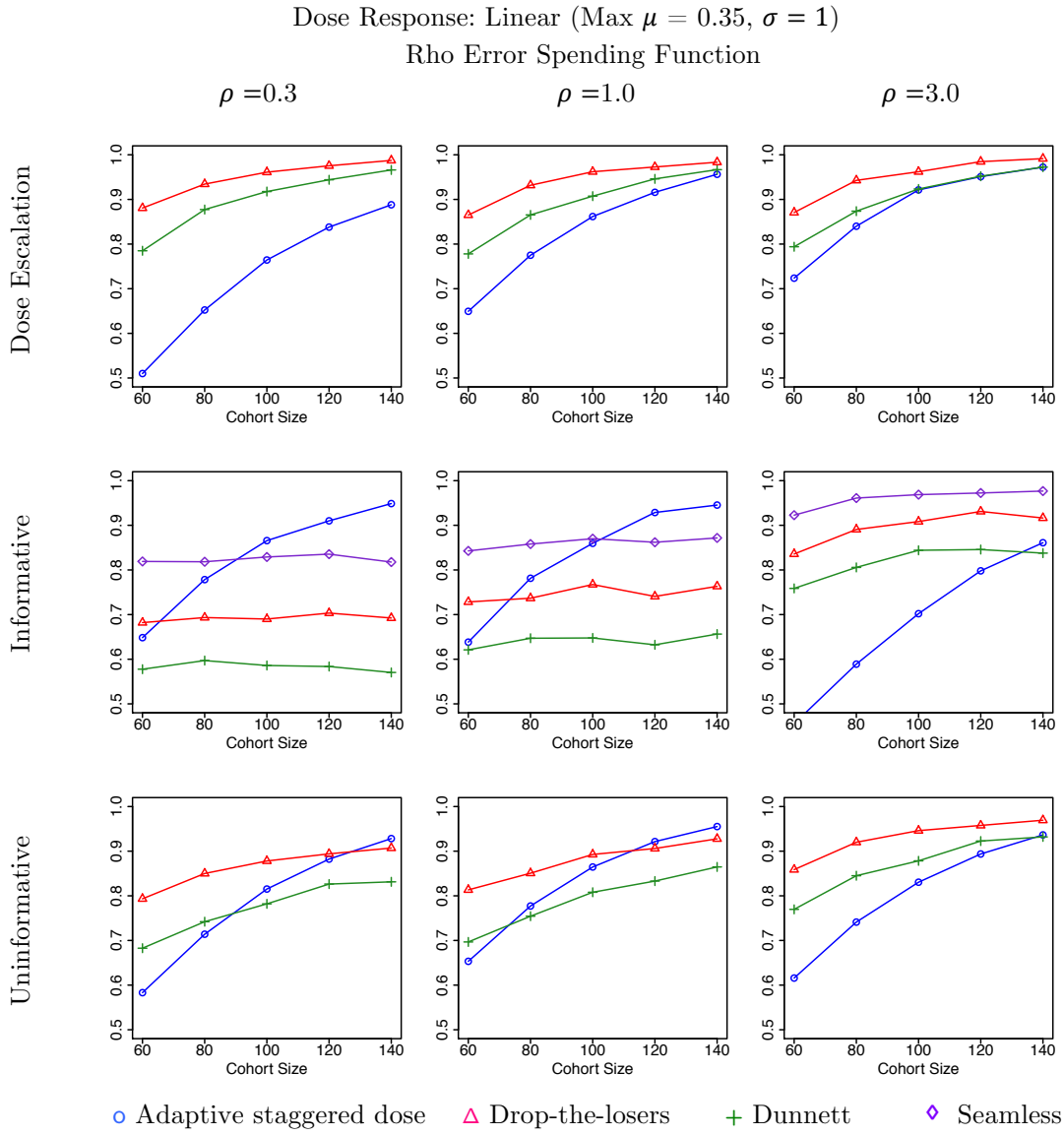


Figure 3.4: Comparison of statistical power under Emax dose response model (3,000 simulated trials)

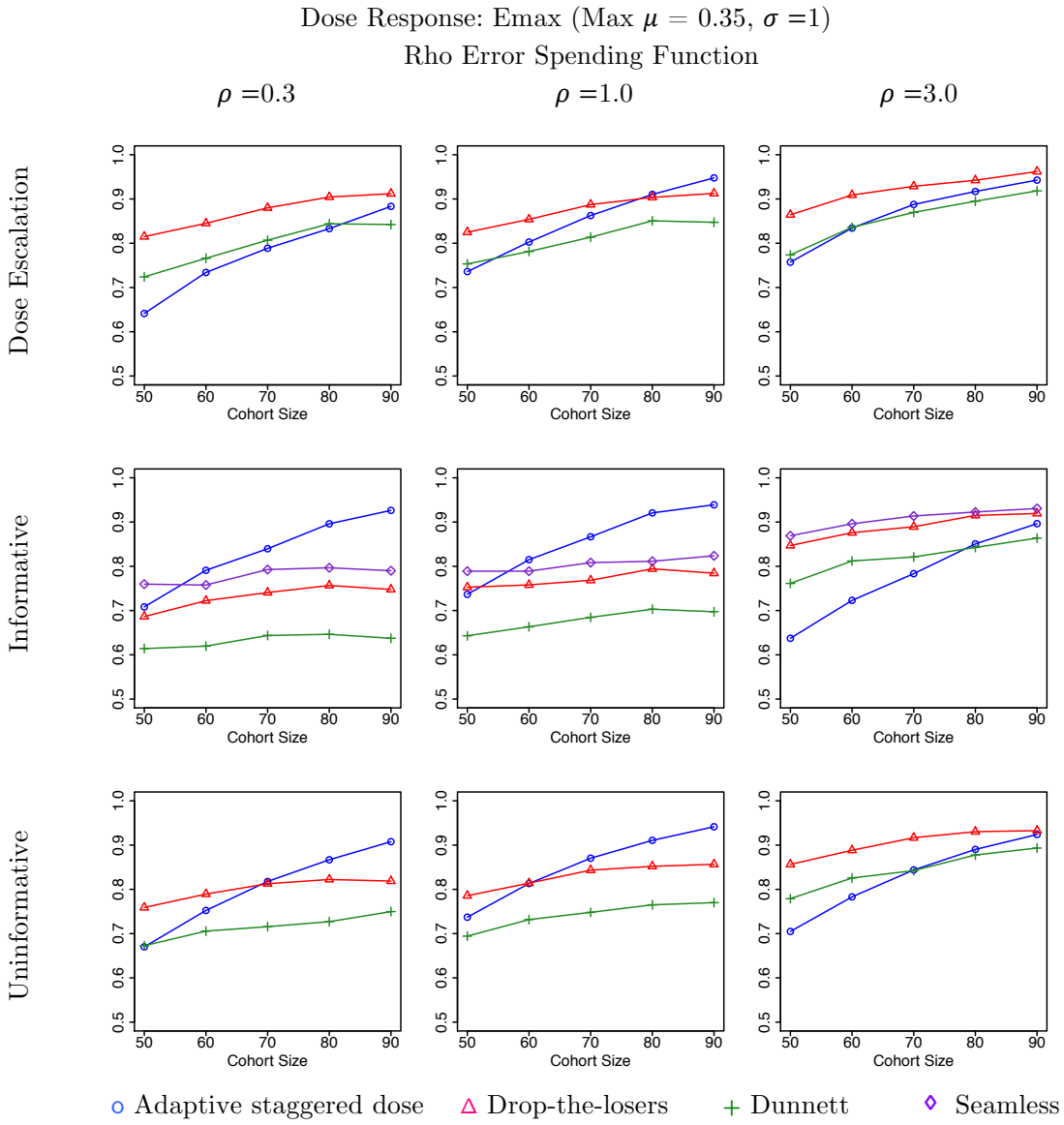
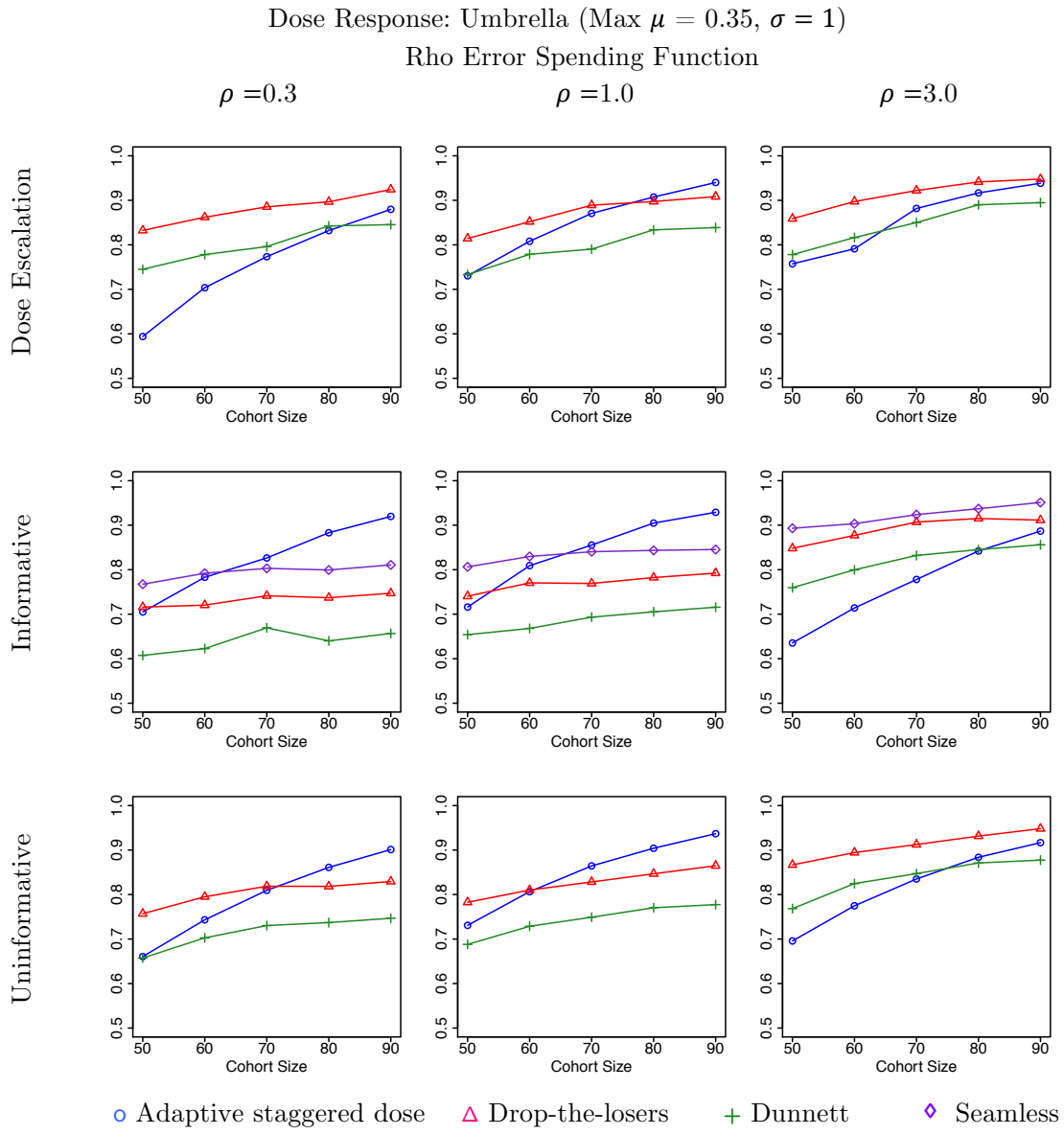


Figure 3.5: Comparison of statistical power under umbrella dose response model (3,000 simulated trials)



3.5 Summary and Discussion

Reflecting on the simulation results from the previous section, we can learn some important lessons about this adaptive staggered dose procedure. In order to achieve additional gain in trial efficiency, several conditions must be met. First, it is assumed that the doses chosen for the study are from a dose range of acceptable safety. Second, an informative ordering of the doses necessitates, at a minimum, some prior qualitative knowledge of the dose response relationship such as monotonicity from previous pre-clinical or clinical studies on a similar class of drugs. However, an assumed parametric model for dose response is not necessary. Third, an application of an optimistic error spending functions such as Pocock or Rho with $\rho \leqslant 1$ allows early efficacious doses to be selected.

By randomizing patients in an adaptive staggered dose fashion instead of randomizing to all of the doses at the same time, we can learn about efficacy more quickly from one dose to another. If a dose shows evidence of futility, we can drop it and move onto the next dose, but if it shows evidence of efficacy, we do not need to expose patients to the remaining and potentially inferior doses. As a result, we may be able to save patient resources earlier by stopping for efficacy earlier. For example, in some oncology trials, where the recruitment period is extended and patients enter a trial at a staggered rate, we can learn and make decisions quickly by focusing patient resources on fewer but potentially more efficacious doses first rather than allocating them to many doses, some of which may be potentially inefficacious.

As we have seen from the earlier plots, the power curves for the adaptive procedure are steeper as cohort size increases. Even under unfavorable ordering such as dose escalation ordering, we may be able to use an error spending scheme that favors later stages such as O'Brien & Fleming or Rho with $\rho > 1$ *and* a larger cohort size to offset the loss of statistical power due to the use of dose escalation ordering. However, this may risk selecting a sub-optimal dose too early and failing to explore other doses.

This adaptive procedure, therefore, may have its limitations. First, information on dose-response relationship may not be available before Phase 2B or 3 studies. This design, unlike the other three comparator designs, may not be able to adequately estimate the dose response, as this is not one of the objectives. Therefore, this proposed design may also be suitable to situations when a parametric dose response is not relevant such as in the case when we are comparing and selecting from a set of treatment regimes or schedules, rather than dose levels, of the same agent. For example, a Phase 3 trial is planned to select and confirm a regime out of several for a new biological product administered subcutaneously to treat rheumatoid arthritis (RA). The regimes differ by the dosage, frequency of injections, and the inclusion of Methotrexate (MTX), an immunosuppressant, or not. In this case, the regimes are ordered in such a way that the assumed better treatment regimes are placed earlier in the ordering.

Second, we have assumed that the J *a priori* doses are within range of acceptable safety and the selection of best dose is based on efficacy alone. In practice, safety issues take a longer time to assess and may happen even after regulatory approval. It can be suggested that, during the conduct of this trial, safety can be qualitatively evaluated through an independent data safety monitoring committee (DSMB) which can provide input on the safety of the selected dose. Further work will be needed if the criterion of selecting the optimal dose is explicitly based on joint assessment of efficacy and toxicity or risk-to-benefit ratio.

Third, it is also likely to randomize more patients to the control arm than to the new test treatment. This is particularly undesirable if the trial continues to later stages due to unsuccessful earlier doses. One way to overcome this undesirable condition, while maintaining a blinded randomization, is to change the randomization ratio using a higher R such as $1:3$ or $1:4$. These ratios may reduce the overall power. Therefore, a trade-off may exist in reducing the number of patients allocated to the control and maximizing statistical power. In addition, if one insists on using the dose escalation ordering because of safety issues, this adaptive staggered dose procedure may not be efficient compared to other designs. There-

fore, this adaptive design may take longer time to complete as doses are studied one after the other, especially if earlier doses do not show evidence of efficacy.

In section 3.3.3, we have mentioned the use of separate alpha spending functions for each of the doses instead of using one single alpha spending function. We can have better control over how the alpha is spent on each of the doses. As a result, the application of this adaptive staggered dose procedure can be flexibly modified depending on the objectives of a specific trial. Therefore, if a drug trial is to use this adaptive staggered dose procedure for dose selection, it is strongly suggested that simulations are performed to understand its operating characteristics under all plausible dose response models. Since the operating characteristics depend on the chosen design parameters, by trying different sets of these design parameters we can better understand their relative performance in dose selection. Investigators can choose the trial among the ones studied that can optimize the trial efficiency under the most probable scenario. Using simulations, the team can also determine the cohort sizes and plan the trial sample sizes to achieve their target statistical power accordingly.

As we have discussed, this adaptive procedure has demonstrated some desirable features in increasing trial efficiency. To further improve this design and to overcome some of the potential drawbacks identified earlier, several possible solutions can be proposed. For example, by increasing R, we can randomize fewer patients to the control arm and more to the experimental doses. This will further reduce the overall expected sample size but may affect the statistical power to some extent. Another solution is to use $D = 2$, that is, to use a design that looks at two concurrent experimental doses at the same time. In this case, if $J = 4$, we can conduct the experiment by randomizing patients to the control and the first two experimental doses given the dose ordering. When either one or both of these two doses do not show evidence of efficacy, the next dose or the remaining two doses can be added to the experiment for randomization in order to maintain two concurrent doses. In this case, we randomize fewer patients to the control dose. Another modification in the trial design is to use different number of maximum per-dose stages M. However, higher M

means additional interim analyses which can be costly. Therefore, the investigators should carefully consider the availability of resources when choosing the design parameters.

3.6 Appendix

3.6.1 Stage-wise Type I Error $\psi_{j,k}$ for $D = 1, M = 2$

In this section, we want to show the result in equation (3.3.2).

Proof. We represent the probability of rejecting the null hypothesis H_{j0} for dose d_j at the kth interim given that it is true as $\psi_{j,k}$ for $j = 1, 2, .., J$ and $k = 1, 2, ..., K$. We let a_k be the futility boundary and b_k the efficacy boundary for the kth interim. Under condition of no futility analysis, $a_k = -\infty$, the probability that dose d_j will be dropped due to futility when $m = 1$ is $P(Z_{jm} \leqslant a_k) = 0$. As a result, dose d_j is added to the trial only if the previous doses $d_1, d_2, ..., d_{j-1}$ do not show evidence of efficacy and dose d_j will always go through all $M = 2$ per-dose stages. Therefore, for $j = 2, ..., J$ and when $m = 1$,

$$\psi_{j,k} = \left(\prod_{i=1}^{\frac{k-1}{2}} P(Z_1 \leqslant b_{2i-1}, Z_2 \leqslant b_{2i}) \right) P(Z > b_k),$$

and when $m = 2$,

$$\psi_{j,k} = \left(\prod_{i=1}^{\frac{k-2}{2}} P(Z_1 \leqslant b_{2i-1}, Z_2 \leqslant b_{2i}) \right) P(Z_1 \leqslant b_{k-1}, Z_2 > b_k)$$

where Z and (Z_1, Z_2) follow the distribution in 3.3.1. We can see that $\psi_{j,k}$ can be similarly generalized to any M with $D = 1$. □

3.6.2 Stage-wise Statistical Power $\xi_{j,k}$ for $D = 1, M = 2$

This section will derive the analytical form of the stage-wise statistical power, which is also know as boundary crossing probability. Based on the distributions of the standardized test statistics in (3.2.4), we can derive the stage-wise statistical power, denoted as $\xi_{j,k}$, for dose d_j at the kth global stage. Under the alternative hypothesis of a dose response model, $\mu_j = f(d_j)$, the stage-wise statistical power can be shown as

$$\xi_{j,k} = \begin{cases} \Phi\left(-\omega_{j,k}\right) & \text{if } j = 1,\ k = 1 \\ \int_{\omega_{j,k}}^{\infty} \Phi\left(\sqrt{2}\omega_{j,k-1} - t\right)\phi(t)dt & \text{if } j = 1,\ k = 2 \\ \left\{\prod_{i=1}^{\frac{k-1}{2}} \left(\int_{-\infty}^{\omega_{i,2i}} \Phi\left(\sqrt{2}\omega_{i,2i-1} - t\right)\phi(t)dt\right)\right\}\Phi\left(-\omega_{j,k}\right) & \text{if } j > 1,\ k = 2j-1 \\ \left\{\prod_{i=1}^{\frac{k-2}{2}} \left(\int_{-\infty}^{\omega_{i,2i}} \Phi\left(\sqrt{2}\omega_{i,2i-1} - t\right)\phi(t)dt\right)\right\} & \\ \left(\int_{\omega_{j,k}}^{\infty} \Phi\left(\sqrt{2}\omega_{j,k-1} - t\right)\phi(t)dt\right) & \text{if } j > 1,\ k = 2j \end{cases}$$

where

$$\omega_{j,k} = b_k - \frac{f(d_j) - \mu_0}{\sqrt{\frac{R+1}{c}}} \text{ when } k = 2j-1,$$

$$\omega_{j,k} = b_k - \frac{f(d_j) - \mu_0}{\sqrt{\frac{R+1}{2c}}} \text{ when } k = 2j,$$

and Φ and ϕ are the cumulative density and the probability density functions of the standard normal distribution.

Proof. When $j = 1$ and $k = 1$, we need $P(Z_{11} > b_1)$,

$$\begin{aligned} P(Z_{11} > b_1) &= P\left(Z_{11} - \frac{f(d_1) - \mu_0}{\sqrt{\frac{R+1}{c}}} > b_1 - \frac{f(d_1) - \mu_0}{\sqrt{\frac{R+1}{c}}}\right) \\ &= P\left(Z > b_1 - \frac{f(d_1) - \mu_0}{\sqrt{\frac{R+1}{c}}}\right) \\ &= 1 - \Phi\left(\omega_{1,1}\right) \\ &= \Phi\left(-\omega_{1,1}\right) \end{aligned}$$

When $j = 1$ and $k = 2$, we need $P(Z_{11} \leqslant b_1, Z_{12} > b_2)$. In general, since the joint distribution $P(Z_{j1}, Z_{j2})$ for the jth dose can also be written as $P(Z_{j1}|Z_{j2})P(Z_{j2})$. It is known that the conditional distribution of $Z_{j1}|Z_{j2}$ follows

$$Z_{j1}|Z_{j2} = z_{j2} \sim N\left(\frac{f(d_j) - \mu_0}{\sqrt{\frac{R+1}{c}}} + \frac{1}{\sqrt{2}}\left(z_{j2} - \frac{f(d_j) - \mu_0}{\sqrt{\frac{R+1}{2c}}}\right), \frac{1}{2}\right).$$

Therefore,

$$\begin{aligned} P(Z_{j1} \leqslant b_{k-1}|Z_{j2} = z_{j2}) &= \Phi\left(\sqrt{2}\left(b_{k-1} - \frac{f(d_j) - \mu_0}{\sqrt{\frac{R+1}{c}}}\right) - \left(z_{j2} - \frac{f(d_j) - \mu_0}{\sqrt{\frac{R+1}{2c}}}\right)\right) \\ &= \Phi\left(\sqrt{2}\omega_{j,k-1} - \left(z_{j2} - \frac{f(d_j) - \mu_0}{\sqrt{\frac{R+1}{2c}}}\right)\right) \end{aligned}$$

Suppose we want the probability $P(Z_{j1} \leqslant b_{k-1}, Z_{j2} > b_k)$, and it can be re-written as

$$\begin{aligned} & P(Z_{j1} \leqslant b_{k-1}, Z_{j2} > b_k) \\ = & \int_{b_k}^{\infty} \Phi\left(\sqrt{2}\omega_{j,k-1} - \left(z_{j2} - \frac{f(d_j) - \mu_0}{\sqrt{\frac{R+1}{2c}}}\right)\right) P(Z_{j2} = z_{j2})\, dz_{j2} \\ = & \int_{\omega_{j,k}}^{\infty} \Phi\left(\sqrt{2}\omega_{j,k-1} - t\right) \phi(t)\, dt \end{aligned}$$

where the last equality follows from a transformation of $t = z_{j2} - \frac{f(d_j)-\mu_0}{\sqrt{\frac{R+1}{2c}}}$. □

The remaining boundary crossing probabilities can be derived by inserting the correct limits into the integral. For a given set of stopping boundaries and cohort size c, the stage-wise power can be evaluated using these probabilities. It can be verified that $\xi = \sum_{k=1}^{K} \xi_{j,k}$ is monotonically increasing in c.

3.6.3 Expected Stages $E(\mathcal{K})$

In this section, we want to show the result in equation (3.3.6). We denote the boundary crossing probability under the alternative hypothesis for the jth dose at the kth stage as $\xi_{j,k}$.

Proof. For $M = 2$, the sample space for the stopping stage, $\mathcal{K}$, is $(1, 2, ..., K = 2J)$, therefore its expectation is given by

$$\begin{aligned} & E(\mathcal{K}) \\ = \; & \xi_{1,1} + 2\xi_{1,2} + 3\xi_{2,3} + ... + (2J-1)\xi_{J,2J-1} + 2J\left(1 - \sum_{j=1}^{J-1}(\xi_{j,2j-1} + \xi_{j,2j}) - \xi_{J,2J-1}\right) \\ = \; & 2J - \left[\sum_{j=1}^{J-1}((2J-2j+1)\xi_{j,2j-1} + (2J-2j)\xi_{j,2j})\right] - \xi_{J,2J-1}. \end{aligned}$$

□

We can see that $E(\mathcal{K}) < 2J$ under alternative hypothesis.

3.6.4 Strong Control of Type I Error

Proposition 1. *Given an adaptive staggered dose procedure with the following parameters: $D = 1$, $M = 2$, and without stopping for futility, the α-level efficacy stopping boundaries under global null hypothesis are sufficient to provide control of family-wise type I error rate (FWER) under α in the strong sense.*

Proof. Suppose we have the stopping boundaries $\boldsymbol{b} = (b_1, b_2, ..., b_k, .., b_K)$ calculated under the global null hypothesis

$$H_{G0} = \bigcap_{j=1}^{J} H_{j0}$$

using the null distributions in (3.3.1), the forms of $\psi_{j,k}$ in (3.3.2) and alpha spending as in

(3.3.4) that preserves the family-wise type I error under α. We need to show that this set of boundary values $\boldsymbol{b}$ can also keep the type I error rate under α for any other intersection hypotheses which contain any subsets of all null hypotheses. In other words, if we let $\mathcal{J} \subset \{1, 2, ..., J\}$ be any subset of doses with true null hypotheses $\mathcal{J} = \{j : H_{j0} : \mu_j \leqslant \mu_0 \text{ is true}\}$, we want to show that this set of boundary values $\boldsymbol{b}$ can also control the type I error for the following intersection hypothesis

$$H_{0\mathcal{J}} = \bigcap_{j \in \mathcal{J}} H_{0j}.$$

When only a subset of doses $\mathcal{J}$ have true null hypotheses, we can let $\mathcal{J}'$ be the set of the remaining doses whose alternative hypotheses are true, such that $\mathcal{J}' \bigcap \mathcal{J} = \phi$ and $\mathcal{J}' \bigcup \mathcal{J} = \{1, 2, ..., J\}$. Therefore, we have $\mathcal{J}' = \{j : H_{ja} : \mu_j > \mu_0 \text{ is true}\}$. We can see that for $j' \in \mathcal{J}'$,

$$P(Z_{j'1} < b_k, Z_{j'2} < b_{k+1}) < P(Z_1 < b_k, Z_2 < b_{k+1})$$

where $(Z_{j'1}, Z_{j'2})$ follows the alternative distributions in (3.2.4) while (Z_1, Z_2) follows the null distributions in (3.3.1). It follows from the fact that $\mu_{j'} - \mu_0 > 0$, and therefore given the same stopping boundaries (b_k, b_{k+1}), the probability on the left hand side is smaller than the one on the right hand side. As a result, any dose $j \in \mathcal{J}$ that follows a dose $j' \in \mathcal{J}'$ in the pre-specified dose ordering will have smaller type I error $\psi_{j,k}$ as in (3.3.2) as it depends on the outcomes of the previous doses. Therefore, the probability of rejecting the intersection null hypothesis stated above is smaller than α given this set of stopping boundary values. Since the probability of rejecting any intersection null hypothesis, when it is true, is always controlled at α, therefore for any dose d_j with true null hypothesis H_{j0}, all intersection null hypotheses containing it are tested at α. We can claim that this boundary set computed under this staggered dose testing procedure has strong control of family-wise type I error according to the closed testing principle (Marcus, Peritz, and Gabriel, 1976). □

3.7 R Codes

3.7.1 Function Codes

```
pocockesf <- function(alpha, t) {
  if (t < 1) {at <- alpha*log(1+(exp(1)-1)*t)} else {at <- alpha}
  return(at)
}

obfesf <- function(alpha, t) {
  if (t < 1) {at <- 2*(1-pnorm((qnorm(1-alpha/2))/sqrt(t)))} else {at <- alpha}
  return(at)
}

rhoesf <- function(alpha, rho, t) {
  if (t < 1) {at <- alpha*(t^rho)} else {at <- alpha}
  return(at)
}

alphasegments <- function(type, alpha, K, rho=1) {
  alphaseg <- rep(NA,K)
  if (type==1) {
    for (i in 1:K) {
      alphaseg[i] <- pocockesf(alpha=alpha, t=(i/K))
    }
  } else if (type==2) {
    for (i in 1:K) {
      alphaseg[i] <- obfesf(alpha=alpha, t=(i/K))
    }
  } else if (type==3) {
    for (i in 1:K) {
      alphaseg[i] <- rhoesf(alpha=alpha, rho=rho, t=(i/K))
    }
  }
  alphaseg2 <- alphaseg-c(0, alphaseg[1:(K-1)])
  return(alphaseg2)
}

library(mvtnorm)
# p1 = P(Z < x)
p1 <- function(x) {pnorm(x)}
# p2 = P(Z > x)
p2 <- function(x) {pnorm(x, lower.tail=F)}
# p3 = P(x1 < Z1 < x2, Z2 < x3)
p3 <- function(x1, x2, x3) {prob <- pmvnorm(lower=c(x1,-Inf), upper=c(x2,x3), mean=rep(0,2),
corr=matrix(c(1,1/sqrt(2),1/sqrt(2),1), c(2,2))); return(prob)}
# p4 = P(x1 < Z1 < x2, Z2 > x3)
p4 <- function(x1, x2, x3) {prob <- pmvnorm(lower=c(x1,x3), upper=c(x2,Inf), mean=rep(0,2),
corr=matrix(c(1,1/sqrt(2),1/sqrt(2),1), c(2,2))); return(prob)}

effboundaryJ4D1M2 <- function(alphaseg, a, limits=c(1,3)) {
  result <- list()
  K <- length(alphaseg)
  b <- rep(NA, K)
  accuracy <- rep(NA, K)
  tryb <- seq(limits[1], limits[2], by=0.001)
  n <- length(tryb)
  trypsi <- rep(NA, n)
```

```
w <- 1
# find b[1]
if (w > K) {stop} else {
  for (i in 1:n) {
    trypsi[i] <- p2(tryb[i])
  }
  accuracy[1] <- min(abs(trypsi-alphaseg[1]))
  b[1] <- tryb[which(abs(trypsi-alphaseg[1])==min(abs(trypsi-alphaseg[1])))]; w <- w+1}
# find b[2]
if (w > K) {stop} else {
  for (i in 1:n) {
    trypsi[i] <- p4(a[1],b[1],tryb[i])+p1(a[1])*p2(tryb[i])
  }
  accuracy[2] <- min(abs(trypsi-alphaseg[2]))
  b[2] <- tryb[which(abs(trypsi-alphaseg[2])==min(abs(trypsi-alphaseg[2])))]; w<- w+1}
# find b[3]
if (w > K) {stop} else {
  for (i in 1:n) {
    trypsi[i] <- p3(a[1],b[1],b[2])*p2(tryb[i])+p1(a[1])*p4(a[2],b[2],tryb[i])+p1(a[1])*p1(a[2])
    *p2(tryb[i])
  }
  accuracy[3] <- min(abs(trypsi-alphaseg[3]))
  b[3] <- tryb[which(abs(trypsi-alphaseg[3])==min(abs(trypsi-alphaseg[3])))]; w <- w+1}
# find b[4]
if (w > K) {stop} else {
  for (i in 1:n) {
    trypsi[i] <- p3(a[1],b[1],b[2])*p4(a[3],b[3],tryb[i])+p3(a[1],b[1],b[2])*p1(a[3])
    *p2(tryb[i])+p1(a[1])*p3(a[2],b[2],b[3])*p2(tryb[i])+p1(a[1])*p1(a[2])*p4(a[3],b[3],tryb[i])+
    p1(a[1])*p1(a[2])*p1(a[3])*p2(tryb[i])
  }
  accuracy[4] <- min(abs(trypsi-alphaseg[4]))
  b[4] <- tryb[which(abs(trypsi-alphaseg[4])==min(abs(trypsi-alphaseg[4])))]; w <- w+1}
# find b[5]
if (w > K) {stop} else {
  for (i in 1:n) {
    trypsi[i] <- p3(a[1],b[1],b[2])*p3(a[3],b[3],b[4])*p2(tryb[i])+p3(a[1],b[1],b[2])*p1(a[3])
    *p4(a[4],b[4],tryb[i])+p1(a[1])*p3(a[2],b[2],b[3])*p4(a[4],b[4],tryb[i])+p3(a[1],b[1],b[2])
    *p1(a[3])*p1(a[4])*p2(tryb[i])+p1(a[1])*p3(a[2],b[2],b[3])*p1(a[4])*p2(tryb[i])+p1(a[1])
    *p1(a[2])*p3(a[3],b[3],b[4])*p2(tryb[i])+p1(a[1])*p1(a[2])*p1(a[3])*p4(a[4],b[4],tryb[i])
  }
  accuracy[5] <- min(abs(trypsi-alphaseg[5]))
  b[5] <- tryb[which(abs(trypsi-alphaseg[5])==min(abs(trypsi-alphaseg[5])))]; w <- w+1}
# find b[6]
if (w > K) {stop} else {
  for (i in 1:n) {
    trypsi[i] <- p3(a[1],b[1],b[2])*p3(a[3],b[3],b[4])*p4(a[5],b[5],tryb[i])+p3(a[1],b[1],b[2])
    *p3(a[3],b[3],b[4])*p1(a[5])*p2(tryb[i])+p3(a[1],b[1],b[2])*p1(a[3])*p3(a[4],b[4],b[5])
    *p2(tryb[i])+p3(a[1],b[1],b[2])*p1(a[3])*p1(a[4])*p4(a[5],b[5],tryb[i])+p1(a[1])*
    p3(a[2],b[2],b[3])*p3(a[4],b[4],b[5])*p2(tryb[i])+p1(a[1])*p3(a[2],b[2],b[3])*p1(a[4])
    *p4(a[5],b[5],tryb[i])+p1(a[1])*p1(a[2])*p3(a[3],b[3],b[4])*p4(a[5],b[5],tryb[i])
  }
  accuracy[6] <- min(abs(trypsi-alphaseg[6]))
  b[6] <- tryb[which(abs(trypsi-alphaseg[6])==min(abs(trypsi-alphaseg[6])))]; w <- w+1}
# find b[7]
if (w > K) {stop} else {
  for (i in 1:n) {
    trypsi[i] <- p3(a[1],b[1],b[2])*p3(a[3],b[3],b[4])*p3(a[5],b[5],b[6])*p2(tryb[i])+
    p3(a[1],b[1],b[2])*p3(a[3],b[3],b[4])*p1(a[5])*p4(a[6],b[6],tryb[i])+p3(a[1],b[1],b[2])
    *p1(a[3])*p3(a[4],b[4],b[5])*p4(a[6],b[6],tryb[i])+p1(a[1])*p3(a[2],b[2],b[3])
    *p3(a[4],b[4],b[5])*p4(a[6],b[6],tryb[i])
  }
  accuracy[7] <- min(abs(trypsi-alphaseg[7]))
  b[7] <- tryb[which(abs(trypsi-alphaseg[7])==min(abs(trypsi-alphaseg[7])))]; w <- w+1}
# find b[8]
```

```
  if (w > K) {stop} else {
    for (i in 1:n) {
      trypsi[i] <- p3(a[1],b[1],b[2])*p3(a[3],b[3],b[4])*p3(a[5],b[5],b[6])*p4(a[7],b[7],tryb[i])
    }
    accuracy[8] <- min(abs(trypsi-alphaseg[8]))
    b[8] <- tryb[which(abs(trypsi-alphaseg[8])==min(abs(trypsi-alphaseg[8])))]; w <- w+1}
  result$accuracy <- accuracy
  result$effboundary <- b
  return(result)
}

dose <- c(0,2,4,6,8)
maxmu <- 0.35
maxd <- dose[5]
ed50 <- dose[2]
maxumb <- dose[4]
midd <- dose[3]; alog <- -0.015
flat <- function(x) {y <- (x!=0)*maxmu; return(y)}
linear <- function(x) {y <- (maxmu/maxd)*x; return(y)}
emax <- function(x) {y <- (maxmu*(maxd+ed50)/maxd)*(x/(ed50+x)); return(y)}
umbrella <- function(x) {y <- (2*maxmu/maxumb)*x-(maxmu/(maxumb)^2)*x^2
                return(y)}
logistic <- function(x) {b <- (((maxmu/2)-alog)*2); c <- (log(-1-(b/alog)))/4
                y <- alog+b/(1+exp(c*(midd-x))) ; return(y)}

drcurve <- function(model, dose) {
  ndoses <- length(dose);
  if (model==0) {
    return(rep(0,ndoses))
  } else if (model==1) {
    return(flat(dose))
  } else if (model==2) {
    return(linear(dose))
  } else if (model==3) {
    return(emax(dose))
  } else if (model==4) {
    return(logistic(dose))
  } else if (model==5) {return(umbrella(dose))}
}

onetrialD1 <- function(J=4, m=2, R, K, c, model, dose, ordering, stdev, delta=0, a, b) {
  result <- list()
  Nmax <- m*c
  result$ordering <- ordering
  result$dosereordered <- c(dose[1], dose[2:(J+1)][ordering])
  mu <- drcurve(model=model, dose=dose)
  result$muoriginal <- mu
  mu <- c(mu[1],mu[2:(J+1)][ordering])
  result$mureordered <- mu
  x <- matrix(rnorm(J*Nmax, mean=mu[2:(J+1)], sd=stdev), c(J, Nmax), byrow=F)
  result$simulatedmeans <- c(apply(x, 1, mean))
  k <- 1
  i <- 1
  j <- 1
  w <- 0
  Nj <- 0
  success <- 0
  result$success <- success
  ztrace <- rep(NA,K)
  mtrace <- rep(NA,K)
  atrace <- rep(NA,K)
  jtrace <- rep(NA,K)
  ktrace <- rep(NA,K)
  while (j < (J+1) & k < (K+1)) {
```

```
    if (w==0) {x0 <- rnorm(Nmax/R, mean=mu[1], sd=stdev)}
    w <- w+1
    Nj <- Nj+c
    jtrace[i] <- j
    ktrace[i] <- k
    z <- (mean(x[j,1:(w*c)])-mean(x0[1:(w*(c/R))])-delta)/(stdev*sqrt((R+1)/(w*c)))
    ztrace[i] <- z
    mtrace[i] <- mean(x[j,1:(w*c)])
    a_k <- a[k]; b_k <- b[k]
    if (z > b_k) {ind <- 1 # reject null
    } else if (z < a_k) {ind <- 2 # accept null
    } else if (a_k <= z & z <= b_k ) {ind <- 3} # continue to next cohort
    atrace[i] <- ind
    if (ind==1) { success <- 1; result$success <- success; result$selecteddose <- j
    result$lastinterim <- k
    result$summary <- paste("success=", success, ", current dose=", j, ",
    current stage=", k, sep=""); break
    } else if (ind==2) {j <- j+1; k <- k+1; w <- 0; Nj <- 0
    } else if (ind==3 & Nj < Nmax) {j <- j; k <- k+1
    } else if (ind==3 & Nj >= Nmax) {j <- j+1; k <- k+1; w <- 0; Nj <- 0}
    i <- i+1
  }
  # note that the selecteddose is the j corresponding to the new order
  if (result$success==0) { result$selecteddose <- NA; result$lastinterim <- k-1
  result$summary <- "unsuccessful trial" }
  trace <- cbind(ktrace, jtrace, mtrace, ztrace, atrace)
  colnames(trace) <- c("interim","dose","samplemean","zstat","decision")
  result$trace <- trace
  return(result)
}

rejectnullprD1 <- function(J=4, m=2, R, K, c, model, dose, ordering, stdev,
delta=0, a, b, Nsim) {
  sim <- list()
  type1 <- rep(NA, Nsim); stage <- rep(NA, Nsim); selected <- rep(NA, Nsim)
  for (g in 1:Nsim) {
    result <- onetrialD1(J=4, m=2, R=R, K=K, c=c, model=model, dose=dose,
    ordering=ordering, stdev=stdev, delta=delta, a=a, b=b)
    type1[g] <- result$success
    stage[g] <- result$lastinterim
    if (result$success==1) {
      selected[g] <- result$selecteddose
    } else if (result$success==0) { selected[g] <- NA }
  }
  sim$avgtype1 <- mean(type1)
  analyse <- cbind(type1,stage)
  analyse <- analyse[analyse[,1]==1,]
  counts <- rep(NA,K)
  for (w in 1:K) {counts[w] <- sum(analyse[,2]==w)}
  sim$type1bystage <- counts/Nsim
  selected2 <- selected[complete.cases(selected)]
  n <- length(selected2)
  counts <- rep(NA,J)
  for (w in 1:J) { counts[w] <- sum(selected2==w) }
  # note the doseprop refers to the doses in new order, not original order
  sim$doseprop <- counts/n
  sim$avgk <- mean(stage)
  sim$summary <- paste("probability of rejecting null =", sim$avgtype1, ",",
  "average interim stage =", sim$avgk, sep=" ")
  return(sim)
}

findcohortc3 <- function(targetpower, possiblec, J=4, m=2, R, K=8, model, dose,
ordering, stdev=1, delta=0, b) {
```

```
  meandose <- drcurve(model=model,dose=dose)
  reordereddose <- meandose[2:5][ordering]
  posspower <- rep(NA,length(possiblec))
  findpower <- function(ccc) {
    w1 <- b[1]-((reordereddose[1]-0)/(sqrt((R+1)/(ccc))))
    w2 <- b[2]-((reordereddose[1]-0)/(sqrt((R+1)/(2*ccc))))
    w3 <- b[3]-((reordereddose[2]-0)/(sqrt((R+1)/(ccc))))
    w4 <- b[4]-((reordereddose[2]-0)/(sqrt((R+1)/(2*ccc))))
    w5 <- b[5]-((reordereddose[3]-0)/(sqrt((R+1)/(ccc))))
    w6 <- b[6]-((reordereddose[3]-0)/(sqrt((R+1)/(2*ccc))))
    w7 <- b[7]-((reordereddose[4]-0)/(sqrt((R+1)/(ccc))))
    w8 <- b[8]-((reordereddose[4]-0)/(sqrt((R+1)/(2*ccc))))
    power1 <- p2(w1)
    power2 <- p4(-Inf,w1,w2)
    power3 <- p3(-Inf,w1,w2)*p2(w3)
    power4 <- p3(-Inf,w1,w2)*p4(-Inf,w3,w4)
    power5 <- p3(-Inf,w1,w2)*p3(-Inf,w3,w4)*p2(w5)
    power6 <- p3(-Inf,w1,w2)*p3(-Inf,w3,w4)*p4(-Inf,w5,w6)
    power7 <- p3(-Inf,w1,w2)*p3(-Inf,w3,w4)*p3(-Inf,w5,w6)*p2(w7)
    power8 <- p3(-Inf,w1,w2)*p3(-Inf,w3,w4)*p3(-Inf,w5,w6)*p4(-Inf,w7,w8)
    stagewise <- list()
    stagewise$foundpower <- power1+power2+power3+power4+power5+power6+
    power7+power8
    stagewise$stagewisepower <- c(power1, power2, power3, power4, power5,
    power6, power7, power8)
    return(stagewise)
  }
  for (i in 1:length(possiblec)) {
    ccc <- findpower(possiblec[i])
    posspower[i] <- ccc$foundpower
  }
  result <- list()
  result$possiblepower <- posspower
  position <- which(posspower > targetpower)
  position <- min(position)
  result$position <- position
  targetc <- possiblec[position]
  result$targetc <- targetc
  final <- findpower(targetc)
  result$stagewisepower <- final$stagewisepower
  check <- final$stagewisepower
  result$expectedstage <- 8-(7*check[1]+6*check[2])-(5*check[3]+4*check[4])-
  (3*check[5]+2*check[6])-check[7]
  originalorder <- c(which(ordering==1),which(ordering==2),which(ordering==3),
  which(ordering==4))
  doseprob <- c(check[1]+check[2],check[3]+check[4],check[5]+check[6],check[7]+
  check[8])/sum(check)
  result$doseprob_originorder <- doseprob[originalorder]
  # return the dose selection probabilities in original order
  return(result)
}

library(multcomp)
onetrialdunnett <- function(J=4, ss, model, dose, stdev, delta=0, alternative="greater",
alpha=0.05) {
  output <- list()
  if ((J+1)!=length(dose)) {print("The length of dose do not match (J+1).", quote=F)}
  outcome <- NULL
  factor <- NULL
  mu <- drcurve(model=model,dose=dose)
  for (i in 1:(J+1)) {
    outcome <- c(outcome, rnorm(ss, mu[i], sd=stdev));
    factor <- c(factor, rep((i-1),ss))
  }
```

```
  trial <- cbind(outcome,factor)
  colnames(trial) <- c("outcome","treatment")
  trial <- data.frame(trial)
  trial$treatment <- as.factor(trial$treatment)
  modelmc <- aov(outcome~treatment, data=trial)
  result1 <- glht(model=modelmc, linfct = mcp(treatment = "Dunnett"),
  alternative=alternative)
  ci95 <- confint(result1, level=(1-alpha))
  confidence <- ci95[[10]]
  output$success <- (sum(confidence[,2]>0)>0)*1
  output$means <- confidence[,1]
  output$significantdoses <- which(confidence[,2]>0)
  return(output)
}

rejectnullprdunnett <- function(J=4, ss, model, dose, stdev, delta=0, alternative="greater",
alpha=0.05, Nsim) {
  sim <- list()
  type1 <- rep(NA, Nsim)
  for (m in 1:Nsim) {
    result <- onetrialdunnett(J=4, ss=ss, model=model, dose=dose, stdev=stdev, delta=0,
    alternative="greater", alpha=0.05)
    type1[m] <- result$success
  }
  sim$avgtype1 <- mean(type1)
  sim$summary <- paste("probability of rejecting null =", sim$avgtype1, sep=" ")
  return(sim)
}

onetrialdroploser2 <- function(J=4, c, model, dose, stdev=1, delta=0, a, b) {
  result <- list()
  mu <- drcurve(model=model, dose=dose)
  x0 <- rnorm(2*c, mean=mu[1], sd=stdev)
  x <- NULL
  for (j in 1:J) {
    x <- rbind(x, rnorm(2*c, mean=mu[j+1], sd=stdev))
  }
  result$simulatedmeans <- c(mean(x0),apply(x,1,mean))
  # first stage
  result$success <- 0
  mean1 <- apply(x[,1:c],1,mean)
  whichj <- which(mean1==max(mean1))
  result$selecteddose <- whichj
  # second stage
  z <- (mean(x[whichj,1:(2*c)])-mean(x0[1:(2*c)])-delta)/(stdev*sqrt(2/(2*c)))
  result$z <- z
  if (z > b) { result$success <- 1;
               result$summary <- paste("success=", result$success, ", selected dose=",
               whichj, sep="")
  } else if (z <= b) {result$summary <- "Unsuccessful trial"}
  return(result)
}

rejectnullprdroploser <- function(J=4, c, model, dose, stdev=1, delta=0, a, b, Nsim=1000) {
  sim <- list()
  type1 <- rep(NA, Nsim); selected <- rep(NA, Nsim)
  for (g in 1:Nsim) {
    result <- onetrialdroploser2(J=4, c=c, model=model, dose=dose, stdev=stdev,
    delta=delta, a=a, b=b)
    type1[g] <- result$success
    if (result$success==1) {
      selected[g] <- result$selecteddose
    } else if (result$success==0) { selected[g] <- NA }
  }
```

```
  sim$avgtype1 <- mean(type1)
  selected2 <- selected[complete.cases(selected)]
  n <- length(selected2)
  counts <- rep(NA,J)
  for (w in 1:J) { counts[w] <- sum(selected2==w) }
  sim$doseprop <- counts/n
  sim$summary <- paste("probability of rejecting null =", sim$avgtype1,  sep=" ")
  return(sim)
}

effbounddroploser <- function(J=4, m=2, c, model, dose, stdev=1, delta=0, a=0,
Nsim=1000, decimal=2, limits=c(1.5,2.5), alpha=0.05) {
  tryb <- seq(limits[1], limits[2], by=10^(-decimal))
  n <- length(tryb)
  type1error <- rep(NA, n)
  for (i in 1:n) {
    temp <- rejectnullprdroploser(J=4, c=c, model=model, dose=dose, stdev=stdev,
    delta=delta, a=a, b=tryb[i], Nsim=Nsim)
    type1error[i] <- temp$avgtype1
  }
  smooth <- lowess(tryb, type1error)
  posit <- which(smooth$y < alpha)
  posit <- min(posit)
  oneb <- smooth$x[posit]
  return(oneb)
}
# For J=4, under null, boundary is 2.0588

bretzfinddose <- function(model, datain) {
  if (model==2) {
    nlsmodel <- nls(response ~ beta*dose,
                    data=datain,
                    start=list(beta=0.1))
    coef(nlsmodel)
    nlslinear <- function(x) {y <- (coef(nlsmodel))*x; return(y)}
    estmu <- nlslinear(dose[c(2,3,4,5)])
    whichj <- which(estmu==max(estmu))
    return(whichj)
  } else if (model==3) {
    nlsmodel <- nls(response ~ beta*(dose/(2+dose)),
                    data=datain,
                    start=list(beta=0.1))
    coef(nlsmodel)
    nlsemax <- function(x) {y <- coef(nlsmodel)*(x/(2+x)); return(y)}
    estmu <- nlsemax(dose[c(2,3,4,5)])
    whichj <- which(estmu==max(estmu))
    return(whichj)
  } else if (model==5) {
    nlsmodel <- nls(response ~ beta1*dose + beta2*(dose^2),
                    data=datain,
                    start=list(beta1=0.1, beta2=0.1))
    coef(nlsmodel)
    nlsumbrella <- function(x) {y <- (coef(nlsmodel)[1])*x+(coef(nlsmodel)[2])*x^2; return(y)}
    estmu <- nlsumbrella(dose[c(2,3,4,5)])
    whichj <- which(estmu==max(estmu))
    return(whichj)
  }
}

onetrialbretz <- function(J=4, c, model, dose, stdev, delta=0, b) {
  result <- list()
  result$model <- model
  result$success <- 0
  mu <- drcurve(model=model, dose=dose)
```

```
  datafinal <- NULL
  for (i in 1:(J+1)) {
    data <- NULL
    data <- rnorm(c, mean=mu[i], sd=stdev)
    dim(data) <- c(c,1)
    data <- cbind(data, rep(dose[i], c), rep((i-1), c))
    datafinal <- rbind(datafinal, data) }
  datafinal2 <- data.frame(datafinal)
  names(datafinal2) <- c("response", "dose", "which")
  # first stage
  whichj <- bretzfinddose(model=model, datain=datafinal2)
  result$selecteddose <- whichj
  # second stage
  z <- (  mean(c(datafinal[datafinal[,3]==whichj,1], rnorm(c, mu[whichj+1], sd=stdev)))
          - mean(c(datafinal[datafinal[,3]==0,1], rnorm(c, mu[1], sd=stdev)))
          - delta)/(stdev*sqrt(2/(2*c)))
  result$z <- z
  if (z > b) { result$success <- 1;
               result$summary <- paste("success=", result$success, ", selected dose=",
               whichj, sep="")
  } else if (z <= b) {result$summary <- "Unsuccessful trial"}
  return(result)
}

rejectnullprbretz <- function(J, c, model, dose, stdev, delta=0, b, Nsim) {
  sim <- list()
  type1 <- rep(NA, Nsim); selected <- rep(NA, Nsim)
  for (g in 1:Nsim) {
    result <- onetrialbretz(J=J, c=c, model=model, dose=dose, stdev=stdev,
    delta=delta, b=b)
    type1[g] <- result[[2]]
    if (result[[2]]==1) {
      selected[g] <- result[[3]]
    } else if (result[[2]]==0) { selected[g] <- NA }
  }
  sim$avgtype1 <- mean(type1)
  selected2 <- selected[complete.cases(selected)]
  n <- length(selected2)
  counts <- rep(NA,J)
  for (w in 1:J) { counts[w] <- sum(selected2==w) }
  sim$doseprop <- counts/n
  sim$summary <- paste("probability of rejecting null =", sim$avgtype1,  sep=" ")
  return(sim)
}
```

3.7.2 Analysis Codes

```
D <- 1
J <- 4
m <- 2
K <- 8
R <- 2

dose <- c(0,2,4,6,8)
maxmu <- 0.35
rho <- c(0.3, 0.5, 1, 2, 3)
drc <- c("flat","linear","emax","logistic","umbrella")
```

```
numericalb <- rep(2.4422,K)
pocockb <- c(2.337, 2.291, 2.451, 2.399, 2.534, 2.480, 2.599, 2.544)
obfb <- c(5.421, 3.750, 3.015, 2.600, 2.426, 2.220, 2.232, 2.078)
rho0.3b <- c(1.930, 2.277, 2.619, 2.613, 2.755, 2.728, 2.837, 2.802)
rho0.5b <- c(2.104, 2.273, 2.526, 2.493, 2.625, 2.577, 2.685, 2.632)
rho1.0b <- c(2.498, 2.407, 2.493, 2.402, 2.489, 2.397, 2.484, 2.392)
rho2.0b <- c(3.163, 2.797, 2.659, 2.474, 2.451, 2.289, 2.310, 2.151)
rho3.0b <- c(3.725, 3.190, 2.901, 2.638, 2.512, 2.289, 2.236, 2.016)
effbounds <- rbind(numericalb, pocockb, obfb, rho0.3b, rho0.5b, rho1.0b, rho2.0b, rho3.0b)
droploser <- 2.0588

ordering1 <- 1:4
orderingflat <- 1:4
orderinglinear <- 4:1
orderingemax <- 4:1
orderinglogistic <- 4:1
orderingumbrella <- c(3,2,4,1)
informordering <- rbind(orderingflat, orderinglinear, orderingemax,
orderinglogistic, orderingumbrella)
informordering
permute <- cbind(rep(1:4,each=6), c(2,2,3,3,4,4,1,1,3,3,4,4,1,1,2,2,4,4,1,1,2,2,3,3),
                 c(3,4,2,4,2,3,3,4,1,4,1,3,2,4,1,4,1,2,2,3,1,3,1,2),
                 c(4,3,4,2,3,2,4,3,4,1,3,1,4,2,4,1,2,1,3,2,3,1,2,1))
permute

# dose escalation ordering
check <- NULL; p <- 1
cohort8 <- NULL; cohort9 <- NULL
simpower8 <- NULL; simpower9 <- NULL
expstages8 <- NULL; expstages9 <- NULL
proportions8 <- NULL; proportions9 <- NULL
for (i in 1:8) {
  for (j in 1:5) {
    usethis1 <- findcohortc3(targetpower=0.8, possiblec=seq(30,130), J=4, m=2,
    R=2, K=8, model=j, dose=c(0,2,4,6,8), ordering=ordering1, stdev=1, delta=0, b=effbounds[i,])
    usethis2 <- findcohortc3(targetpower=0.9, possiblec=seq(50,160), J=4, m=2,
    R=2, K=8, model=j, dose=c(0,2,4,6,8), ordering=ordering1, stdev=1, delta=0, b=effbounds[i,])
    for (k in 1:4) {
      if (k==1) { cohort8 <- rbind(cohort8, usethis1[[3]])
                  cohort9 <- rbind(cohort9, usethis2[[3]])
      } else if (k==2) { simpower8 <- rbind(simpower8, usethis1[[1]][usethis1[[2]]])
                         simpower9 <- rbind(simpower9, usethis2[[1]][usethis2[[2]]])
      } else if (k==3) { proportions8 <- rbind(proportions8, matrix(usethis1[[6]],c(1,J)))
                         proportions9 <- rbind(proportions9, matrix(usethis2[[6]],c(1,J)))
      } else if (k==4) { expstages8 <- rbind(expstages8, usethis1[[5]])
                         expstages9 <- rbind(expstages9, usethis2[[5]])}
    }
    check <- matrix(c(i,j,p),c(1,3))
    write.table(check, "check1.txt", append=T, quote=F, row.names=F, col.names=F)
  }
}
dim(cohort8) <- c(5,8); cohort8; write.table(cohort8, "8cohortorderesca.txt",
append=T, quote=F, row.names=F, col.names=F)
dim(cohort9) <- c(5,8); cohort9; write.table(cohort9, "9cohortorderesca.txt",
append=T, quote=F, row.names=F, col.names=F)

plannedss8 <- cohort8*K*((1/R)+1); plannedss8; write.table(plannedss8,
"8plannedssorderesca.txt", append=T, quote=F, row.names=F, col.names=F)
plannedss9 <- cohort9*K*((1/R)+1); plannedss9; write.table(plannedss9,
"9plannedssorderesca.txt", append=T, quote=F, row.names=F, col.names=F)

dim(simpower8) <- c(5,8); simpower8; write.table(simpower8,
"8simpowerorderesca.txt", append=T, quote=F, row.names=F, col.names=F)
dim(simpower9) <- c(5,8); simpower9; write.table(simpower9,
```

```
"9simpowerorderesca.txt", append=T, quote=F, row.names=F, col.names=F)

dim(expstages8) <- c(5,8); expstages8; write.table(expstages8,
"8expstageorderesca.txt", append=T, quote=F, row.names=F, col.names=F)
dim(expstages9) <- c(5,8); expstages9; write.table(expstages9,
"9expstageorderesca.txt", append=T, quote=F, row.names=F, col.names=F)

expectedss8 <- cohort8*expstages8*((1/R)+1); expectedss8; write.table(expectedss8,
"8expssorderesca.txt", append=T, quote=F, row.names=F, col.names=F)
expectedss9 <- cohort9*expstages9*((1/R)+1); expectedss9; write.table(expectedss9,
"9expssorderesca.txt", append=T, quote=F, row.names=F, col.names=F)

proportions8; write.table(proportions8, "8proporderesca.txt", append=T, quote=F,
row.names=F, col.names=F)
proportions9; write.table(proportions9, "9proporderesca.txt", append=T, quote=F,
row.names=F, col.names=F)

# informative ordering
check <- NULL; p <- 2
cohort8 <- NULL; cohort9 <- NULL
simpower8 <- NULL; simpower9 <- NULL
expstages8 <- NULL; expstages9 <- NULL
proportions8 <- NULL; proportions9 <- NULL
for (i in 1:8) {
  for (j in 1:5) {
    usethis1 <- findcohortc3(targetpower=0.8, possiblec=seq(30,150), J=4, m=2,
    R=2, K=8, model=j, dose=c(0,2,4,6,8), ordering=informordering[j,], stdev=1, delta=0,
    b=effbounds[i,])
    usethis2 <- findcohortc3(targetpower=0.9, possiblec=seq(50,180), J=4, m=2,
    R=2, K=8, model=j, dose=c(0,2,4,6,8), ordering=informordering[j,], stdev=1, delta=0,
    b=effbounds[i,])
    for (k in 1:4) {
      if (k==1) { cohort8 <- rbind(cohort8, usethis1[[3]])
                  cohort9 <- rbind(cohort9, usethis2[[3]])
      } else if (k==2) { simpower8 <- rbind(simpower8, usethis1[[1]][usethis1[[2]]])
                         simpower9 <- rbind(simpower9, usethis2[[1]][usethis2[[2]]])
      } else if (k==3) { proportions8 <- rbind(proportions8, matrix(usethis1[[6]],c(1,J)))
                         proportions9 <- rbind(proportions9, matrix(usethis2[[6]],c(1,J)))
      } else if (k==4) { expstages8 <- rbind(expstages8, usethis1[[5]])
                         expstages9 <- rbind(expstages9, usethis2[[5]])}
    }
    check <- matrix(c(i,j,p),c(1,3))
    write.table(check, "check2.txt", append=T, quote=F, row.names=F, col.names=F)
  }
}
dim(cohort8) <- c(5,8); cohort8; write.table(cohort8, "8cohortorderinform.txt", append=T,
quote=F, row.names=F, col.names=F)
dim(cohort9) <- c(5,8); cohort9; write.table(cohort9, "9cohortorderinform.txt", append=T,
quote=F, row.names=F, col.names=F)

plannedss8 <- cohort8*K*((1/R)+1); plannedss8; write.table(plannedss8,
"8plannedssorderinform.txt", append=T, quote=F, row.names=F, col.names=F)
plannedss9 <- cohort9*K*((1/R)+1); plannedss9; write.table(plannedss9,
"9plannedssorderinform.txt", append=T, quote=F, row.names=F, col.names=F)

dim(simpower8) <- c(5,8); simpower8; write.table(simpower8, "8simpowerorderinform.txt",
append=T, quote=F, row.names=F, col.names=F)
dim(simpower9) <- c(5,8); simpower9; write.table(simpower9, "9simpowerorderinform.txt",
append=T, quote=F, row.names=F, col.names=F)

dim(expstages8) <- c(5,8); expstages8; write.table(expstages8, "8expstageorderinform.txt",
append=T, quote=F, row.names=F, col.names=F)
dim(expstages9) <- c(5,8); expstages9; write.table(expstages9, "9expstageorderinform.txt",
append=T, quote=F, row.names=F, col.names=F)
```

```
expectedss8 <- cohort8*expstages8*((1/R)+1); expectedss8; write.table(expectedss8,
"8expssorderinform.txt", append=T, quote=F, row.names=F, col.names=F)
expectedss9 <- cohort9*expstages9*((1/R)+1); expectedss9; write.table(expectedss9,
"9expssorderinform.txt", append=T, quote=F, row.names=F, col.names=F)

proportions8; write.table(proportions8, "8proporderinform.txt", append=T, quote=F,
row.names=F, col.names=F)
proportions9; write.table(proportions9, "9proporderinform.txt", append=T, quote=F,
row.names=F, col.names=F)

# uninformative ordering
check <- NULL
cohort8 <- NULL; cohort9 <- NULL
cohort8.2 <- NULL; cohort9.2 <- NULL
simpower8 <- NULL; simpower9 <- NULL
simpower8.2 <- NULL; simpower9.2 <- NULL
expstages8 <- NULL; expstages9 <- NULL
expstages8.2 <- NULL; expstages9.2 <- NULL
proportions8 <- NULL; proportions9 <- NULL
proportions8.2 <- matrix(rep(0,5*8*J),c(5*8,J))
proportions9.2 <- matrix(rep(0,5*8*J),c(5*8,J))
for (p in 1:24) { # ordering
  for (i in 1:8) { # error spending
    for (j in 1:5) { # dose response model
      usethis1 <- findcohortc3(targetpower=0.8, possiblec=seq(30,300), J=4, m=2, R=2,
      K=8, model=j, dose=c(0,2,4,6,8), ordering=permute[p,], stdev=1, delta=0,
      b=effbounds[i,])
      usethis2 <- findcohortc3(targetpower=0.9, possiblec=seq(30,300), J=4, m=2, R=2,
      K=8, model=j, dose=c(0,2,4,6,8), ordering=permute[p,], stdev=1, delta=0,
      b=effbounds[i,])
      for (k in 1:4) {
        if (k==1) { cohort8 <- rbind(cohort8, usethis1[[3]])
                    cohort9 <- rbind(cohort9, usethis2[[3]])
        } else if (k==2) { simpower8 <- rbind(simpower8, usethis1[[1]][usethis1[[2]]])
                           simpower9 <- rbind(simpower9, usethis2[[1]][usethis2[[2]]])
        } else if (k==3) { proportions8 <- rbind(proportions8, matrix(usethis1[[6]],c(1,J)))
                           proportions9 <- rbind(proportions9, matrix(usethis2[[6]],c(1,J)))
        } else if (k==4) { expstages8 <- rbind(expstages8, usethis1[[5]])
                           expstages9 <- rbind(expstages9, usethis2[[5]])}
      }
      check <- matrix(c(i,j,p),c(1,3))
      write.table(check, "check3.txt", append=T, quote=F, row.names=F, col.names=F)
    }
  }
  dim(cohort8) <- c(1,40); dim(cohort9) <- c(1,40)
  write.table(cohort8, "cohort8.txt", append=T, quote=F, row.names=F, col.names=F)
  write.table(cohort9, "cohort9.txt", append=T, quote=F, row.names=F, col.names=F)
  cohort8.2 <- rbind(cohort8.2, cohort8)
  cohort9.2 <- rbind(cohort9.2, cohort9)
  cohort8 <- NULL; cohort9 <- NULL

  dim(simpower8) <- c(1,40); dim(simpower9) <- c(1,40)
  write.table(simpower8, "simpower8.txt", append=T, quote=F, row.names=F, col.names=F)
  write.table(simpower9, "simpower9.txt", append=T, quote=F, row.names=F, col.names=F)
  simpower8.2 <- rbind(simpower8.2, simpower8)
  simpower9.2 <- rbind(simpower9.2, simpower9)
  simpower8 <- NULL; simpower9 <- NULL

  dim(expstages8) <- c(1,40); dim(expstages9) <- c(1,40)
  write.table(expstages8, "expstages8.txt", append=T, quote=F, row.names=F, col.names=F)
  write.table(expstages9, "expstages9.txt", append=T, quote=F, row.names=F, col.names=F)
  expstages8.2 <- rbind(expstages8.2, expstages8)
  expstages9.2 <- rbind(expstages9.2, expstages9)
```

```
  expstages8 <- NULL; expstages9 <- NULL

  write.table(proportions8, "proportions8.txt", append=T, quote=F, row.names=F, col.names=F)
  write.table(proportions9, "proportions9.txt", append=T, quote=F, row.names=F, col.names=F)
  proportions8.2 <- proportions8.2+proportions8
  proportions9.2 <- proportions9.2+proportions9
  proportions8 <- NULL; proportions9 <- NULL
}

cohort8.2
cohort8.3 <- apply(cohort8.2, 2, mean)
write.table(matrix(cohort8.3,c(5,8)), "8cohortorderuninf.txt", append=T, quote=F,
row.names=F, col.names=F)
dim(cohort8.3) <- c(5,8)
cohort8.3

cohort9.2
cohort9.3 <- apply(cohort9.2, 2, mean)
write.table(matrix(cohort9.3,c(5,8)), "9cohortorderuninf.txt", append=T, quote=F,
row.names=F, col.names=F)
dim(cohort9.3) <- c(5,8)
cohort9.3

plannedss8 <- cohort8.3*K*((1/R)+1); plannedss8; write.table(plannedss8,
"8plannedssorderuninf.txt", append=T, quote=F, row.names=F, col.names=F)
plannedss9 <- cohort9.3*K*((1/R)+1); plannedss9; write.table(plannedss9,
"9plannedssorderuninf.txt", append=T, quote=F, row.names=F, col.names=F)

simpower8.2
simpower8.3 <- apply(simpower8.2, 2, mean)
write.table(matrix(simpower8.3,c(5,8)), "8simpowerorderuninf.txt", append=T, quote=F,
row.names=F, col.names=F)
dim(simpower8.3) <- c(5,8)
simpower8.3

simpower9.2
simpower9.3 <- apply(simpower9.2, 2, mean)
write.table(matrix(simpower9.3,c(5,8)), "9simpowerorderuninf.txt", append=T, quote=F,
row.names=F, col.names=F)
dim(simpower9.3) <- c(5,8)
simpower9.3

expstages8.2
expstages8.3 <- apply(expstages8.2, 2, mean)
write.table(matrix(expstages8.3,c(5,8)), "8expstageorderuninf.txt", append=T, quote=F,
row.names=F, col.names=F)
dim(expstages8.3) <- c(5,8)
expstages8.3

expstages9.2
expstages9.3 <- apply(expstages9.2, 2, mean)
write.table(matrix(expstages9.3,c(5,8)), "9expstageorderuninf.txt", append=T, quote=F,
row.names=F, col.names=F)
dim(expstages9.3) <- c(5,8)
expstages9.3

expectedss8 <- cohort8.3*expstages8.3*((1/R)+1); expectedss8; write.table(expectedss8,
"8expssorderuninf.txt", quote=F, row.names=F, col.names=F)
expectedss9 <- cohort9.3*expstages9.3*((1/R)+1); expectedss9; write.table(expectedss9,
"9expssorderuninf.txt", quote=F, row.names=F, col.names=F)

proportions8.3 <- proportions8.2/factorial(4); proportions8.3
proportions9.3 <- proportions9.2/factorial(4); proportions9.3
write.table(proportions8.3, "8proporderuninf.txt", append=T, quote=F, row.names=F, col.names=F)
```

```
write.table(proportions9.3, "9proporderuninf.txt", append=T, quote=F, row.names=F, col.names=F)

# set number of simulated trials
Nsim4power <- 10000

# parallel group design with dunnett's adjustment
powerdunnett8 <- matrix(rep(NA,5*8), c(5,8))
powerdunnett9 <- matrix(rep(NA,5*8), c(5,8))
p <- 8
for (i in 1:8) {
  for (j in 1:5) {
    usethis <- rejectnullprdunnett(J=4, ss=ceiling(expectedss8[j,i]/(J+1)), model=j,
    dose=c(0,2,4,6,8), stdev=1, delta=0, alternative="greater", alpha=0.05, Nsim=Nsim4power)
    powerdunnett8[j,i] <- usethis[[1]]
    check <- matrix(c(i,j,p),c(1,3))
    write.table(check, "check4.txt", append=T, quote=F, row.names=F, col.names=F)
  }
}
powerdunnett8; write.table(powerdunnett8, "powerdunnett8.txt", quote=F, col.names=F,
row.names=F)

p <- 9
for (i in 1:8) {
  for (j in 1:5) {
    usethis <- rejectnullprdunnett(J=4, ss=ceiling(expectedss9[j,i]/(J+1)), model=j,
    dose=c(0,2,4,6,8), stdev=1, delta=0, alternative="greater", alpha=0.05, Nsim=Nsim4power)
    powerdunnett9[j,i] <- usethis[[1]]
    check <- matrix(c(i,j,p),c(1,3))
    write.table(check, "check4.txt", append=T, quote=F, row.names=F, col.names=F)
  }
}
powerdunnett9; write.table(powerdunnett9, "powerdunnett9.txt", quote=F, col.names=F,
row.names=F)

# drop-the-losers design
droploser <- rep(NA, 10)
for (i in 1:10) {
  droploser[i] <- effbounddroploser(J=4, m=2, c=20, model=0, dose=c(0,2,4,6,8), stdev=1,
  delta=0, a=0, Nsim=5000, decimal=3, limits=c(1.8,2.2), alpha=0.05)
}
droploser
droploser <- mean(droploser) # 2.0588
droploser
write.table(droploser, "dtleffbound.txt", quote=F, col.names=F, row.names=F)

powerdroploser8 <- matrix(rep(NA,5*8), c(5,8))
powerdroploser9 <- matrix(rep(NA,5*8), c(5,8))
proportions8 <- NULL
proportions9 <- NULL

p <- 8
for (i in 1:8) {
  for (j in 1:5) {
    usethis <- rejectnullprdroploser(J=4, c=ceiling(expectedss8[j,i]/(4+3)), model=j,
    dose=c(0,2,4,6,8), stdev=1, delta=0, a=0, b=droploser, Nsim=Nsim4power)
    powerdroploser8[j,i] <- usethis[[1]]
    proportions8 <- rbind(proportions8, usethis[[2]])
    check <- matrix(c(i,j,p),c(1,3))
    write.table(check, "check5.txt", append=T, quote=F, row.names=F, col.names=F)
  }
}
powerdroploser8; write.table(powerdroploser8, "powerdroploser8.txt", quote=F, col.names=F,
row.names=F)
proportions8; write.table(proportions8, "droplosersprop8.txt", quote=F, col.names=F, row.names=F)
```

```
p <- 9
for (i in 1:8) {
  for (j in 1:5) {
    usethis <- rejectnullprdroploser(J=4, c=ceiling(expectedss9[j,i]/(4+3)), model=j,
    dose=c(0,2,4,6,8), stdev=1, delta=0, a=0, b=droploser, Nsim=Nsim4power)
    powerdroploser9[j,i] <- usethis[[1]]
    proportions9 <- rbind(proportions9, usethis[[2]])
    check <- matrix(c(i,j,p),c(1,3))
    write.table(check, "check5.txt", append=T, quote=F, row.names=F, col.names=F)
  }
}
powerdroploser9; write.table(powerdroploser9, "powerdroploser9.txt", quote=F, col.names=F,
row.names=F)
proportions9; write.table(proportions9, "droplosersprop9.txt", quote=F, col.names=F, row.names=F)


# dose-response informed seamless design
forbretz8 <- read.table("8expssorderinform.txt")
forbretz8 <- forbretz8[c(-1,-4),] # remove flat and logistic expected sample sizes
forbretz9 <- read.table("9expssorderinform.txt")
forbretz9 <- forbretz8[c(-1,-4),] # remove flat and logistic expected sample sizes

modelno <- c(2,3,5)
powerbretz8 <- matrix(rep(NA,3*8), c(3,8))
powerbretz9 <- matrix(rep(NA,3*8), c(3,8))
propbretz8 <- NULL
propbretz9 <- NULL

p <- 10
for (i in 1:8) {
  for (j in 1:3) {
    usethis <- rejectnullprbretz(J=4, c=ceiling(forbretz8[j,i]/(4+3)), model=modelno[j],
    dose=dose, stdev=1, delta=0, b=1.645, Nsim=Nsim4power)
    powerbretz8[j,i] <- usethis[[1]]
    propbretz8 <- rbind(propbretz8, usethis[[2]])
    check <- matrix(c(i,j,p),c(1,3))
    write.table(check, "check6.txt", append=T, quote=F, row.names=F, col.names=F)
  }
}
powerbretz8; write.table(powerbretz8, "powerbretz8.txt", quote=F, col.names=F, row.names=F)
propbretz8; write.table(propbretz8, "bretzprop8.txt", quote=F, col.names=F, row.names=F)

p <- 10
for (i in 1:8) {
  for (j in 1:3) {
    usethis <- rejectnullprbretz(J=4, c=ceiling(forbretz9[j,i]/(4+3)), model=modelno[j],
    dose=dose, stdev=1, delta=0, b=1.645, Nsim=Nsim4power)
    powerbretz9[j,i] <- usethis[[1]]
    propbretz9 <- rbind(propbretz9, usethis[[2]])
    check <- matrix(c(i,j,p),c(1,3))
    write.table(check, "check6.txt", append=T, quote=F, row.names=F, col.names=F)
  }
}
powerbretz9; write.table(powerbretz9, "powerbretz9.txt", quote=F, col.names=F, row.names=F)
propbretz9; write.table(propbretz9, "bretzprop9.txt", quote=F, col.names=F, row.names=F)

# end of chapter code
```

Chapter 4

Variants of Adaptive Staggered Dose Design

4.1 Introduction

In Chapter 3, we have proposed an adaptive staggered dose design for a normal endpoint and have fully investigated its operating characteristics via simulation and numerical studies. We have also discussed its strengths as well as limitations. In this chapter, we are mainly interested in varying four of the design parameters as in Table 3.1 to further characterize its performance in treatment selection and confirmation. In the current context, the objective of the clinical trial remains the same: selecting one or two efficacious doses and confirming the efficacy and safety of these selected doses. As we know, a clinical development is a very costly process and requires scrupulous attention to its design and conduct. Therefore, the number of interim monitoring stages in a sequential trial should be carefully planned. The more interim analyses are planned in a trial, the more costly it may become and the more likely operational bias may be unintentionally introduced into the trial. Therefore, a trial with fewer interim stages may be desirable. As for randomization ratio, Peto (1978) has advocated the use of unequal allocation such as 2:1 of intervention to control. The rationale for such allocation is that the study may gain more information about subjects responses to the new treatment, especially if adverse event rate is low. Also, as the risk-to-benefit profile for the control treatment has been established in the past, more subjects may be able to

benefit from a potentially more effective new intervention. However, unequal allocation has been known to decrease overall statistical power, and so a delicate balance exists between unequal allocation and ethical considerations (Peto *et al.*, 1976). Finally, alpha spending function has been used to monitor group sequential clinical trials. When multiple arms are involved, Chen, DeMets, and Lan (2010) suggested using marginal monitoring with each dose-to-control comparison monitored separately by its own alpha spending function with a pre-specified nominal alpha level $\alpha_j (j = 1, 2, ..., J)$ where J is the number of experimental doses. In this case, they suggested applying a Bonferroni adjustment such that $\alpha_j = \alpha/J$. Many of these proposals can be used for improving the efficiency of multi-arm clinical trials.

In the following sections, we will look at several variations of the adaptive staggered dose designs. Just like the specific version discussed in Chapter 3, the following setting will apply to all of these variants, (1) J doses or arms of the experimental treatment and one control dose, (2) dose ordering denoted by $\{d_1, d_2, ..., d_J\}$ and control dose denoted by d_0, (3) normal model for the endpoint as in (3.2.1), (4) one-sided hypotheses as in (3.2.2), (5) c as the cohort size per dose and stage for the experimental dose, (6) $1 : R$ as the randomization ratio between the control and experimental dose, (7) D as the number of doses being studied per stage, (8) M per-dose stages and K global stages of the trial, (9) standardized test statistics as in (3.2.3), (10) no futility analysis, (11) four dose response curves $\mu_j = f(d_j)$ as in Table 3.2, and (12) $\alpha(t)$ as the alpha spending function where t is the information fraction. The distributions of the test statistics under the null hypotheses are given in (3.3.1) while those under the alternative hypotheses are given in (3.2.4).

We consider four variations of the design and compare them with the specific version in Chapter 3 ($D = 1, M = 2, R = 2$, global $\alpha(t)$) in their operating characteristics. In Section 4.2, we look at a design that studies two concurrent doses at each stage ($D = 2$). In Section 4.3, we investigate a design that allows only one stage per dose ($M = 1$). In Section 4.4, we explore a design with each dose assigned its own marginal alpha spending function ($\alpha_j(t)$). In Section 4.5, we vary the randomization ratio to $R = 3$. We want to assess the impact of

these design parameters on the overall performance of the staggered dose design. Finally, in Section 4.7, we will discuss how this adaptive staggered dose design can be extended to binary and time-to-event endpoints.

4.2 Two Concurrent Doses $D = 2$

In this section, we want to consider a variant of the adaptive staggered dose design that tests the hypotheses of two concurrent doses with the control dose at each interim stage. That means, we are interested in $D = 2$ (see Table 3.1), while keeping other design parameters constant at $M = 2$ and $R = 2$. When we allow each dose to have a maximum of $M = 2$ *per-dose* stages, the total number of *global* stages will be $K = J$ stages if J is even, but $K = J + 1$ if J is odd. The decision rule is that if at least one of the two null hypotheses of the two doses is rejected, then the trial will stop to declare efficacy. Therefore, if at the kth $(k = 1, ..., K)$ interim stage, we are testing two doses, d_{j_1} and d_{j_2}, we can denote the probability of rejecting either of the null hypotheses, $\{H_{j_1 0}, H_{j_2 0}\}$ or both of them by $\psi_{(j_1,j_2),k}$ when they are both true, where $(j_1, j_2) \in \{(1, 2), (3, 4), ...\}$. Using the distributions of the test statistics Z and (Z_1, Z_2) in (3.3.1) under the global null hypothesis, and when no futility analysis is performed, the probability of rejecting either $H_{j_1 0}$, or $H_{j_2 0}$, or both can be given by

$$
\begin{aligned}
&\psi_{(j_1,j_2),k} \\
&= \begin{cases}
\left(P(Z > b_k)\right)^2 + 2P(Z \leqslant b_k)P(Z > b_k) & k = 1 \\
\left(P(Z_1 \leqslant b_{k-1}, Z_2 > b_k)\right)^2 + 2P(Z_1 \leqslant b_{k-1}, Z_2 \leqslant b_k)P(Z_1 \leqslant b_{k-1}, Z_2 > b_k) & k = 2 \\
\left[\prod_{i=1}^{\frac{k-1}{2}} \left(P(Z_1 \leqslant b_{2i-1}, Z_2 \leqslant b_{2i})\right)^2\right] \left[\left(P(Z > b_k)\right)^2 + 2P(Z \leqslant b_k)P(Z > b_k)\right] & k = j_1 \\
\left[\prod_{i=1}^{\frac{k-2}{2}} \left(P(Z_1 \leqslant b_{2i-1}, Z_2 \leqslant b_{2i})\right)^2\right] \times & \\
\left[\left(P(Z_1 \leqslant b_{k-1}, Z_2 > b_k)\right)^2 + 2P(Z_1 \leqslant b_{k-1}, Z_2 \leqslant b_k)P(Z_1 \leqslant b_{k-1}, Z_2 > b_k)\right] & k = j_2.
\end{cases}
\end{aligned}
$$

We have assumed conservatively here that the test statistics of the two concurrent doses against the control dose are independent and we have also applied the Sĭdák's adjustment

in (4.2.1). Stopping boundary values b_k's can be evaluated accordingly when an alpha spending function is specified. If one assumes that these two test statistics are not independent due to the use of the same control in the test statistics, one can use Dunnett's adjustment to calculate the stopping boundary values. In this case, the test statistics can be re-parameterized in terms of independent stochastic increments and stopping boundaries values can be evaluated on this re-parameterized scale. The overall type I error under the global null hypothesis for this trial is given by

$$\psi = \sum_{i=1}^{J/2} (\psi_{(2i-1,2i),2i-1} + \psi_{(2i-1,2i),2i})$$

if J is even. This overall type I error may be smaller than the overall type I error when independence of test statistics is not assumed. The goal is to keep this overall type I error under a target level of α. Based on this setting, we can also obtain the form of the stage-wise statistical power or boundary crossing probabilities, $\xi_{(j_1,j_2),k}$ under the alternative hypotheses as follows:

$$\xi_{(j_1,j_2),k}$$

$$= \begin{cases} \Phi(-\omega_{j_1,k})\Phi(-\omega_{j_2,k}) + \Phi(-\omega_{j_1,k})\Phi(\omega_{j_2,k}) + \Phi(-\omega_{j_2,k})\Phi(\omega_{j_1,k}) & k = 1 \\ \\ \left(\int_{\omega_{j_1,k}}^{\infty} \Phi\left(\sqrt{2}\omega_{j_1,k-1} - t\right)\phi(t)dt\right)\left(\int_{\omega_{j_2,k}}^{\infty} \Phi\left(\sqrt{2}\omega_{j_2,k-1} - t\right)\phi(t)dt\right) & \\ + \left(\int_{\omega_{j_1,k}}^{\infty} \Phi\left(\sqrt{2}\omega_{j_1,k-1} - t\right)\phi(t)dt\right)\left(\int_{-\infty}^{\omega_{j_2,k}} \Phi\left(\sqrt{2}\omega_{j_2,k-1} - t\right)\phi(t)dt\right) & \\ + \left(\int_{\omega_{j_2,k}}^{\infty} \Phi\left(\sqrt{2}\omega_{j_2,k-1} - t\right)\phi(t)dt\right)\left(\int_{-\infty}^{\omega_{j_1,k}} \Phi\left(\sqrt{2}\omega_{j_1,k-1} - t\right)\phi(t)dt\right) & k = 2 \\ \\ \left[\prod_{i=1}^{\frac{k-1}{2}} \left(\int_{-\infty}^{\omega_{i,i+1}} \Phi\left(\sqrt{2}\omega_{i,i} - t\right)\phi(t)dt\right)\left(\int_{-\infty}^{\omega_{i+1,i+1}} \Phi\left(\sqrt{2}\omega_{i+1,i} - t\right)\phi(t)dt\right)\right] & \\ \times [\Phi(-\omega_{j_1,k})\Phi(-\omega_{j_2,k}) + \Phi(-\omega_{j_1,k})\Phi(\omega_{j_2,k}) + \Phi(-\omega_{j_2,k})\Phi(\omega_{j_1,k})] & k = j_1 \\ \\ \left[\prod_{i=1}^{\frac{k-2}{2}} \left(\int_{-\infty}^{\omega_{i,i+1}} \Phi\left(\sqrt{2}\omega_{i,i} - t\right)\phi(t)dt\right)\left(\int_{-\infty}^{\omega_{i+1,i+1}} \Phi\left(\sqrt{2}\omega_{i+1,i} - t\right)\phi(t)dt\right)\right] & \\ \times \left[\left(\int_{\omega_{j_1,k}}^{\infty} \Phi\left(\sqrt{2}\omega_{j_1,k-1} - t\right)\phi(t)dt\right)\left(\int_{\omega_{j_2,k}}^{\infty} \Phi\left(\sqrt{2}\omega_{j_2,k-1} - t\right)\phi(t)dt\right)\right. & \\ + \left(\int_{\omega_{j_1,k}}^{\infty} \Phi\left(\sqrt{2}\omega_{j_1,k-1} - t\right)\phi(t)dt\right)\left(\int_{-\infty}^{\omega_{j_2,k}} \Phi\left(\sqrt{2}\omega_{j_2,k-1} - t\right)\phi(t)dt\right) & \\ \left. + \left(\int_{\omega_{j_2,k}}^{\infty} \Phi\left(\sqrt{2}\omega_{j_2,k-1} - t\right)\phi(t)dt\right)\left(\int_{-\infty}^{\omega_{j_1,k}} \Phi\left(\sqrt{2}\omega_{j_1,k-1} - t\right)\phi(t)dt\right)\right] & k = j_2 \end{cases}$$

since doses (d_{j_1}, d_{j_2}) are explored at interim stages $k = j_1$ and $k = j_2 = j_1 + 1$ and

$$\omega_{j_1,k} = b_k - \frac{f(d_{j_1}) - \mu_0}{\sqrt{\frac{R+1}{c}}},\ \omega_{j_2,k} = b_k - \frac{f(d_{j_2}) - \mu_0}{\sqrt{\frac{R+1}{c}}} \text{ when } k = j_1,$$

$$\omega_{j_1,k} = b_k - \frac{f(d_{j_1}) - \mu_0}{\sqrt{\frac{R+1}{2c}}},\ \omega_{j_2,k} = b_k - \frac{f(d_{j_2}) - \mu_0}{\sqrt{\frac{R+1}{2c}}} \text{ when } k = j_2 = j_1 + 1.$$

In addition, the expected stage for stopping for efficacy, if J is even and therefore $J = K$, is given by

$$\begin{aligned} E(\mathcal{K}) &= \xi_{(1,2),1} + 2\xi_{(1,2),2} + 3\xi_{(3,4),3} + ... + J\left(1 - \xi_{(1,2),1} - ... - \xi_{(J-1,J),J-1}\right) \\ &= J - (J-1)\xi_{(1,2),1} - (J-2)\xi_{(1,2),2} - ... - \xi_{(J-1,J),J-1} \\ &= J - \left[\sum_{i=1}^{J/2-1} (J - 2i + 1)\,\xi_{(2i-1,2i),2i-1} + (J-2i)\xi_{(2i-1,2i),2i}\right] - \xi_{(J-1,J),J-1} \end{aligned}$$

and the expected trial sample size is given by

$$E(\mathcal{S}) = c\left(\frac{1}{R} + 2\right) E(\mathcal{K}).$$

It can be seen that $E(\mathcal{K}) < J$ if stopping probabilities are large. Also, for the case when J is odd, the expected stage and sample size can also be similarly derived.

In this simulation, we want to compare the cohort sizes, expected stages, and expected sample sizes between designs with $D = 1$ and $D = 2$ under the four dose response models: flat, linear, emax, and umbrella models, given dose escalation and informative and uninformative orderings with the alpha spending plans of Pocock, O'Brien-Fleming, $\rho = 0.3, 1, 3$ to attain statistical power of 90%. In addition, we also want to characterize the relationship between statistical power and expected sample size for these two designs. For these designs, we assume $J = 4$, $R = 2$, $M = 2$, and $K = J = 4$ since J is even.

Table 4.1 gives the stopping boundary values that keep the type I error under one-sided

α-level of 0.05. We can see that Pocock-type boundary and Rho with $\rho = 0.3$ tend to favor earlier stages and hence the earlier doses. For Rho function with $\rho = 1.0$, error spending is similar across stages. Table 4.2 displays the cohort sizes, expected stages, expected sample sizes under this variant design while Figures 4.1, 4.2, 4.3, and 4.4 plot the statistical power of the designs across expected trial sample size. As expected, we can see the variant design with $D = 2$ achieves better power than the original design with $D = 1$ given the same expected sample size. This is observed almost unanimously under different dose response curves, alpha spending plans, and dose orderings. This variant design, again like the original design, performs best when informative dose ordering is used.

However, under the linear dose response model with informative ordering and especially Rho function with $\rho = 0.3$, the variant design appears to perform slightly worse than the original design when expected sample size increases. As observed, the expected sample size is not monotonic with the overall statistical power, due to the fact that expected sample size depends on (1) the cohort size and (2) expected stopping stage, as in (3.3.7). It can be verified that statistical power is monotonic with cohort size (see Section 3.6.2), but as c increases, under informative ordering where doses with better effects are placed earlier, the expected stage in stopping the trial $E(\mathcal{K})$ can decrease substantially. Therefore, the increase in c can be offset by greater decrease in $E(\mathcal{K})$, resulting in decreased expected sample size $E(\mathcal{S})$ even when power increases. This explains the curved plot and the non-monotonicity.

The operating characteristics of this variant design using $D = 2$ is presented, and the expected sample size given the same power is lower than that of the original design with $D = 1$. It is possible that since we have two concurrent doses in each interim stage, this may increase the probability of at least one dose is declared efficacious at each interim stage. In addition, fewer subjects are needed for the control dose as the number of global stages decreases. This variant design also shortens the duration of the trial. However, in case when two doses are selected simultaneously due to joint statistical significance, it is still unsure which dose of the two should be used for the indication.

Table 4.1: Five alpha spending schemes and their corresponding stage-wise efficacy stopping boundaries (b_k) and errors (α_k) for $J = 4, D = 2, M = 2, R = 2$, and $K = J = 4$. Family-wise type I error is controlled at one-sided $\alpha = 0.05$. No futility is adopted, $a_k = -\infty$ for all k.

Error Spending Scheme	k	1	2	3	4
1. Pocock-type boundary	b_k	2.37	2.37	2.56	2.53
	α_k	0.0179	0.0131	0.010	0.0086
2. O'Brien & Fleming-type boundary	b_k	3.92	2.78	2.37	2.13
	α_k	0.0001	0.0055	0.0180	0.0264
3. Rho error spending $\rho = 0.3$	b_k	2.13	2.48	2.78	2.78
	α_k	0.0330	0.0076	0.0052	0.0041
4. Rho error spending $\rho = 1.0$	b_k	2.50	2.41	2.49	2.40
	α_k	0.0125	0.0125	0.0125	0.0125
5. Rho error spending $\rho = 3.0$	b_k	3.36	2.76	2.44	2.11
	α_k	0.0008	0.0055	0.0148	0.0289

Figure 4.1: Comparison of statistical power under flat dose response model for variant design $D = 2$.

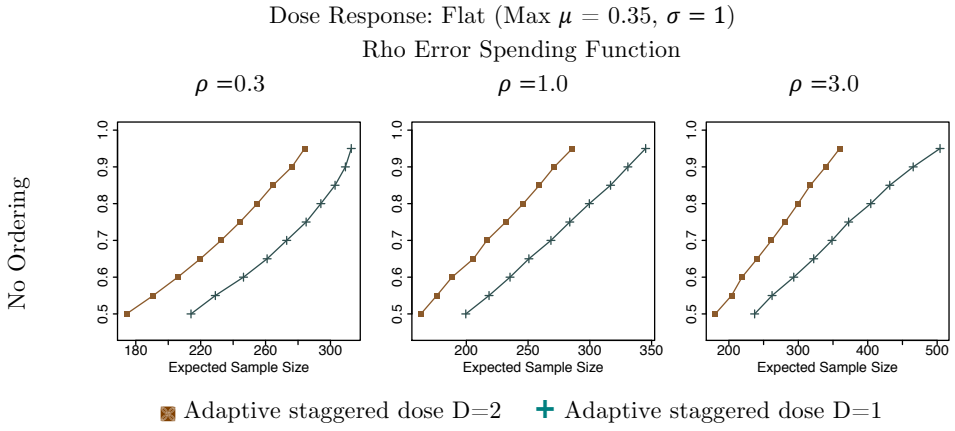


Table 4.2: Cohort size per stage (c), expected global stage to stop for efficacy ($E(\mathcal{K})$), and expected trial sample size $\left(\frac{c(2+R)}{R}E(\mathcal{K})\right)$ for attaining statistical power of 90% for variant design with $D = 2$.

Dose		Pocock-type boundary			O'Brien & Fleming boundary			—		
Response		Escalation	Informative	Uninformative	Escalation	Informative	Uninformative	-	-	-
1. Flat	cohort size	42	42	42	45	45	45	-	-	-
	expected stage	2.5	2.5	2.5	3.1	3.1	3.1	-	-	-
	expected sample size	266	266	266	349	349	349	-	-	-
2. Linear	cohort size	98	88	85	90	100	89	-	-	-
	expected stage	3.1	2.0	2.6	3.4	2.6	3.1	-	-	-
	expected sample size	758	436	542	767	654	678	-	-	-
3. Emax	cohort size	58	56	57	60	61	60	-	-	-
	expected stage	2.8	2.3	2.5	3.2	2.9	3.1	-	-	-
	expected sample size	400	323	356	487	448	642	-	-	-
4. Umbrella	cohort size	59	56	58	61	62	61	-	-	-
	expected stage	2.7	2.4	2.5	3.2	3.0	3.1	-	-	-
	expected sample size	401	330	363	491	459	472	-	-	-

Dose		Rho error spending ρ=0.3			Rho error spending ρ=1.0			Rho error spending ρ=3.0		
Response		Escalation	Informative	Uninformative	Escalation	Informative	Uninformative	Escalation	Informative	Uninformative
1. Flat	cohort size	48	48	48	41	41	41	44	44	44
	expected stage	2.3	2.3	2.3	2.6	2.6	2.6	3.1	3.1	3.1
	expected sample size	277	277	277	272	272	272	340	340	340
2. Linear	cohort size	111	97	95	93	87	83	88	97	86
	expected stage	3.0	1.8	2.4	3.2	2.1	2.7	3.4	2.6	3.0
	expected sample size	830	429	564	737	455	549	756	620	655
3. Emax	cohort size	66	62	64	56	54	55	58	59	58
	expected stage	2.6	2.1	2.3	2.9	2.4	2.7	3.2	2.9	3.1
	expected sample size	424	325	368	401	330	362	472	430	446
4. Umbrella	cohort size	66	63	65	57	55	56	59	60	60
	expected stage	2.5	2.1	2.7	2.8	2.5	x	3.2	2.9	3.1
	expected sample size	420	335	375	403	340	369	476	441	456

Figure 4.2: Comparison of statistical power under linear dose response model for variant design $D = 2$.

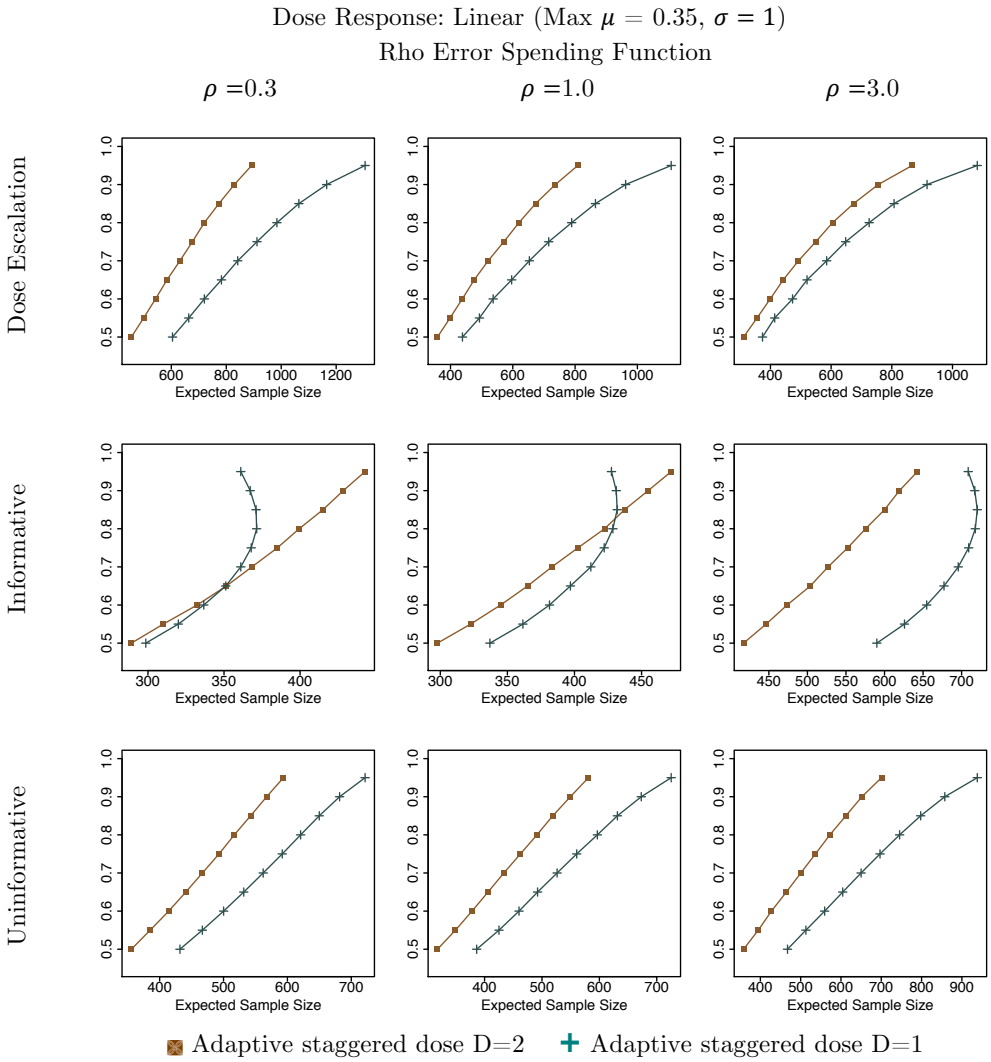


Figure 4.3: Comparison of statistical power under emax dose response model for variant design $D = 2$.

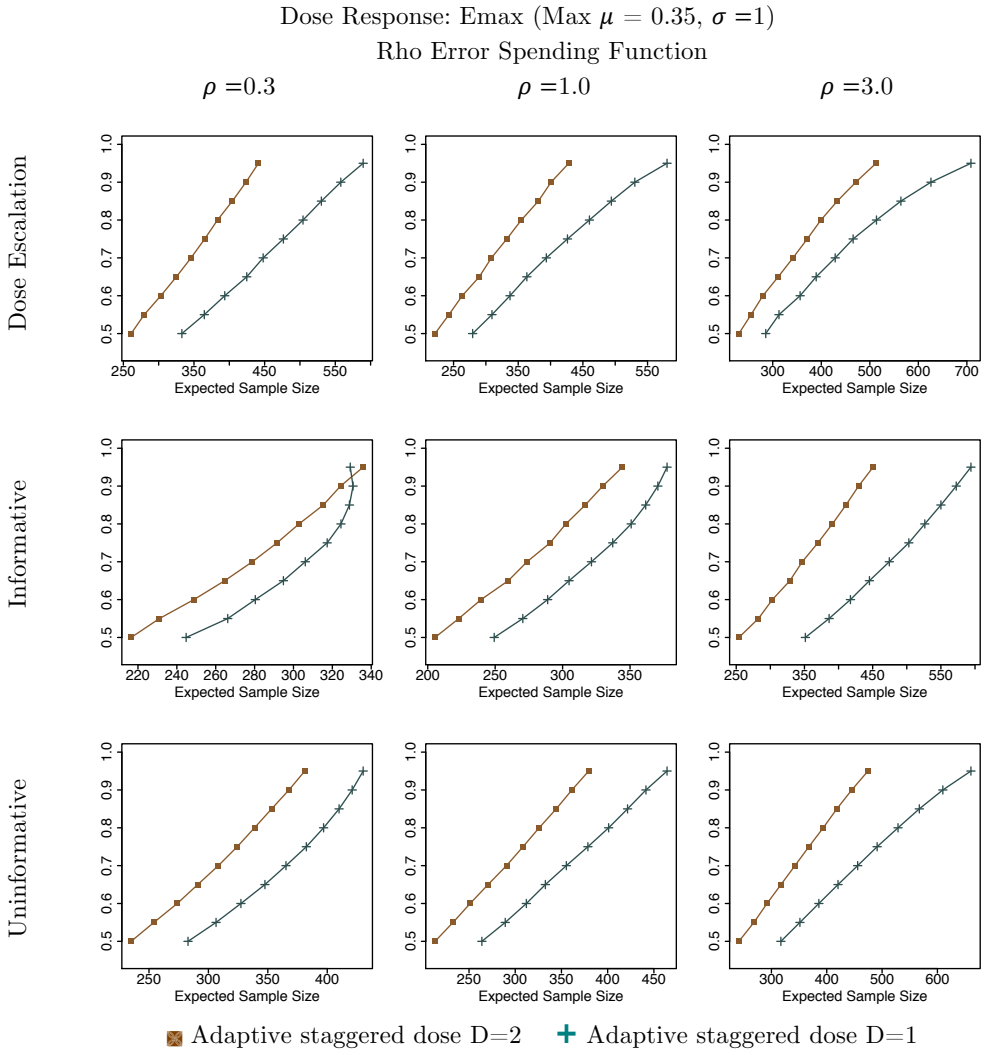
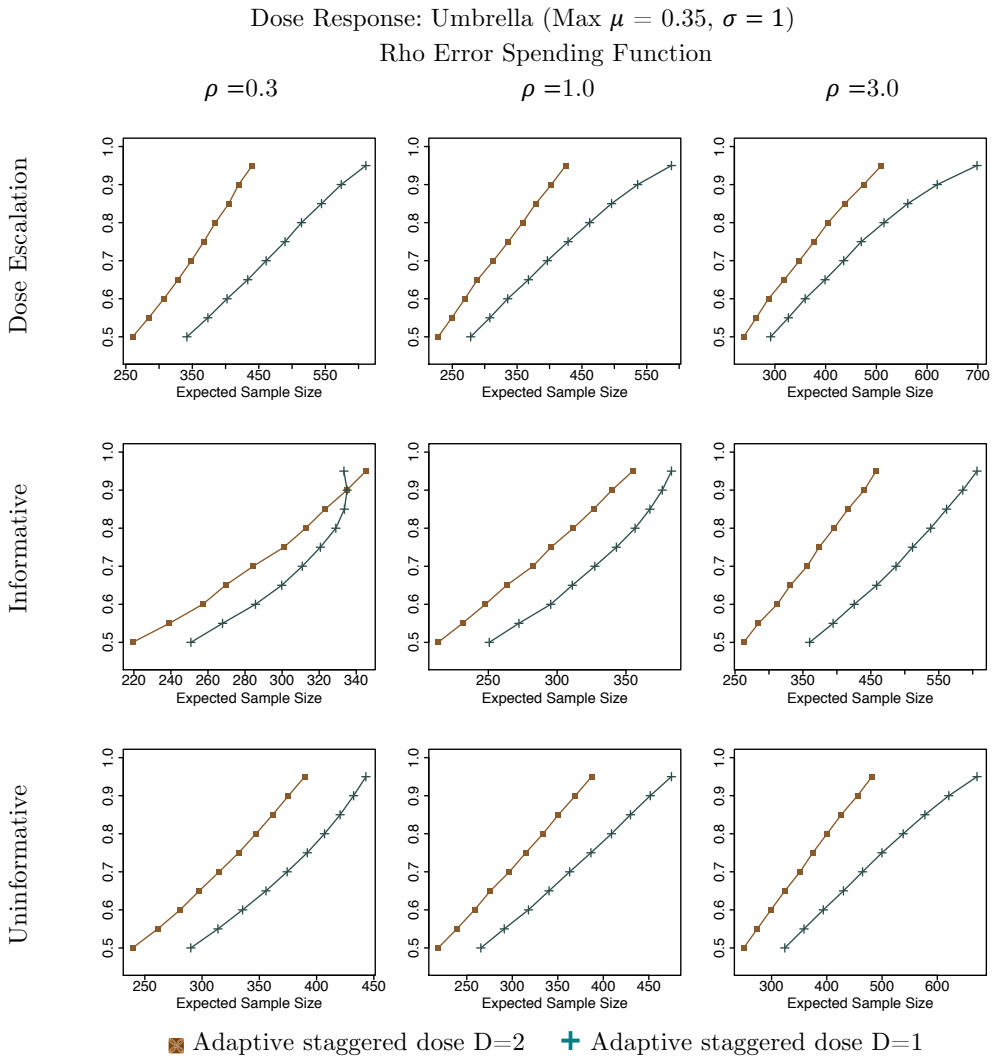


Figure 4.4: Comparison of statistical power under umbrella dose response model for variant design $D = 2$.



4.3 One Stage Per Dose $M = 1$

In this section, we want to explore another variant of the adaptive staggered dose design that only allows one stage per dose ($M = 1$) rather than two stages ($M = 2$) as in the original design. This variant design continues to stagger the doses according to their presumed effects and keeps other design parameters constant at $D = 1$ and $R = 2$. Therefore, in this

design, the total number of *global* stages is equal to the number of doses ($K = J$). At the kth interim stage, we are testing the dose $d_j (j = k)$ against the control dose, and we can represent the probability of rejecting the null hypothesis of this dose, H_{j0} when it is true, by ψ_j where $j = 1, 2, ..., J = K$. The stage-wise type I error can simply be given by

$$\psi_j = \begin{cases} P(Z > b_j) & \text{if } j = 1 \\ \left[\prod_{i=1}^{j-1} P(Z \leqslant b_i)\right] P(Z > b_j) & \text{if } j > 1. \end{cases}$$

Therefore, the family-wise type I error for this trial is simply the sum of all dose-wise type I errors as in

$$\psi = \sum_{i=1}^{J} \psi_i$$

and we are interested in preserving it under target α level. The corresponding boundary crossing probabilities ξ_j can be given by

$$\xi_j = \begin{cases} \Phi(-\omega_j) & \text{if } j = 1 \\ \left[\prod_{i=1}^{j-1} \Phi(\omega_i)\right] \Phi(-\omega_j). & \text{if } j > 1 \end{cases}$$

where

$$\omega_j = b_j - \frac{f(d_j) - \mu_0}{\sqrt{\frac{R+1}{c}}}.$$

Under this setting, the expected stage for stopping early for efficacy is

$$\begin{aligned} E(\mathcal{K}) &= \xi_1 + 2\xi_2 + ... + (J-1)\xi_{J-1} + J\left(1 - \xi_1 - ... - \xi_{J-1}\right) \\ &= J - \left[\sum_{j=1}^{J-1} (J-j)\xi_j\right] \end{aligned}$$

and it can be noted that $E(\mathcal{K}) < J$. The expected trial sample size is thus equal to $c\left(1/R + 1\right) E(\mathcal{K})$.

Given an alpha spending plan, we can calculate the stopping boundary values b_j's under this variant design. Our objective is to compare the cohort sizes, expected stages, and

expected sample sizes between the designs where $M = 1$ and $M = 2$, under different dose response models, dose orderings, and error spending plans such as Pocock, O'Brien-Fleming, Rho with $\rho = 0.3, 1, 3$ for the power of 90%. In addition, we also want to characterize the relationship between statistical power and expected sample size for these two designs. Like before, for these two designs, we assume $J = 4$, $R = 2$, $D = 1$, and $K = J = 4$ except for M.

Table 4.3 gives the stopping boundary values. As before, we find that the Pocock-type and Rho with $\rho = 0.3$ spending plans provide an opportunity to stop the trial at earlier stages if an informative dose ordering is applied, while the O'Brien-Fleming-type and Rho with $\rho = 3.0$ favor later stages and doses. Rho with $\rho = 1.0$ offers almost equal error spending across the staggered doses with similar stopping boundary values. Table 4.4 displays the results of the simulation. Figures 4.5, 4.6, 4.7, and 4.8 show the plots of statistical power of the designs across expected trial sample size.

We have presented another variant design that sets $M = 1$ and allows one stage for each dose, but doses are still entering the trial sequentially if the previous ones fail to show efficacy. This design however performs better than the original design with $M = 2$ only when (1) informative ordering is used, or (2) escalation ordering with $\rho \leqslant 1.0$ is adopted. A practical advantage of this variant design is that it reduces the number of interim analyses performed and hence the administration cost. If the dose ordering is favoring early efficacious doses, and the cost of performing monitoring analyses and meetings is high, this design is preferred to the design with $M = 2$.

Table 4.3: Five alpha spending schemes and their corresponding stage-wise efficacy stopping boundaries (b_j) and errors (α_j) for $J = 4, D = 1, M = 1, R = 2$, and $K = J = 4$. Family-wise type I error is controlled at one-sided $\alpha = 0.05$. No futility is adopted, $a_k = -\infty$ for all k.

Error Spending Scheme	k	1	2	3	4
1. Pocock-type boundary	b_j	2.10	2.22	2.30	2.37
	α_j	0.0179	0.0131	0.010	0.0086
2. O'Brien & Fleming-type boundary	b_j	3.75	2.54	2.09	1.93
	α_j	0.0001	0.0055	0.0180	0.0264
3. Rho error spending $\rho = 0.3$	b_j	1.84	2.41	2.54	2.62
	α_j	0.0330	0.0076	0.0052	0.0041
4. Rho error spending $\rho = 1.0$	b_j	2.24	2.24	2.23	2.22
	α_j	0.0125	0.0125	0.0125	0.0125
5. Rho error spending $\rho = 3.0$	b_j	3.16	2.54	2.17	1.89
	α_j	0.0008	0.0055	0.0148	0.0289

Figure 4.5: Comparison of statistical power under flat dose response model for variant design $M = 1$.

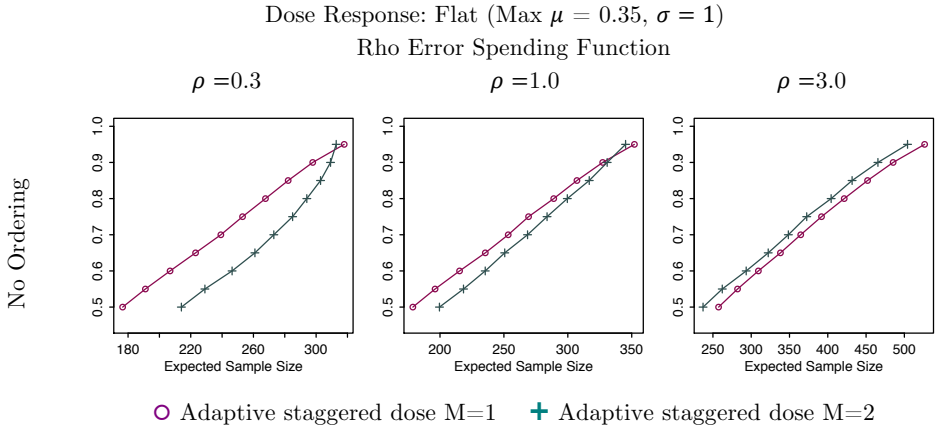


Table 4.4: Cohort size per stage (c), expected global stage to stop for efficacy ($E(\mathcal{K})$), expected trial sample size $\left(\frac{c(1+R)}{R}E(\mathcal{K})\right)$, and probabilities of dose selection for attaining statistical power of 90% for variant design with $M = 1$.

Dose		Pocock-type boundary			O'Brien & Fleming boundary			—		
Response		Escalation	Informative	Uninformative	Escalation	Informative	Uninformative	-	-	-
1. Flat	cohort size	105	105	105	129	129	129	-	-	-
	expected stage	1.96	1.97	1.96	2.71	2.71	2.71	-	-	-
	expected sample size	310	310	310	526	526	526	-	-	-
	dose selection prob	(0.542, 0.251, 0.130, 0.077)	(0.542, 0.251, 0.130, 0.076)	(0.250, 0.250, 0.250, 0.250)	(0.081, 0.413, 0.356, 0.150)	(0.081, 0.413, 0.356, 0.150)	(0.250, 0.250, 0.250, 0.250)	-	-	-
2. Linear	cohort size	284	184	225	268	311	303	-	-	-
	expected stage	2.76	1.53	2.06	3.09	2.00	2.52	-	-	-
	expected sample size	1178	421	696	1243	928	1144	-	-	-
	dose selection prob	(0.118, 0.302, 0.415, 0.166)	(0.024, 0.029, 0.126, 0.821)	(0.056, 0.145, 0.317, 0.482)	(0.002, 0.207, 0.586, 0.205)	(0.067, 0.108, 0.351, 0.473)	(0.050, 0.116, 0.312, 0.522)	-	-	-
3. Emax	cohort size	149	127	142	157	169	177	-	-	-
	expected stage	2.31	1.81	1.98	2.90	2.55	2.68	-	-	-
	expected sample size	516	345	421	683	646	710	-	-	-
	dose selection prob	(0.320, 0.345, 0.225, 0.111)	(0.057, 0.087, 0.223, 0.633)	(0.137, 0.236, 0.295, 0.331)	(0.017, 0.364, 0.446, 0.174)	(0.127, 0.276, 0.451, 0.145)	(0.125, 0.230, 0.300, 0.345)	-	-	-
4. Umbrella	cohort size	136	124	145	139	160	183	-	-	-
	expected stage	2.42	1.86	1.99	2.91	2.65	2.67	-	-	-
	expected sample size	494	346	432	607	637	729	-	-	-
	dose selection prob	(0.238, 0.394, 0.249, 0.119)	(0.067, 0.203, 0.621, 0.109)	(0.115, 0.273, 0.338, 0.273)	(0.008, 0.369, 0.449, 0.173)	(0.146, 0.385, 0.129, 0.340)	(0.103, 0.272, 0.353, 0.272)	-	-	-

Dose		Rho error spending ρ=0.3			Rho error spending ρ=1.0			Rho error spending ρ=3.0		
Response		Escalation	Informative	Uninformative	Escalation	Informative	Uninformative	Escalation	Informative	Uninformative
1. Flat	cohort size	113	113	113	106	106	106	129	129	129
	expected stage	1.76	1.76	1.76	2.06	2.06	2.06	2.51	2.51	2.51
	expected sample size	298	298	298	328	328	328	486	486	486
	dose selection prob	(0.689, 0.166, 0.088, 0.057)	(0.689, 0.166, 0.088, 0.057)	(0.250, 0.250, 0.250, 0.250)	(0.484, 0.274, 0.154, 0.087)	(0.484, 0.274, 0.155, 0.087)	(0.250, 0.250, 0.250, 0.250)	(0.214, 0.359, 0.294, 0.133)	(0.213, 0.359, 0.294, 0.133)	(0.250, 0.250, 0.250, 0.250)
2. Linear	cohort size	333	178	238	275	194	229	282	279	288
	expected stage	2.65	1.44	1.93	2.83	1.56	2.12	3.06	1.77	2.39
	expected sample size	1324	384	686	1168	455	728	1297	741	1031
	dose selection prob	(0.200, 0.259, 0.383, 0.158)	(0.015, 0.016, 0.076, 0.894)	(0.065, 0.163, 0.316, 0.456)	(0.089, 0.293, 0.445, 0.173)	(0.028, 0.035, 0.141, 0.795)	(0.052, 0.138, 0.317, 0.493)	(0.011, 0.218, 0.569, 0.202)	(0.050, 0.073, 0.228, 0.648)	(0.045, 0.110, 0.312, 0.533)
3. Emax	cohort size	168	132	151	147	129	143	163	165	175
	expected stage	2.11	1.64	1.79	2.41	1.89	2.07	2.80	2.30	2.48
	expected sample size	530	325	407	532	367	444	685	570	651
	dose selection prob	(0.466, 0.263, 0.177, 0.095)	(0.041, 0.056, 0.142, 0.761)	(0.152, 0.241, 0.289, 0.318)	(0.265, 0.357, 0.256, 0.122)	(0.067, 0.106, 0.248, 0.579)	(0.036, 0.126, 0.372, 0.466)	(0.067, 0.362, 0.407, 0.164)	(0.108, 0.214, 0.361, 0.317)	(0.116, 0.225, 0.304, 0.354)
4. Umbrella	cohort size	156	130	155	133	126	147	145	157	180
	expected stage	2.24	1.67	1.81	2.51	1.94	2.07	2.84	2.40	2.47
	expected sample size	525	326	418	502	368	456	618	567	667
	dose selection prob	(0.367, 0.320, 0.207, 0.105)	(0.049, 0.127, 0.753, 0.071)	(0.130, 0.272, 0.325, 0.272)	(0.191, 0.400, 0.280, 0.129)	(0.077, 0.224, 0.566, 0.132)	(0.109, 0.273, 0.344, 0.273)	(0.039, 0.376, 0.418, 0.166)	(0.125, 0.314, 0.293, 0.267)	(0.094, 0.271, 0.362, 0.272)

Figure 4.6: Comparison of statistical power under linear dose response model for variant design $M = 1$.

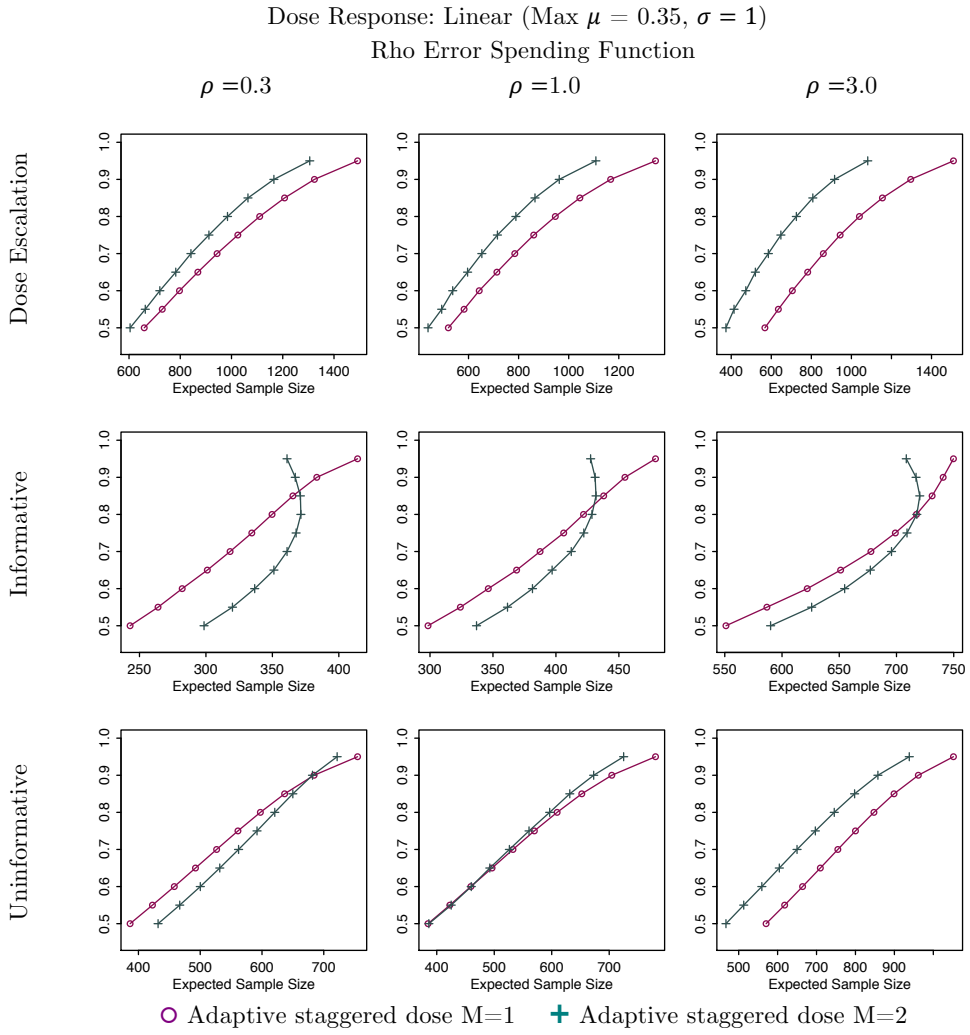


Figure 4.7: Comparison of statistical power under emax dose response model for variant design $M = 1$.

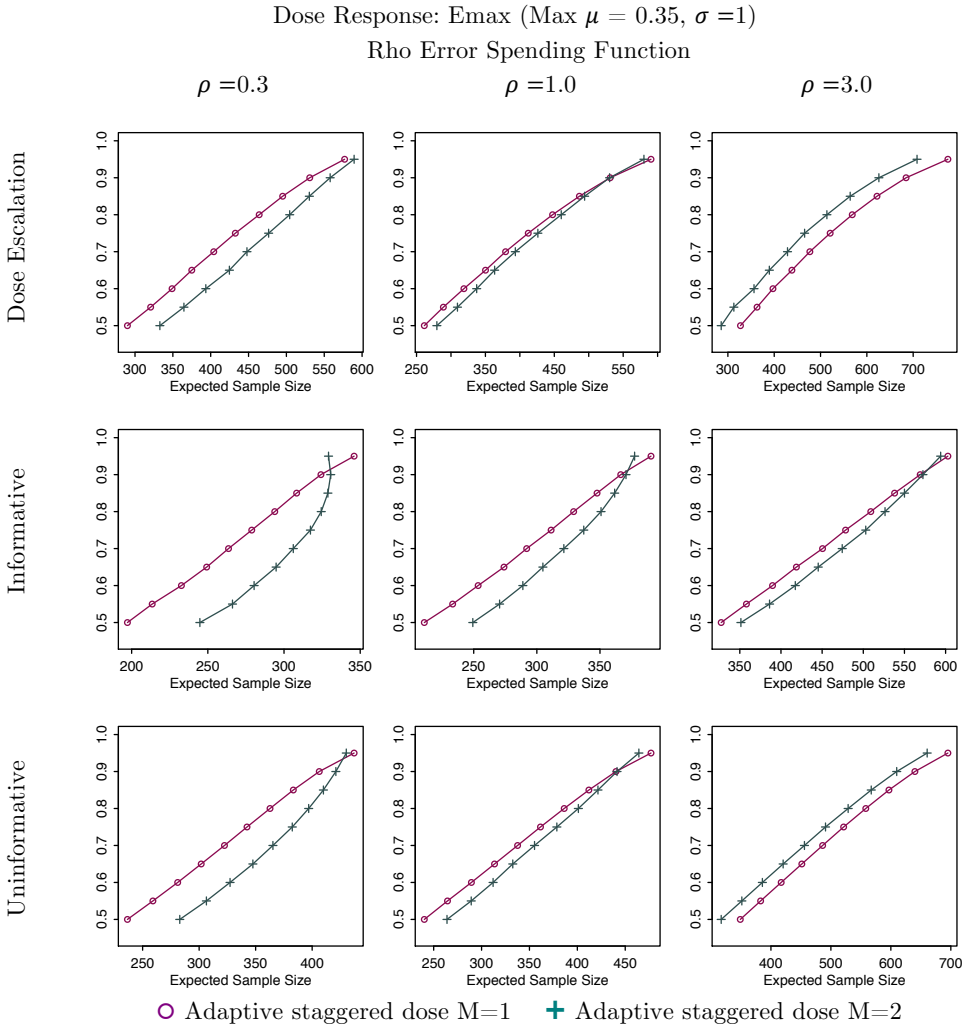
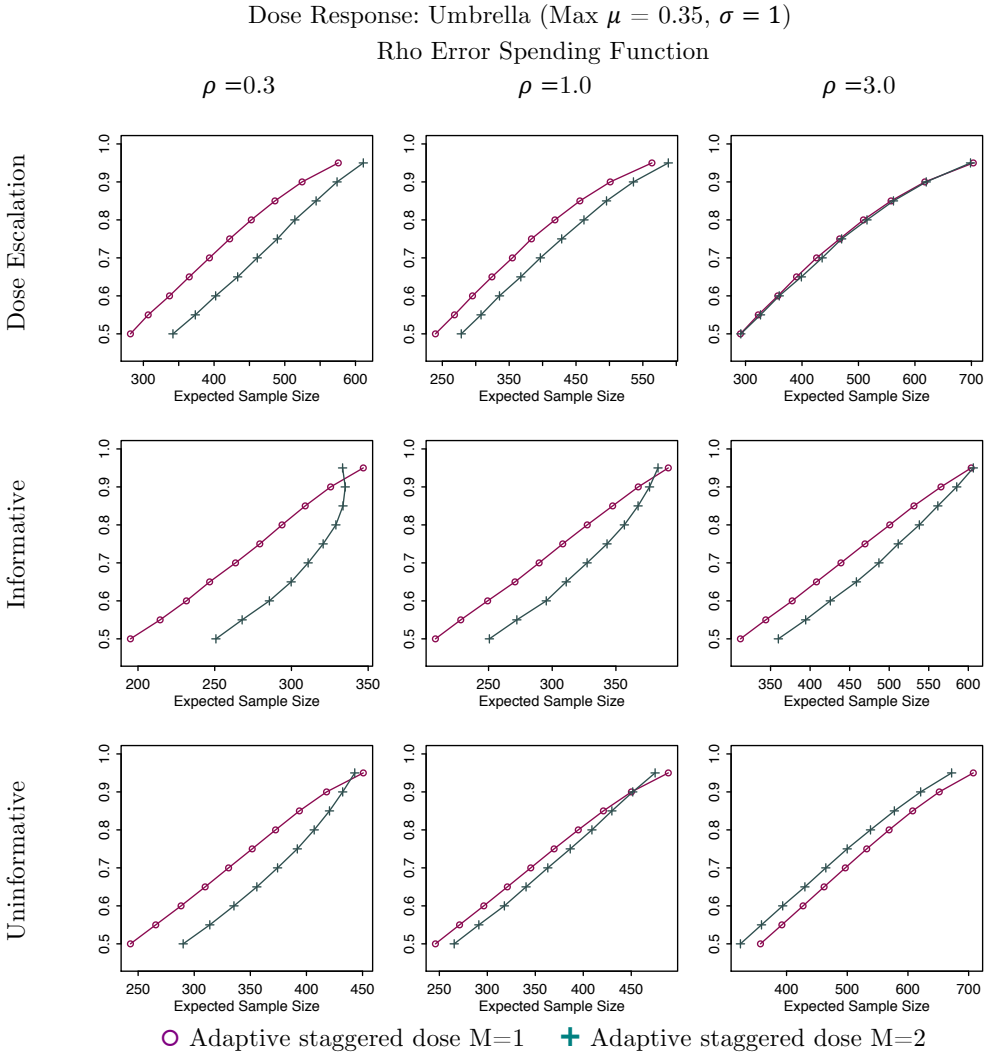


Figure 4.8: Comparison of statistical power under umbrella dose response model for variant design $M = 1$.



4.4 Use of Marginal Alpha Spending Functions

In Chapter 3, we have demonstrated the use of an alpha spending function to globally monitor the type I error of a trial across the stages and doses. However, this adaptive staggered dose can allow for additional flexibility in specifying marginal alpha spending functions to the doses. Therefore, each dose or set of doses can have its or their own alpha

spending functions, depending on how optimistic or conservative the investigators think regarding the dose ordering. We will illustrate this concept using two different variant designs that employ marginal alpha spending plans. These two variant designs have the same design setting as in the original design in Chapter 3, that is, $D = 1$, $M = 2$, $R = 2$, and $K = 2J$.

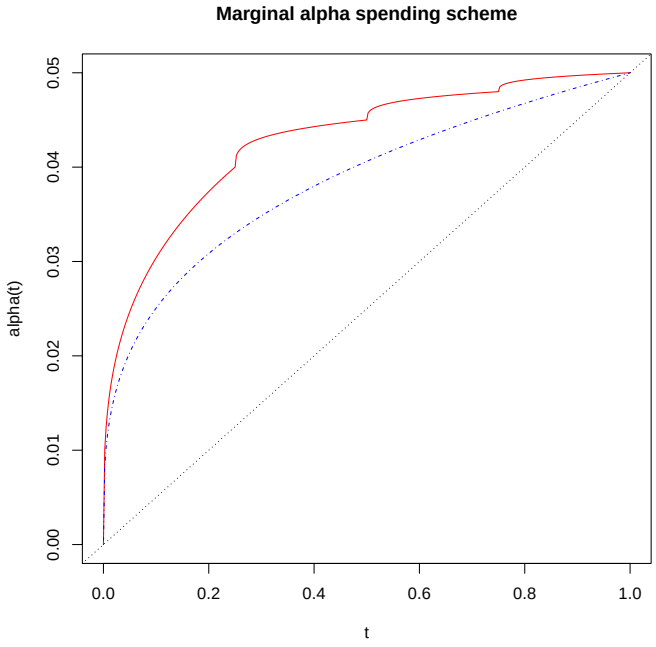


Figure 4.9: Alpha spending plan when each dose has its own $\alpha_{j,\rho_j}(t)$ for $J = 4$ and $\rho_j = 0.3$, represented by red solid line. Global $\alpha(t)$ with $\rho = 0.3$ by blue dashed line.

In the first variant design, we specify for each dose its own marginal alpha spending function. We can represent the function for dose d_j $(j = 1, 2, ..., J)$ by $\alpha_{j,\rho_j}(t) = \alpha_j t^{\rho_j}$ where ρ_j refers to the dose-specific parameter for Rho spending function. We also assign α_j as the nominal alpha level for dose d_j such that $\sum_{j=1}^{J} \alpha_j = \alpha$. We only use the Rho spending function for illustration here because of its flexibility, but the spending function is not restricted to Rho function. Under the scheme that $M = 2$ and if $\rho_j = 0.3$ for all j, then the sequence of

stage-wise alphas α'_k's across the stages will be

$$\begin{aligned}
\alpha'_1 &= \alpha_{1,\rho_1=0.3}(1/2) \\
\alpha'_2 &= \alpha_{1,\rho_1=0.3}(1) - \alpha_{1,\rho_1=0.3}(1/2) \\
\alpha'_3 &= \alpha_{2,\rho_2=0.3}(1/2) \\
\vdots &= \vdots \\
\alpha'_{K-1} &= \alpha_{J,\rho_J=0.3}(1/2) \\
\alpha'_K &= \alpha_{J,\rho_J=0.3}(1) - \alpha_{J,\rho_J=0.3}(1/2)
\end{aligned}$$

since the cohort size c is constant across stage. Based on this sequence of α'_k's where $\sum_{k=1}^{K} \alpha'_k = \sum_{j=1}^{J} \alpha_j = \alpha$, and given the null distribution of the test statistics as in (3.3.2), we can evaluate the stopping boundary values b_k's using numerical technique. Figure 4.9 shows the an example of marginal alpha spending plan for $J = 4$.

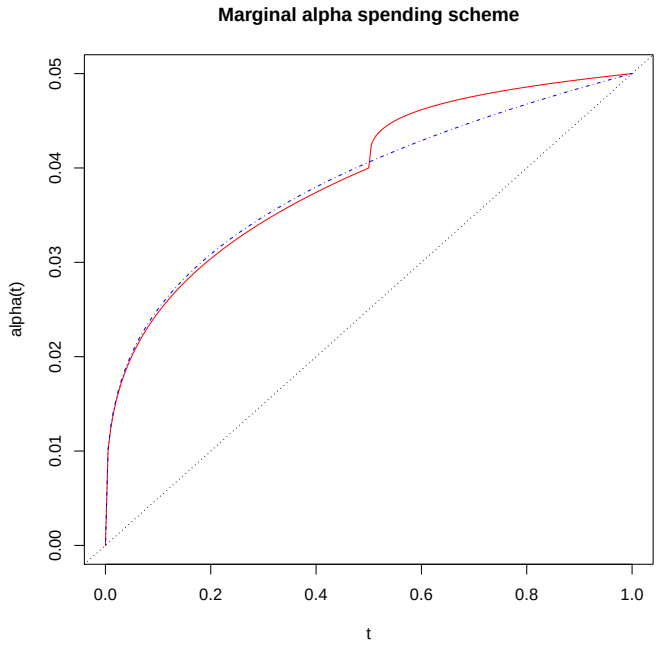


Figure 4.10: Alpha spending plan when each two adjacent doses have their own $\alpha_{(j_1,j_2),\rho_{(j_1,j_2)}}(t)$ for $J = 4$ and $\rho_{(j_1,j_2)} = 0.3$, represented by red solid line. Global $\alpha(t)$ with $\rho = 0.3$ by blue dashed line.

The second variant design will have one alpha spending function specified for each set of two adjacent doses. We can represent the function for doses d_{j_1} and d_{j_2} by $\alpha_{(j_1,j_2),\rho_{(j_1,j_2)}}(t) = \alpha_{(j_1,j_2)}t^{\rho_{(j_1,j_2)}}$ where $(j_1, j_2) = \{(1,2),(3,4),...\}$. For example, if $J = 4$, then there will be two alpha spending functions: $\alpha_{(1,2),\rho_{(1,2)}}(t) = \alpha_{(1,2)}t^{\rho_{(1,2)}}$ and $\alpha_{(3,4),\rho_{(3,4)}}(t) = \alpha_{(3,4)}t^{\rho_{(3,4)}}$. Under the scheme that $M = 2$, and $\rho_{(j_1,j_2)} = 0.3$, the sequence of α'_k's spent across the stages will be

$$\begin{aligned}
\alpha'_1 &= \alpha_{(1,2),\rho_{(1,2)}=0.3}(1/4) \\
\alpha'_2 &= \alpha_{(1,2),\rho_{(1,2)}=0.3}(2/4) - \alpha_{(1,2),\rho_{(1,2)}=0.3}(1/4) \\
\vdots &= \vdots \\
\alpha'_7 &= \alpha_{(3,4),\rho_{(3,4)}=0.3}(3/4) - \alpha_{(3,4),\rho_{(3,4)}=0.3}(2/4) \\
\alpha'_8 &= \alpha_{(3,4),\rho_{(3,4)}=0.3}(4/4) - \alpha_{(3,4),\rho_{(3,4)}=0.3}(3/4).
\end{aligned}$$

Again, the stopping boundary values can be computed as usual. Figure 4.10 shows an example of this marginal alpha spending plan for $J = 4$. Our objective is to compare the cohort sizes, expected stages, and expected sample sizes between these designs under the specification of $\rho = 0.3$.

In this section, we will explore the use of marginal alpha spending plan and specifically, the two examples in Figures 4.9 and 4.10. For the one with one function for each dose, we have $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.040, 0.005, 0.003, 0.002)$ and for the one with one function for each two adjacent doses, we have $(\alpha_{(1,2)}, \alpha_{(3,4)}) = (0.04, 0.01)$. We also want to compare these two spending plans to the original one we used in Chapter 3, which is Rho with $\rho = 0.3$. Table 4.5 tabulates the stopping boundary values under the three alpha spending plans.

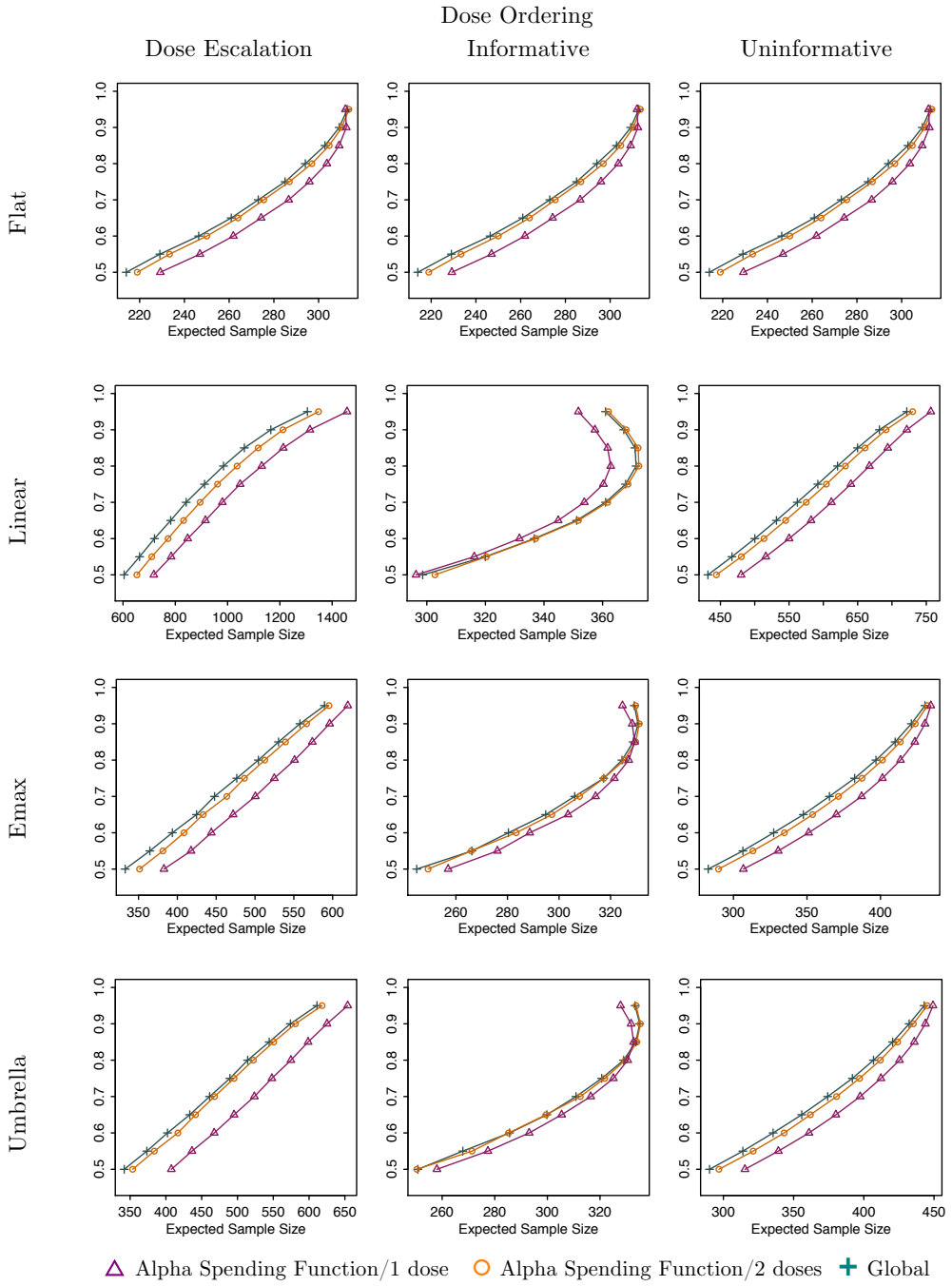
Figure 4.11 displays the plots of statistical power against the expected sample size for the three alpha spending plans. As seen from the plots, the alpha spending plan using one function for two adjacent doses appears to have slightly better or similar performance as the one using a global alpha spending function in most of the scenarios. However, the plan

using one function for each dose shows the worse performance, except under linear dose response model and using informative dose ordering where it offers a small advantage over the global alpha spending function.

Table 4.5: Three alpha spending schemes and their corresponding stage-wise efficacy stopping boundaries (b_k) and errors (α_k) for $J = 4, D = 1, M = 2, R = 2$, and $K = 2J = 8$. Family-wise type I error is controlled at one-sided $\alpha = 0.05$. No futility is adopted, $a_k = -\infty$ for all k.

Error Spending Scheme	k	1	2	3	4	5	6	7	8
1. Global $\alpha(t)$, $\rho = 0.3$	b_k	1.930	2.277	2.619	2.613	2.755	2.728	2.837	2.802
	α_k	0.0268	0.0062	0.0043	0.0034	0.0028	0.0024	0.0022	0.0020
2. Marginal $\alpha_{(1,2)} = 0.04, \alpha_{(3,4)} = 0.01$	b_k	1.937	2.284	2.624	2.619	2.464	2.811	3.000	3.000
$\rho_{(1,2)} = \rho_{(3,4)} = 0.3$	α_k	0.0264	0.0061	0.0042	0.0033	0.0066	0.0015	0.0013	0.0010
3. Marginal $\alpha_1 = 0.04, \alpha_2 = 0.005$	b_k	1.845	2.191	2.633	2.978	2.800	3.000	2.928	3.000
$\alpha_3 = 0.003, \alpha_4 = 0.002, \rho_j = 0.3$	α_k	0.0325	0.0075	0.0041	0.0009	0.0024	0.0009	0.0016	0.0010

Figure 4.11: Comparison of statistical power under different dose response models for marginal alpha spending functions.



4.5 Randomization Ratio $1 : R$

In this section, we want to assess the impact of changing the randomization ratio. Specifically, we want to investigate the impact of increasing R from 2 to 3 and randomizing more subjects to the experimental doses on the statistical power given the same expected sample size.

In this variant design, we keep $J = 4$, $D = 1$, $M = 2$, $K = 8$, but vary $R = 2, 3$ under the usual four dose response models, three dose orderings, and alpha spending plans of Pocock, O'Brien-Fleming, and Rho with $\rho = 0.3, 1, 3$ for power of 90%. Since the null distributions in (3.3.1) do not depend on R, the same set of stopping boundaries as in Table 3.3 can be used. When R increases, we can see that $\omega_{j,k}$ will also increase, and therefore, $\xi_{j,k}$ is expected to decrease, and hence the overall statistical power. In other words, if we want to keep the same power when increasing R, then the cohort size needs to increase to offset the loss of power and therefore, the expected sample size will increase as well. This clear impact can be verified in Table 4.6 and Figures 4.12, 4.13, 4.14, and 4.15.

Figure 4.12: Comparison of statistical power under flat dose response model for variant design $R = 3$.

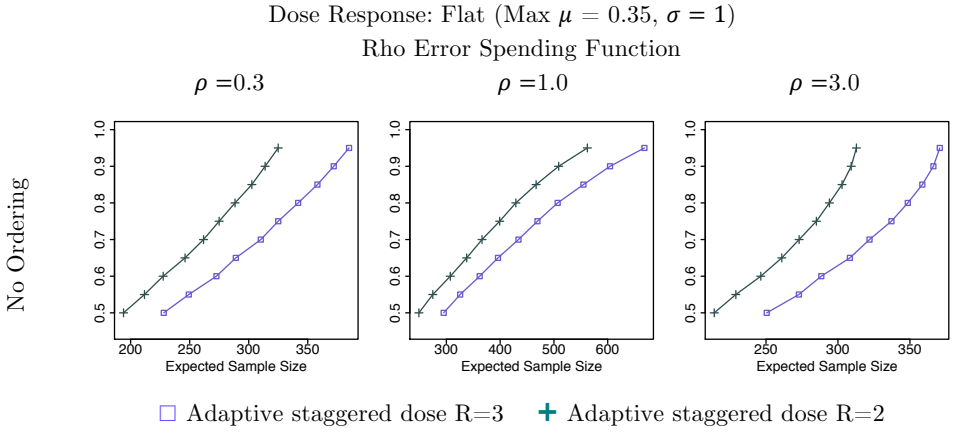


Table 4.6: Cohort size per stage (c), expected global stage to stop for efficacy ($E(\mathcal{K})$), expected trial sample size $\left(\frac{c(1+R)}{R}E(\mathcal{K})\right)$, and probabilities of dose selection for attaining statistical power of 90% for variant design with $R = 3$.

Dose		Pocock-type boundary			O'Brien & Fleming boundary			—		
Response		Escalation	Informative	Uninformative	Escalation	Informative	Uninformative	-	-	-
1. Flat	cohort size	80	80	80	87	87	87	-	-	-
	expected stage	3.5	3.5	3.5	5.2	5.2	5.2	-	-	-
	expected sample size	372	372	372	604	604	604	-	-	-
	dose selection prob	(0.551, 0.252, 0.128, 0.070)	(0.551, 0.252, 0.128, 0.070)	(0.250, 0.250, 0.250, 0.250)	(0.083, 0.403, 0.344, 0.170)	(0.083, 0.402, 0.344, 0.170)	(0.250, 0.250, 0.250, 0.250)	-	-	-
2. Linear	cohort size	160	143	152	129	235	170	-	-	-
	expected stage	5.6	2.5	3.8	6.3	3.5	5.1	-	-	-
	expected sample size	1197	471	773	1086	1086	1154	-	-	-
	dose selection prob	(0.093, 0.231, 0.371, 0.305)	(0.005, 0.025, 0.118, 0.851)	(0.036, 0.123, 0.297, 0.543)	(0.001, 0.135, 0.463, 0.400)	(0.019, 0.083, 0.323, 0.574)	(0.024, 0.093, 0.259, 0.623)	-	-	-
3. Emax	cohort size	108	102	105	100	131	115	-	-	-
	expected stage	4.3	3.0	3.6	5.7	4.5	5.2	-	-	-
	expected sample size	623	410	500	754	791	797	-	-	-
	dose selection prob	(0.306, 0.342, 0.228, 0.123)	(0.025, 0.080, 0.219, 0.675)	(0.124, 0.232, 0.300, 0.344)	(0.015, 0.331, 0.426, 0.227)	(0.074, 0.246, 0.481, 0.199)	(0.107, 0.221, 0.305, 0.366)	-	-	-
4. Umbrella	cohort size	109	104	107	102	134	118	-	-	-
	expected stage	4.3	3.0	3.6	5.4	4.5	5.2	-	-	-
	expected sample size	629	417	510	739	810	813	-	-	-
	dose selection prob	(0.248, 0.419, 0.250, 0.082)	(0.019, 0.197, 0.685, 0.098)	(0.100, 0.273, 0.352, 0.274)	(0.010, 0.396, 0.441, 0.152)	(0.056, 0.437, 0.208, 0.298)	(0.083, 0.270, 0.377, 0.270)	-	-	-

Dose		Rho error spending ρ=0.3			Rho error spending ρ=1.0			Rho error spending ρ=3.0		
Response		Escalation	Informative	Uninformative	Escalation	Informative	Uninformative	Escalation	Informative	Uninformative
1. Flat	cohort size	89	89	89	78	78	78	84	84	84
	expected stage	3.1	3.1	3.2	3.8	3.8	3.8	4.9	4.9	4.9
	expected sample size	366	366	366	392	392	392	552	552	552
	dose selection prob	(0.656, 0.189, 0.098, 0.057)	(0.656, 0.189, 0.098, 0.057)	(0.250, 0.250, 0.250, 0.250)	(0.484, 0.274, 0.155, 0.087)	(0.484, 0.274, 0.155, 0.087)	(0.250, 0.250, 0.250, 0.250)	(0.200, 0.334, 0.292, 0.174)	(0.200, 0.334, 0.292, 0.174)	(0.250, 0.250, 0.250, 0.250)
2. Linear	cohort size	189	148	169	148	149	149	128	206	162
	expected stage	5.5	2.2	3.5	5.8	2.6	4.0	6.4	3.1	4.9
	expected sample size	1380	436	791	1141	512	798	1086	850	1046
	dose selection prob	(0.145, 0.197, 0.347, 0.311)	(0.003, 0.017, 0.082, 0.898)	(0.044, 0.133, 0.298, 0.524)	(0.068, 0.215, 0.388, 0.328)	(0.007, 0.031, 0.133, 0.828)	(0.032, 0.114, 0.292, 0.561)	(0.007, 0.128, 0.432, 0.432)	(0.019, 0.061, 0.206, 0.713)	(0.022, 0.087, 0.262, 0.629)
3. Emax	cohort size	125	112	118	103	103	103	99	123	111
	expected stage	4.0	2.6	3.2	4.6	3.2	3.8	5.6	4.1	4.9
	expected sample size	661	392	497	630	439	524	738	679	726
	dose selection prob	(0.410, 0.283, 0.194, 0.112)	(0.017, 0.058, 0.158, 0.766)	(0.134, 0.234, 0.260, 0.335)	(0.248, 0.347, 0.260, 0.144)	(0.033, 0.097, 0.244, 0.626)	(0.118, 0.229, 0.302, 0.351)	(0.056, 0.306, 0.393, 0.245)	(0.075, 0.197, 0.361, 0.367)	(0.101, 0.219, 0.307, 0.373)
4. Umbrella	cohort size	128	113	120	105	105	106	102	126	113
	expected stage	4.0	2.6	3.2	4.5	3.2	3.8	5.4	4.1	4.9
	expected sample size	681	398	510	635	446	535	732	693	740
	dose selection prob	(0.345, 0.363, 0.220, 0.072)	(0.013, 0.141, 0.773, 0.073)	(0.111, 0.272, 0.343, 0.273)	(0.200, 0.423, 0.281, 0.100)	(0.025, 0.221, 0.636, 0.118)	(0.094, 0.273, 0.359, 0.273)	(0.041, 0.375, 0.416, 0.167)	(0.057, 0.324, 0.380, 0.238)	(0.077, 0.270, 0.383, 0.270)

Figure 4.13: Comparison of statistical power under linear dose response model for variant design $R = 3$.

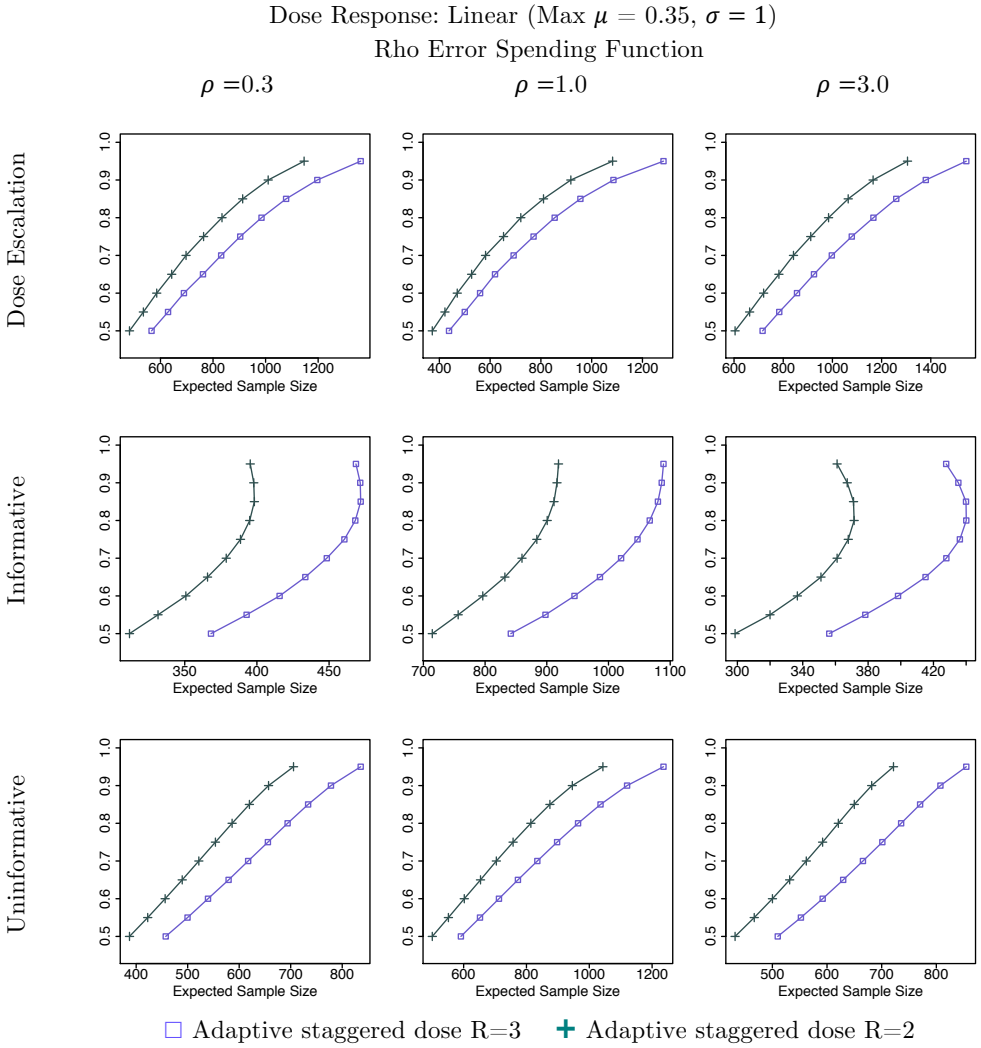


Figure 4.14: Comparison of statistical power under emax dose response model for variant design $R = 3$.

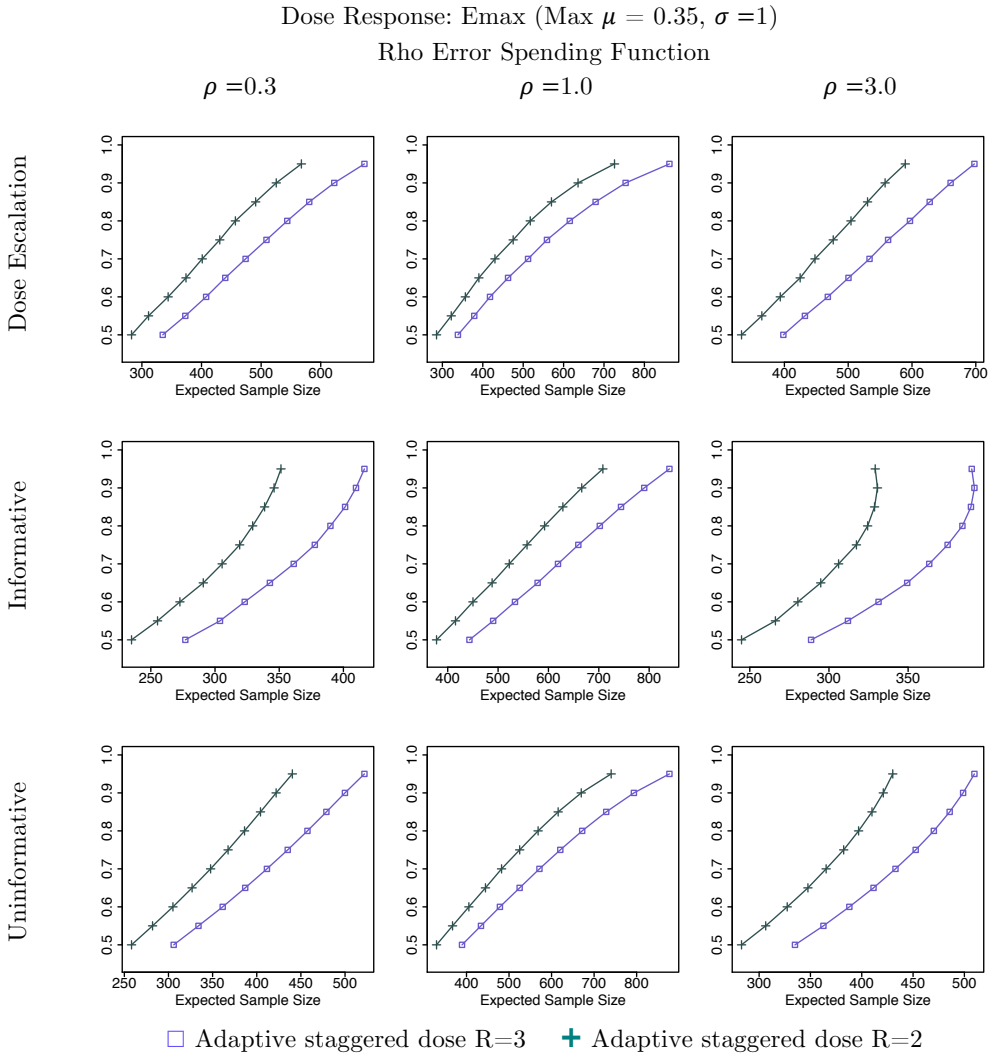
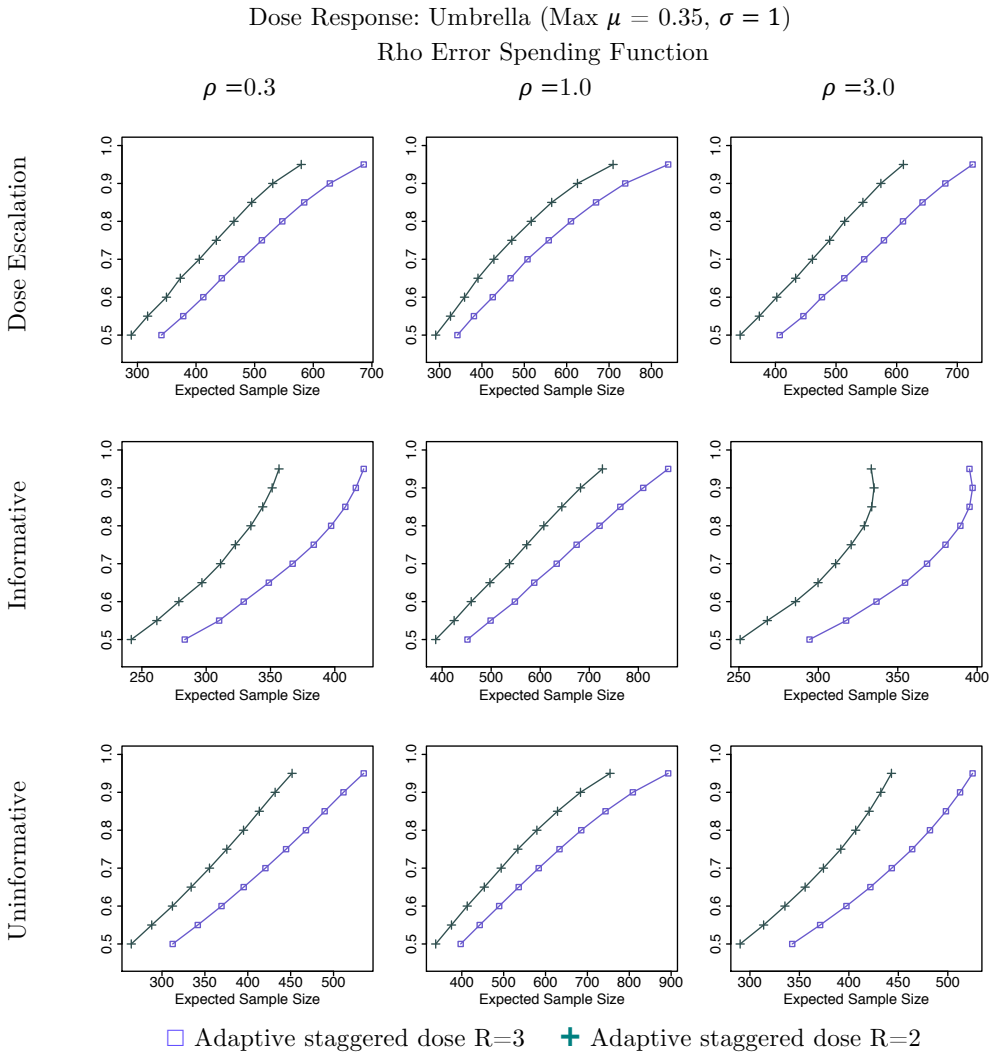


Figure 4.15: Comparison of statistical power under umbrella dose response model for variant design $R = 3$.



4.6 Summary and Discussion

We have presented four different variations of the staggered dose design and described their operating characteristics comparing to the original design as the reference design in Chapter 3. In summary, the design that simultaneously explores two concurrent doses will shorten the duration of the trial and likely select more than one dose. However, if the objective is

to select one dose only, then uncertainty still remains at the end of the trial as to which one of the two selected doses is better. The use of one stage per dose $M = 1$ is only desirable if the dose ordering is based on informative dose response in order to obtain additional gain in statistical power and reduction in expected sample size. However, marginal alpha spending function will only reap small improvement if informative dose ordering is plausible. Lastly, increasing R for the randomization ratio $1 : R$ is not recommended as it reduces overall statistical power. Therefore, when investigators think that this staggered dose design is applicable to their clinical development, simulation and numerical techniques should be used to evaluate and compare the performance of the design under different variations.

4.7 Binary and Time-to-Event Endpoints

4.7.1 Binary Endpoint

In some clinical trials, the proof of efficacy or safety is based on a binary outcome. The proposed staggered dose design can also be applied with a simple adjustment. An example of a binary endpoint is whether a subject responds to a given treatment or if this subject experiences a specific toxic side effect. This binary outcome is usually coded as $Y_{ji} = 0$ if a clinical response fails to take place, but $Y_{ji} = 1$ if it is observed, where j represents the jth treatment for the ith subject $(i = 1, 2, ...)$ receiving this treatment. Therefore, the model is

$$Y_{ji} \sim Bernoulli(\pi_j) \tag{4.7.1}$$

where π_j is the probability of a clinical response. If there are J different experimental treatments and a control treatment in the trial, then we can have the following set of one-sided hypotheses:

$$H_{j0} : \pi_j \leqslant \pi_0, \;\; H_{ja} : \pi_j > \pi_0 \tag{4.7.2}$$

where π_0 is the probability of response in the control arm. Using the same approach as in (3.2.3), we define the sample means as $\overline{Y}_{jm} = \sum_{i=1}^{cm} Y_{ji}/(cm)$ and $\overline{Y}_{0m} = \sum_{i=1}^{(cm)/R} Y_{0i}/(\frac{cm}{R})$ for $m = 1, 2$ like before. Under the null hypothesis, H_{j0}, we assume a common probability as $\pi_j = \pi_0 = \tilde{\pi}_j$ and asymptotic normality under large sample condition,

$$\overline{Y}_{jm} - \overline{Y}_{0m} \xrightarrow{d} N\left(0, \frac{\tilde{\pi}_j(1-\tilde{\pi}_j)}{\frac{cm}{R+1}}\right) \tag{4.7.3}$$

and therefore, we define the test statistics as

$$Z_{jm} = \frac{\overline{Y}_{jm} - \overline{Y}_{0m}}{\sqrt{\frac{\tilde{\pi}_j(1-\tilde{\pi}_j)(R+1)}{cm}}} \tag{4.7.4}$$

and $Z_{jm} \xrightarrow{d} N(0,1)$. As a result, under the null hypotheses, the null distributions of the above test statistics Z_{j1} and $(Z_{j1}, Z_{j2})'$ also converge to the distributions in (3.3.1). However, since we do not know the true value of $\tilde{\pi}_j$, we can estimate it using the combined data $\hat{\tilde{\pi}}_j = (\sum_{i=1}^{cm} Y_{ji} + \sum_{i=1}^{(cm)/R} Y_{0i})/(cm + cm/R)$. In this case, the stopping boundary values b_k's can be computed using the usual method.

Under the alternative hypotheses, we can assume a dose response model such that $\pi_j = g(d_j)$ where $g(d)$ is the dose response function and $\pi_0 = g(d_0)$ and $0 \leqslant g(d) \leqslant 1$. Commonly, the function $g(d)$ is assumed to be monotonic increasing, but non-monotonicity is also observed such as a downturn model discussed earlier. We can denote the probability difference as $\theta_j = \pi_j - \pi_0 = g(d_j) - g(d_0)$. Using this notation, we can see that

$$\begin{aligned} \overline{Y}_{jm} - \overline{Y}_{0m} &\xrightarrow{d} N\left(\theta_j, \frac{(\pi_0+\theta_j)(1-\pi_0-\theta_j)}{cm} + \frac{R\pi_0(1-\pi_0)}{cm}\right) \\ &= N\left(g(d_j) - g(d_0), \frac{g(d_j)(1-g(d_j))}{cm} + \frac{Rg(d_0)(1-g(d_0))}{cm}\right) \end{aligned} \tag{4.7.5}$$

and thus

$$\begin{aligned} Z_{jm} &= \frac{\overline{Y}_{jm} - \overline{Y}_{0m}}{\sqrt{\frac{g(d_j)(1-g(d_j))}{cm} + \frac{Rg(d_0)(1-g(d_0))}{cm}}} \\ &\xrightarrow{d} N\left(\frac{g(d_j) - g(d_0)}{\sqrt{\frac{g(d_j)(1-g(d_j))}{cm} + \frac{Rg(d_0)(1-g(d_0))}{cm}}}, 1\right) \end{aligned} \tag{4.7.6}$$

and $cov(Z_{j1}, Z_{j2})$ can be shown to be $1/\sqrt{2}$ like before. Therefore, the cohort size c used to attain a specified statistical power for a given dose response model can be evaluated using the same form in Section 3.6.2, but this time,

$$\omega_{j,k} = b_k - \frac{g(d_j) - g(d_0)}{\sqrt{\frac{g(d_j)(1-g(d_j))}{c} + \frac{Rg(d_0)(1-g(d_0))}{c}}}$$

when $k = 2j - 1$, and

$$\omega_{j,k} = b_k - \frac{g(d_j) - g(d_0)}{\sqrt{\frac{g(d_j)(1-g(d_j))}{2c} + \frac{Rg(d_0)(1-g(d_0))}{2c}}}$$

when $k = 2j$.

In some other instances, one can also model the comparison in term of odds ratio, OR_{j0} such that

$$OR_{j0} = \frac{\pi_j(1-\pi_0)}{\pi_0(1-\pi_j)}$$

and that we are interested in testing if the logarithm of OR_{j0}, is equal to 0 or greater than 0. In this case, the logarithm of the Mantel-Haenszel estimator can be used with the assumption of normality under large sample condition (Mantel and Haenszel, 1959). This can be left for future work.

4.7.2 Time-to-Event Endpoint

In some other clinical trials such as an oncology trial, the evidence of efficacy is based on a survival endpoint such as overall survival (OS) or progression-free survival (PFS). This type of outcome is also known as time-to-failure. The proposed staggered dose design can also be applied with some adjustment. We can denote the *time elapsed since the start of treatment until the observation of failure before or at the calendar time of analysis* (time-to-failure) as $T_{ji}\,(T_{ji} > 0)$ for the ith subject receiving the jth treatment where $i = 1, 2, ...$ and $j = 1, 2, ..., J$. If at the calendar time of analysis, the event of failure is already observed, then T_{ji} is considered uncensored with $s_{ji} = 0$, but right-censored with $s_{ji} = 1$; and if the event of interest is not observed, then T_{ji} is in fact the time elapsed since the start of treatment until calendar time of analysis. We can assume a model like

$$T_{ji} \sim f_j(T_{ji}) \tag{4.7.7}$$

where $f_j(T_{ji})$ is a probability density function with unknown parametric form. We can also represent the survival function as $S_j(T_{ji}) = \int_{T_{ji}}^{\infty} f_j(u)du$ and therefore, the hazard function as $h_j(T_{ji}) = f_j(T_{ji})/S_j(T_{ji})$. If we denote $\lambda_{j0} = h_j(T_{ji})/h_0(T_{0i})$ as the hazard ratio between the dose d_j and the control dose d_0, then we are interested in the following set of one-sided hypotheses:

$$H_{j0} : \lambda_{j0} \geqslant 1, \;\; H_{ja} : \lambda_{j0} < 1. \tag{4.7.8}$$

In this case, under the proportional hazard assumption, the test of choice is the optimal non-parametric log-rank test. First of all, we denote the calendar times of analysis as l_{jm} for d_j at mth per-dose interim ($m = 1, 2$). It is sometimes more practical to pre-specify these calendar times at the design of the trial. In the staggered dose design with $D = 1$, these calendar times are in this order: $l_{11}, l_{12}, l_{21}, ..., l_{J1}, l_{J2}$. We further let q_{jm} be the number of failures combined for both dose d_j and control dose observed at calendar time l_{jm} with corresponding times to failures as $\tau_1(l_{jm}) < \tau_2(l_{jm}) < ... < \tau_i(l_{jm}) < ... < \tau_{q_{jm}}(l_{jm})$ for

$m = 1, 2$. At each of these q_{jm} distinct times, we can count the number of failures and number of subjects at risk by treatment dose as below:

$$\begin{aligned}
e_j(\tau_i(l_{jm})) &= \text{no. of failures for } d_j \text{ at } \tau_i(l_{jm}) \text{ for calendar time } l_{jm} \\
r_j(\tau_i(l_{jm})) &= \text{no. of subjects at risk for failure for } d_j \text{ at } \tau_i(l_{jm}) \text{ for calendar time } l_{jm} \\
e_0(\tau_i(l_{jm})) &= \text{no. of failures for } d_0 \text{ at } \tau_i(l_{jm}) \text{ for calendar time } l_{jm} \\
r_0(\tau_i(l_{jm})) &= \text{no. of subjects at risk for failure for } d_0 \text{ at } \tau_i(l_{jm}) \text{ for calendar time } l_{jm}
\end{aligned}$$

for $i = 1, 2, ..., q_{jm}$. For example, if there are no ties at failure time $\tau_i(l_{jm})$, then if this failure happens for d_j, then $e_j(\tau_i(l_{jm})) = 1$ and correspondingly, $e_0(\tau_i(l_{jm})) = 0$; else if it happens for the control dose d_0, then the numbers are reversed. The log-rank score statistic for dose d_j at mth per dose stage is given by

$$S_{jm} = -\sum_{i=1}^{q_{jm}} \left\{ e_j(\tau_i(l_{jm})) - r_j(\tau_i(l_{jm})) \left(\frac{e_0(\tau_i(l_{jm})) + e_j(\tau_i(l_{jm}))}{r_0(\tau_i(l_{jm})) + r_j(\tau_i(l_{jm}))} \right) \right\}. \tag{4.7.9}$$

If the distributions are the same ($f_j = f_0$) or, in other words, under the null hypothesis of $\lambda_{j0} = 1$, then the following convergence in distribution holds

$$S_{jm} \xrightarrow{d} N(0, var(S_{jm})). \tag{4.7.10}$$

If we make a further assumption of proportional hazard, that is, the ratio λ_{j0} is constant over time. In this case,

$$S_{jm} \xrightarrow{d} N\left(-\frac{ln(\lambda_{j0}) D(l_{jm}) R}{(R+1)^2}, \frac{D(l_{jm}) R}{(R+1)^2} \right) \tag{4.7.11}$$

where $D(l_{jm}) = \sum_{i=1}^{q_{jm}} (e_0(\tau_i(l_{jm})) + e_j(\tau_i(l_{jm})))$ is the total combined number of failures observed by calendar time l_{jm} and R is our usual definition of randomization ratio (see Table 3.1). Under the null hypothesis, $S_{jm} \xrightarrow{d} N\left(0, \frac{D(l_{jm})R}{(R+1)^2}\right)$. Tsiatis (1981) proved that S_{j1}, S_{j2} have independent increments, and so $cov(S_{j1}, S_{j2}) = var(S_{j1})$. For the special case

when $R = 1$ in balanced allocation, under the null hypothesis, S_{jm} is sometimes known to follow $N(0, D(l_{jm})/4)$.

For our staggered dose design, we can standardize it as

$$Z_{jm} = \frac{S_{jm}}{\sqrt{\frac{D(l_{jm})R}{(R+1)^2}}}$$

so that $Z_{jm} \sim N(0,1)$ under the null hypothesis and $cov(Z_{j1}, Z_{j2}) = \sqrt{\frac{D(l_{j1})}{D(l_{j2})}}$. It is important to notice that instead of cohort size, the Fisher Information is the number of failures which depends on the calendar times, l_{jm}, and if the calendar times of analysis are equally spaced, then approximately, $cov(Z_{j1}, Z_{j2}) = 1/\sqrt{2}$. We can then resort to using the distributions in (3.3.1) to calculate the stopping boundaries. If we assume a dose response model q such that $\lambda_{j0} = q(d_j, d_0)$, we can calculate, instead of cohort size c, the number of failures D for a specified statistical power. We will also compute the expected stages, expected trial number of failures, and hence the calendar time required to observe these failures. This can be left for future work.

4.8 R Codes

4.8.1 Function Codes

```
# Only the variant design with D=2 is presented here
# Codes for other designs can be available on request
findboundsJ4D2 <- function(J=4, alphaseg) {
  b <- NULL
  # b1
  bseq <- seq(2,4, by=0.01)
  alphaseq <- rep(NA, length(bseq))
  for (i in 1:length(bseq)) {
    alphaseq[i] <- ((1-pnorm(bseq[i]))^2) + 2*(pnorm(bseq[i])*(1-pnorm(bseq[i])))
  }
  minposition <- min(which(alphaseq < alphaseg[1]))
  bound1 <- bseq[minposition]
  b <- c(b, bound1)
  # b2
  alphaseq <- rep(NA, length(bseq))
```

```
  for (i in 1:length(bseq)) {
    alphaseq[i] <- ((p4(-Inf, bound1, bseq[i]))^2) + 2*(p3(-Inf, bound1, bseq[i]))
    *(p4(-Inf, bound1, bseq[i]))
  }
  minposition <- min(which(alphaseq < alphaseg[2]))
  bound2 <- bseq[minposition]
  b <- c(b, bound2)
  # b3
  alphaseq <- rep(NA, length(bseq))
  for (i in 1:length(bseq)) {
    alphaseq[i] <- (p3(-Inf, bound1, bound2)^2)*(((1-pnorm(bseq[i]))^2) +
    2*(pnorm(bseq[i])*(1-pnorm(bseq[i]))))
  }
  minposition <- min(which(alphaseq < alphaseg[3]))
  bound3 <- bseq[minposition]
  b <- c(b, bound3)
  # b4
  alphaseq <- rep(NA, length(bseq))
  for (i in 1:length(bseq)) {
    alphaseq[i] <- (p3(-Inf, bound1, bound2)^2)*(((p4(-Inf, bound3, bseq[i]))^2) +
    2*(p3(-Inf, bound3, bseq[i]))*(p4(-Inf, bound3, bseq[i])))
  }
  minposition <- min(which(alphaseq < alphaseg[4]))
  bound4 <- bseq[minposition]
  b <- c(b, bound4)
  return(b)
}

findcohortcD2 <- function(targetpower, possiblec, J=4, m=2, R, K=4, model, dose,
ordering, stdev=1, delta=0, b) {
  meandose <- drcurve(model=model,dose=dose)
  reordereddose <- meandose[2:5][ordering]
  posspower <- rep(NA,length(possiblec))
  findpower <- function(ccc) {
    w11 <- b[1]-((reordereddose[1]-0)/(sqrt((R+1)/(ccc))))
    w21 <- b[1]-((reordereddose[2]-0)/(sqrt((R+1)/(ccc))))

    w12 <- b[2]-((reordereddose[1]-0)/(sqrt((R+1)/(2*ccc))))
    w22 <- b[2]-((reordereddose[2]-0)/(sqrt((R+1)/(2*ccc))))

    w33 <- b[3]-((reordereddose[3]-0)/(sqrt((R+1)/(ccc))))
    w43 <- b[3]-((reordereddose[4]-0)/(sqrt((R+1)/(ccc))))

    w34 <- b[4]-((reordereddose[3]-0)/(sqrt((R+1)/(2*ccc))))
    w44 <- b[4]-((reordereddose[4]-0)/(sqrt((R+1)/(2*ccc))))

    power1 <- (1-pnorm(w11))*(1-pnorm(w21)) + (pnorm(w11)*(1-pnorm(w21))) +
    (pnorm(w21)*(1-pnorm(w11)))
    power2 <- (p4(-Inf, w11, w12))*(p4(-Inf, w21, w22)) + (p3(-Inf, w11, w12))
    *(p4(-Inf, w21, w22)) + (p3(-Inf, w21, w22))*(p4(-Inf, w11, w12))
    power3 <- p3(-Inf, w11, w12)*p3(-Inf, w21, w22)*((1-pnorm(w33))*(1-pnorm(w43)) +
    (pnorm(w33)*(1-pnorm(w43))) + (pnorm(w43)*(1-pnorm(w33))))
    power4 <- p3(-Inf, w11, w12)*p3(-Inf, w21, w22)*(p4(-Inf, w33, w34))
    *(p4(-Inf, w43, w44)) + (p3(-Inf, w33, w34))*(p4(-Inf, w43, w44)) +
     (p3(-Inf, w43, w44))*(p4(-Inf, w33, w34))

    stagewise <- list()
    stagewise$foundpower <- power1+power2+power3+power4
    stagewise$stagewisepower <- c(power1, power2, power3, power4)
    return(stagewise)
  }
  for (i in 1:length(possiblec)) {
    ccc <- findpower(possiblec[i])
    posspower[i] <- ccc$foundpower
```

```
  }
  result <- list()
  result$possiblepower <- posspower
  position <- which(posspower > targetpower)
  position <- min(position)
  result$position <- position
  targetc <- possiblec[position]
  result$targetc <- targetc
  final <- findpower(targetc)
  result$stagewisepower <- final$stagewisepower
  result$simpower <- sum(final$stagewisepower)
  check <- final$stagewisepower
  result$expectedstage <- 1*check[1]+2*check[2]+3*check[3]+4*(1-check[1]-
  check[2]-check[3])
  return(result)
}
```

4.8.2 Analysis Codes

```
# Only the variant design with D=2 is presented here
# Codes for other designs can be available on request
D <- 2
J <- 4
m <- 2
K <- 4
R <- 2

segpocock <- alphasegments(type=1, alpha=0.05, K=4)
segpocock
sum(segpocock)
boundpocock <- findboundsJ4D2(J=4, alphaseg=segpocock)

segobf <- alphasegments(type=2, alpha=0.05, K=4)
segobf
sum(segobf)
boundobf <- findboundsJ4D2(J=4, alphaseg=segobf)

segrho0.3 <- alphasegments(type=3, alpha=0.05, K=4, rho=0.3)
segrho0.3
sum(segrho0.3)
boundrho0.3 <- findboundsJ4D2(J=4, alphaseg=segrho0.3)

segrho1.0 <- alphasegments(type=3, alpha=0.05, K=4, rho=1.0)
segrho1.0
sum(segrho1.0)
boundrho1.0 <- findboundsJ4D2(J=4, alphaseg=segrho1.0)

segrho3.0 <- alphasegments(type=3, alpha=0.05, K=4, rho=3.0)
segrho3.0
sum(segrho3.0)
boundrho3.0 <- findboundsJ4D2(J=4, alphaseg=segrho3.0)

effbounds <- rbind(boundpocock,
                   boundobf,
                   boundrho0.3,
                   boundrho1.0,
                   boundrho3.0)
effbounds; write.table(effbounds, "effboundsD2.txt", append=F, quote=F, row.names=T,
```

```
col.names=F)

ordering1 <- 1:4
orderingflat <- 1:4
orderinglinear <- 4:1
orderingemax <- 4:1
orderinglogistic <- 4:1
orderingumbrella <- c(3,2,4,1)
informordering <- rbind(orderingflat, orderinglinear, orderingemax, orderinglogistic,
orderingumbrella)
informordering
permute <- cbind(rep(1:4,each=6), c(2,2,3,3,4,4,1,1,3,3,4,4,1,1,2,2,4,4,1,1,2,2,3,3),
                 c(3,4,2,4,2,3,3,4,1,4,1,3,2,4,1,4,1,2,2,3,1,3,1,2),
                 c(4,3,4,2,3,2,4,3,4,1,3,1,4,2,4,1,2,1,3,2,3,1,2,1))
permute

# dose escalation ordering
check <- NULL; p <- 1
cohort9 <- NULL
simpower9 <- NULL
expstages9 <- NULL
proportions9 <- NULL
for (i in 1:5) { # 5 error spending plans
  for (j in 1:5) { # 5 dose response models
    usethis <- findcohortcD2(targetpower=0.9, possiblec=seq(20,250), J=4, m=2,
    R=2, K=4, model=j, dose=c(0,2,4,6,8), ordering=ordering1, stdev=1, delta=0, b=effbounds[i,])
    cohort9 <- rbind(cohort9, usethis[[3]])
    simpower9 <- rbind(simpower9, usethis[[5]])
    expstages9 <- rbind(expstages9, usethis[[6]])
    check <- matrix(c(i,j,p),c(1,3))
    write.table(check, "check1.txt", append=T, quote=F, row.names=F, col.names=F)
  }
}
cohort9
simpower9
expstages9

dim(cohort9) <- c(5,5); cohort9; write.table(cohort9, "9cohortorderesca.txt", append=T,
quote=F, row.names=F, col.names=F)

plannedss9 <- cohort9*K*((1/R)+2); plannedss9; write.table(plannedss9,
"9plannedssorderesca.txt", append=T, quote=F, row.names=F, col.names=F)

dim(simpower9) <- c(5,5); simpower9; write.table(simpower9, "9simpowerorderesca.txt",
append=T, quote=F, row.names=F, col.names=F)

dim(expstages9) <- c(5,5); expstages9; write.table(expstages9, "9expstageorderesca.txt",
append=T, quote=F, row.names=F, col.names=F)

expectedss9 <- cohort9*expstages9*((1/R)+2); expectedss9; write.table(expectedss9,
"9expssorderesca.txt", append=T, quote=F, row.names=F, col.names=F)

# informative ordering
check <- NULL; p <- 2
cohort9 <- NULL
simpower9 <- NULL
expstages9 <- NULL
for (i in 1:5) {
  for (j in 1:5) {
    usethis <- findcohortcD2(targetpower=0.9, possiblec=seq(20,250), J=4, m=2, R=2, K=4,
    model=j, dose=c(0,2,4,6,8), ordering=informordering[j,], stdev=1, delta=0, b=effbounds[i,])
    cohort9 <- rbind(cohort9, usethis[[3]])
    simpower9 <- rbind(simpower9, usethis[[5]])
    expstages9 <- rbind(expstages9, usethis[[6]])
```

```
    check <- matrix(c(i,j,p),c(1,3))
    write.table(check, "check2.txt", append=T, quote=F, row.names=F, col.names=F)
  }
}
dim(cohort9) <- c(5,5); cohort9; write.table(cohort9, "9cohortorderinform.txt", append=T, quote=F,
row.names=F, col.names=F)

plannedss9 <- cohort9*K*((1/R)+2); plannedss9; write.table(plannedss9,
"9plannedssorderinform.txt", append=T, quote=F, row.names=F, col.names=F)

dim(simpower9) <- c(5,5); simpower9; write.table(simpower9, "9simpowerorderinform.txt",
append=T, quote=F, row.names=F, col.names=F)

dim(expstages9) <- c(5,5); expstages9; write.table(expstages9, "9expstageorderinform.txt",
append=T, quote=F, row.names=F, col.names=F)

expectedss9 <- cohort9*expstages9*((1/R)+2); expectedss9; write.table(expectedss9,
"9expssorderinform.txt", append=T, quote=F, row.names=F, col.names=F)

# uninformative ordering
check <- NULL
cohort9 <- NULL
cohort9.2 <- NULL
simpower9 <- NULL
simpower9.2 <- NULL
expstages9 <- NULL
expstages9.2 <- NULL

for (p in 1:24) { # ordering
  for (i in 1:5) { # error spending
    for (j in 1:5) { # dose response model
      usethis <- findcohortcD2(targetpower=0.9, possiblec=seq(20,300), J=4, m=2, R=2, K=4,
      model=j, dose=c(0,2,4,6,8), ordering=permute[p,], stdev=1, delta=0, b=effbounds[i,])
      cohort9 <- rbind(cohort9, usethis[[3]])
      simpower9 <- rbind(simpower9, usethis[[5]])
      expstages9 <- rbind(expstages9, usethis[[6]])
      check <- matrix(c(i,j,p),c(1,3))
      write.table(check, "check3.txt", append=T, quote=F, row.names=F, col.names=F)
    }
  }
  dim(cohort9) <- c(1,25)
  write.table(cohort9, "cohort9.txt", append=T, quote=F, row.names=F, col.names=F)
  cohort9.2 <- rbind(cohort9.2, cohort9)
  cohort9 <- NULL

  dim(simpower9) <- c(1,25)
  write.table(simpower9, "simpower9.txt", append=T, quote=F, row.names=F, col.names=F)
  simpower9.2 <- rbind(simpower9.2, simpower9)
  simpower9 <- NULL

  dim(expstages9) <- c(1,25)
  write.table(expstages9, "expstages9.txt", append=T, quote=F, row.names=F,
  col.names=F)
  expstages9.2 <- rbind(expstages9.2, expstages9)
  expstages9 <- NULL

}

cohort9.2
cohort9.3 <- apply(cohort9.2, 2, mean)
write.table(matrix(cohort9.3,c(5,5)), "9cohortorderuninf.txt", append=T, quote=F,
row.names=F,
col.names=F)
dim(cohort9.3) <- c(5,5)
```

```
cohort9.3

plannedss9 <- cohort9.3*K*((1/R)+2); plannedss9; write.table(plannedss9,
"9plannedssorderuninf.txt", append=T, quote=F, row.names=F, col.names=F)

simpower9.2
simpower9.3 <- apply(simpower9.2, 2, mean)
write.table(matrix(simpower9.3,c(5,5)), "9simpowerorderuninf.txt", append=T, quote=F,
row.names=F, col.names=F)
dim(simpower9.3) <- c(5,5)
simpower9.3

expstages9.2
expstages9.3 <- apply(expstages9.2, 2, mean)
write.table(matrix(expstages9.3,c(5,5)), "9expstageorderuninf.txt", append=T, quote=F,
row.names=F, col.names=F)
dim(expstages9.3) <- c(5,5)
expstages9.3

expectedss9 <- cohort9.3*expstages9.3*((1/R)+2); expectedss9; write.table(expectedss9,
"9expssorderuninf.txt", quote=F, row.names=F, col.names=F)

# end of chapter code
```

Chapter 5

Bayesian Hierarchical Bias Model for Establishing Biosimilarity

5.1 Introduction

The concept of biosimilarity has received increasing popularity within the scientific community recently. One big motivation to explore biosimilar products is the unprecedented opportunity gradually opened up by numerous soon-to-be expiring licenses of major biological products. In 2010, the passage of the Biologics Price Competition and Innovation Act (BPCI) created an abbreviated licensure pathway in section 351(k) of the Public Health Service Act (PHS). This new law allows for an expeditious approval process for a generic follow-on biological product shown to be biosimilar to a licensed reference biological product. Section 351(i) of the PHS Act defines biosimilarity to mean "that the biological product is highly similar to the reference product notwithstanding minor differences in clinically inactive components" and that "there are no clinically meaningful differences between the biological product and the reference product in terms of the safety, purity, and potency of the product." Due to their large and complex molecular structures, biological products are fundamentally disparate from small synthetic drugs, and so are their mechanisms of action (see Figure 5.1). Traditional statistical methods used to test for bioequivalence as

in a generic drug development may not be the most efficient way to apply to biosimilarity (Kang and Chow, 2012).

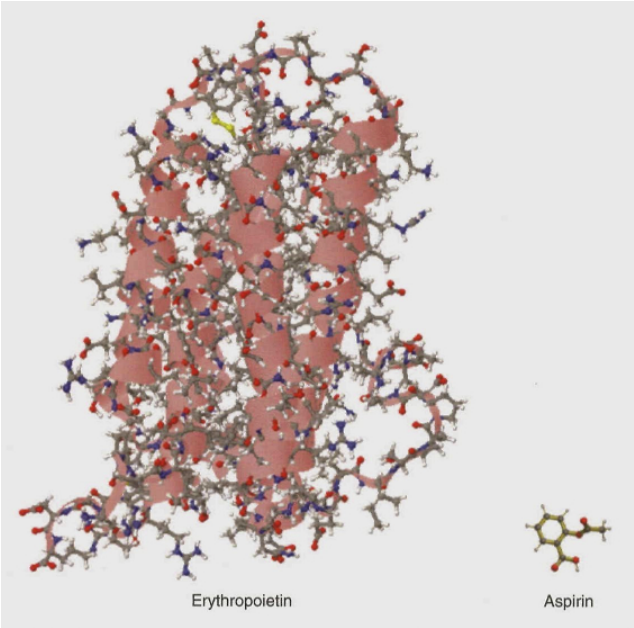


Figure 5.1: Structure of erythropoietin and aspirin, illustrating the larger and small complex structure of biological products compared with traditional small molecule therapeutics. (Reprint with permission from Springer. Calvo and Zuñiga, 2012)

Many of the recently proposed methods to establish biosimilarity between an innovator reference biological product and a generic follow-on biological product primarily borrowed ideas from average bioequivalence (ABE) trials. An ABE trial using a 2×2 crossover design is the standard approach suggested by the U.S. Food and Drug Administration (FDA) to test for the equivalence between a reference drug and a new generic drug. Often before the conduct of a bioequivalence trial, if the dissolution profiles of the reference and the new generic drugs are proved to be similar, the need to conduct a clinical bioequivalence trial will be waived (Saranadasa and Krishnamoorthy, 2005). This can save both time and cost in conducting a clinical study. Therefore, statistical methodologies also exist to show similarity between dissolution profiles. Before discussing their suitability to biosimilarity studies, it is important to review and compare the current approaches to both average bioequivalence and dissolution profile studies.

When estimating a drug dissolution profile, the mean dissolution concentration is measured across p specified time points. Testing for dissolution profile similarity requires a multivariate test of similarity. The current FDA standard approach is to assess if the similarity factor f_2 defined as

$$f_2 = 50\, log_{10}\left\{100\left(1+\frac{\boldsymbol{\mu}'\boldsymbol{\mu}}{p}\right)\right\}$$

falls within similarity margins, where $\boldsymbol{\mu} = \boldsymbol{\mu}_T - \boldsymbol{\mu}_R$ is the p-dimensional vector of mean differences between the generic test drug and the reference drug and p is the total number of time points (Moore and Flanner, 1996). Saranadasa and Krishnamoorthy (2005) assumed that $\boldsymbol{\mu}_T - \boldsymbol{\mu}_R = \delta \boldsymbol{e}$ where $\boldsymbol{e}$ is a unit vector and developed a test that shows the scalar parameter δ is within equivalence margins $\pm\delta_0$. The methods just described rely on reducing a multivariate testing problem into a univariate testing problem. Other methods assuming a profile model are also proposed. A common example in these model-dependent approaches is the auto-regressive time series model (Tsong *et al.*, 1997; Chow and Ki, 1997). When directly approaching this multivariate problem, Berger and Hsu (1996) provided an α-level test that rejects dissimilar dissolution profiles if

$$Max\left\{|d_i| + c\left(\frac{S_i^2}{\nu}\right)^{\frac{1}{2}}\right\} < \delta_0$$

among all time points where $i = 1, 2, ..., p$, $|d_i|$ is the observed absolute mean difference, S_i^2 is the pooled variance at ith time point, $\nu = n_T + n_R - 2$ is the degree of freedom, and c is the $(1-\alpha)100\%$ percentile of the Student's t distribution.

For testing average bioequivalence, the standard approach is the two one-sided test (TOST) developed by Schuirmann (1987) for the following equivalence hypotheses:

$$H_0 : \mu_T - \mu_R \geqslant \delta \qquad \text{or} \qquad \mu_T - \mu_R \leqslant -\delta \qquad \text{versus} \qquad H_A : -\delta < \mu_T - \mu_R < \delta$$

where μ_T and μ_R represent the mean bioavailability measures for the test and reference drugs respectively on the logarithmic scale. The three major bioavailability measures are T_{max}, the

time until the maximum concentration of the drug in plasma is reached, C_{max}, the maximum concentration, and AUC, the area under the concentration curve from dose administration to final observation time. The most accepted rule stated in the regulatory guideline is the 80%/125% margins which are $\pm\delta = \pm 0.22314$ on the logarithmic scale. Under normality assumption, the two sets of one-sided hypotheses will be tested with ordinary one-sided t-tests. It can be concluded that μ_R and μ_T are equivalent, for a balanced study ($n = n_R = n_T$), if

$$t_1 = \frac{(\overline{x}_T - \overline{x}_R) - \delta}{s\sqrt{2/n}} \geqslant t_{1-\alpha(\nu)} \qquad \text{and} \qquad t_2 = \frac{\delta - (\overline{x}_T - \overline{x}_R)}{s\sqrt{2/n}} \geqslant t_{1-\alpha(\nu)}$$

where ν is the degree of freedom. This is statistically equivalent to the confidence interval approach developed by Westlake (1981). Recently, the concepts of population bioequivalence (PBE) and individual bioequivalence (IBE) are proposed and they are related to drug prescribability and switchability respectively. PBE refers to not only the equivalence of population means in bioavailability between reference and generic drugs, but also the equivalence of their population variances in bioavailability. When these two formulations are equivalent in both means and variances, they are considered as equally prescribable to a new patient. IBE refers to the equivalence of bioavailability when a patient switches from the reference formulation to the generic formulation, or vice versa and so within-subject variances for both formulations will be accounted for. Additional criteria were proposed to test for these equivalence concepts. Dragalin *et al.* (2003) introduced the symmetric Kullback-Leibler divergence (KLD), a distance metric, as a new criterion for both PBE and IBE,

$$\triangle(f,g) = \left[\int \left(f(x) - g(x)) log\left(\frac{f(x)}{g(x)}\right)\right) dx\right]^{1/2}$$

where f and g are probability density functions of the outcome variables corresponding to the two formulations, and $\triangle(f,g)$ represents their difference. They derived the criteria using KLD for the parametric exponential family distributions. The advantages of using KLD are that it is invariant to monotonic transformation and that it can be easily extended

to the multivariate cases. Chervoneva, Hyslop, and Hauck (2007) extended the univariate population equivalence criterion to the multivariate case by looking into the trace of the multivariate expected difference between reference and generic drugs.

Biological products are fundamentally different from drug compounds. They are large polypeptide molecules with a much larger molecular weight than small-molecule synthetic drugs. Therefore, they tend to have a longer half-life and require a longer wash-out period. In this case, the standard crossover design normally used for bioequivalence trial may not be efficient if applied to biosimilarity trials. A more appropriate design would be the parallel group design. Various biosimilarity criteria have been suggested and they depend on the study designs and objectives. For a parallel three-arm trial with two of the arms for the reference product from two different manufacturing lots and the other one for the follow-on biological product, Kang and Chow (2012) proposed the relative distance rd as a biosimilarity criterion

$$rd = \left| \frac{\mu_T - \mu_R}{\mu_{R_1} - \mu_{R_2}} \right|$$

where μ_T is the population mean of the efficacy outcome for test biosimilar product T, μ_{R_1} and μ_{R_2} are those of reference product R_1 and R_2 from two different manufacturing lots, and finally $\mu_R = (\mu_{R_1} + \mu_{R_2})/2$. The authors developed a test that assumes asymptotic distribution of its maximum likelihood estimator (MLE). Lin *et al.* (2012) presented the parallel line assay design that requires two dose-response trials for both the reference and follow-on biological products. Under the assumption of the parallel line bioassay, they assumed a linear relationship between the binary efficacy endpoint and the mean dose-dependent product characteristic. The biosimilarity criterion in this case is the relative potency defined as

$$\triangle = \frac{\alpha_T - \alpha_R}{\beta_c}$$

where α_T is the intercept term of the linear regression for new generic biologics, α_R for reference biologics and β_c is the assumed common slope. Earlier, Chow and Liu (2010) also presented the same criterion but the efficacy endpoint is normally distributed. In order

to avert the distributional assumption of normality and developing a potentially insufficient aggregated criterion, Lei and Olson (2010) introduced the use of non-parametric tests comparing the distributions of the efficacy endpoint between reference and follow-on biologics. They performed a simulation study to compare the performance of the TSOT, the Kolmogorov-Smimov (KS) test, and the overlap coefficient test. TSOT is sensitive to the magnitude of the variance but the two non-parametric tests are not. However, the non-parametric tests are sensitive to group differences in variance and centrality. When the two distributions are the same, the KS test has a more stable probability of claiming equivalence without requiring bigger sample size.

This chapter is motivated by the need to develop an innovative statistical method for proving biosimilarity. Here is the organization of the subsequent sections. Section 5.2 introduces the composite endpoint of interest and describes the proposed clinical study for the demonstration of biosimilarity that uses this composite endpoint. This section also provides the rationale for a non-inferiority testing framework and a Bayesian inferential approach to achieve the study's objectives. Section 5.3 describes the details of a simulation plan to examine the operating characteristics of this proposed method and also summarizes the simulation results with comparison to the frequentist approach. Section 5.3.4 discusses the impact of different prior densities on the operating characteristics. Section 5.3.3 describes an adaptive two-stage design that has an interim assessment based on predictive probability and briefly studies its characteristics. Section 5.4 discusses the overall results and proposes further work in this area.

5.2 Biosimilarity Using Composite Endpoint

Although in the past few years some statistical methods have been proposed for the case of a single primary efficacy endpoint, some biological products are designed to treat medical conditions with improvement measured by several endpoints. For example, rheumatoid

arthritis (RA) is a disease of the immune system that leads to the inflammation in the joints. It causes a myriad of symptoms such as pain, joint swelling, fatigue, weakness, and stiffness. It also leads to loss of physical function and permanent joint damage. The exact cause of RA is unknown, but it is believed that patients with RA have changing immune and inflammatory system and an over-abundance of tumor necrosis factor (TNF). An increased level of TNF is responsible for joint inflammation. Treatments of mild cases of RA include disease-modifying anti-rheumatic drugs (DMARDs). However, for moderate to severe cases of RA that do not respond well to DMARDs, intervention using biological products such as TNF blockers may be helpful in slowing down RA progression.

In clinical trials studying RA, the current standard measure of efficacy is the ACR20 criteria recommended by the American College of Rheumatology (ACR) Committee. For each individual patient in a trial, it measures if this patient has experienced a clinical response of overall improvement by evaluating the percentage of improvement in a core set of variables during the trial. Generally speaking, if a patient experiences at least 20% improvement from baseline in multiple variables simultaneously, this patient is defined as having satisfied the definition of a clinical response. Therefore, ACR20 is a composite criterion and has served as a working model to other disorders that currently require multiple primary endpoints (Offen *et al.*, 2007). The percent change in each of these variables is also measured at different time points such as 3, 6 and 12 months and one of the time points is used to establish primary efficacy. Table 5.1 summarizes the ACR20 improvement criteria (Felson *et al.*, 1993). Other more stringent measures such as ACR50 or ACR70 (i.e. $\geqslant$ 50% or $\geqslant$ 70% improvement on the same set of endpoints) and other validated scales such as the Disease Activity Score in 28 joints (DAS28) criteria are also adopted.

5.2.1 Study Design and Non-Inferiority Hypotheses

We can consider a clinical development program to show an experimental generic biological product is biosimilar to a licensed reference biological product, and that these products

Table 5.1: ACR20 Improvement Criteria

Quantitative criterion	Endpoints
percent reduction $\geqslant$ 20% improvement in	1. Tender joint count, and 2. Swollen joint count, and At least 3 of the following: 3. Physician global assessment of disease activity 4. Patient global assessment of disease activity 5. Patient assessment of pain (e.g. Visual Analog Scale) 6. Physical disability or functionality 7. Inflammatory marker: erythrocyte sedimentation rate (ESR) or C-reactive protein (CRP)

are used to treat a medical condition with symptom improvement defined by a composite endpoint. Since biological products require a much longer washout period, a clinical cross-over design may not be efficient, and therefore we suggest using a parallel two-arm randomized clinical design. As implied in the FDA guidance document for the industry titled "Scientific Considerations in Demonstrating Biosimilarity to a Reference Product", when accumulated evidence on biosimilarity is clear from previous molecular, functional, and pre-clinical studies, a smaller clinical study can be convened to confirm biosimilarity.

Motivated by the need to decrease sample size for a clinical study, we propose a non-inferiority framework to test for biosimilarity using the clinical data. This non-inferiority trial design allows the current biosimilarity trial to meaningfully connect to any similarly conducted historical trials that have evaluated the effect of the licensed reference biological product. Since a standard treatment is already available for the medical condition, including a placebo arm in the current trial will not be ethical. We can use k to index the biological product with $k = 1$ representing the innovator reference product and $k = 2$ the proposed follow-on biological product. In this case, we are interested in testing if, based on the composite endpoint, the proposed biological product is not inferior to the licensed biological product. This non-inferiority design may reduce the number of sample subjects needed.

In this two-arm design, patients are randomized to either the original reference or the follow-on generic biological product. For each patient, outcomes on J multiple endpoints will be measured at pre-specified follow-up times. These J endpoints can be generally considered as

independent measures. We can use x_{kji} to denote the jth endpoint ($j = 1, 2, ..., J$) observed in the ith patient receiving the product k. In this case, we can assume that it is normally distributed as

$$x_{kji} \sim N(\mu_{kj}, \sigma_k^2) \tag{5.2.1}$$

where $k = 1$ or 2, and $i = 1, 2, ..., n_k$. The fixed randomization ratio is therefore equal to $R = n_2/n_1$. μ_{kj} is the mean response for the jth endpoint and σ_k^2 is the variance which is assumed to be the same for all J endpoints but different between the products. In addition, we want to consider combining these J endpoints into a single composite binary efficacy endpoint y_{ki} which can be generally defined as

$$y_{ki} = \begin{cases} 1 & \boldsymbol{x}_{ki} \geqslant \boldsymbol{\omega} \\ 0 & \text{otherwise} \end{cases} \tag{5.2.2}$$

where $\boldsymbol{x}_{ki} = (x_{k1i}, x_{k2i}, ..., x_{kJi})'$ is the random vector of outcomes for the ith patient and $\boldsymbol{\omega} = (\omega_1, \omega_2, ..., \omega_J)'$ is a J-dimensional vector of cutoff points for the endpoints common to both biological products, assuming that higher values of x_{kji}'s are desirable. If we denote the probability of a response on the composite endpoint for product k as p_k, then $p_k = P(y_{ki} = 1) = P(\boldsymbol{x}_{ki} \geqslant \boldsymbol{\omega})$, and our non-inferiority (NI) hypotheses of interest can be constructed as

$$H_0 : p_2 - p_1 \leqslant -\delta \qquad \text{versus} \qquad H_A : p_2 - p_1 > -\delta \tag{5.2.3}$$

where $\delta(\delta > 0)$ is the pre-specified non-inferiority margin for the difference between the two probabilities.

5.2.2 Bayesian Approach

There are several advantages of approaching this biosimilarity hypothesis testing problem from a Bayesian perspective. First, if the efficacy profile of the licensed reference biological

product comparing to a placebo control has been established in a historical trial, and if there are additional historical trials evaluating the effect of this product, then we can incorporate these sources of information into the analysis of the current trial of biosimilarity. These historical trials can be further assumed to be exchangeable. Also, since multiple endpoints are considered in developing the composite endpoint, the Bayesian approach allows for the borrowing of estimative strength between the J multiple endpoints on the precision parameters in addition to the borrowing between historical trials. This also means that fewer subjects may be needed for the reference product and more subjects can be randomized to the new and potentially biosimilar product. This is a realization of the FDA guidance regarding its suggestion to use smaller clinical studies and to convene them based on results from previously conducted studies. Furthermore, the composite endpoint can be defined by criteria on the multiple endpoints which provide clinically meaningful interpretation. According to (5.2.2), p_k will be defined as a function of the parameters such that $p_k = f(\mu_{k1}, \mu_{k2}, ..., \mu_{kJ}, \sigma_k^2)$ for $k = 1$ or 2.

5.2.3 Hierarchical Bias Model

In this proposed hierarchical bias model, we allow the inclusion of any number of historical trials for the licensed reference product. For example, if there are H historical trials available before the conduct of the current biosimilarity trial, we can let x_{1hji} be the value of the jth endpoint observed for the ith patient receiving the original reference product $k = 1$ in the hth historical trial such that

$$x_{1hji} \sim N(\mu_{1hj}, \sigma_1^2) \tag{5.2.4}$$

where $i = 1, 2, ..., n_{1h}$, $h = 1, 2, ..., H$, and $j = 1, 2, ..., J$. In the above model, we assume that these H trials and the current biosimilarity trial share the same within-study variance parameter σ_1^2 and it is also assumed to be constant across all J endpoints. This assumption allows borrowing between the historical trials and also between the J endpoints. In addition, we can represent the jth sample mean as $\overline{x}_{1hj}$ which is equal to $(\sum_{i=1}^{n_{1h}} x_{1hji})/n_{1h}$. Other

sample means can be similarly defined.

Additionally, under exchangeability, we consider the mean parameters, μ_{1j} of the current biosimilarity trial and μ_{1hj} of the hth historical trial, for the original reference product, come from the same distribution as

$$\mu_{1j}, \mu_{1hj} \sim N(\mu_{1j}^{o}, \sigma_{1b}^{2}) \tag{5.2.5}$$

where $h = 1, 2, ..., H$. μ_{1j}^{o} is the overall mean and σ_{1b}^{2} is the between-trial variance parameter, which is assumed to be the same across the J endpoints. Hierarchical modeling is a logical way of combining historical data when exchangeability between parameters is highly plausible, and as in the current problem, these historical trials used an efficacy response defined by the same criterion. This hierarchical structure implies heterogeneity of the mean endpoints. Modeling the mean endpoints hierarchically recognizes that these historical trials, although using the same licensed reference product, may exhibit slightly different mean endpoints due to possible but small differences in the conduct of the trials or in the study populations they enrolled to these trials. This variation will be captured by the between-trial variance parameter. This ensures that the current biosimilarity trial is validly connected to the historical trials via the assumption of exchangeability.

For the new generic follow-on product in the current biosimilarity trial, we think of its mean response on the j endpoint, μ_{2j}, as having a bias term from that of the mean endpoint of the original product, μ_{1j}. Pocock (1976) discussed the parameterizing a bias term to model the difference between mean of historical control and the same control but in the current trial. Therefore, we propose this relationship

$$\mu_{1j} = \mu_{2j} + \xi_{j} \tag{5.2.6}$$

where ξ_j represents the bias of μ_{2j} from μ_{1j}. If ξ_j is equal to 0, then $\mu_{2j} = \mu_{1j}$ meaning that the follow-on product has the same mean as the licensed reference product on the jth

endpoint. If $\xi_j < 0$, then it means the follow-on product exhibits a better effect than the reference product in the jth endpoint, and the opposite interpretation follows if $\xi_j > 0$. Since we do not know the true value of ξ_j, we can assume a model for this bias parameter as

$$\xi_j \sim N(\theta, \sigma_\xi^2) \tag{5.2.7}$$

where $j = 1, 2, ..., J$. We center the expectation of ξ_j skeptically at the null hypothesis, θ, to allow the data to reflect and influence its true direction and magnitude away from the null value. The null value θ is the margin on the scale of individual endpoints, such that when this margin is uniformly subtracted from all of the mean responses, the probability of the binary composite endpoint will decrease by exactly the amount of δ as in $f(\mu_{k1} - \theta, \mu_{k2} - \theta, ..., \mu_{kJ} - \theta, \sigma_k^2) - f(\mu_{k1}, \mu_{k2}, ..., \mu_{kJ}, \sigma_k^2) = -\delta$. This relationship between δ and θ is one-on-one. Therefore, centering the mean of ξ_j on θ also suggests that μ_{1j} and μ_{2j} are dissimilar to begin with. We also assume that the variance parameter σ_ξ^2 to be the same across all J endpoints but a large σ_ξ^2 will suggest that this distribution is only weakly informative. Figure 5.2 displays the graphical representation of this model with each circle representing a random node, a single-line arrow representing the dependent stochastic relationship and a double-line arrow representing a logical relationship.

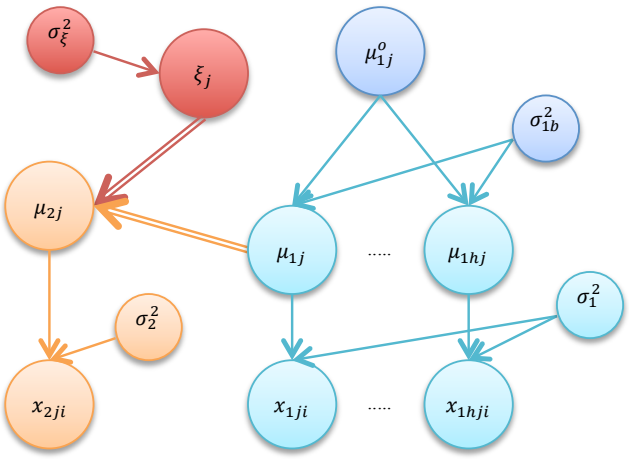


Figure 5.2: Graphical representation of the proposed Bayesian hierarchical bias model, $j = 1, 2, ..., J$.

Now that we have completely specified the hierarchical bias model, we can consider the following prior distributions for the parameters. For μ_{1j}^o, we can assume a flat non-informative prior as $P(\mu_{1j}^o) \propto 1$ for $j = 1, 2, ..., J$ since we have no prior information regarding these overall mean parameters. We can elicit Jeffrey's prior distributions for the remaining variance parameters such that $P(\sigma_1^2) \propto 1/\sigma_1^2$, $P(\sigma_2^2) \propto 1/\sigma_2^2$, $P(\sigma_{1b}^2) \propto 1/\sigma_{1b}^2$, and $P(\sigma_\xi^2) \propto 1/\sigma_\xi^2$. As a result, the joint posterior distribution of all parameters will be given by the product of all likelihoods and specified densities of the parameters as

$$
\begin{aligned}
& P(\boldsymbol{\mu}_1, \boldsymbol{\mu}_{11}, ..., \boldsymbol{\mu}_{1h}, ..., \boldsymbol{\mu}_{1H}, \boldsymbol{\mu}_1^o, \boldsymbol{\mu}_2, \boldsymbol{\xi}, \sigma_1^2, \sigma_{1b}^2, \sigma_2^2, \sigma_\xi^2 | \boldsymbol{x}_{1j}, \boldsymbol{x}_{11j}, ..., \boldsymbol{x}_{1Hj}, \boldsymbol{x}_{2j}, j = 1, ..., J) \\
\propto & \left[\prod_{j=1}^{J} \left\{ L(\mu_{1j}, \sigma_1^2 | \boldsymbol{x}_{1j}) \left(\prod_{h=1}^{H} L(\mu_{1hj}, \sigma_1^2 | \boldsymbol{x}_{1hj}) \right) L(\mu_{2j} = \mu_{1j} - \xi_j, \sigma_2^2 | \boldsymbol{x}_{2j}) P(\mu_{1j} | \mu_{1j}^o, \sigma_{1b}^2) \right. \right. \\
& \left. \left. \left(\prod_{h=1}^{H} P(\mu_{1hj} | \mu_{1j}^o, \sigma_{1b}^2) \right) P(\xi_j | \theta, \sigma_\xi^2) \right\} P(\mu_{1j}^o) \right] \left(\frac{1}{\sigma_1^2 \sigma_{1b}^2 \sigma_2^2 \sigma_\xi^2} \right) \qquad (5.2.8)
\end{aligned}
$$

where $\boldsymbol{\mu}$'s and $\boldsymbol{\xi} \in \mathbb{R}^J$, $\boldsymbol{x}_{1j}$ representing the data on all n_1 subjects, $\boldsymbol{x}_{1hj}$ representing the data on all n_{1h} subjects, and $\boldsymbol{x}_{2j}$ representing the data on all n_2 subjects for the jth endpoint with $j = 1, 2, ..., J$.

Based on the joint density in (5.2.8), we can find the conditional posterior distributions for the parameters. For the hth historical trial, the conditional posterior distribution for μ_{1hj} is given by

$$\mu_{1hj}|\boldsymbol{x}_{1hj} \sim N\left(\tilde{\sigma}^2_{1h}\left(\frac{n_{1h}\overline{x}_{1hj}}{\sigma_1^2}+\frac{\mu^o_{1j}}{\sigma^2_{1b}}\right), \tilde{\sigma}^2_{1h}=\left(\frac{n_{1h}}{\sigma_1^2}+\frac{1}{\sigma^2_{1b}}\right)^{-1}\right) \tag{5.2.9}$$

where $h = 1, 2, ..., H$ and $j = 1, 2, ..., J$. The mean of the posterior distribution is a weighted average of the sample mean $\overline{x}_{1hj}$ and μ^o_{1j}. As for the original reference product in the current trial, the mean parameters will have conditional posterior distributions as

$$\mu_{1j}|\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j} \sim N\left(\tilde{\sigma}^2_1\left(\frac{n_1\overline{x}_{1j}}{\sigma_1^2}+\frac{n_2(\overline{x}_{2j}+\xi_j)}{\sigma_2^2}+\frac{\mu^o_{1j}}{\sigma^2_{1b}}\right), \tilde{\sigma}^2_1=\left(\frac{n_1}{\sigma_1^2}+\frac{n_2}{\sigma_2^2}+\frac{1}{\sigma^2_{1b}}\right)^{-1}\right) \tag{5.2.10}$$

The mean of the distribution is given by the weighted average of the sample mean $\overline{x}_{1j}$, μ^o_{1j}, and $(\overline{x}_{2j} + \xi_j)$. The derivation of this analytical form is given in Section 5.5.1 in the Appendix. The overall mean parameter μ^o_{1j} will therefore have conditional posterior distribution given by

$$\mu^o_{1j} \sim N\left(\frac{\mu_{1j}+\sum_{h=1}^{H}\mu_{1hj}}{1+H}, \tilde{\sigma}^2_o=\frac{\sigma^2_{1b}}{1+H}\right). \tag{5.2.11}$$

The bias parameters will take on the following conditional posterior distributions

$$\xi_j \sim N\left(\tilde{\sigma}^2_\xi\left(\frac{n_2(\mu_{1j}-\overline{x}_{2j})}{\sigma_2^2}+\frac{\theta}{\sigma^2_\xi}\right), \tilde{\sigma}^2_\xi=\left(\frac{n_2}{\sigma_2^2}+\frac{1}{\sigma^2_\xi}\right)^{-1}\right). \tag{5.2.12}$$

However, the mean parameters for the generic follow-on product is completely specified by both μ_{1j} and ξ_j as in (5.2.6), therefore, its posterior distribution will be given by their respective posterior distributions

$$\mu_{2j}|\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j} = \mu_{1j}|\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j} - \xi_j, \qquad j = 1, 2, ..., J. \tag{5.2.13}$$

As for the within-study variance parameter for the licensed reference product, using the Jeffrey's prior, we get the conditional posterior distribution as the inverse-gamma distribution with shape and scale parameters given by

$$\begin{aligned} & \sigma_1^2|\boldsymbol{x}_{1j}, \boldsymbol{x}_{11j}, ..., \boldsymbol{x}_{1Hj}, j = 1, 2, ..., J \\ \sim \quad & IG\left(\frac{J(n_1 + \sum_{h=1}^{H} n_{1h})}{2}, \frac{1}{2}\left[\sum_{j=1}^{J}\sum_{i=1}^{n_1}(x_{1ji} - \mu_{1j})^2 + \sum_{h=1}^{H}\sum_{j=1}^{J}\sum_{i=1}^{n_{1h}}(x_{1hji} - \mu_{1hj})^2\right]\right) \end{aligned} \tag{5.2.14}$$

The derivation of this analytical form is given in Section 5.5.2 in the Appendix. The between-study variance parameter for the licensed reference product will also follow a conditional posterior distribution as an inverse-gamma distribution given by

$$\sigma_{1b}^2 \sim IG\left(\frac{J(1+H)}{2}, \frac{1}{2}\left[\sum_{j=1}^{J}(\mu_{1j} - \mu_{1j}^o)^2 + \sum_{h=1}^{H}\sum_{j=1}^{J}(\mu_{1hj} - \mu_{1j}^o)^2\right]\right). \tag{5.2.15}$$

Additionally, the variance parameter for the follow-on product has conditional posterior distribution as

$$\sigma_2^2|\boldsymbol{x}_{2j}, j = 1, 2, ..., J \sim IG\left(\frac{Jn_2}{2}, \frac{1}{2}\sum_{j=1}^{J}\sum_{i=1}^{n_2}(x_{2ji} - (\mu_{1j} - \xi_j))^2\right). \tag{5.2.16}$$

Lastly, the variance parameter for the bias term σ_ξ^2 can be shown to be an inverse-gamma distribution

$$\sigma_\xi^2 \sim IG\left(\frac{J}{2}, \frac{1}{2}\sum_{j=1}^{J}(\xi_j - \theta)^2\right). \tag{5.2.17}$$

Using the conditional posterior distributions in (5.2.9), (5.2.10), (5.2.13), (5.2.12), (5.2.14), (5.2.15), (5.2.16), and (5.2.17), we can find the marginal posterior distributions of μ_{1j}, σ_1^2, μ_{2j}, and σ_2^2, and hence those of $p_k = f(\mu_{k1}, \mu_{k2}, ..., \mu_{kJ}, \sigma_k^2)$ for $k = 1$ or 2 and perform posterior inference based on the distribution of $p_2 - p_1$. There are two ways to find the marginal posterior densities. One way is to integrate out the conditioning parameters using their conditional posterior densities, but this can be highly intractable. A

more viable way is to use Markov chain Monte Carlo (MCMC) methods to simulate the marginal posterior densities from the conditional posterior densities. In this case, since we have close forms for the conditional posterior distributions, we can use Gibbs sampling, one of the widely used MCMC techniques. Using the same Gibbs sampling, we can also simulate the marginal posterior distribution of $p_2 - p_1$. We can directly estimate the posterior probability $P(p_2 - p_1 > -\delta|\boldsymbol{x}_{1j}, \boldsymbol{x}_{11j}, ..., \boldsymbol{x}_{1Hj}, \boldsymbol{x}_{2j}, j = 1, 2, ..., J) = E[I(p_2 - p_1 > -\delta)|\boldsymbol{x}_{1j}, \boldsymbol{x}_{11j}, ..., \boldsymbol{x}_{1Hj}, \boldsymbol{x}_{2j}, j = 1, 2, ..., J]$. The decision rule is to reject the null hypothesis when this posterior probability is greater than a critical probability p_c which can be pre-specified as high as 95% or 97.5% depending on the clinical significance.

5.2.4 Determination of Bayesian Non-Inferiority Margin

Another major challenge in a non-inferiority trial design is to determine the NI margin δ and hence its corresponding θ for each of the individual endpoints. One way to specify δ is to mirror the fixed margin method in the frequentist paradigm in the current Bayesian paradigm (Gamalo, Wu, Tiwari, 2012; Gamalo, Tiwari, LaVange, 2013). In the frequentist paradigm, the NI margin is set to be the lower bound of the $100\%(1 - \alpha)$ confidence interval for the effect $p_{1h'} - p_{0h'}$ in a selected historical placebo-controlled trial h', where $0h'$ represents the placebo arm and $1h'$ represents the innovator reference product arm in this trial. This historical placebo-controlled trial h' was usually a trial that led to its first FDA approval.

If we assume a similar model as in (5.2.4) for the placebo and treatment arms in this placebo-controlled trial

$$x_{kh'ji} \sim N(\mu_{kh'j}, \sigma^2_{kh'})$$

where $k = 0$ or 1 and elicit a flat non-informative prior for $\mu_{kh'j}$ as in $P(\mu_{kh'j}) \propto 1$ and a

Jeffery's prior for the variance $\sigma^2_{kh'}$ as in $P(\sigma^2_{kh'}) \propto 1/\sigma^2_{kh'}$, then

$$\begin{aligned}
\mu_{kh'j}|\boldsymbol{x}_{kh'j}, j = 1, 2, ..., J &\sim N\left(\overline{x}_{kh'j}, \frac{\sigma^2_{kh'}}{n_{kh'}}\right) \\
\sigma^2_{kh'}|\boldsymbol{x}_{kh'j}, j = 1, 2, ..., J &\sim IG\left(\frac{Jn_{kh'}}{2}, \frac{1}{2}\sum_{j=1}^{J}\sum_{i=1}^{n_{kh'}}(x_{kh'ji} - \mu_{kh'j})^2\right). \quad (5.2.18)
\end{aligned}$$

We can use Gibbs sampling to simulate for $p_{1h'} - p_{0h'}$ and solve for δ as the lower bound of the $100\%(1-\alpha)$ credibility interval such that

$$P(p_{1h'} - p_{0h'} > \delta|\boldsymbol{x}_{0h'j}, \boldsymbol{x}_{1h'j}, j = 1, 2, ..., J) \geqslant 1 - \frac{\alpha}{2}. \quad (5.2.19)$$

In addition, we want to explore a slightly more conservative margin $\delta_\lambda = (1-\lambda)\delta$ where $0 < \lambda < 1$. This margin δ_λ can represent the clinically relevant effect that the follow-on generic product should not be worse than the innovator reference product. Examples of λ are 0% (full margin: $\delta_0 = \delta$), 25%, or 50% (half of the margin: $\delta_{0.5} = \delta/2$).

5.3 Simulation Study

5.3.1 Simulation Objectives and Plan

In order to characterize the operating characteristics of this Bayesian non-inferiority design for biosimilarity, we will conduct a simulation study that is motivated by our previous example of rheumatoid arthritis (RA). The primary composite efficacy endpoint is the ACR20 at 6 months (or 24 weeks) although ACR50, ACR70, or ACR20 at other time points can be secondary endpoints for generating future hypotheses. There is no safety endpoint in this study and it can be assumed that doses higher than the recommended dose do not create safety concerns.

The ACR20 has seven components and they represent separate categories of symptoms as in Table 5.1. These components are generally assumed to be independent measures. Therefore,

J is equal to 7 such that μ_{k1} and μ_{k2} are at least 20% and at least 3 of $\{\mu_{k3}, ..., \mu_{k7}\}$ are at least 20% where $k = 0h', 1h', 1h, 1, 2$ and $h = 1, 2, ..., H$. The objectives of this simulation study are (1) to assess the type I error in the Bayesian paradigm under the null hypothesis and to compare it with that in the frequentist paradigm, (2) to evaluate the statistical power in the Bayesian paradigm under the alternative hypothesis given overall sample size n and randomization ratio R as well as to compare it with that in the frequentist paradigm, and (3) to characterize the impact of different λ as in δ_λ and p_c on the aforementioned characteristics. The following delineate the simulation steps.

1. As a real-life motivating example, we conducted a literature search on historical trials on Etanercept. Etanercept is a TNF receptor (p75) fusion protein, linked to the Fc portion of human IgG1. We found five published studies: (1) Moreland *et al.* (1997), (2) Moreland *et al.* (1999), (3) Weinblatt *et al.* (1999), (4) Bathon *et al.* (2000), and (5) Klareskog *et al.* (2004). Among these studies, only one of them (Moreland *et al.*, 1999) was a confirmatory placebo-controlled trial for the monotherapy of Etanercept (25mg/mL) while the other trials studied either combined therapies of Etanercept or lower doses of Etanercept. Etanercept (25mg/mL) was administered subcutaneously twice a week and the primary efficacy endpoint is ACR20 at 6 months (or 24 weeks). This trial led to its FDA approval for RA in 1998. Therefore, $H = 1$, and we will use this historical trial to determine the NI margin as well as including it in the hierarchical bias model ($h' = h$). Table 5.2 below summarizes the partial result from this historical trial. A positive percent change is interpreted as a reduction in the corresponding symptom component while a negative percent change means an increase. Table 5.3 describes the simulation setting for the follow-on biological product in the current proposed biosimilarity study. We can specify $\mu_{0h'j}, \mu_{1h'j}, \mu_{1j}, \mu_{2j}, \sigma^2_{0h'}, \sigma^2_{1h'}, \sigma^2_1, \sigma^2_2, n_{0h'}, n_{1h'}, n, R, \lambda$, and p_c for $j = 1, 2, ..., 7$ using the results from this trial.

2. Since we do not have patient-level data on the chosen historical trial, except reported

sample mean values on the endpoints in Table 5.2, we will use these reported sample means as true values for the parameters. We simulate data on $x_{0h'ji} \sim N(\mu_{0h'j}, \sigma^2_{0h'})$ for $i = 1, 2, ..., n_{0h'}$ and $x_{1h'ji} \sim N(\mu_{1h'j}, \sigma^2_{1h'})$ for $i = 1, 2, ..., n_{1h'}$.

3. Use Gibbs sampling to generate posterior samples for $\mu_{0h'j}, \mu_{1h'j}, \sigma^2_{0h'}$, and $\sigma^2_{1h'}$ using conditional posterior distributions in (5.2.18). Using Gelman-Rubin-Brooks plots, trace plots, and auto-correlation plots with five chains of sampling, determine the number of MCMC iterations N sufficient for the convergence of the MCMC chains after 10% burn-in (Gelman and Rubin, 1992; Plummer *et al.*, 1992).

4. During each successive iteration of the same Gibbs sampling in step (3), compute $p_{0h'}$ and $p_{1h'}$ according to the definition of the composite endpoint ACR20 as follow:

$$\begin{aligned} p_k &= f(\mu_{k1}, \mu_{k2}, ..., \mu_{k7}, \sigma^2_k) \\ &= \prod_{j=1}^{2} \Phi\left(-\frac{20-\mu_{kj}}{\sigma_k}\right) \left[P(\{S_3\}) + P(\{S_4\}) + \prod_{j=3}^{7} \Phi\left(-\frac{20-\mu_{kj}}{\sigma_k}\right)\right] \quad (5.3.1) \end{aligned}$$

where $k = 0h'$ or $1h'$. Φ is the cumulative density function of the standardized normal distribution. S_3 is the event that any three of the 3rd to 7th endpoints are equal to or greater than 20% while the remaining two endpoints are less than 20% and similarly S_4 is the event that any four of them are equal to or greater than 20% while the remaining one is less than 20%.

5. Using the posterior samples of $p_{0h'}$ and $p_{1h'}$ after 10% burn-in from step (4), estimate δ, the lower bound of the credibility interval of $p_{1h'} - p_{0h'}$, as in (5.2.19) using an α level of 0.05. For a given λ, we can specify the NI margin as $\delta_\lambda = (1-\lambda)\delta$. Find the corresponding θ_λ such that $\mu_{2j} - \mu_{1j} = -\theta_\lambda (\theta_\lambda > 0)$ for $j = 1, 2, ..., J$ and $p_2 - p_1 = -\delta_\lambda$ according to (5.3.1). A strictly (one-on-one) increasing relationship exists between δ_λ and θ_λ since there is a one-on-one relationship between θ and $\Phi(-(20-(\mu_k-\theta))/\sigma_k)$

and hence p_k, therefore as

$$\begin{aligned} p_2 - p_1 &= f(\mu_{21}, \mu_{22}, ..., \mu_{27}, \sigma_2^2) - f(\mu_{11}, \mu_{12}, ..., \mu_{17}, \sigma_1^2) \\ &= f(\mu_{11} - \theta_\lambda, \mu_{12} - \theta_\lambda, ..., \mu_{17} - \theta_\lambda, \sigma_2^2) - f(\mu_{11}, \mu_{12}, ..., \mu_{17}, \sigma_1^2) \\ &= -\delta_\lambda, \end{aligned}$$

then θ_λ and δ_λ are also one-on-one.

6. If there is more than one historical trial ($H > 1$), then we can simulate data on $x_{1hji} \sim N(\mu_{1hj}, \sigma_{1h}^2)$ for $h = 1, 2, ..., H$ using the reported sample means for μ_{1hj}. However, since we only have one historical trial ($H = 1$) and it is the same as the placebo-controlled trial ($h = h'$), we will simply use the data simulated in step (2). For the current trial of biosimilarity, we simulate data using $\mu_{1j} = \mu_{1h'j} + \triangle$ where $\triangle$ represents the difference of the means of the reference product in the current and historical trials, uniform across J endpoints. If $\triangle = 0$, it means the effect is the same, i.e. constancy assumption is met. If $\triangle > 0$, it means the reference product is performing better in the current trial than in the historical trial; and then the opposite argument is true when $\triangle < 0$. Then we simulate data on $x_{1ji} \sim N(\mu_{1j}, \sigma_1^2)$ for $i = 1, 2, ..., n_1$ and $x_{2ji} \sim N(\mu_{2j}, \sigma_2^2)$ for $i = 1, 2, ..., n_2$ where $n_1 = n/(1 + R)$ and $n_2 = nR/(1 + R)$.

7. Use Gibbs sampling to generate posterior samples of size N after 10% burn-in for the following parameters: $\mu_{1hj}, \mu_{1j}, \mu_{1j}^o, \mu_{2j}, \xi_j, \sigma_1^2, \sigma_{1b}^2, \sigma_2^2$ and σ_ξ^2 using conditional posterior distributions in (5.2.9), (5.2.10), (5.2.11), (5.2.13), (5.2.12), (5.2.14), (5.2.15), (5.2.16), and (5.2.17).

8. During each successive iteration of the same Gibbs sampling, use the posterior samples of these parameters, μ_{1j}, σ_1^2, μ_{2j}, and σ_2^2 in step (7) to compute p_1 and p_2 according to the definition of the composite endpoint ACR20.

9. Using the posterior samples of p_1 and p_2 after 10% burn-in from step (8), estimate

the posterior probability, $P(p_2 - p_1 > -\delta_\lambda | \boldsymbol{x}_{1j}, \boldsymbol{x}_{1h'j}, \boldsymbol{x}_{2j}, j = 1, 2, ..., J)$ and if this posterior probability is greater than p_c as the Bayesian criterion or decision rule, then we can conclude that this non-inferiority biosimilarity trial is a success in term of its efficacy in ACR20.

10. In order to assess the Bayesian type I error, 10,000 simulated trials will be used to estimate the probability of rejecting the null hypothesis when $\mu_{2j} - \mu_{1j} = -\theta_\lambda (\theta_\lambda > 0)$ for $j = 1, 2, ..., J$ such that $p_2 - p_1 = -\delta_\lambda$ according to (5.3.1).

11. In order to assess the Bayesian power given n, R, and $\mu_{2j} - \mu_{1j} = -\theta_a > -\theta_\lambda$ for $j = 1, 2, ..., J$, 10,000 simulated trials will also be used to estimate the probability of rejecting the null hypothesis. This will estimate the Bayesian power.

12. In order to compare the Bayesian type I error with the frequentist type I error, use the simulated data from steps (2) and (6) to estimate the probabilities $p_k = P(y_{ki} = 1), k = 0h', 1h', 1, 2$. These estimators can be denoted as $\hat{p}_k$. We conclude that in the current biosimilarity trial, the follow-on biologic is non-inferior to the reference biologic if the following is true:

$$
\begin{aligned}
&(\hat{p}_2 - \hat{p}_1) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2} + \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} \\
&> -(1-\lambda) \left[(\hat{p}_{1h'} - \hat{p}_{0h'}) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_{1h'}(1-\hat{p}_{1h'})}{n_{1h'}} + \frac{\hat{p}_{0h'}(1-\hat{p}_{0h'})}{n_{0h'}}} \right] \quad (5.3.2)
\end{aligned}
$$

where $\alpha/2$ is set to be 0.025. Using the data from the same 10,000 trials in steps (10) and (11), the probability of rejecting the null hypothesis is estimated and can be compared to that from the Bayesian paradigm.

13. All of the statistical programming will be conducted in the open-source $\mathcal{R}$ software.

Table 5.2: Historical trial on monotherapy of Etanercept (25mg/mL) at 6 months - Moreland *et al.*, 1999

k	Treatment	n_k	$P(ACR20 = 1)$ or $\hat{p}_k$	$\hat{\mu}_{k1}$	$\hat{\mu}_{k2}$	$\hat{\mu}_{k3}$	$\hat{\mu}_{k4}$	$\hat{\mu}_{k5}$	$\hat{\mu}_{k6}$	$\hat{\mu}_{k7}$	σ_k^2
$0h'$	Placebo	80	11%	6%	-7%	2%	-3%	-22%	2%	-207%	1600%
$1h'$	Etanercept (25mg/mL)	78	59%	56%	47%	44%	46%	53%	39%	31%	1600%

Note: For $\hat{\mu}_{k7}$, CRP is used instead of ESR. The variance was reported in Moreland *et al.*, 1997 and was assumed to be 1,600% for the calculation of sample size.

Table 5.3: Simulation setting for the current non-inferiority biosimilarity trial.

Parameter	Values	Description
λ	0, 0.25, 0.5	Sizing factor for non-inferiority margin
$\delta_\lambda = (1-\lambda)\delta$	$\delta_0 = \delta, \delta_{0.25} = 0.75\delta, \delta_{0.5} = 0.5\delta$	Re-sized non-inferiority margin for δ
θ_λ	$\theta_0, \theta_{0.25}, \theta_{0.5}$	Re-sized non-inferiority margin for θ
$\triangle$	-2, 0, 2	Impact of constancy assumption
μ_{1j}, σ_1^2	$\mu_{1j} = \mu_{1h'j} + \triangle$, $\sigma_{1h'}^2$	Use historical trial on Etanercept (25mg/ML) arm in Table 5.2
μ_{2j}, σ_2^2	$\mu_{1j} - \theta_\lambda, \sigma_{1h'}^2$	For assessing Bayesian type I error
μ_{2j}, σ_2^2	$\mu_{1j} - \theta_a$ $(\theta_a = 0, \theta_\lambda/2), \sigma_{1h'}^2$	For assessing Bayesian power
n	60, 120	Overall trial sample size
R	1, 2	Fixed randomization ratio
p_c	95%, 97.5%	Critical probability
N	determined by simulation	Number of posterior Gibbs samples after 10% burn-in

5.3.2 Simulation Results

Using the estimated means from Table 5.2 and the simulation setting laid out in Table 5.3, we simulated patient-level data for the selected historical trial. Using this hypothetical patient-level data, we generated Gibbs samplings on the parameters, $\mu_{0h'j}, \mu_{1h'j}, \sigma_{0h'}^2$, and $\sigma_{1h'}^2$. In order to determine the optimal number of iterations for posterior convergence, Figures 5.3, 5.4, and 5.5 for $k = 1h'$ are used to assess if $N = 3,000$ is sufficient for convergence. As seen from the Gelman-Rubin-Brooks plots (Figure 5.3), the shrink factors are well close to 1, suggesting convincing evidence of convergence. Therefore, before 10% burn-in, $N = 3,000$ can be a sufficient choice. This is also confirmed by the trace plots (Figure 5.5) and auto-correlation plots (Figure 5.4) which indicate minimal lags between simulations. Figure 5.6 shows the kernel plots of the posterior samples after 10% burn-in. Using these chains of sampling, we derived the posterior samples of the probability of

clinical response ACR20, $p_{0h'}$ and $p_{1h'}$, and hence their difference, $p_{1h'} - p_{0h'}$. The kernel distribution of the posterior probability difference is given in Figure 5.7. The lower bound of the 95% credibility interval is estimated to be 0.4604. Therefore, under different pre-specified sizing factors λ, we can state the different NI margins for subsequent simulation: $\lambda = 0$ will give $\delta_0 = 0.4604$, $\lambda = 0.25$ will give $\delta_{0.25} = 0.3453$, and finally $\lambda = 0.5$ will give $\delta_{0.5} = 0.2302$. Using the equation in (5.3.1), we can find their corresponding θ: $\theta_0 = 34.7$, $\theta_{0.25} = 22.2$, and $\theta_{0.5} = 14.0$.

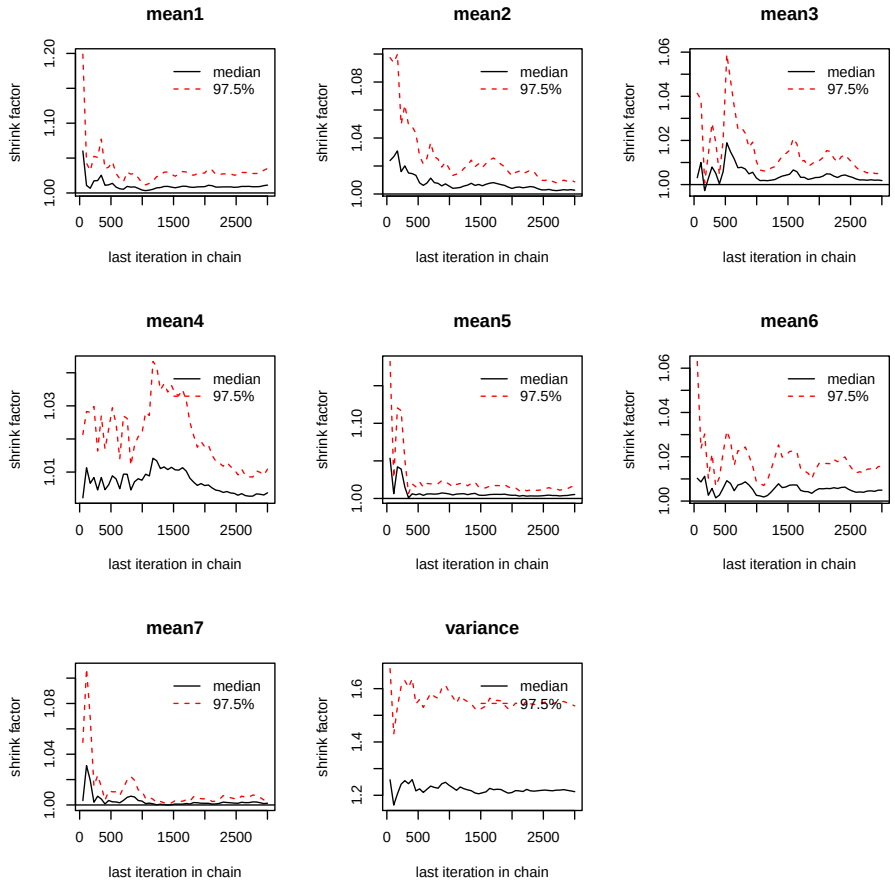


Figure 5.3: Gelman-Rubin-Brooks plots of posterior simulations for mean and variance parameters, $\mu_{1h'j}$ $(j = 1, 2, ..., 7)$ and $\sigma^2_{1h'}$.

Using the same patient-level data, we also calculated the probabilities of clinical response ACR20: $\hat{p}_{0h'} = 0$ and $\hat{p}_{1h'} = 0.5641$ under the frequentist perspective. The estimate for the

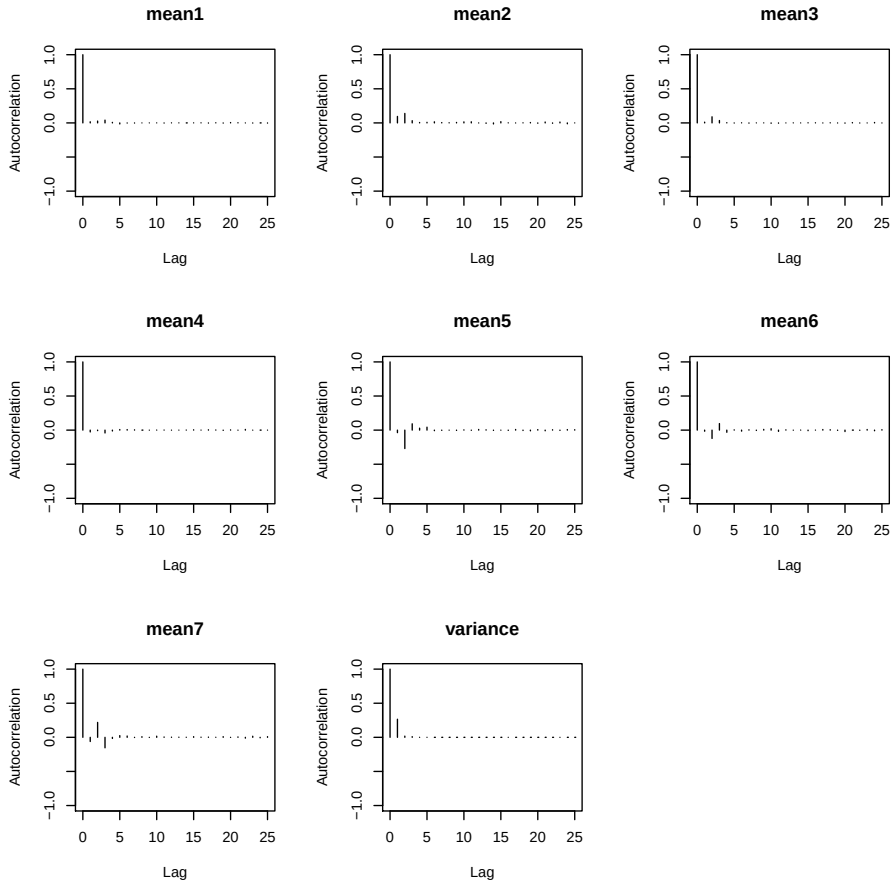


Figure 5.4: Autocorrelation plots of posterior simulations for mean and variance parameters, $\mu_{1h'j}$ $(j = 1, 2, ..., 7)$ and $\sigma^2_{1h'}$.

treatment arm is not far from the one reported in this historical trial (59% in Table 5.2), but the estimate for the placebo arm is under-estimated (11% in Table 5.2). The lower bound of the 95% confidence interval is therefore estimated to be 0.4019. The corresponding re-sized NI margins will be 0.4019, 0.3014, and 0.2009. These are somewhat smaller than the corresponding ones estimated in the Bayesian method above. Based on the same simulation plan as described in Table 5.3, we conducted subsequent simulation using 10,000 simulated identical trials. The same simulated two-arm trial data will be used to determine if the trial is a success separately for the proposed Bayesian method and the standard frequentist method. Table 5.4 shows the result of the simulated type I error under both analytical

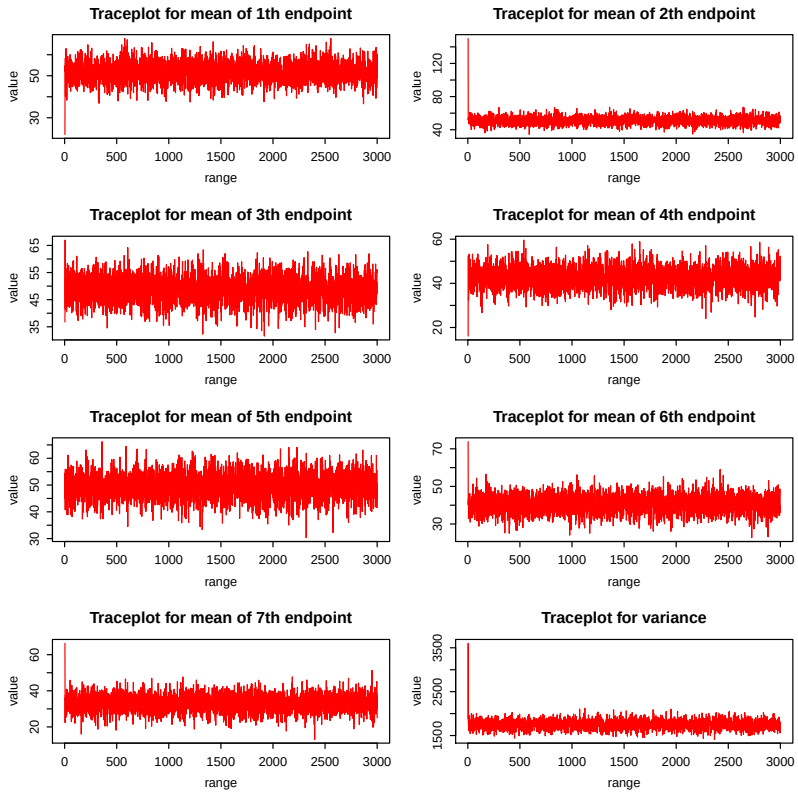


Figure 5.5: Trace plots of posterior simulations for mean and variance parameters, $\mu_{1h'j}(j = 1, 2, ..., 7)$ and $\sigma^2_{1h'}$.

paradigms and Table 5.5 displays the result of the simulated statistical power.

In Table 5.4, we observe general preservation of type I error under 0.025 when $\triangle = \mu_{1j} - \mu_{1hj} = 0$ or 2, but inflated type I error when $\triangle = -2$. That is, when the reference product is performing identically or better in the current biosimilarity trial than in the reference historical trial, the type I error is controlled under the target size. However, if it performs worse in the current trial than in the historical trial, the type I error is inflated, both for the Bayesian and the frequentist methods. However, as λ increases to 0.5 when the NI margin is smaller, the type I error inflation is only seen in the frequentist approach but not the proposed Bayesian approach. In fact, the type I error under the Bayesian method is well-controlled under 0.01 even when the reference product is doing worse in the current trial. Figure 5.8 explores the relationship between the type I error rate and $\triangle = \mu_{1j} - \mu_{1hj}$

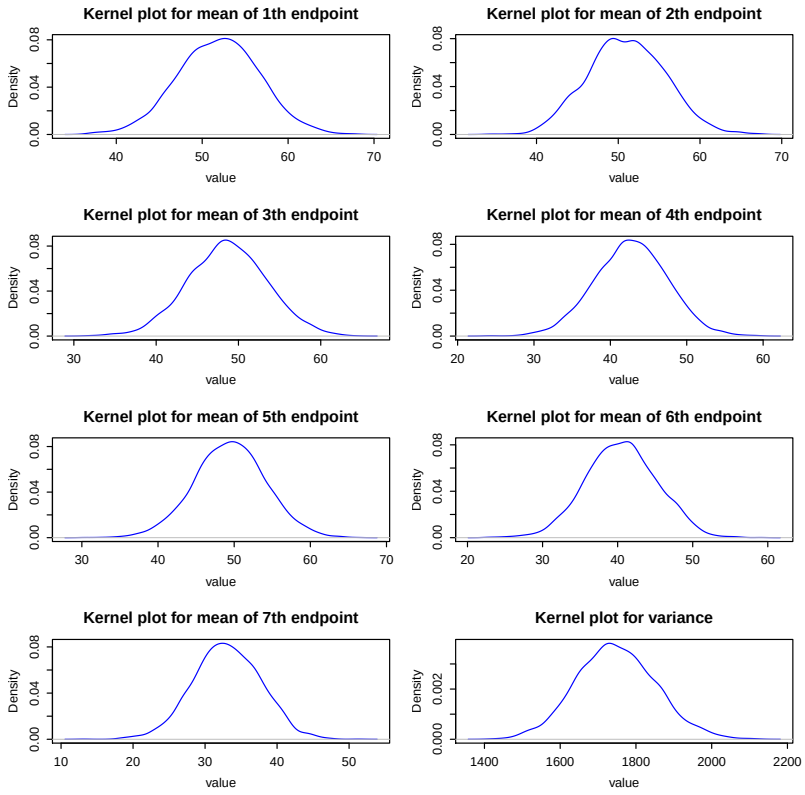


Figure 5.6: Kernel plots of posterior simulations for mean and variance parameters, $\mu_{1h'j}$ $(j = 1, 2, ..., 7)$ and $\sigma^2_{1h'}$.

for $n = 60, R = 1$ and 0.95. As $\triangle$ falls below zero, indicating the reference product is doing worse in the current trial than in the reference historical trial, type I error rate increases in both Bayesian and frequentist approaches, with the Bayesian approach having a steeper increase than the frequentist approach. Both methods are able to preserve the type I error at 0.025 when $\triangle = 0$, that is when the effect of the reference product is constant in both trials. The inflation of type I error, when reference product is doing worse in the current trial, is possibly due to the larger lower bound of 95% credibility interval in the historical trial as related to the reduced effect size of the reference product comparing to the putative placebo, which does not exist in the current trial. However, as λ increases, the re-sized NI margin narrows, due to the influence of the skeptical prior for ξ_j, the proposed Bayesian method is able to protect the inflation of type I error, even when $\triangle < 0$ but the frequentist

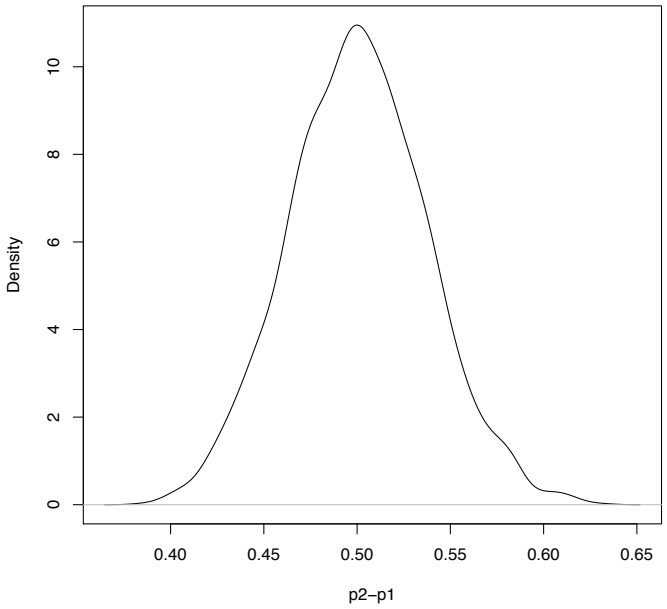


Figure 5.7: Kernel plot showing the posterior distribution of probability difference $p_{1h'} - p_{0h'}$ based on simulated hypothetical patient-level data.

approach cannot.

In Table 5.5, we can see that when $\lambda = 0$ and $\triangle = 0, 2$, that is, when the full NI margin is used, the statistical power of the Bayesian method is unanimously higher than that of the frequentist method. As for $\lambda = 0.25$ and $\triangle = 0, 2$, statistical power of the Bayesian method is smaller than that of the frequentist method only when sample size is small as in $n = 60$ and when the alternative is at $\delta_a = \delta_{0.25}/2$. Other than that, the power of the Bayesian method is superior to the frequentist method. As λ decreases to 0.5 and when $\triangle = 0, 2$, the NI margin narrows down to $\delta_{0.5} = 0.2302$, when $n = 60$ and the alternative is either at $\delta_a = \delta_{0.5}/2 = 0.1151$ or at $\delta_a = 0$, statistical power is very low in both the Bayesian and frequentist approaches with the Bayesian method suffering more loss of power due to the strong influence of the skeptical prior on ξ_j within the smaller margin. However, an increase in sample size to $n = 120$ seems to promise a much better improvement in statistical power when $\delta_a = 0$ than the improvement in frequentist power. This is mainly due to the increasing influence of the data over the skeptical null prior, resulting in improved Bayesian

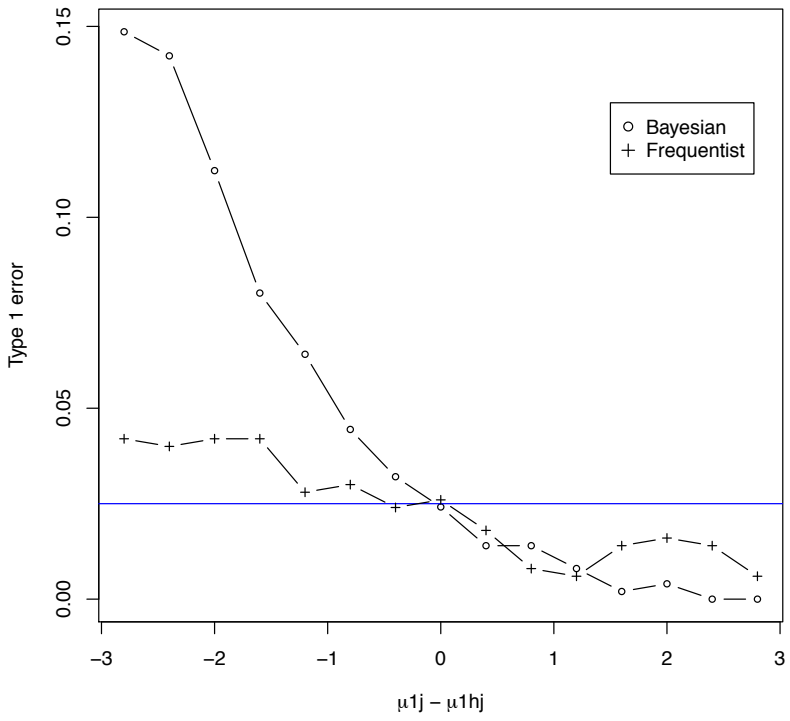


Figure 5.8: Plot of type I error against value of $\triangle = \mu_{1j} - \mu_{1hj}$ for all j. Setting is $n = 60$, $R = 1$, and $p_c = 0.95$

Table 5.4: Simulation result on Bayesian and frequentist type I error rate using 10,000 simulated trials.

				$n = 60$				$n = 120$			
				$R = 1$		$R = 2$		$R = 1$		$R = 2$	
λ	δ_λ	θ_λ	$\triangle$	$p_c = 0.95$	$p_c = 0.975$	$p_c = 0.95$	$p_c = 0.975$	$p_c = 0.95$	$p_c = 0.975$	$p_c = 0.95$	$p_c = 0.975$
0	0.4604	34.7	-2	0.102, 0.037	0.051, 0.037	0.096, 0.045	0.042, 0.045	0.159, 0.041	0.086, 0.041	0.139, 0.039	0.071, 0.040
0	0.4604	34.7	0	0.021, 0.024	0.009, 0.023	0.023, 0.025	0.006, 0.025	0.020, 0.020	0.010, 0.019	0.017, 0.021	0.006, 0.021
0	0.4604	34.7	2	0.003, 0.011	0.002, 0.012	0.004, 0.018	0.001, 0.016	0.001, 0.008	0.001, 0.009	0.002, 0.011	0.000, 0.011
0.25	0.3453	22.2	-2	0.038, 0.027	0.014, 0.028	0.029, 0.029	0.011, 0.034	0.051, 0.027	0.018, 0.025	0.041, 0.033	0.014, 0.032
0.25	0.3453	22.2	0	0.010, 0.020	0.003, 0.020	0.007, 0.026	0.002, 0.029	0.010, 0.019	0.003, 0.019	0.006, 0.023	0.002, 0.019
0.25	0.3453	22.2	2	0.003, 0.018	0.001, 0.019	0.002, 0.022	0.000, 0.023	0.003, 0.016	0.001, 0.016	0.001, 0.017	0.001, 0.017
0.5	0.2302	14.0	-2	0.009, 0.028	0.002, 0.023	0.006, 0.032	0.002, 0.033	0.009, 0.022	0.003, 0.024	0.006, 0.027	0.002, 0.027
0.5	0.2302	14.0	0	0.008, 0.023	0.002, 0.024	0.003, 0.030	0.001, 0.025	0.007, 0.023	0.002, 0.019	0.003, 0.025	0.001, 0.022
0.5	0.2302	14.0	2	0.007, 0.024	0.002, 0.023	0.003, 0.025	0.000, 0.025	0.007, 0.022	0.002, 0.024	0.002, 0.024	0.001, 0.023

Note: For each cell, the left number refers to the Bayesian type I error rate and the right number to frequentist type I error rate.

Table 5.5: Simulation result on Bayesian and frequentist statistical power using 10,000 simulated trials.

						$n = 60$				$n = 120$			
						$R = 1$		$R = 2$		$R = 1$		$R = 2$	
λ	δ_λ	θ_λ	δ_a	θ_a	$\triangle$	$p_c = 0.95$	$p_c = 0.975$	$p_c = 0.95$	$p_c = 0.975$	$p_c = 0.95$	$p_c = 0.975$	$p_c = 0.95$	$p_c = 0.975$
0	0.4604	34.7	0.2302	14.0	-2	0.794, 0.490	0.706, 0.486	0.700, 0.390	0.640, 0.402	0.998, 0.744	0.980, 0.772	0.976, 0.668	0.962, 0.626
0	0.4604	34.7	0.2302	14.0	0	0.748, 0.450	0.692, 0.464	0.706, 0.388	0.636, 0.374	0.984, 0.726	0.984, 0.714	0.986, 0.658	0.958, 0.628
0	0.4604	34.7	0.2302	14.0	2	0.736, 0.450	0.688, 0.460	0.682, 0.382	0.584, 0.350	0.992, 0.748	0.982, 0.696	0.988, 0.686	0.964, 0.658
0	0.4604	34.7	0	0	-2	1.000, 0.970	1.000, 0.964	1.000, 0.932	1.000, 0.892	1.000, 1.000	1.000, 0.994	1.000, 1.000	1.000, 0.996
0	0.4604	34.7	0	0	0	1.000, 0.944	1.000, 0.940	1.000, 0.932	1.000, 0.934	1.000, 0.998	1.000, 1.000	1.000, 1.000	1.000, 0.998
0	0.4604	34.7	0	0	2	1.000, 0.960	1.000, 0.946	1.000, 0.936	1.000, 0.944	1.000, 1.000	1.000, 1.000	1.000, 1.000	1.000, 1.000
0.25	0.3453	22.2	0.1726	10.4	-2	0.182, 0.310	0.116, 0.250	0.134, 0.280	0.068, 0.222	0.534, 0.522	0.430, 0.466	0.476, 0.438	0.401, 0.406
0.25	0.3453	22.2	0.1726	10.4	0	0.184, 0.268	0.086, 0.258	0.148, 0.248	0.104, 0.276	0.528, 0.434	0.430, 0.426	0.478, 0.432	0.414, 0.420
0.25	0.3453	22.2	0.1726	10.4	2	0.176, 0.262	0.116, 0.254	0.124, 0.270	0.102, 0.246	0.558, 0.434	0.480, 0.470	0.522, 0.470	0.398, 0.412
0.25	0.3453	22.2	0	0	-2	0.906, 0.746	0.862, 0.742	0.900, 0.734	0.850, 0.714	1.000, 0.974	1.000, 0.966	0.996, 0.956	1.000, 0.952
0.25	0.3453	22.2	0	0	0	0.940, 0.756	0.888, 0.730	0.876, 0.720	0.858, 0.694	1.000, 0.956	1.000, 0.958	1.000, 0.940	0.998, 0.936
0.25	0.3453	22.2	0	0	2	0.914, 0.764	0.884, 0.770	0.894, 0.754	0.854, 0.752	1.000, 0.960	1.000, 0.964	1.000, 0.974	1.000, 0.946
0.5	0.2302	14.0	0.1151	6.9	-2	0.040, 0.154	0.018, 0.158	0.016, 0.150	0.014, 0.128	0.128, 0.232	0.064, 0.262	0.074, 0.222	0.046, 0.226
0.5	0.2302	14.0	0.1151	6.9	0	0.036, 0.144	0.012, 0.122	0.022, 0.114	0.008, 0.130	0.106, 0.224	0.072, 0.210	0.082, 0.214	0.056, 0.222
0.5	0.2302	14.0	0.1151	6.9	2	0.038, 0.152	0.018, 0.154	0.044, 0.158	0.012, 0.146	0.134, 0.220	0.070, 0.220	0.104, 0.200	0.056, 0.218
0.5	0.2302	14.0	0	0	-2	0.314, 0.476	0.246, 0.456	0.264, 0.380	0.218, 0.380	0.790, 0.674	0.756, 0.692	0.744, 0.604	0.716, 0.662
0.5	0.2302	14.0	0	0	0	0.342, 0.452	0.226, 0.432	0.254, 0.380	0.214, 0.432	0.844, 0.726	0.802, 0.684	0.798, 0.668	0.690, 0.634
0.5	0.2302	14.0	0	0	2	0.370, 0.474	0.286, 0.496	0.324, 0.406	0.254, 0.416	0.838, 0.710	0.800, 0.736	0.820, 0.666	0.716, 0.652

Note: For each cell, the left number refers to the Bayesian power and the right number to frequentist power.

power.

5.3.3 Adaptive Two-Stage Bayesian Design

Additionally, we want to propose a design that is comprised of two stages, with an interim assessment that is based on predictive probability to decide if the trial can be stopped for early efficacy. Predictive distribution has been proposed as a criterion for clinical trial monitoring (Spiegelhalter and Freedman, 1986; Dmitrienko and Wang, 2006) and has been successfully applied to clinical trials (Lee and Liu, 2008). In this Bayesian adaptive design, we can assign $n_{(1)}$ subjects to the first stage and $n_{(2)}$ subjects to the second stage so that $n_{(1)}+n_{(2)} = n$. As a result, the innovator reference product will receive $n_{(s)1} = n_{(s)}R/(1+R)$ subjects in the sth stage ($s = 1, 2$) such that $n_{(1)1} + n_{(2)1} = n_1$ and generic follow-on product will have $n_{(s)2} = n_{(s)}/(1 + R)$ subjects randomized in the sth stage such that $n_{(1)2} + n_{(2)2} = n_2$.

We also let $x_{(s)kji}$ be the outcome for the jth endpoint for products $k = 1, 2$, collected in the sth stage where $i = 1, 2, ..., n_{(s)k}$ and that it follows the same model in (5.2.1). Interim inference will use the predictive probability of rejecting the null hypothesis P^* based on the interim data $\boldsymbol{x}_{(1)kji}$ observed so far as well as future samples $\boldsymbol{x}^*_{(2)kji}$ that is based on their predictive distributions. In this problem, we have multiple parameters and therefore the estimation of predictive distributions will rely on Gibbs sampling using their respective conditional predictive distributions. For example, the conditional predictive distributions of future sample means for the reference product for the second stage $\overline{x}^*_{(2)1j}$ given the interim

data of the first stage collected so far can be shown to be

$$\begin{aligned}
&\overline{x}^*_{(2)1j}|\boldsymbol{x}_{(1)1j}, \boldsymbol{x}_{(1)2j}, \boldsymbol{x}_{1hj} \\
\sim\ & N\left(\left(\frac{n_{(1)1}}{\sigma_1^2}+\frac{n_{(1)2}}{\sigma_2^2}+\frac{1}{\sigma_{1b}^2}\right)^{-1}\left(\frac{n_{(1)1}\overline{x}_{(1)1j}}{\sigma_1^2}+\frac{n_{(1)2}(\overline{x}_{(1)2j}+\xi_j)}{\sigma_2^2}+\frac{\mu_{1j}^o}{\sigma_{1b}^2}\right),\right. \\
&\left.\tilde{\sigma}_1^{2*}=\frac{\sigma_1^2}{n_{(2)1}}+\left(\frac{n_{(1)1}}{\sigma_1^2}+\frac{n_{(1)2}}{\sigma_2^2}+\frac{1}{\sigma_{1b}^2}\right)^{-1}\right).
\end{aligned} \tag{5.3.3}$$

The close forms of the conditional predictive distributions of future sample means for the follow-on product for the second stage $\overline{x}^*_{(2)2j}$ are not shown here since they depend on those of ξ_j^*. If we denote all of the interim data as $\boldsymbol{x}_{(1)kj}$ based on $n_{(1)k}$ and predictive future data as $\boldsymbol{x}^*_{(2)kj}$ based on $n_{(2)k}$, the predictive probability of trial success at the end of the first interim stage is defined as

$$P^* = \int I\left\{P(p_2 - p_1 > -\delta|\boldsymbol{x}_{(1)kj}, \boldsymbol{x}^*_{(2)kj}, k=1,2; j=1,2,...,J) > \gamma\right\} P\left(\boldsymbol{x}^*_{(2)kj}\right) d\boldsymbol{x}^*_{(2)kj} \tag{5.3.4}$$

where $I\{.\}$ is an indicator function, $P\left(\boldsymbol{x}^*_{(2)kj}\right)$ is the joint predictive probability distribution of the future samples $\boldsymbol{x}^*_{(2)kj}$, and γ can be set to as high as 0.90. P^* is essentially a weighted average of the indicator functions conditioned on all possible predictive samples. Additionally, we can set p_{c1} as the stopping boundary for the first stage such that we will reject the null hypothesis if $P^* > p_{c1}$. However, obtaining exact form of P^* is intractable and therefore, we will rely on the Gibbs sampling. If the trial fails to reject the null hypothesis based on the interim data, the trial will continue by randomizing the remaining planned $n_{(2)}$ samples and final inference will be based on the original posterior distribution using the combined data. The stopping boundary is set at p_{c2}. The additional simulation steps for this two-stage design are listed as follows.

1. Use the same simulation setting as in Table 5.3, but this time we only focus on $\lambda = 0.5$,

$n = 120$ with $n_{(1)} = 90$ and $n_{(2)} = 30$. γ is set at 0.90. Stopping probability for the first stage is set at $p_{c1} = 0.98$ and for the second stage $p_{c2} = 0.95$.

2. When simulating data for a current trial of biosimilarity, we simulate only the first stage data: $x_{1ji}(i = 1, 2, ..., n_{(1)1})$ and $x_{2ji}(i = 1, 2, ..., n_{(1)2})$.

3. Use Gibbs sampling to generate posterior samples of size N with 10% burn-in on the parameters using the conditional posterior distributions.

4. During each iteration of the Gibbs sampling in the previous step, we also generate predictive samples of size $n_{(2)}$: $x^*_{1ji}(i = 1, 2, ..., n_{(2)1})$ and $x^*_{2ji}(i = 1, 2, ..., n_{(2)2})$ for both arms using the Gibbs samples of parameters. Theoretically, these predictive samples will converge to the predictive distributions as the posterior samples of the parameters converge to their posteriors. These predictive future samples will be appended to the interim data simulated in step (2). Generate 1,000 corresponding posterior samples on the parameters now conditioning on the combined interim trial data and predicted future data.

5. Use the 1,000 posterior samples of the parameters generated in the previous step to calculate the corresponding posterior samples of p_k, and hence determine if the estimated posterior probability $P(p_2 - p_1 > -\delta | \boldsymbol{x}_{(1)kj}, \boldsymbol{x}^*_{(2)kj}, k = 1, 2; j = 1, 2, ..., J)$ is greater than γ or not. If yes, then for this Gibbs iteration, we assign an indicator value of 1, otherwise 0. The predictive probability P^* will be empirically estimated using the Gibbs sampler of size N after 10% burn-in of these indicator values. If this empirical predictive probability is greater than p_{c1}, then we stop the trial and conclude it is a success based on the composite efficacy endpoint; otherwise, we will continue to the second stage.

6. If interim analysis fails to reject the null hypothesis, additionally simulate trial data on the remaining $n_{(2)}$ data. Follow the same previous steps and obtain the estimate of posterior probability. If this posterior probability is greater than p_{c2}, then the trial

is a success, otherwise, the trial cannot claim success.

7. Based on the above steps, we can simulate 10,000 identical trials and assess the type I error, statistical power, and expected sample size based on this Bayesian two-stage design. Additionally, we will also compute the type I error and statistical power based on the original Bayesian fixed sample design and the frequentist approach.

Table 5.6 shows the result of this simulation on the Bayesian two-stage design. Generally, we can see that type I error is preserved under the level of 0.025. The statistical power using the two-stage design is comparable to the one using the fixed sample design in all scenarios. However, there is a noticeable decrease in expected trial sample size to around 105 for cases where $\delta_a = 0$, that is when the follow-on product has identical effect as the reference product, indicating possible early trial termination based on interim evidence of efficacy.

5.3.4 Sensitivity Analysis

In order to assess the robustness of the results we obtain in Tables 5.4 and 5.5, we want to conduct sensitivity analysis to further investigate if these results are sensitive to alternative specifications of the prior density. We have seen earlier that the operating characteristics of this Bayesian design can be influenced by the informative yet skeptical prior density for the bias term $\xi_j (j = 1, 2, ..., J)$ as well as the prior density specified for its variance parameter, σ_ξ^2. Equivalently, we can also view it as the precision parameter defined as $\tau_\xi^2 = 1/\sigma_\xi^2$. In general, if we assume a gamma distribution such as $G(a, b)$ as prior density for τ_ξ^2 where a is the shape parameter and b is the rate parameter, then the variance parameter will have inverse-gamma density $IG(a, b)$ as prior density where a is still the shape parameter, but b is called the scale parameter. By changing the pre-specified hyper-parameters: decreasing the shape parameter a or increasing the scale parameter b, one may adjust the influence of the skeptical prior because the strength of the skeptical prior for the bias term is parameterized by the variance parameter. If the variance parameter is distributed at higher values or

Table 5.6: Simulation result on Bayesian fixed stage, Bayesian two-stage, and frequentist operating characteristics (Type I error, statistical power, and expected sample trial size). 10,000 simulated trials are used.

							$n = 120$					
							$R = 1$			$R = 2$		
λ	δ_λ	θ_λ	$\triangle$				Bayesian fixed $p_c = 0.95$	Bayesian 2-stage $n_{(1)} = 90, n_{(2)} = 30$ $p_{c1} = 0.98, p_{c2} = 0.95$	Frequentist $\alpha/2 = 0.025$	Bayesian fixed $p_c = 0.95$	Bayesian 2-stage $n_{(1)} = 90, n_{(2)} = 30$ $p_{c1} = 0.98, p_{c2} = 0.95$	Frequentist $p_c = 0.95$
0.5	0.2302	14.0	-2			type I	0.009	0.006	0.022	0.006	0.004	0.027
0.5	0.2302	14.0	0			type I	0.007	0.004	0.023	0.003	0.004	0.023
0.5	0.2302	14.0	2			type I	0.007	0.002	0.022	0.002	0.002	0.022
λ	δ_λ	θ_λ	δ_a	θ_a	$\triangle$		Bayesian fixed $p_c = 0.95$	Bayesian 2-stage $n_{(1)} = 90, n_{(2)} = 30$ $p_{c1} = 0.98, p_{c2} = 0.95$	Frequentist $\alpha/2 = 0.025$	Bayesian fixed $p_c = 0.95$	Bayesian 2-stage $n_{(1)} = 90, n_{(2)} = 30$ $p_{c1} = 0.98, p_{c2} = 0.95$	Frequentist $p_c = 0.95$
0.5	0.2302	14.0	0.1151	6.9	-2	power	0.128	0.118	0.232	0.064	0.096	0.262
						exp. s.s.	120	119.3	120	120	119.8	120
0.5	0.2302	14.0	0.1151	6.9	0	power	0.106	0.098	0.224	0.082	0.096	0.214
						exp. s.s.	120	119.6	120	120	119.6	120
0.5	0.2302	14.0	0.1151	6.9	2	power	0.134	0.164	0.220	0.104	0.126	0.200
						exp. s.s.	120	0.119.0	120	120	119.5	120
0.5	0.2302	14.0	0.0	0.0	-2	power	0.790	0.858	0.674	0.756	0.788	0.692
						exp. s.s.	120	106.4	120	120	109.5	120
0.5	0.2302	14.0	0.0	0.0	0	power	0.844	0.852	0.726	0.798	0.796	0.668
						exp. s.s.	120	107.3	120	120	107.7	120
0.5	0.2302	14.0	0.0	0.0	2	power	0.838	0.862	0.710	0.820	0.802	0.666
						exp. s.s.	120	105.2	120	120	107.4	120

precision parameter at lower values, this will weaken the skeptical prior. If the variability or uncertainty of the variance or precision parameter is high, this may further weaken the skeptical prior.

In the design, we have specified a Jeffrey's prior for this variance parameter $P(\sigma_\xi^2) \propto 1/\sigma_\xi^2$ which is often thought of as having $a = 0$ and $b = 0$. Jeffrey's prior densities have been known to be non-informative and invariant to transformations of the parameter, but they occasionally give *improper* posterior densities. In our case, the posterior density is a proper density given by (5.2.17). Other non-informative prior densities have been suggested in the literature such as the "just proper" inverse-gamma prior density of $P(\sigma_\xi^2) = IG(0.001, 0.001)$ or a Uniform prior density $P(\sigma_\xi^2) \propto constant$ (Gelman *et al.*, 1995). When the $IG(0.001, 0.001)$ is used as prior density, the posterior density will be given by

$$\sigma_\xi^2 \sim IG\left(\frac{J}{2} + 0.001, \frac{1}{2}\sum_{j=1}^{J}(\xi_j - \theta)^2 + 0.001\right).$$

When the Uniform prior is used, it assumes a locally uniform distribution and therefore, the posterior density will be given by

$$\sigma_\xi^2 \sim IG\left(\frac{J}{2} - 1, \frac{1}{2}\sum_{j=1}^{J}(\xi_j - \theta)^2\right).$$

In addition, we also want to consider another inverse-gamma prior $IG(0.001, 1)$ which has a larger prior variability or uncertainty and is distributed more at larger values for the variance parameter than the Jeffery's prior and $IG(0.001, 0.001)$ prior do (see Figure 5.9). Figure 5.10 displays examples of posterior distributions of the variance parameter σ_ξ^2 after the likelihood of the data is combined with the four different prior densities. Jeffrey's prior and $IG(0.001, 0.001)$ prior appear to result in similar posterior distributions. $IG(0.001, 1)$ prior gives a posterior distribution that allows slightly higher values of the variance parameter than Jeffery's and $IG(0.001, 0.001)$, and finally Uniform prior results in much more variable and higher values in the posterior distribution.

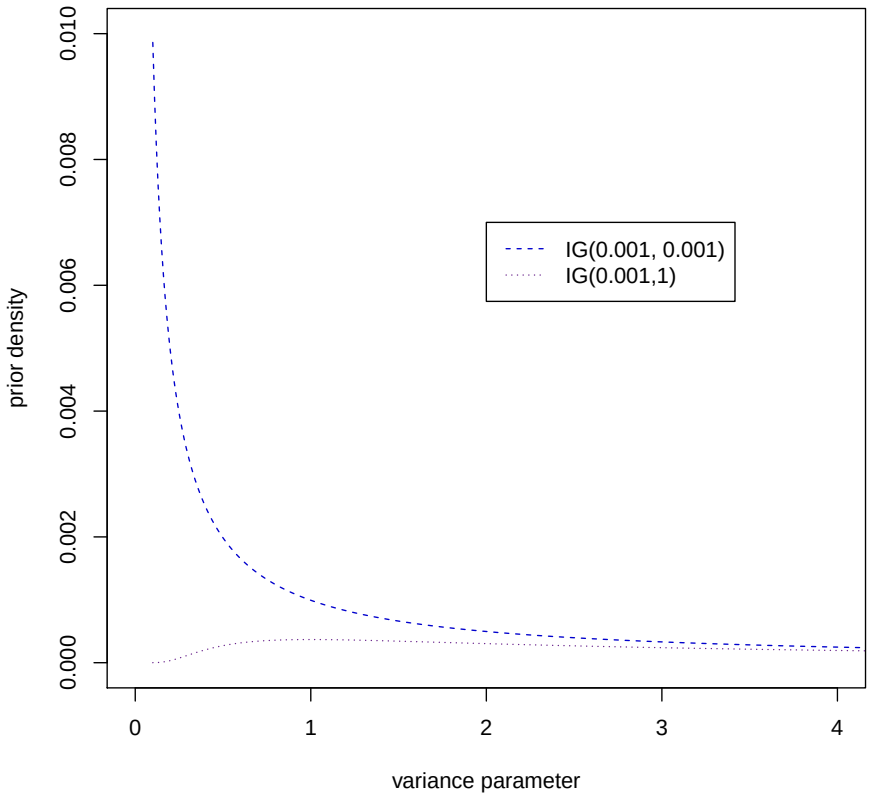


Figure 5.9: Prior densities for the variance parameter σ_ξ^2 for two IG.

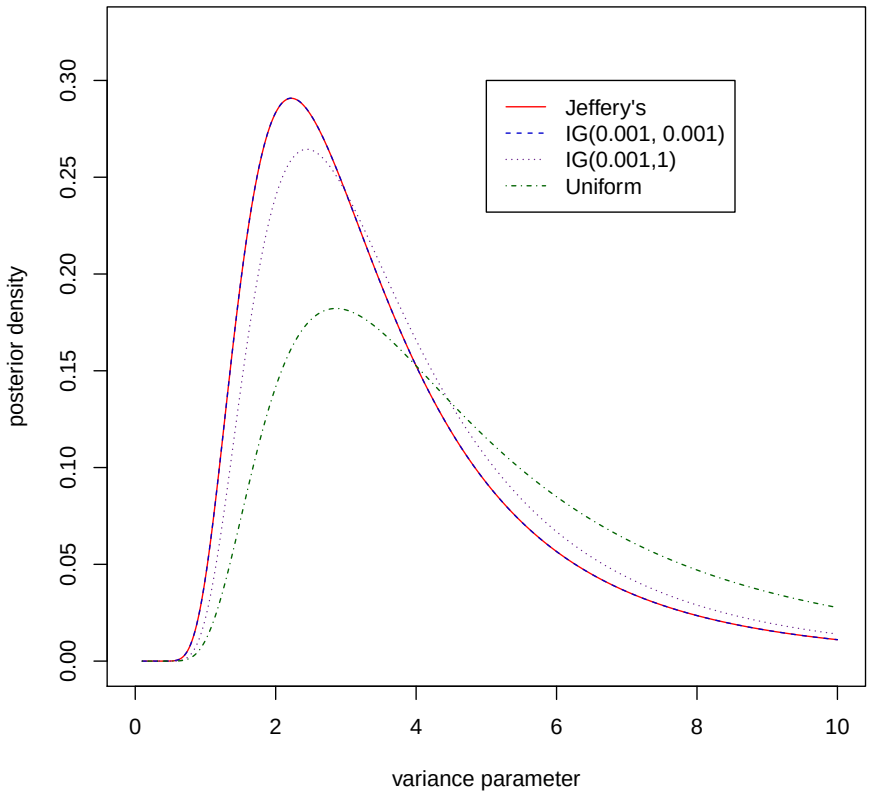


Figure 5.10: Posterior densities for the variance parameter σ_ξ^2 under different prior density specifications.

Table 5.7: Simulated Bayesian type I error rate on different prior density specifications using 10,000 simulated trials.

				$p_c = 0.95$				$p_c = 0.975$				
λ	δ_λ	θ_λ	$\triangle$	Jeffery's	IG(0.001, 0.001)	IG(0.001, 1)	Uniform	Jeffery's	IG(0.001, 0.001)	IG(0.001, 1)	Uniform	Frequentist
0	0.4604	34.7	-2	0.159	0.155	0.139	0.133	0.086	0.085	0.070	0.070	0.041
0	0.4604	34.7	0	0.020	0.024	0.022	0.034	0.010	0.009	0.009	0.014	0.020
0	0.4604	34.7	2	0.001	0.002	0.002	0.007	0.001	0.001	0.001	0.003	0.009
0.25	0.3453	22.2	-2	0.051	0.040	0.027	0.033	0.018	0.015	0.008	0.018	0.027
0.25	0.3453	22.2	0	0.010	0.008	0.007	0.015	0.003	0.002	0.002	0.007	0.019
0.25	0.3453	22.2	2	0.003	0.001	0.002	0.009	0.001	0.001	0.001	0.003	0.016
0.5	0.2302	14.0	-2	0.009	0.006	0.003	0.009	0.003	0.001	0.001	0.004	0.024
0.5	0.2302	14.0	0	0.007	0.003	0.003	0.009	0.002	0.001	0.001	0.002	0.023
0.5	0.2302	14.0	2	0.007	0.004	0.003	0.008	0.002	0.001	0.001	0.003	0.024

Table 5.8: Simulated Bayesian power on different prior density specifications using 10,000 simulated trials.

						$p_c = 0.95$				$p_c = 0.975$				
λ	δ_λ	θ_λ	δ_a	θ_a	$\triangle$	Jeffery's	IG(0.001, 0.001)	IG(0.001, 1)	Uniform	Jeffery's	IG(0.001, 0.001)	IG(0.001, 1)	Uniform	Frequentist
0	0.4604	34.7	0.2302	14.0	-2	0.998	0.993	0.995	0.998	0.980	0.989	0.987	0.997	0.744
0	0.4604	34.7	0.2302	14.0	0	0.984	0.991	0.993	0.997	0.984	0.985	0.986	0.995	0.726
0	0.4604	34.7	0.2302	14.0	2	0.992	0.990	0.992	0.996	0.982	0.981	0.980	0.993	0.748
0	0.4604	34.7	0	0	-2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0	0.4604	34.7	0	0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998
0	0.4604	34.7	0	0	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.25	0.3453	22.2	0.1726	10.4	-2	0.534	0.574	0.691	0.869	0.430	0.491	0.573	0.787	0.522
0.25	0.3453	22.2	0.1726	10.4	0	0.528	0.560	0.665	0.863	0.430	0.475	0.558	0.770	0.434
0.25	0.3453	22.2	0.1726	10.4	2	0.558	0.566	0.665	0.853	0.480	0.482	0.551	0.762	0.434
0.25	0.3453	22.2	0	0	-2	1.000	0.999	1.000	1.000	1.000	0.999	0.999	1.000	0.974
0.25	0.3453	22.2	0	0	0	1.000	0.999	0.999	1.000	1.000	0.999	0.999	1.000	0.956
0.25	0.3453	22.2	0	0	2	1.000	0.999	1.000	1.000	1.000	0.999	0.999	1.000	0.960
0.5	0.2302	14.0	0.1151	6.9	-2	0.128	0.106	0.169	0.364	0.064	0.063	0.100	0.245	0.232
0.5	0.2302	14.0	0.1151	6.9	0	0.106	0.111	0.180	0.371	0.072	0.063	0.108	0.244	0.224
0.5	0.2302	14.0	0.1151	6.9	2	0.134	0.128	0.207	0.397	0.070	0.074	0.118	0.262	0.220
0.5	0.2302	14.0	0	0	-2	0.790	0.821	0.892	0.966	0.756	0.749	0.817	0.928	0.674
0.5	0.2302	14.0	0	0	0	0.844	0.836	0.913	0.969	0.802	0.774	0.830	0.930	0.726
0.5	0.2302	14.0	0	0	2	0.838	0.855	0.913	0.968	0.800	0.786	0.843	0.937	0.710

good control of type I error rate at 0.025 in all scenarios where $\triangle \geqslant 0$.

Table 5.8 tabulates the simulated Bayesian powers based on different prior densities. Again, Jeffery's prior and $IG(0.001, 0.001)$ perform similarly. For prior of $IG(0.001, 1)$, there is an increase in statistical power comparing to Jeffery's and $IG(0.001, 0.001)$ priors. However, for scenarios where $\lambda = 0.5$, $\delta_\lambda = 0.2302$, and alternative hypothesis at $\delta_a = 0.1151$, all three priors, Jeffery's, $IG(0.001, 0.001)$, and $IG(0.001, 1)$ cannot surpass the frequentist power except for Uniform prior. The design using the Uniform prior generally has the greatest statistical power comparing to other priors as well as the frequentist approach in all the scenarios under study. Given this result, we see that the operating characteristics display some degree of sensitivity to the hyper-parameters of the inverse-gamma prior density of σ_ξ^2. In fact, given a very small value of the shape hyper-parameter such as $a = 0.001$, one can adjust the value of the scale hyper-parameter b to a suitable value while preserving the type I error at 0.025. Increasing the value of b, has the effect of increasing the posterior mean or median of σ_ξ^2 as well as its variability, thus weakening the skeptical prior on ξ_j in the direction to favor the alternative hypothesis. The control of type I error using the Bayesian approach with different prior specifications is still better than the frequentist approach, particularly when $\triangle = -2$, that is, when the reference product is performing worse in the current trial than in the historical trial. Therefore, it is strongly recommended to conduct simulation to assess the influence of different candidate prior densities when deciding which one to use with regulatory agencies.

5.4 Summary and Discussion

In this chapter, we have presented a Bayesian method to assess biosimilarity between a licensed reference biological product and a generic follow-on (also known as a subsequent-entry) biological product. This approach adopts a non-inferiority testing framework that connects the current trial of biosimilarity to historical trials of the reference product. The

proposed Bayesian analytical approach recognizes that the reference product was approved for license in the past and that information in these historical trials can be meaningfully incorporated in the analysis of the current trial. However, due to changing clinical practices and improvement in the overall delivery of care over time, the effect of a medicinal product may not be always constant. This is, in the context of a non-inferiority clinical trial, sometimes known as the constancy assumption, the historical difference between the original product and placebo is assumed to hold in the current setting of the new trial if a placebo is in place (D'Agostino, Massaro, and Sullivan, 2003). Therefore, we presented the hierarchical model to incorporate historical trials while accounting for the potential lack of biosimilarity via a bias parameter. In this model, non-informative priors are elicited for most parameters except for the bias parameter which assumes a skeptical prior with expectation centered on the null hypothesis. We also characterized the sensitivity of this design to different prior specifications for the variance parameter for the bias term. As most biological products are meant to treat illnesses with improvement in multiple endpoints, we illustrate the application of this method to studying rheumatoid arthritis that uses a composite efficacy endpoint known as ACR20.

Simulation studies have demonstrated that the Bayesian method usually has type I error preserved under the α-level of 0.025, comparable to a typical level assumed in a one-sided non-inferiority trial. This is made possible with the placement of the skeptical prior on the bias parameter, even when a more relaxed critical probability $p_c = 0.95$ is used under different prior specifications. When the reference product performs worse in the current trial, due to potential violation of the constancy assumption, the NI margin that is based on its historical trial appears to be wider, thus inflating its type I error. Both Bayesian and frequentist methods have no immunity to this inflation, however, the Bayesian method is able to cancel out this inflation by tapping into the influence of the skeptical null prior as NI margin narrows, therefore offering some protection even when constancy assumption is slightly violated in the negative direction. It is important to emphasize that this type I error is an error rate conditional on the outcomes of the historical trial selected. Under

this hierarchical model, we presume in (5.2.5) that both μ_{1j} and μ_{1hj} come from the same underlying distribution, therefore the difference $\triangle = \mu_{1j} - \mu_{1hj}$ follows the normal distribution, $N(0, 2\sigma_{1b}^2)$. Another way to look at the type I error is the average type I error rate over all possible values of $\triangle$. Further simulation can be useful in characterizing this average type I error over all possible trial performance for the reference product in historical and current trials. It is important that, prior to the design of the biosimilarity trial, a thorough literature search should be made to assess if the effect of the reference product is consistent in the historical trials and if the design and conduct of these studies are not too dissimilar. If such large variability in estimation is observed, sources of this inconsistency should be investigated.

As for statistical power, it somewhat suffers when NI margin is small. However, as sample size increases from $n = 60$ to 120 under smaller margins and as the follow-on product is truly biosimilar to the reference product, Bayesian statistical power starts to outperform the frequentist approach. Adopting an equal randomization ratio of $R = 1$ offers only a slight advantage in the overall statistical power. In the Bayesian two-stage adaptive design using predictive probability as an interim stopping criterion, reduced expected sample size is observed especially in cases when a follow-on product is biosimilar to the reference product without compromising its statistical power. In our example, as much as a 12.5% reduction in expected sample size is observed.

Another possibility of using hierarchical modeling is that we may be able to include other historical trials which perhaps studied different doses of the reference product or were conducted under systematically different trial-specific circumstances. If such characteristics can be assumed to be linearly related to the efficacy parameters, their inclusion into the model may help increase the precision of the estimation, and hence the inference.

In this chapter, we have illustrated the method using a composite endpoint that has several separate endpoints combined into a single one. When we directly model the component endpoints, it is likely that instead of the global null hypothesis, some of the component

endpoints may have inferior means such that for some j, $\mu_{2j} \leqslant \mu_{1j} - \theta$ but not the others, and this trial can still claim success based on the predictive or posterior probability. Composite endpoint may present different null configurations which may warrant further study. In our example, we have only presented the global null configuration using θ as the non-inferiority margin across all component endpoints. In other cases, a single endpoint or multiple endpoints are used to establish efficacy. For example, for studying psoriasis, a common chronic inflammatory skin disease characterized by thick red flaky patches called scales, there are two major endpoints: proportion of subjects who achieved at least 75% reduction in PASI score (PASI75) and treatment success on the Physician's Global Assessment (PGA). This Bayesian hierarchical bias approach can still be similarly applied and final inference may be based on the joint posterior probabilities that these endpoints are greater than their respective non-inferiority margins.

5.5 Appendix

5.5.1 Conditional Posterior Distribution of μ_{1j}

In this section, we want to show how to obtain the conditional posterior distribution for μ_{1j} as in (5.2.10). A similar algebraic algorithm can be employed to obtain the other conditional posterior densities for normal mean parameters.

Proof. Based on the complete likelihood function in (5.2.8), the parameter μ_{1j} depends on

the following components

$$\begin{aligned}
\mu_{1j}|\boldsymbol{x}_{1j},\boldsymbol{x}_{2j} &\propto L(\mu_{1j},\sigma_1^2|\boldsymbol{x}_{1j})L(\mu_{2j}=\mu_{1j}-\xi_j,\sigma_2^2|\boldsymbol{x}_{2j}P(\mu_{1j}|\mu_{1j}^o,\sigma_{1b}^2) \\
&\propto exp\left\{-\frac{1}{2}\left(\frac{\sum_{i=1}^{n_1}(\mu_{1j}-x_{1ji})^2}{\sigma_1^2}+\frac{\sum_{i=1}^{n_2}(\mu_{1j}-(\xi_j+x_{2ji}))^2}{\sigma_2^2}+\frac{(\mu_{1j}-\mu_{1j}^o)^2}{\sigma_{1b}^2}\right)\right\} \\
&\propto exp\left\{-\frac{1}{2}\left(\frac{n_1\mu_{1j}^2-2\mu_{1j}\sum_{i=1}^{n_1}x_{1ji}}{\sigma_1^2}+\frac{n_2\mu_{1j}^2-2\mu_{1j}(n_2\xi_j+\sum_{i=1}^{n_2}x_{2ji})}{\sigma_2^2}+\frac{\mu_{1j}^2-2\mu_{1j}\mu_{1j}^o}{\sigma_{1b}^2}\right)\right\} \\
&\propto exp\left\{-\frac{1}{2}\left(\frac{n_1}{\sigma_1^2}+\frac{n_2}{\sigma_2^2}+\frac{1}{\sigma_{1b}^2}\right)\left(\mu_{1j}^2-2\frac{\left(\frac{n_1\overline{x}_{1j}}{\sigma_1^2}+\frac{n_2(\overline{x}_{2j}+\xi_j)}{\sigma_2^2}+\frac{\mu_{1j}^o}{\sigma_{1b}^2}\right)}{\left(\frac{n_1}{\sigma_1^2}+\frac{n_2}{\sigma_2^2}+\frac{1}{\sigma_{1b}^2}\right)}\mu_{1j}\right)\right\} \\
&\propto exp\left\{-\frac{1}{2}\left(\frac{n_1}{\sigma_1^2}+\frac{n_2}{\sigma_2^2}+\frac{1}{\sigma_{1b}^2}\right)\left(\mu_{1j}-\frac{\left(\frac{n_1\overline{x}_{1j}}{\sigma_1^2}+\frac{n_2(\overline{x}_{2j}+\xi_j)}{\sigma_2^2}+\frac{\mu_{1j}^o}{\sigma_{1b}^2}\right)}{\left(\frac{n_1}{\sigma_1^2}+\frac{n_2}{\sigma_2^2}+\frac{1}{\sigma_{1b}^2}\right)}\right)^2\right\}.
\end{aligned}$$

We can see that it belongs to a normal distribution with mean and variance given by (5.2.10). □

5.5.2 Conditional Posterior Distribution of σ_1^2

In this section, we want to show how to obtain the conditional posterior distribution for σ_1^2 as in (5.2.14). A similar algebraic algorithm can be employed to obtain the posterior densities for other variance parameters.

Proof. Based on the complete likelihood function in (5.2.8), the parameter σ_1^2 depends on the following components

$$\begin{aligned}
\sigma_1^2|\boldsymbol{x}_{1j},\boldsymbol{x}_{2j} &\propto \prod_{j=1}^{J}\left(L(\mu_{1j},\sigma_1^2|\boldsymbol{x}_{1j})\left(\prod_{h=1}^{H}L(\mu_{1hj},\sigma_1^2|\boldsymbol{x}_{1hj})\right)\right)\frac{1}{\sigma_1^2} \\
&\propto \frac{1}{(\sigma_1^2)^{\frac{J\left(n_1+\sum_{h=1}^{H}n_{1h}\right)}{2}+1}}exp\left\{-\frac{1}{2}\sum_{j=1}^{J}\left(\frac{\sum_{i=1}^{n_1}(\mu_{1j}-x_{1ji})^2}{\sigma_1^2}+\sum_{h=1}^{H}\left(\frac{\sum_{i=1}^{n_{1h}}(\mu_{1hj}-x_{1hji})^2}{\sigma_1^2}\right)\right)\right\} \\
&\propto (\sigma_1^2)^{-\frac{J\left(n_1+\sum_{h=1}^{H}n_{1h}\right)}{2}-1}exp\left\{-\frac{\frac{1}{2}\sum_{j=1}^{J}\left(\sum_{i=1}^{n_1}(x_{1ji}-\mu_{1j})^2+\sum_{h=1}^{H}\sum_{i=1}^{n_{1h}}(x_{1hji}-\mu_{1hj})^2\right)}{\sigma_1^2}\right\}.
\end{aligned}$$

It can be seen that this posterior distribution takes on the form of an inverse-gamma dis-

tribution. Simulation from this inverse-gamma distribution can be performed by simulating from its corresponding gamma distribution with shape parameter given by $\frac{J\left(n_1+\sum_{h=1}^{H} n_{1h}\right)}{2}$ and rate parameter given by $\frac{1}{2}\sum_{j=1}^{J}\left(\sum_{i=1}^{n_1}(x_{1ji}-\mu_{1j})^2+\sum_{h=1}^{H}\sum_{i=1}^{n_{1h}}(x_{1hji}-\mu_{1hj})^2\right)$ and then taking reciprocal of the simulated value. □

5.6 R Codes

5.6.1 Function Codes

```
myvariance <- function(x){
  sum((x-.Internal(mean(x)))^2)/(length(x)-1)
}

plotconvgmatrix <- function(gibbssample, range, outfile) {
  pdf(outfile)
  par(mfrow=c(4,2), mex=0.7)
  for (i in 1:ncol(gibbssample)) {
    if (i < 8) { plot(range, gibbssample[range, i], type="l", col="red", ylab="value",
    main=paste("Traceplot for mean of ", i, "th endpoint", sep=""))
    } else if (i==8) { plot(range, gibbssample[range, i], type="l", col="red", ylab="value",
    main=paste("Traceplot for variance")) }}
  dev.off()
}

plotkernelmatrix <- function(gibbssample, range, outfile) {
  pdf(outfile)
  par(mfrow=c(4,2), mex=0.7)
  for (i in 1:ncol(gibbssample)) {
    if (i < 8) { plot(density(gibbssample[range, i]), col="blue", xlab="value",
    main=paste("Kernel plot for mean of ", i, "th endpoint", sep=""))
    } else if (i==8) { plot(density(gibbssample[range, i]), col="blue", xlab="value",
    main=paste("Kernel plot for variance")) }}
  dev.off()
}

gibbs <- function(xdata, nsim, outfile) {

  inits <- c(rnorm(7, 0, 1000), runif(1, 0, 100000))
  nn <- ncol(xdata)
  writeinits <- inits
  dim(writeinits) <- c(1,8)

  file.con <- file(outfile,"w")
  write.table(writeinits, file.con, append=T, row.names=F, col.names=F, quote=F)
  lastsim <- inits

  for (i in 1:nsim) {
    newsim <- rnorm(7, rowMeans(xdata), sd=sqrt(lastsim[8]/nn))
    newsim[8] <- 1/rgamma(1, shape=(7*nn/2), rate=((1/2)
        *sum(apply((xdata - lastsim[1:7])^2, 1, sum))))
    writenewsim <- newsim
    dim(writenewsim) <- c(1, 8)
```

```
    write.table(writenewsim, file.con, append=T, row.names=F, col.names=F, quote=F)
    lastsim <- newsim
  }
  close(file.con)
}

calculatep <- function(meanandvar) {
  probvecge <- pnorm(as.numeric(-(20-meanandvar[1:7])/(sqrt(meanandvar[8]))))
  probvecst <- pnorm(as.numeric((20-meanandvar[1:7])/(sqrt(meanandvar[8]))))

  pS3 <- prod(probvecge[c(3,4,5)])*prod(probvecst[c(6,7)])
         + prod(probvecge[c(3,4,6)])*prod(probvecst[c(5,7)])
         + prod(probvecge[c(3,4,7)])*prod(probvecst[c(5,6)])
         + prod(probvecge[c(3,5,6)])*prod(probvecst[c(4,7)])
         + prod(probvecge[c(3,5,7)])*prod(probvecst[c(4,6)])
         + prod(probvecge[c(4,5,6)])*prod(probvecst[c(3,7)])
         + prod(probvecge[c(4,5,7)])*prod(probvecst[c(3,6)])
         + prod(probvecge[c(5,6,7)])*prod(probvecst[c(3,4)])
         + prod(probvecge[c(3,6,7)])*prod(probvecst[c(4,5)])
         + prod(probvecge[c(4,6,7)])*prod(probvecst[c(3,5)])

  pS4 <- prod(probvecge[c(4,5,6,7)])*prod(probvecst[3])
            + prod(probvecge[c(3,5,6,7)])*prod(probvecst[4])
            + prod(probvecge[c(3,4,6,7)])*prod(probvecst[5])
            + prod(probvecge[c(3,4,5,7)])*prod(probvecst[6])
            + prod(probvecge[c(3,4,5,6)])*prod(probvecst[7])

  ptotal <- probvecge[1]*probvecge[2]*(pS3 + pS4 + prod(probvecge[3:7]))
  return(ptotal)
}

findtheta <- function(delta, meanvec, vari) {

  result <- list()
  kk <- length(delta)
  theta <- vector(mode="numeric",length=kk)
  thetaerror <- vector(mode="numeric",length=kk)

  theta_lambda <- seq(1, 60, by=0.1)
  nn <- length(theta_lambda)
  delta_lambda <- vector(mode="numeric",length=nn)

  for (k in 1:nn) {
    delta_lambda[k] <-  calculatep(meanandvar=c(meanvec, vari)) -
    calculatep(meanandvar=c(meanvec-theta_lambda[k], vari))
  }
  for (i in 1:kk) {
    error <- abs(delta_lambda-delta[i])
    theta[i] <- theta_lambda[which(error == min(error))]
    thetaerror[i] <- min(error)
  }

  result$theta <- theta
  result$thetaerror <- thetaerror
  result$theta_lambda <- theta_lambda
  result$delta_lambda <- delta_lambda
  return(result)
}

acr20 <- function(vec) {
  return(ifelse((vec[1]>20 & vec[2]>20 & ((vec[3]>20)+(vec[4]>20)+(vec[5]>20)
  +(vec[6]>20)+(vec[7]>20))>=3), 1, 0))
}
```

```
getpk <- function(xdata) {
  txdata <- t(xdata)
  return(mean(apply(txdata, 1, acr20)))
}

# for H=1 only
onetrialhier <- function(x1h, n, R, pc, triangle, lambdaf, deltaf, thetaf, p0h, p1h, n0h,
meanvec1h, thetafactual, var1, var2, nsim, plotconvg) {

  result <- list()
  result$true.x1h.means <- meanvec1h
  meanfor1 <- meanvec1h + triangle
  result$true.x1.means <- meanfor1
  meanfor2 <- meanfor1 - thetafactual
  result$true.x2.means <- meanfor2
  result$delta <- deltaf
  result$thetas <- c(thetaf, thetafactual)

  n1 <- n/(1+R)
  n2 <- n*R/(1+R)
  n1h <- ncol(x1h)
  range <- (ceiling(0.1*nsim)):nsim
  result$no.iterations.used <- length(range)

  x1 <- matrix(rnorm(7*n1, meanfor1, sqrt(var1)), c(7,n1), byrow=F)
  result$observed.x1.means <- rowMeans(x1)
  freqp1 <- getpk(x1)
  result$observed.freq.p1 <- freqp1
  x2 <- matrix(rnorm(7*n2, meanfor2, sqrt(var2)), c(7,n2), byrow=F)
  result$observed.x2.means <- rowMeans(x2)
  freqp2 <- getpk(x2)
  result$observed.freq.p2 <- freqp2

  mu1ogibbs <- matrix(data=NA, nrow=nsim+1, ncol=7)
  mu1hgibbs <- matrix(data=NA, nrow=nsim+1, ncol=7)
  mu1gibbs <-  matrix(data=NA, nrow=nsim+1, ncol=7)
  biasgibbs <- matrix(data=NA, nrow=nsim+1, ncol=7)
  vargibbs <-  matrix(data=NA, nrow=nsim+1, ncol=4)

  lastmu1o <- rnorm(7, 0, 1000)
  lastmu1h <- rnorm(7, 0, 1000)
  lastmu1 <-  rnorm(7, 0, 1000)
  lastbias <- rnorm(7, 0, 1000)
  lastvar <-  runif(4, 0, 100000)

  mu1ogibbs[1,] <- lastmu1o
  mu1hgibbs[1,] <- lastmu1h
  mu1gibbs[1,] <- lastmu1
  biasgibbs[1,] <- lastbias
  vargibbs[1,] <- lastvar

  x1hmean <- rowMeans(x1h)
  x1mean  <- rowMeans(x1)
  x2mean  <- rowMeans(x2)

  for (i in 1:nsim) {

    newmu1o <- rnorm(7, mean=(lastmu1+lastmu1h)/2, sd=sqrt(lastvar[3]/2))
    newmu1h <- rnorm(7, mean=(1/((n1h/lastvar[1])+(1/lastvar[3])))
                        *((n1h*x1hmean/lastvar[1])+(lastmu1o/lastvar[3])),
                        sd=sqrt(1/((n1h/lastvar[1])+(1/lastvar[3]))))
    newmu1  <- rnorm(7, mean=(1/((n1/lastvar[1])+(n2/lastvar[2])+(1/lastvar[3])))
                        *((n1*x1mean/lastvar[1])+((n2*(x2mean+lastbias))/lastvar[2])
                        +(lastmu1o/lastvar[3])),
```

```
                    sd=sqrt(1/((n1/lastvar[1])+(n2/lastvar[2])+(1/lastvar[3]))))
  newbias  <- rnorm(7, mean=(1/((n2/lastvar[2])+(1/lastvar[4])))
                    *(((n2*(lastmu1-x2mean))/lastvar[2])+(thetaf/lastvar[4])),
                    sd=sqrt(1/((n2/lastvar[2])+(1/lastvar[4]))))

  shapes <- c(7*(n1+n1h)/2, 7*n2/2, 7*2/2, 7/2)
  rates <-  c(0.5*(sum(rowSums((x1h-lastmu1h)^2))+sum(rowSums((x1-lastmu1)^2))),
              0.5*sum(rowSums((x2-(lastmu1-lastbias))^2)),
              0.5*(sum((lastmu1-lastmu1o)^2)+sum((lastmu1h-lastmu1o)^2)),
              0.5*sum((lastbias-thetaf)^2))
  newvar <- 1/rgamma(4, shape=shapes, rate=rates)

  mu1ogibbs[i+1,] <- newmu1o
  mu1hgibbs[i+1,] <- newmu1h
  mu1gibbs[i+1,] <- newmu1
  biasgibbs[i+1,] <- newbias
  vargibbs[i+1,] <- newvar

  lastmu1o <- newmu1o
  lastmu1h <- newmu1h
  lastmu1 <- newmu1
  lastbias <- newbias
  lastvar <- newvar
}
mu2gibbs   <- mu1gibbs - biasgibbs
forp1gibbs <- cbind(mu1gibbs, vargibbs[,1])
forp2gibbs <- cbind(mu2gibbs, vargibbs[,2])

result$gibbs.mu1o.means <- colMeans(mu1ogibbs[range,])
result$gibbs.var.1o <- apply(mu1ogibbs[range,], 2, myvariance)
result$gibbs.mu1.means <- colMeans(mu1gibbs[range,])
result$gibbs.mu2.means <- colMeans(mu2gibbs[range,])
result$gibbs.bias.means <- colMeans(biasgibbs[range,])
result$gibbs.var1.var2.var1b.varbias.means <- colMeans(vargibbs[range,])

# optional assessment of convergence for mu2gibbs, j=1
if (plotconvg==1) {
  plot(range, mu2gibbs[range, 1], type="l")
  plot(density(mu2gibbs[range, 1], na.rm=T)) }

p1 <- apply(forp1gibbs, 1, calculatep)
p2 <- apply(forp2gibbs, 1, calculatep)
Ip2.p1 <- ifelse(p2-p1 > -deltaf, 1, 0)
p1 <- p1[range]
p2 <- p2[range]
result$bay.p1.mean <- .Internal(mean(p1))
result$bay.p2.mean <- .Internal(mean(p2))
result$bay.p2.p1.diff.mean <- .Internal(mean(p2-p1))

obspc <- .Internal(mean(Ip2.p1[range]))
result$obs.prob.decision <- obspc
if (is.na(obspc)==0 & obspc > pc) {baysuccess <- 1
} else if (is.na(obspc)==0 & obspc <= pc) {baysuccess <- 0
} else if (is.na(obspc)==1) {baysuccess <- NA}
result$bay.success <- baysuccess

# only explore alpha/2 = 0.025 for one-sided test
if ((freqp2-freqp1-qnorm(0.975)*sqrt((freqp2*(1-freqp2)/n2)+(freqp1*(1-freqp1)/n1))) >
   -(1-lambdaf)*(p1h-p0h-qnorm(0.975)*sqrt((p1h*(1-p1h)/n1h)+(p0h*(1-p0h)/n0h)))) {freqsuccess <- 1
   } else {freqsuccess <- 0}
result$freq.success <- freqsuccess

return(result)
}
```

```
probrejectnullhier <- function(x1h, n, R, pc, triangle, lambdaf, deltaf, thetaf, p0h, p1h, n0h,
                               meanvec1h, thetafactual, var1, var2, nsim, plotconvg, nsimtrial) {
  result <- list()
  obspc <- rep(NA, nsimtrial)
  baytype1 <- rep(NA, nsimtrial)
  freqtype1 <- rep(NA, nsimtrial)
  for (k in 1:nsimtrial) {
    onetrial <- onetrialhier(x1h=x1h, n=n, R=R, pc=pc, triangle=triangle,
                           lambdaf=lambdaf, deltaf=deltaf, thetaf=thetaf,
                           p0h=p0h, p1h=p1h, n0h=n0h,
                           meanvec1h=meanvec1h, thetafactual=thetafactual,
                           var1=var1, var2=var2, nsim=nsim, plotconvg=0)
    obspc[k] <- onetrial$obs.prob.decision
    baytype1[k] <- onetrial$bay.success
    freqtype1[k] <- onetrial$freq.success
  }
  result$obspc <- obspc
  result$baytype1 <- baytype1
  result$freqtype1 <- freqtype1
  result$baysuccesssum <- sum(baytype1, na.rm=T)
  result$baytype1error <- mean(baytype1, na.rm=T)
  result$freqtype1error <- mean(freqtype1, na.rm=T)
  return(result)
}

# for H=1 only
onetrialhier2stage <- function(x1h, ns1, ns2, R, pc1, pc2, gammaf=0.9, triangle, lambdaf, deltaf,
thetaf, p0h, p1h, n0h, meanvec1h, thetafactual, var1, var2, nsim, plotconvg) {

  result <- list()
  result$true.x1h.means <- meanvec1h
  meanfor1 <- meanvec1h + triangle
  result$true.x1.means <- meanfor1
  meanfor2 <- meanfor1 - thetafactual
  result$true.x2.means <- meanfor2
  result$delta <- deltaf
  result$thetas <- c(thetaf, thetafactual)

  ns11 <- ns1/(1+R)
  ns12 <- ns1*R/(1+R)
  ns21 <- ns2/(1+R)
  ns22 <- ns2*R/(1+R)
  n1h <- ncol(x1h)
  range <- (ceiling(0.1*nsim)):nsim
  result$no.iterations.used <- length(range)

  x1 <- matrix(rnorm(7*(ns11+ns21), meanfor1, sqrt(var1)), c(7, (ns11+ns21)), byrow=F)
  result$observed.x1.means <- rowMeans(x1)
  freqp1 <- getpk(x1)
  result$observed.freq.p1 <- freqp1
  result$obs.freq.p1.stg1.stg2 <- c(getpk(x1[,1:ns11]), freqp1)

  x2 <- matrix(rnorm(7*(ns12+ns22), meanfor2, sqrt(var2)), c(7, (ns12+ns22)), byrow=F)
  result$observed.x2.means <- rowMeans(x2)
  freqp2 <- getpk(x2)
  result$observed.freq.p2 <- freqp2
  result$obs.freq.p2.stg1.stg2 <- c(getpk(x2[,1:ns12]), freqp2)

  lastmu1o <- rnorm(7, 0, 1000)
  lastmu1h <- rnorm(7, 0, 1000)
  lastmu1 <- rnorm(7, 0, 1000)
  lastbias <- rnorm(7, 0, 1000)
  lastvar <- runif(4, 0, 1000)
```

```
YNpred <- rep(NA, nsim)
x1hmean <- rowMeans(x1h)
x1mean  <- rowMeans(x1[,1:ns11])
x2mean  <- rowMeans(x2[,1:ns12])

for (i in 1:nsim) {

  newmu1o <- rnorm(7, mean=(lastmu1+lastmu1h)/2, sd=sqrt(lastvar[3]/2))
  newmu1h <- rnorm(7, mean=(1/((n1h/lastvar[1])+(1/lastvar[3])))
                    *((n1h*x1hmean/lastvar[1])+(lastmu1o/lastvar[3])),
                    sd=sqrt(1/((n1h/lastvar[1])+(1/lastvar[3]))))
  newmu1 <-  rnorm(7, mean=(1/((ns11/lastvar[1])+(ns12/lastvar[2])+(1/lastvar[3])))
                   *((ns11*x1mean/lastvar[1])+((ns12*(x2mean+lastbias))/lastvar[2])
                   +(lastmu1o/lastvar[3])),
                   sd=sqrt(1/((ns11/lastvar[1])+(ns12/lastvar[2])+(1/lastvar[3]))))
  newbias <- rnorm(7, mean=(1/((ns12/lastvar[2])+(1/lastvar[4])))
                  *(((ns12*(lastmu1-x2mean))/lastvar[2])+(thetaf/lastvar[4])),
                  sd=sqrt(1/((ns12/lastvar[2])+(1/lastvar[4]))))
  newmu2 <- newmu1 - newbias
  shapes <- c(7*(ns11+n1h)/2, 7*ns12/2, 7*2/2, 7/2)
  rates <- c( 0.5*(sum(rowSums((x1h-lastmu1h)^2))+sum(rowSums((x1[,1:ns11]-lastmu1)^2))),
                0.5*sum(rowSums((x2[,1:ns12]-(lastmu1-lastbias))^2)),
                0.5*(sum((lastmu1-lastmu1o)^2)+sum((lastmu1h-lastmu1o)^2)),
                0.5*sum((lastbias-thetaf)^2))
  newvar <- 1/rgamma(4, shape=shapes, rate=rates)

  newmu1vec <- sapply(newmu1, function(x) rep(x, ns21))
  x1predict <- matrix(rnorm(7*ns21, newmu1vec, sqrt(newvar[1])), c(7, ns21), byrow=T)
  x1comb <- cbind(x1[,1:ns11], x1predict)

  newmu2vec <- sapply(newmu2, function(x) rep(x, ns22))
  x2predict <- matrix(rnorm(7*ns22, newmu2vec, sqrt(newvar[2])), c(7, ns22), byrow=T)
  x2comb <- cbind(x2[,1:ns12], x2predict)

  pnewmu1vec <- (1/(((ns11+ns21)/newvar[1])+((ns12+ns22)/newvar[2])+(1/newvar[3])))
                        *(((ns11+ns21)*rowMeans(x1comb)/newvar[1])+(((ns12+ns22)
                        *(rowMeans(x2comb)+newbias))/newvar[2])+(newmu1o/newvar[3]))
  pnewmu1 <- t(matrix(rnorm(7*1000, pnewmu1vec, sqrt(1/(((ns11+ns21)/newvar[1])
                  +((ns12+ns22)/newvar[2])+(1/newvar[3])))), c(7, 1000), byrow=F))
  pnewbiasvec <- (1/(((ns12+ns22)/newvar[2])+(1/newvar[4])))
                       *((((ns12+ns22)*(newmu1-rowMeans(x2comb)))/newvar[2])+(thetaf/newvar[4]))
  pnewbias <- t(matrix(rnorm(7*1000, pnewbiasvec, sqrt(1/(((ns12+ns22)/newvar[2])
                  +(1/newvar[4])))), c(7, 1000), byrow=F))
  pnewmu2 <- pnewmu1 - pnewbias
  pnewvar <- matrix(rep(NA, 2*1000), nrow=1000)
  pnewvar[,1] <- 1/rgamma(1000, shape=(7*(ns11+ns21+n1h)/2),
                        rate=(1/2)*(sum(rowSums((x1h-newmu1h)^2))+ sum(rowSums((x1comb-newmu1)^2))))
  pnewvar[,2] <- 1/rgamma(1000, shape=(7*(ns12+ns22)/2),
                        rate=(1/2)*(sum(rowSums((x2comb-(newmu1-newbias))^2))))

  lastmu1o <- newmu1o
  lastmu1h <- newmu1h
  lastmu1 <- newmu1
  lastbias <- newbias
  lastvar <- newvar
  lastmu2 <- newmu2

  predp1 <- apply(cbind(pnewmu1, pnewvar[,1]), 1, calculatep)
  predp2 <- apply(cbind(pnewmu2, pnewvar[,2]), 1, calculatep)
  predIp2.p1 <- ifelse(predp2-predp1 > -deltaf, 1, 0)
  YNpred[i] <- mean(predIp2.p1, na.rm=T) > gammaf
}
```

```
obspc1 <- mean(YNpred, na.rm=T)
result$pred.obs.prob.decision <- obspc1
if (is.na(obspc1)==0 & obspc1 > pc1) {baysuccess1 <- 1
} else if (is.na(obspc1)==0 & obspc1 <= pc1) {baysuccess1 <- 0
} else if (is.na(obspc1)==1) {baysuccess1 <- NA}
result$bay.success1 <- baysuccess1

mu1ogibbs <- matrix(data=NA, nrow=nsim+1, ncol=7)
mu1hgibbs <- matrix(data=NA, nrow=nsim+1, ncol=7)
mu1gibbs <-  matrix(data=NA, nrow=nsim+1, ncol=7)
biasgibbs <- matrix(data=NA, nrow=nsim+1, ncol=7)
vargibbs <-  matrix(data=NA, nrow=nsim+1, ncol=4)

lastmu1o <- rnorm(7, 0, 1000)
lastmu1h <- rnorm(7, 0, 1000)
lastmu1 <-  rnorm(7, 0, 1000)
lastbias <- rnorm(7, 0, 1000)
lastvar <-  runif(4, 0, 100000)

mu1ogibbs[1,] <- lastmu1o
mu1hgibbs[1,] <- lastmu1h
mu1gibbs[1,] <- lastmu1
biasgibbs[1,] <- lastbias
vargibbs[1,] <- lastvar

x1hmean <- rowMeans(x1h)
x1mean  <- rowMeans(x1)
x2mean  <- rowMeans(x2)

for (i in 1:nsim) {

  newmu1o <- rnorm(7, mean=(lastmu1+lastmu1h)/2, sd=sqrt(lastvar[3]/2))
  newmu1h <- rnorm(7, mean=(1/((n1h/lastvar[1])+(1/lastvar[3])))*((n1h*x1hmean/lastvar[1])
                  +(lastmu1o/lastvar[3])), sd=sqrt(1/((n1h/lastvar[1])+(1/lastvar[3]))))
  newmu1  <- rnorm(7, mean=(1/(((ns11+ns21)/lastvar[1])+((ns12+ns22)/lastvar[2])+(1/lastvar[3])))
                   *(((ns11+ns21)*x1mean/lastvar[1])+(((ns12+ns22)*(x2mean+lastbias))/lastvar[2])
                   +(lastmu1o/lastvar[3])), sd=sqrt(1/(((ns11+ns21)/lastvar[1])
                   +((ns12+ns22)/lastvar[2])+(1/lastvar[3]))))
  newbias  <- rnorm(7, mean=(1/(((ns12+ns22)/lastvar[2])+(1/lastvar[4])))*((((ns12+ns22)
                   *(lastmu1-x2mean))/lastvar[2])+(thetaf/lastvar[4])),
                   sd=sqrt(1/(((ns12+ns22)/lastvar[2])+(1/lastvar[4]))))

  shapes <- c(7*(ns11+ns21+n1h)/2, 7*(ns12+ns22)/2, 7*2/2, 7/2)
  rates <- c( 0.5*(sum(rowSums((x1h-lastmu1h)^2))+sum(rowSums((x1-lastmu1)^2))),
              0.5*sum(rowSums((x2-(lastmu1-lastbias))^2)),
              0.5*(sum((lastmu1-lastmu1o)^2)+sum((lastmu1h-lastmu1o)^2)),
              0.5*sum((lastbias-thetaf)^2))
  newvar <- 1/rgamma(4, shape=shapes, rate=rates)

  mu1ogibbs[i+1,] <- newmu1o
  mu1hgibbs[i+1,] <- newmu1h
  mu1gibbs[i+1,] <- newmu1
  biasgibbs[i+1,] <- newbias
  vargibbs[i+1,] <- newvar

  lastmu1o <- newmu1o
  lastmu1h <- newmu1h
  lastmu1 <- newmu1
  lastbias <- newbias
  lastvar <- newvar
}

mu2gibbs   <- mu1gibbs - biasgibbs
forp1gibbs <- cbind(mu1gibbs, vargibbs[,1])
```

```
  forp2gibbs <- cbind(mu2gibbs, vargibbs[,2])

  result$gibbs.mu1o.means <- colMeans(mu1ogibbs[range,])
  result$gibbs.var.1o <- apply(mu1ogibbs[range,], 2, myvariance)
  result$gibbs.mu1.means <- colMeans(mu1gibbs[range,])
  result$gibbs.mu2.means <- colMeans(mu2gibbs[range,])
  result$gibbs.bias.means <- colMeans(biasgibbs[range,])
  result$gibbs.var1.var2.var1b.varbias.means <- colMeans(vargibbs[range,])

  if (plotconvg==1) {
    plot(range, mu2gibbs[range, 1], type="l")
    plot(density(mu2gibbs[range, 1], na.rm=T))}

  p1 <- apply(forp1gibbs, 1, calculatep)
  p2 <- apply(forp2gibbs, 1, calculatep)
  Ip2.p1 <- ifelse(p2-p1 > -deltaf, 1, 0)
  p1 <- p1[range]
  p2 <- p2[range]
  result$p1.mean <- .Internal(mean(p1))
  result$p2.mean <- .Internal(mean(p2))
  result$p2.p1.diff.mean <- .Internal(mean(p2-p1))

  obspc2 <- mean(Ip2.p1[range])
  result$obs.prob.decision <- obspc2
  if (is.na(obspc2)==0 & obspc2 > pc2) {baysuccess2 <- 1
  } else if (is.na(obspc2)==0 & obspc2 <= pc2) {baysuccess2 <- 0
  } else if (is.na(obspc2)==1) {baysuccess2 <- NA}
  result$bay.success2 <- baysuccess2

  if ((freqp2-freqp1-qnorm(0.975)*sqrt((freqp2*(1-freqp2)/(ns12+ns22))
     +(freqp1*(1-freqp1)/(ns11+ns21)))) > -(1-lambdaf)*(p1h-p0h-qnorm(0.975)
     *sqrt((p1h*(1-p1h)/n1h)+(p0h*(1-p0h)/n0h)))) {freqsuccess <- 1
     } else {freqsuccess <- 0}
  result$freq.success <- freqsuccess

  return(result)
}

probrejectnull2stage <- function(x1h, ns1, ns2, R, pc1, pc2, gammaf=0.9, triangle, lambdaf,
deltaf, thetaf, p0h, p1h, n0h, meanvec1h, thetafactual, var1, var2, nsim, plotconvg, nsimtrial) {

  result <- list()
  baytype1stg1 <- rep(NA, nsimtrial)
  baytype1stg2 <- rep(NA, nsimtrial)
  freqtype1 <- rep(NA, nsimtrial)
  for (k in 1:nsimtrial) {
    onetrial <- onetrialhier2stage(x1h=x1h, ns1=ns1, ns2=ns2, R=R, pc1=pc1, pc2=pc2,
    gammaf=0.9, triangle=triangle, lambdaf=lambdaf, deltaf=deltaf, thetaf=thetaf, p0h=p0h,
    p1h=p1h, n0h=n0h, meanvec1h=meanvec1h, thetafactual=thetafactual, var1=var1, var2=var2,
    nsim=nsim, plotconvg=0)
    baytype1stg1[k] <- onetrial$bay.success1
    baytype1stg2[k] <- onetrial$bay.success2
    freqtype1[k] <- onetrial$freq.success
  }
  result$baytype1stg1 <- baytype1stg1
  result$baytype1stg2 <- baytype1stg2
  result$baytype1sum <- baytype1stg1+baytype1stg2
  result$baytype1error <- mean(baytype1stg1+baytype1stg2 > 0, na.rm=T)
  result$expsamplesize <- (mean(baytype1stg1, na.rm=T))*ns1
  +(1-(mean(baytype1stg1, na.rm=T)))*(ns1+ns2)
  result$freqtype1error <- mean(freqtype1, na.rm=T)
  return(result)
}
```

5.6.2 Analysis Codes

```
# simulation setting
mu0h <- c(6, -7, 2, -3, -22, 2, -207)
mu1h <- c(56, 47, 44, 46, 53, 39, 31)
var0h <- 1600
var1h <- 1600
var1 <- 1600
var2 <- 1600
n0h <- 80
n1h <- 78
n <- c(60, 120)
R <- c(1, 2)
lambda <- c(0, 0.25, 0.5)
nsim <- 3000
nsimtrial <- 10000
range <- (ceiling(0.1*nsim)):nsim

# simulate historical trial
x0h <- matrix(rnorm(7*n0h, mu0h, sqrt(var0h)), c(7, n0h), byrow=F)
write.table(x0h, "x0h.txt", row.names=F, col.names=F, quote=F)
rowMeans(x0h)
px0h <- getpk(x0h)
px0h

x1h <- matrix(rnorm(7*n1h, mu1h, sqrt(var1h)), c(7, n1h), byrow=F)
write.table(x1h, "x1h.txt", row.names=F, col.names=F, quote=F)
rowMeans(x1h)
px1h <- getpk(x1h)
px1h

freqnimargin <- (px1h-px0h-qnorm(0.975)*sqrt((px1h*(1-px1h)/n1h)+(px0h*(1-px0h)/n0h)))
write.table(cbind(px0h, px1h, freqnimargin), "px0hpx1h.txt", quote=F, row.names=F, col.names=T)
px0hpx1h <- read.table("px0hpx1h.txt", header=T)

# determining non-inferiority margins, deltas
gibbs(xdata=x0h, nsim=nsim, outfile="x0hgibbs.txt")
gibbs(xdata=x1h, nsim=nsim, outfile="x1hgibbs.txt")

# producing trace plots and kernel plots
x0hgibbs <- read.table("x0hgibbs.txt", header=F); dim(x0hgibbs)
plotconvgmatrix(gibbssample=x0hgibbs, range=3:3000, outfile="traceplots0h.pdf")
plotkernelmatrix(gibbssample=x0hgibbs, range=range, outfile="kernelplots0h.pdf")

x1hgibbs <- read.table("x1hgibbs.txt", header=F); dim(x1hgibbs)
plotconvgmatrix(gibbssample=x1hgibbs, range=3:3000, outfile="traceplots1h.pdf")
plotkernelmatrix(gibbssample=x1hgibbs, range=range, outfile="kernelplots1h.pdf")

x0hgibbsmean <- colMeans(x0hgibbs[range,]); x0hgibbsmean
x1hgibbsmean <- colMeans(x1hgibbs[range,]); x1hgibbsmean

x1hmeans <- rowMeans(x1h)
x1hvar <- ((1/2)*sum(rowSums((x1h - x1hmeans)^2)))/(((7*n1h)/2)-1)
x1hgibbsmean2 <- c(x1hmeans, x1hvar)
x1hgibbsmean2
dim(x1hgibbsmean2) <- c(1,8)
write.table(x1hgibbsmean2, "x1hgibbsmean.txt", quote=F, col.names=F, row.names=F)

p0h <- apply(x0hgibbs[range,], 1, calculatep); mean(p0h)
p1h <- apply(x1hgibbs[range,], 1, calculatep); mean(p1h)
```

```
diffp01h <- p1h - p0h
mean(diffp01h)
pdf("histoprobdiff.pdf")
plot(density(diffp01h), main="Kernel plot of historial prob diff", xlab="p2-p1")
dev.off()

delta <- quantile(diffp01h, 0.025)
delta

nimargins <- delta*(1-lambda)
nimargins

# find corresponding thetas
gettheta <- findtheta(delta=nimargins, meanvec=mu1h, vari=var1h)
gettheta$theta
gettheta$thetaerror

margins <- cbind(rep(delta, length(lambda)), lambda, nimargins, gettheta$theta)
margins <- data.frame(margins)
names(margins) <- c("lowbound", "lambda", "delta", "theta")
margins
write.table(margins, "margins.txt", quote=F, col.names=T)

# figure of type 1 error vs triangle
n <- rep(60, 15)
R <- rep(1, 15)
pc <- rep(0.95, 15)
better <- c(-seq(2.8, 0.4, length=7), 0, seq(0.4, 2.8, length=7))
lambdaf <- rep(lambda[1], 15)
deltaf <- rep(margins$delta[1], 15)
thetaf <- rep(margins$theta[1], 15)

frame4inflat <- cbind(n, R, pc, better, lambdaf, deltaf, thetaf)
frame4inflat
write.table(frame4inflat, "frame4inflat.txt", quote=F, col.names=T, row.names=F)
frame4inflat <- read.table("frame4inflat.txt", header=T)
class(frame4inflat)

baytype1err <- rep(NA, length=nrow(frame4inflat))
freqtype1err <- rep(NA, length=nrow(frame4inflat))
file.con <- file("checkinflat.txt", "w")
for (s in 1:nrow(frame4inflat)) {
  value <- probrejectnullhier(x1h=x1h, n=frame4inflat[s,1], R=frame4inflat[s,2],
                              pc=frame4inflat[s,3], triangle=frame4inflat[s,4],
                              lambdaf=frame4inflat[s,5], deltaf=frame4inflat[s,6],
                              thetaf=frame4inflat[s,7], p0h=px0hpx1h[1], p1h=px0hpx1h[2],
                              n0h=n0h, meanvec1h=mu1h, thetafactual=frame4inflat[s,7], var1=1600,
                              var2=1600, nsim=nsim, plotconvg=0, nsimtrial=nsimtrial)
  check <- c(s, value$baytype1error, value$freqtype1error )
  dim(check) <- c(1,3)
  write.table(check, file.con, quote=F, row.names=F, col.names=F, append=T)
  baytype1err[s] <- value$baytype1error
  freqtype1err[s] <- value$freqtype1error
}
close(file.con)

frame4inflat$baytype1error <- baytype1err
frame4inflat$freqtype1error <- freqtype1err
frame4inflat
write.csv(frame4inflat, "frame4inflat.csv", quote=F, row.names=F)

inflat <- read.csv("frame4inflat.csv", header=T)
pdf("inflated.pdf")
plot(inflat$better, inflat$baytype1error, cex=0.7, type="b",
```

```
     main="Type 1 error by mean endpoint difference",
     xlab=expression(paste(mu, "1j - ", mu, "1hj")), ylab="Type 1 error")
abline(h=0.025, col="blue")
lines(inflat$better, inflat$freqtype1error, cex=0.7, type="b", pch=3)
legend(1.5, 0.13, c("Bayesian", "Frequentist"), pch=c(1,3))
dev.off()

# table of type 1 error
n <- c(rep(60, 36), rep(120, 36))
R <- rep(c(rep(1, 18), rep(2, 18)), time=2)
pc <- rep(c(rep(0.95, 9), rep(0.975, 9)), time=4)
triangle <- rep(c(rep(-2, 3), rep(0, 3), rep(2, 3)), time=8)
lambdaf <- rep(lambda, time=24)
deltaf <- rep(margins$delta, time=24)
thetaf <- rep(margins$theta, time=24)

frame4type1 <- cbind(n, R, pc, triangle, lambdaf, deltaf, thetaf)
frame4type1
write.table(frame4type1, "frame4type1.txt", quote=F, col.names=T, row.names=F)
frame4type1 <- read.table("frame4type1.txt", header=T)
class(frame4type1)

baytype1err <- rep(NA, length=nrow(frame4type1))
freqtype1err <- rep(NA, length=nrow(frame4type1))
file.con <- file("checktype1.txt", "w")
for (s in 1:nrow(frame4type1)) {
  value <- probrejectnullhier(x1h=x1h, n=frame4type1[s,1], R=frame4type1[s,2],
                              pc=frame4type1[s,3], triangle=frame4type1[s,4], lambdaf=frame4type1[s,5],
                              deltaf=frame4type1[s,6], thetaf=frame4type1[s,7], p0h=px0hpx1h[1],
                              p1h=px0hpx1h[2], n0h=n0h, meanvec1h=mu1h,
                              thetafactual=frame4type1[s,7], var1=1600, var2=1600, nsim=nsim,
                              plotconvg=0, nsimtrial=nsimtrial)
  check <- c(s, value$baytype1error, value$freqtype1error )
  dim(check) <- c(1,3)
  write.table(check, file.con, quote=F, row.names=F, col.names=F, append=T)
  baytype1err[s] <- value$baytype1error
  freqtype1err[s] <- value$freqtype1error
}
close(file.con)

frame4type1$baytype1error <- baytype1err
frame4type1$freqtype1error <- freqtype1err
frame4type1
write.csv(frame4type1, "frame4type1final.csv", quote=F, row.names=F)


# table of power
# find corresponding thetas for half of the deltas under alternative
getthetahalf <- findtheta(delta=margins$delta*0.5, meanvec=mu1h, vari=var1h)
getthetahalf$thetaerror
getthetahalf <- getthetahalf$theta

n <- c(rep(60, 36), rep(120, 36))
R <- rep(c(rep(1, 18), rep(2, 18)), time=2)
pc <- rep(c(rep(0.95, 9), rep(0.975, 9)), time=4)
triangle <- rep(c(rep(-2, 3), rep(0, 3), rep(2, 3)), time=8)
lambdaf <- rep(lambda, time=24)
deltaf <- rep(margins$delta, time=24)
thetaf <- rep(margins$theta, time=24)
Ha_deltahalf <- rep(margins$delta*0.5, time=24)
Ha_thetahalf <- rep(getthetahalf, time=24)
Ha_deltazero <- rep(rep(0,3), time=24)
Ha_thetazero <- rep(rep(0,3), time=24)
```

```
frame4power <- cbind(n, R, pc, triangle, lambdaf, deltaf, thetaf, Ha_deltahalf,
                     Ha_thetahalf, Ha_deltazero, Ha_thetazero)
frame4power
write.table(frame4power, "frame4power.txt", quote=F, col.names=T, row.names=F)
frame4power <- read.table("frame4power.txt", header=T)
class(frame4power)

baypowerhalf <- rep(NA, length=nrow(frame4power))
freqpowerhalf <- rep(NA, length=nrow(frame4power))
file.con <- file("checkpowerhalf.txt", "w")
for (s in 1:nrow(frame4power)) {
  value <- probrejectnullhier(x1h=x1h, n=frame4power[s,1], R=frame4power[s,2],
                              pc=frame4power[s,3], triangle=frame4power[s,4],
                              lambdaf=frame4power[s,5], deltaf=frame4power[s,6],
                              thetaf=frame4power[s,7], p0h=px0hpx1h[1], p1h=px0hpx1h[2],
                              n0h=n0h, meanvec1h=mu1h, thetafactual=frame4power[s,9], var1=1600,
                              var2=1600, nsim=nsim, plotconvg=0, nsimtrial=nsimtrial)
  check <- c(s, value$baytype1error, value$freqtype1error)
  dim(check) <- c(1,3)
  write.table(check, file.con, quote=F, row.names=F, col.names=F, append=T)
  baypowerhalf[s] <- value$baytype1error
  freqpowerhalf[s] <- value$freqtype1error
}
close(file.con)

frame4power$baypowerhalf <- baypowerhalf
frame4power$freqpowerhalf <- freqpowerhalf
frame4power

baypowerzero <- rep(NA, length=nrow(frame4power))
freqpowerzero <- rep(NA, length=nrow(frame4power))
file.con <- file("checkpowerzero.txt", "w")
for (s in 1:nrow(frame4power)) {
  value <- probrejectnullhier(x1h=x1h, n=frame4power[s,1], R=frame4power[s,2],
                              pc=frame4power[s,3], triangle=frame4power[s,4],
                              lambdaf=frame4power[s,5], deltaf=frame4power[s,6],
                              thetaf=frame4power[s,7], p0h=px0hpx1h[1], p1h=px0hpx1h[2],
                              n0h=n0h, meanvec1h=mu1h, thetafactual=frame4power[s,11],
                              var1=1600, var2=1600, nsim=nsim, plotconvg=0, nsimtrial=nsimtrial)
  check <- c(s, value$baytype1error, value$freqtype1error)
  dim(check) <- c(1,3)
  write.table(check, file.con, quote=F, row.names=F, col.names=F, append=T)
  baypowerzero[s] <- value$baytype1error
  freqpowerzero[s] <- value$freqtype1error
}
close(file.con)

frame4power$baypowerzero <- baypowerzero
frame4power$freqpowerzero <- freqpowerzero
frame4power
write.csv(frame4power, "frame4powerfinal.csv", row.names=F, quote=F)

# table of two-stage
n <- rep(120, 6)
ns1 <- rep(90, 6) # make sure ns1 and ns2 is divisible by (1+R)
ns2 <- rep(30, 6)
R <- rep(c(1,2), time=3)
pc1 <- rep(0.98, 6)
pc2 <- rep(0.95, 6)
triangle <- rep(c(rep(-2, 2), rep(0, 2), rep(2, 2)), time=1)
lambdaf <- rep(lambda[3], time=6)
deltaf <- rep(margins$delta[3], time=6)
thetaf <- rep(margins$theta[3], time=6)
Ha_deltahalf <- rep(margins$delta[3]*0.5,  time=6)
```

```
Ha_thetahalf <- rep(getthetahalf[3], time=6)
Ha_deltazero <- rep(0, time=6)
Ha_thetazero <- rep(0, time=6)

frame <- cbind(n, ns1, ns2, R, pc1, pc2, triangle, lambdaf, deltaf, thetaf,
            Ha_deltahalf, Ha_thetahalf, Ha_deltazero, Ha_thetazero)
frame
write.table(frame, "frame4twostage.txt", quote=F, col.names=T, row.names=F)
frame <- read.table("frame4twostage.txt", header=T)
class(frame)

baytype1err <- rep(NA, length=nrow(frame))
freqtype1err <- rep(NA, length=nrow(frame))
expsstype1err <- rep(NA, length=nrow(frame))
file.con <- file("check2stagetype1.txt", "w")
for (s in 1:nrow(frame)) {
  value <- probrejectnull2stage(x1h=x1h, ns1=frame[s,2], ns2=frame[s,3], R=frame[s,4],
                                pc1=frame[s,5], pc2=frame[s,6], gammaf=0.9, triangle=frame[s,7],
                                lambdaf=frame[s,8], deltaf=frame[s,9], thetaf=frame[s,10],
                                p0h=px0hpx1h[1], p1h=px0hpx1h[2], n0h=n0h, meanvec1h=mu1h,
                                thetafactual=frame[s,10], var1=1600, var2=1600, nsim=nsim,
                                plotconvg=0, nsimtrial=nsimtrial)
  check <- c(s, value$baytype1error, value$freqtype1error, value$expsamplesize)
  dim(check) <- c(1,4)
  write.table(check, file.con, quote=F, row.names=F, col.names=F, append=T)
  baypowerhalf[s] <- value$baytype1error
  freqpowerhalf[s] <- value$freqtype1error
  expsshalf[s] <- value$expsamplesize
}
close(file.con)

frame$baytype1error <- baytype1err
frame$freqtype1error <- freqtype1err
frame$expsstype1err <- expsstype1err
frame
write.csv(frame, "frame2stagetype1final.csv", quote=F, row.names=F)

baypowerhalf <- rep(NA, length=nrow(frame))
freqpowerhalf <- rep(NA, length=nrow(frame))
expsshalf <- rep(NA, length=nrow(frame))
file.con <- file("check2stagehalf.txt", "w")
for (s in 1:nrow(frame)) {
  value <- probrejectnull2stage(x1h=x1h, ns1=frame[s,2], ns2=frame[s,3], R=frame[s,4],
                 pc1=frame[s,5], pc2=frame[s,6], gammaf=0.9, triangle=frame[s,7],
               lambdaf=frame[s,8], deltaf=frame[s,9], thetaf=frame[s,10], p0h=px0hpx1h[1],
               p1h=px0hpx1h[2], n0h=n0h, meanvec1h=mu1h, thetafactual=frame[s,12],
               var1=1600, var2=1600, nsim=nsim, plotconvg=0, nsimtrial=nsimtrial)
  check <- c(s, value$baytype1error, value$freqtype1error, value$expsamplesize)
  dim(check) <- c(1,4)
  write.table(check, file.con, quote=F, row.names=F, col.names=F, append=T)
  baypowerhalf[s] <- value$baytype1error
  freqpowerhalf[s] <- value$freqtype1error
  expsshalf[s] <- value$expsamplesize
}
close(file.con)

frame$baypowerhalf <- baypowerhalf
frame$freqpowerhalf <- freqpowerhalf
frame$expsshalf <- expsshalf
frame

baypowerzero <- rep(NA, length=nrow(frame))
freqpowerzero <- rep(NA, length=nrow(frame))
expsszero <- rep(NA, length=nrow(frame))
```

```
file.con <- file("check2stagezero.txt", "w")
for (s in 1:nrow(frame)) {
  value <- probrejectnull2stage(x1h=x1h, ns1=frame[s,2], ns2=frame[s,3], R=frame[s,4],
                                pc1=frame[s,5], pc2=frame[s,6], gammaf=0.9, triangle=frame[s,7],
                                lambdaf=frame[s,8], deltaf=frame[s,9],
                                thetaf=frame[s,10], p0h=px0hpx1h[1], p1h=px0hpx1h[2], n0h=n0h,
                                meanvec1h=mu1h, thetafactual=frame[s,14], var1=1600, var2=1600,
                                nsim=nsim, plotconvg=0, nsimtrial=nsimtrial)
  check <- c(s, value$baytype1error, value$freqtype1error, value$expsamplesize)
  dim(check) <- c(1,4)
  write.table(check, file.con, quote=F, row.names=F, col.names=F, append=T)
  baypowerzero[s] <- value$baytype1error
  freqpowerzero[s] <- value$freqtype1error
  expsszero[s] <- value$expsamplesize
}
close(file.con)

frame$baypowerzero <- baypowerzero
frame$freqpowerzero <- freqpowerzero
frame$expsszero <- expsszero
frame

write.csv(frame, "frame4twostage.csv", row.names=F, quote=F)

# end of chapter code
```

Chapter 6

Summary and Further Work

6.1 Summary

As the cost of clinical development increases while success rates still remain low, the biopharmaceutical research community is widely embracing new adaptive methods across all phases of the developmental process. Due to the complex dependent nature of the data structure and uncharted properties of many adaptive design, extensive applications are still rare. This dissertation aims to examine the characteristics of current adaptive methodologies used in phase 2 dose-ranging and phase 3 pivotal trials and proposing improved trial designs.

As discussed in Chapters 1 and 2, such adaptive designs are sometimes known as the Seamless phase 2/3 clinical trials (Maca *et al.*, 2006). These are deliberate designs that combine the objectives of two traditionally separate trials into one single trial, such as the selection of an optimal dose in a phase 2 dose-ranging trial and the confirmation of this selected dose in a large-scale phase 3 randomized clinical trial. These designs aim at shortening the wait time between trials and operationally combine these objectives within one single study protocol. If these designs allow for early termination, then they are more efficient with respect to the type I error allocated, compared to conducting two separate trials. Many innovative and interesting approaches that classified themselves as Seamless study designs

were found in the literature, which include, although are not restricted to, (1) incorporating a parametric model (linear or non-linear) to aid the selection of either the minimum effective dose (MED) or the maximum biological dose (MBD) (Huang, Liu, and Hsiao, 2010), (2) using a surrogate endpoint or an early outcome to inform the selection (Shun, Lan, and Soo, 2008; Friede *et al.*, 2011), (3) implementing response-adaptive randomization to narrow down to one or two treatment arms (Wang and Cui, 2007), and others. In addition, Bayesian Seamless 2/3 designs are also proposed and they elicit prior information to help them improve overall performance (Schmidli, Bretz, and Racine-Poon, 2007). The analysis of Seamless designs also presents challenges in controlling type I error. Statistical methods such as different p-value Combination Tests have been proposed and they extended to trials with multiple treatment arms, allowing for multiplicity adjustment (Bauer and Kieser, 1999).

In this dissertation, we have proposed an adaptive staggered dose design for both selecting a dose and confirming the selected dose within a single trial. We recognize that if some qualitative information about the dose response is available such as monotonicity or downturn, we can prioritize and stagger the candidate doses in such a way that the trial can explore the doses of assumed better efficacy first before reaching to doses of uncertain efficacy later. The application of an alpha spending function will be a good choice to favor the earlier doses but this still allows the trial to proceed to the remaining doses if earlier doses do not show evidence of efficacy. The operating characteristics of this design were examined and discussed, and its strengths and weaknesses were also explored. We also discussed that in some situations, we are not comparing and selecting doses of the same agent, but rather different regimes or treatment modalities. In this case, staggered treatment modalities can be explored one after the other with the assumed best one receiving favorable alpha spending. If we can assume that subjects who enter the trial are from a study population with a relatively stable profile of characteristics over the study period, we do not have to continue the control arm to the last arm of the experimental treatment and in this case, we can further reduce the expected sample size. In Chapter 4, we also provide a brief discussion of

how this adaptive staggered treatment design can be flexibly varied in order to optimize its performance in dose selection and confirmation, as well as how it can be extended to binary and survival endpoints. For example, if we are interested in locating the dose that gives the lowest risk of a binary safety endpoint among doses that exhibit comparable efficacy, under large sample condition and if the probability is not too close to either 0 or 1, then we can use normal approximation to obtain the standardized test statistic and its distribution under null hypotheses.

As for the proof of bioisimilarity, we have chosen to approach this testing problem from a Bayesian perspective. As many innovator biological products in the market will have their FDA licenses expiring in the next decade or so, pharmaceutical companies are interested in developing generic versions of the same products. Therefore, many biosimilar products, which are also known as follow-on biologics or subsequent-entry products, need to demonstrate similarity in all aspects of the product's characteristics: molecular structure, functional assay, toxicity, and efficacy, against the backdrop of the original reference product. We advocate for a Bayesian approach because it formally allows the incorporation of historical trial information on the reference product into the analysis of the current biosimilarity. We also choose the non-inferiority framework as it may require smaller sample size. Within the hierarchical bias model proposed in the method, the elicitation of a skeptical prior on the bias parameter allows the method to have good control on the Bayesian type I error even when the constancy assumption does not hold slightly, and as a result, the Bayesian power suffers moderately. However, when sample size increases in the case of true biosimilarity, we confirm that the method can give better statistical power than the frequentist approach. An adaptive two-stage design can shorten the biosimilarity trial and reduce the expected overall trial sample size if biosimilarity is highly plausible. The simulation for the two-stage trial design was computationally intensive as predictive probability was used as a monitoring criterion. As of the writing of this dissertation, no Bayesian method has yet been published for the establishment of biosimilarity.

6.2 Further Work

Previously in Section 6.1, we have discussed some of the strengths and weaknesses of the proposed adaptive methodologies. For the adaptive staggered treatment design, there are at least two statistical properties of the design that may require further characterization.

1. Futility Analysis

   In the proposed method, we did not consider early stopping due to futility. The introduction of futility analysis into the design may allow it to drop inferior doses earlier before they go through all of the pre-assigned M per-dose stages, therefore, saving subjects from being randomizing to these inefficacious doses as well as getting to the more efficacious faster, particularly when the *a priori* dose ordering fails to order the more efficacious doses in earlier positions. In this case, one has to distinguish and decide between the specification of either binding or non-binding futility. Binding futility refers to the condition when the futility boundary is crossed, and the dose has to be jettisoned from the study without affecting the type I error, otherwise, under non-binding futility when the dose can be dropped or retained in the study even if futility boundary is crossed. In this case, the corresponding efficacy boundary will have to be adjusted upward to maintain the overall type I error. Futility analysis is related to type II error and can be implemented using a beta spending function, $\beta(t)$.

2. Maximum Likelihood Estimator (MLE)

   In the usual two-arm group sequential clinical trial using a continuous endpoint, it is well-known that the sampling distribution of the maximum likelihood estimator does not follow exact normal distribution. An exposition of this point estimation problem can be found in chapter 8 of "Group Sequential Methods with Applications to Clinical Trials" by Jennison and Turnbull (2000). In order to investigate this problem, we can look at Figure 6.1 which displays the sampling distributions of the MLE of the four doses.

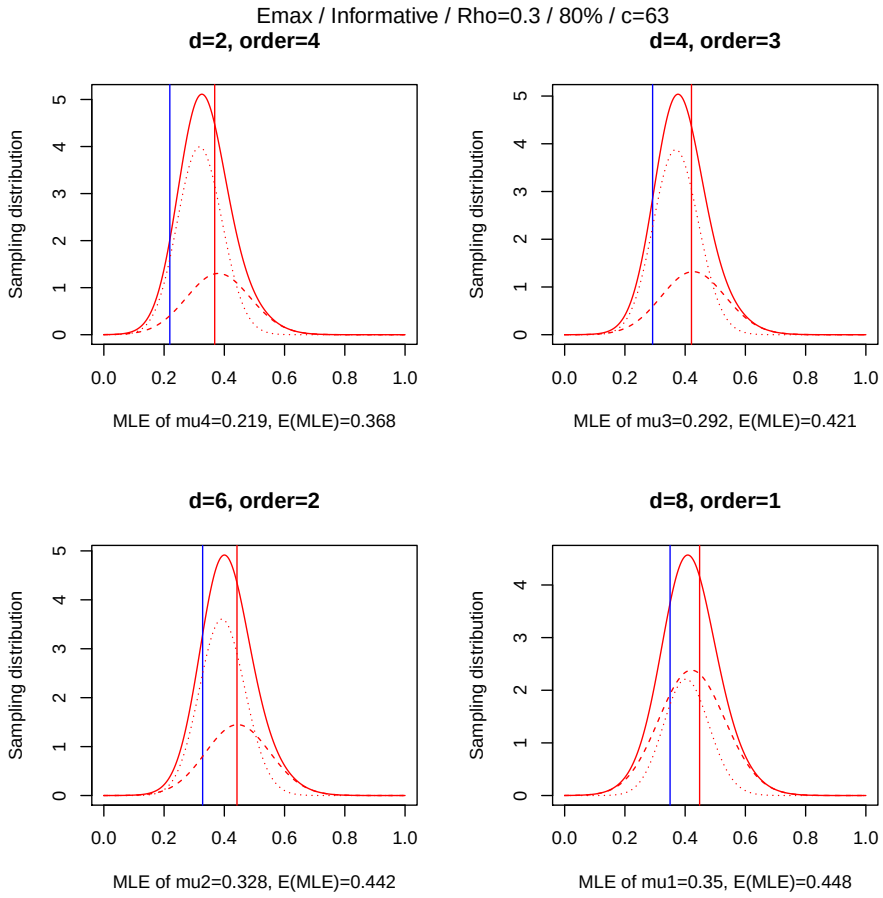


Figure 6.1: The sampling distributions of the four doses for Emax dose response, informative dose ordering, and Rho spending with $\rho = 0.3$. Cohort size c= 63 with power of 80%

We can see in each plot, the dashed curve represents the sampling distribution when the dose crosses efficacy boundary at $m = 1$ and the dotted curve represents that at $m = 2$. We notice that under informative ordering, $d_1 = 8$ has the highest mean outcome of $\mu_1 = 0.350$, and thus sample mean will have to be larger for $m = 1$ conditioned on the event that its value is large enough to allow the rejection of the null hypothesis. The two conditional sampling distributions will be combined to obtain the sampling distribution of the MLE conditioned on the rejection of the null hypothesis for the dose. The blue vertical line represents the true value of the mean, while the red vertical line represents the mean of combined sampling distribution. Therefore, the

MLE itself is biased upward under this staggered dose design. As complete correction of the estimation bias may not be possible, methods of reducing this bias for the group sequential design have been proposed. This can be another area of further research, but it is also important to caution that estimation upon selection will result in bias in our proposed adaptive design and for the group sequential method in general.

For the Bayesian biosimilarity design, there are at least three areas that may warrant further research and investigation.

1. Synthesis of Collective Biosimilarity Evidence

   According to the guidance for the industry document on the scientific and statistical considerations of biosimilarity presented by the FDA, two main principles of proving biosimilarity are succintly advocated, (1) *totality-of-evidence approach*, and (2) *stepwise approach* (FDA, 2012). Although the document did not proceed to explain these principles in detail, the industry is free to interpret them as an approach that the proof of bisimilarity is not simply based on evidence from one or two clinical studies. It is based on a collection of evidence that comprehensively shows that the two polypeptides (reference and follow-on) are similar in major components despite some dissimilarity in minor aspects. In fact, the Bayesian approach is still a desirable method to collectively synthesize the evidence that is accumulated through a series of studies. Perhaps a method that weighs the relative importance of the individual studies (animal, toxicity, molecular, functional, bioavailability, and clinical studies) and synthesizes the evidence collectively, may further help to reduce the sample size needed to conduct the final clinical study.

2. Discounting the Historical Trials

   Instead of modeling the heterogeneity of the historical trials hierarchically in the proposed method, information accrued from historical trials may be discounted due to discrepancies between the design and conduct of the historical trials and the current trial of biosimilarity. Methods of discounting historical information have been pro-

posed such as the power prior density (Ibrahim and Chen, 2000) and commensurate prior density (Hobbs, Carlin, Mandrekar, and Sargent, 2011). In this case, additional power parameter or commensurate parameters will be specified as well as their prior distributions.

3. Dependence Between Individual ACR20 Endpoints

   In the Bayesian method proposed, the seven individual ACR20 endpoints are assumed to be separate and independent; however, in some cases, this assumption may not be true, and therefore the Bayesian method can be updated to allow for the specification of a correlation structure. Composite endpoints are useful and their statistical properties are not fully known. It is important to have studies that can fully describe the properties of a composite endpoint, which is usually defined based on a number of individual endpoints, and can estimate their potential dependency structure. In the Bayesian paradigm, for example, prior density for unstructured correlation matrix can be flexibly modeled via the inverse-Wishart distribution.

Bibliography

[1] Antonijevic, Z., Pinheiro, J., Fardipour, P., Lewis, R.J. (2010). Impact of dose selection strategies used in phase II on the probability of success in phase III. *Statistics in Biopharmaceutical Research.* DOI: 10.1198/sbr 2010.08101.

[2] Armitage, P., McPherson, C.K., Rowe, B.C. (1969). Repeated significance tests on accumulating data. *Journal of Royal Statistical Society, Series A.* Vol. 132, 235-244.

[3] Babb, J., Rogatko, A., Zacks, S. (1998). Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statistical Science.* Vol. 17, 1103-1120.

[4] Bathon, J.M., Martin, R.W., Fleischmann, R.M. et al. (2000). A comparison of etanercept and methotrexate in patients with early rheumatoid arthritis. *The New England Journal of Medicine.* Vol. 343, No. 22, 1586-1593.

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

[6] Bauer, P., Kohne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics.* Vol. 50, No. 4, 1029-1041.

[7] Berger, R.L., Hsu, J.C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science.* Vol. 11, No. 4, 283-302.

[8] Bornkamp, B., Bretz, F., Dmitrienko, A., Enas, G., Gaydos, B., Hsu, C.H., Konig, F., Krams, M., Liu, Q., Neuenschwander, B., Parke, T., Pinheiro, J. (2007). Innova-

tive approaches for designing and analyzing adaptive dose-ranging trials. *Journal of Biopharmaceutical Statistics.* Vol. 17, 965-995.

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

[10] Calvo, B., Zuñiga, L. (2012). The US approach to biosimilars: the long-awaited FDA approval pathway. *Biodrugs.* Vol. 26, No. 6, 357-361.

[11] Chang, M. (2007). Adaptive design method based on sum of p-values. *Statistics in Medicine.* Vol. 26. 2772-2784.

[12] Chen, J., DeMets, D.L., Lan, K.K.G. (2010). Some drop-the-loser designs for monitoring multiple doses. *Statistics in Medicine.* Vol. 29, 1793-1807.

[13] Chervoneva, I., Hyslop, T., Hauck, W.W. (2007). A multivariate test for population bioequivalence. *Statistics in Medicine.* Vol. 26, 1208-1223.

[14] Chow, S.C., Chang, M. (2008). Adaptive design methods in clinical trials - a review. *Orphanet Journal of Rare Diseases..* Vol. 3, 11.

[15] Chow, S.C., Chang, M., Pong, A. (2005). Statistical consideration of adaptive methods in clinical development. *Journal of Biopharmaceutical Statistics.* Vol. 15, 575-591.

[16] Chow, S.C., Ki, F.Y.C. (1997). Statistical comparison between dissolution profiles of drug product. *Journal of Biopharmaceutical Statistics.* Vol. 7, 241-258.

[17] Chow, S.C., Liu, J.P. (2010). Statistical assessment of biosimilar products. *Journal of Biopharmaceutical Statistics.* Vol. 20, 10-30.

[18] Combes, R.D. (1997). Statistical analysis of dose-response data from in vitra assays: an illustration using salmonella mutagenicity data. *Toxicology in Vitro.* Vol. 11, 683-687.

[19] Cui, L., Hung, H.M.J., Wang, S.J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics.* Vol. 55, 853-857.

[20] D'Agostino, R.B., Massaro, J.M., Sullivan, L.M. (2003). Non-inferiority trials: design concepts and issues - encounters of academic consultants in statistics. *Statistics in Medicine*. Vol. 22, 169-186.

[21] Denne, J.S. (2001). Sample size recalculation using conditional power. *Statistics in Medicine*. Vol. 20, 2645-2660.

[22] Dmitrienko, A., Wang, M.D. (2006). Bayesian predictive approach to interim monitoring in clinical trials. *Statistics in Medicine*. Vol. 25, 2178-2195.

[23] Dragalin, V. (2006). Adaptive designs: terminology and classification. *Drug Information Journal*. Vol. 40, 425-435.

[24] Dragalin, V., Fedorov, V., Patterson, S., Jones, B. (2003). Kullback-Leibler divergence for evaluating bioequivalence. *Statistics in Medicine*. Vol. 22, 913-930.

[25] Dunnett, C.W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association.* Vol. 50, No. 272, 1096-1121.

[26] FDA. *Adaptive Design Clinical Trials for Drugs and Biologics.* The United States Food and Drug Administration: Rockville, MD, 2010.

[27] FDA. *Scientific Considerations in Demonstrating Biosimilarity to a Reference Product.* The United States Food and Drug Administration: Rockville, MD, 2012.

[28] Felson D.T., Anderson J.J., Boers M., et al. (1993). The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. *Arthritis Rheumatology*. Vol. 36, 729-40.

[29] Follmann, D.A., Proschan, M.A., Geller, N.L. (1994). Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics.* Vol. 50, No. 2, 325-336.

[30] Friede, T., Parsons, N., Stallard, N., Todd, S., Valdes Marquez, E., Chataway, J., Nicholas, R. (2011). Designing a seamless phase II/III clinical trial using early

outcomes for treatment selection: an application in multiple sclerosis. *Statistics in Medicine.* Vol. 30, 1528-1540.

[31] Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M., Pinheiro, J. (2006). Adaptive designs in clinical drug development - an executive summary of the PhRMA working group. *Journal of Biopharmaceutical Statistics.* Vol. 16, 275-283.

[32] Gallo, P., Anderson, K., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M., Pinheiro, J. (2010). Viewpoints on the FDA adaptive designs guidance from the PhRMA working group. *Journal of Biopharmaceutical Statistics.* Vol. 20, 1115-1124.

[33] Gamalo, M.A., Wu, R., Tiwari, R.C. (2012). Bayesian approach to non-inferiority trials for normal means. *Statistical Methods in Medical Research.* doi: 10.1177/0962280212448723.

[34] Gamalo, M.A., Tiwari, R.C., LaVange, L.M. (2013). Bayesian approach to the design and analysis of non-inferiority trials for anti-infective products. *Pharmaceutical Statistics.* doi: 10.1002/pst. 1588.

[35] Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (1995). *Bayesian Data Analysis.* Chapman and Hall, London.

[36] Gelman, A., Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science.* Vol. 7, 457-511.

[37] Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Hothorn, T. (2011). mvtnorm: Multivariate Normal and t Distributions. URL http://CRAN.R-project.org/package=mvtnorm. *R package version 0.9-9991.*

[38] Haybittle, J.L. (1971). Repeated assessments of results in clinical trials of cancer treatment. *British Journal of Radiology.* Vol. 44 (526), 793-797.

[39] Hobbs, B.P., Carlin, B.P., Mandrekar, S.J., Sargent, D.J. (2011). Hierarchical com-

mensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics.* Vol. 67, 1047-1056.

[40] Hommel, G. (2001). Adaptive modifications of hypotheses after an interim analysis. *Journal of Pharmaceutical Statistics.* Vol. 43, 581-589.

[41] Huang, W.S., Liu, J.P., Hsiao, C.F. (2011). An alternative phase II/III design for continuous endpoints. *Journal of Pharmaceutical Statistics.* Vol. 10, 105-114.

[42] Hwang, I.K., Shih, W.J., De Cani, J.S. (1990). Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine.* Vol. 9(12), 1439-1445.

[43] Ibrahim, J.G., Chen, M.H. (2000). Power prior distributions for regression models. *Statistical Science.* Vol. 15, No. 1, 46-60.

[44] Jennison, C., Turnbull, B.W. (2000). *Group Sequential Methods with Applications to Clinical Trials.* Boca Raton, FL: Chapman & Hall/CRC.

[45] Jenkins, M., Stone, A., Jennison, C. (2011). An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics.* Vol. 10, 347-356.

[46] Jennison, C., Turnbull, B.W. (2007). Adaptive seamless designs: selection and prospective testing of hypothesis. *Journal of Biopharmaceutical Statistics.* Vol. 17, 1135-1161.

[47] Kairalla, J.A., Coffey, C.S., Thomann, M.A., Muller, K.E. (2012). Adaptive trial designs: a review of barriers and opportunities. *Trials.* Vol. 13, 145.

[48] Kang, S.H., Chow, S.C. (2012). Statistical assessment of biosimilarity based on relative distance between follow-on biologics. *Statistics in Medicine.* Vol. 32, 328-392.

[49] Kim, K., DeMets, D.L. (1987). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika.* Vol. 74, No. 1, 149Ð154.

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

[51] Klareskog, L., van der Heijde, D., de Jager, J.P. et al. (2004). Therapeutic effect of the combination of etanercept and methotrexate compared with each treatment alone in patients with rheumatoid arthritis: double-blind randomised controlled trial. *The Lancet.* Vol. 363, 675-681.

[52] Lan, K.K.G., DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika.* Vol. 70, 659-663.

[53] Lan, K.K.G., DeMets, D.L., Halperin, M. (1984). More flexible sequential and non-sequential designs in long-term clinical trials. *Communications in Statistics - Theory and Methods.* Vol. 13, No. 19, 2339-2353.

[54] Lan, K.K.G., Wittes, J. (1988). The B-value: a tool for monitoring data. *Biometrics.* Vol. 44, 579-585.

[55] Lang, T., Auterith, A., Bauer, P. (2000). Trend tests with adaptive scoring. *Biometrical Journal.* Vol. 42, 1007-1020.

[56] Lee, J.J., Liu, D.D. (2008). A predictive probability design for phase II cancer clinical trials. *Clinical Trials.* Vol. 5, 93-106.

[57] Lehmacher, W., Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics.* Vol. 55, 1286-1290.

[58] Lei, L., Olson, K. (2010). Evaluating statistical methods to establish clinical similarity of two biologics. *Journal of Biopharmaceutical Statistics.* Vol. 20, 62-74.

[59] Li, G., Zhu, J., Ouyang, S.P., Xie, J., Deng, L., Law, G. (2009). Adaptive designs for interim dose selection. *Statistics in Biopharmaceutical Research.* Vol. 1, No. 4, 366-376.

[60] Lin, J.R., Chow, S.C., Chang, C.H., Lin, Y.C., Liu, J.P. (2012). Application of the

parallel line assay to assessment of biosimilar products based on binary endpoints. *Statistics in Medicine*. Vol. 32, 449-461.

[61] Maca, J., Bhattacharya, S., Dragalin, V., Gallo, P., Krams, M. (2006). Adaptive seamless phase II/III designs - background, operational aspects, and examples. *Drug Information Journal.* Vol. 40, 463-473.

[62] Mantel, N., Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institute.* Vol. 22, 719-748.

[63] Marcus, R., Peritz, E., Gabriel, K.R. (1976). On closed testing procedure with special reference to ordered analysis of variance. *Biometrika.* Vol. 63, No. 3, 655-660.

[64] Moore, J.W., Flanner, H.H. (1996). Mathematical comparison of dissolution profiles. *Pharmaceutical Technology*. Vol. 20, 64-74.

[65] Moreland, L.W., Baumgartner, S.W., Schiff, M.H. et al. (1997). Treatment of rheumatoid arthritis with a recombinant human tumor necrosis factor receptor (p75)-Fc fusion protein. *The New England Journal of Medicine*. Vol. 337, No. 3, 141-147.

[66] Moreland, L.W., Schiff, M. H., Baumgartner, S.W. et al. (1999). Etanercept therapy in rheumatoid arthritis. *Annals of Internal Medicine*. Vol. 130, 478-486.

[67] O'Brien, P.C., Fleming, T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics*. Vol. 35, No. 3, 549-556.

[68] Offen, W., Chuang-Stein C., Dmitriendko, A., et al. (2007). Multiple co-primary endpoints: medical and statistical solutions. A report from the multiple endpoints expert team of the pharmaceutical research and manufacturers of America. *Drug Information Journal*. Vol. 41, 31-46.

[69] O'Quigley, J., Pepe, M., Fisher, L. (1990). Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*. Vol. 46, No. 1, 33-48.

[70] Peto, R. (1978). Clinical trial methodology. *Biomedicine*. Vol. 28 (special issue), 24-36.

[71] Peto, R., Pike, M.C., Armitage, P., *et al.* (1976). Design and analysis of randomised clincal trials requiring prolonged observation of each patient. 1. Introduction and design. *British Journal of Cancer*. Vol. 34, 585-612.

[72] Plummer, M., Best, N., Cowles, K., Vines, K. (1992). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*. Vol. 6, 7-11.

[73] Pocock, S.J. (1976). The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases.* Vol 29, 175-188.

[74] Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*. Vol. 64, No. 2, 191-199.

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

[76] Proschan, M.A. (2005). Two-stage sample size re-estimation based on a nuisance parameter: a review. *Journal of Biopharmaceutical Statistics*. Vol. 15, 559-574.

[77] Proschan, M.A., Hunsberger, S.A. (1995). Designed extension of studies based on conditional power. *Biometrics*. Vol. 51, 1315-1324.

[78] Reynolds, A.R. (2010). Potential relevance of bell-shaped and U-shaped dose response for the therapeutic targeting of angiogenesis in cancer. *Dose-Response.* Vol. 8, 253-284.

[79] Robins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society.* Vol. 58, 527-536.

[80] Sampson, D.G., Margolin, B.H. (1986). Recursive non-parametric testing for dose response relationships subject to downturns at high doses. *Biometrika.* Vol. 73, No. 3, 589-596.

[81] Sampson, A.R., Sill, M.W. (2005). Drop-the-losers design: normal case. *Biometrial Journal.* Vol. 47, 257-268.

[82] Saranadasa, H., Krishnamoorthy, K. (2005). A multivariate test for similarity of two dissolution profiles. *Journal of Biopharmaceutical Statistics.* Vol. 15, 265-278.

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

[84] Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics.* Vol. 15, 657-680.

[85] Shun, Z.M., Lan, K.K.G., Soo, Y.W. (2008). Interim treatment selection using the normal approximation approach in clinical trials. *Statistics in Medicine.* Vol. 27, 597-618

[86] Spiegelhalter, D.J., Freedman, L.S. (1986). A predictive approach to selecting the size of a clinical trial based on subjective clinical opinion. *Statistics in Medicine.* Vol. 5, 1-13.

[87] Spiegelhalter, D.J., Freedman, L.S., Blackburn, P.R. (1986). Monitoring clinical trials: conditional or predictive power? *Controlled Clinical Trials.* Vol. 7, 8-17.

[88] Spiegelhalter, D.J., Freedman, L.S., Parmar, M.K.B. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society.* Vol. 157, No. 3, 357-416.

[89] Stallard, N., Friede, T. (2008). A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine.* Vol. 27, 6209-6227.

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

[91] Thall, P.F., Simon, R. (1994). Practical Bayesian guidelines for phase IIb clinical trials. *Biometrics*. Vol. 50, No. 2, 337-349.

[92] Tsong, Y., Hammerstrom, T., Sathe, P., Shah, V.P. (1997). Comparing two dissolution data sets for similarity. *American Statsitical Association, 1996 Proceedings of the Biopharmaceutical Section*. 129-134.

[93] Wakana, A., Yoshimura, I., Hamada, C. (2007). A method for therapeutic dose selection in a phase II clinical trial using contrast statistics. *Statistics in Medicine.* Vol. 26, 498-511.

[94] Wang, L., Cui, L. (2007). Seamless phase II/III combination study through response adaptive randomization. *Statistics in Biopharmaceutical Research.* Vol. 17, 1177-1187.

[95] Wang, S. J., Hung, H.M.J. (2005). Adaptive covariate adjustment in clinical trials. *Journal of Biopharmaceutical Statistics*. Vol. 15, 605-611.

[96] Wang, Y., Lan, K.K.G., Li, G., Ouyang, S.P. (2011). A group sequential procedure for interim treatment selection. *Statistics in Biopharmaceutical Research.* Vol. 3, No. 1, 1-13.

[97] Weinblatt, M.E., Kremer, J.M., Bankhurst, A.D. et al. (1999). A trial of etanercept, a recombinant tumor necrosis factor receptor: Fc fusion protein, in patients with rheumatoid arthritis receiving methotrexate. *The New England Journal of Medicine*. Vol. 340, No. 4, 253-259.

[98] Westlake, W.J. (1981). Response to T.B.L. Kirkwood: bioequivalence testing - a need to rethink. *Biometrics*. Vol. 37, 589-594.

[99] Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials.* Revised 2nd ed. Chichester: Wiley.

[100] Woodcock, J., Woosley, R. (2008). Initiative and its influence on new drug development. *The Annual Review of Medicine.* Vol. 59, 1-12.

[101] Zelen, M. (1969). Play-the-winner rule and the controlled clinical trial. *Journal of American Statistical Association.* Vol. 64, 143-146.

Curriculum Vitae

Joseph M.W. Wu
(617) 510-1708
josephwu@bu.edu
40 Taylor Road, Foxboro, MA 02035

EXECUTIVE SUMMARY

Joseph Wu is currently a doctoral candidate at the Boston University Department of Biostatistics, and will shortly defend his thesis on adaptive methods in multi-arm and biosimilarity clinical trials. He is the winner of the 2014 Department of Biostatistics Student Paper Award for his article "An adaptive staggered dose design for a normal endpoint" which has also been accepted for publication in the Journal of Biopharmaceutical Statistics. His chapter "Fitting the dose - adaptive staggered dose design for a normal endpoint" has also been published in the newly released book "Clinical & Statistical Considerations in Personalized Medicine" (edited by Carini, Menon, and Chang). He is the co-author of another upcoming book "Clinical Trials Simulation" (Menon, Wu, and Chang) planned to be released in Spring 2015. He is contributing to five other publications currently under development, one of which is on Bayesian method for establishing biosimilarity for follow-on biological products.

Mr. Wu recently concluded a 1.5 year stint at Prometrika, a clinical trial contract research organization (CRO). He was involved in projects that required meta-analysis, including developing an objective performance criteria (OPC) for a bone stabilization system, a medical device, used to treat humeral fracture, and a meta-analysis of effect sizes of more than 50 medicinal products that use patient-reported outcomes (PRO) as evidence to show the effect of a test drug is comparable to the combined effect and to support its expanded indication.

Mr. Wu has a strong record of academic achievement, currently holding a cumulative GPA of 3.98 in his doctoral program; he also has great proficiency with computer programming, with substantial skills in R, SAS, and LaTeX. His research interests include adaptive methods in clinical trials, biosimilariy, Bayesian inference, computational biology, and analysis of microarray. His teaching experience includes a year of graduate-level biostatistics and SAS computing, as well as prior experience as the sole teaching assistant for a course on statistics and epidemiology.

EDUCATION

- **Boston University**, Boston, MA
  Ph.D. candidate in Biostatistics, 2014. Dissertation topic is "Adaptive methodologies in multi-arm dose response and biosimilarity clinical trials." Primary advisor: Dr. Mark Chang, Vice President of Biometrics, AMAG Pharmaceuticals, Inc. Cumulative GPA is 4.0.
- **Boston University**, Boston, MA
  M.A. in Biostatistics, 2011.
- **Trinity International University**, Deerfield, IL
  M.A. in Counseling Psychology, 2002.
- **The University of Hong Kong**, Hong Kong
  B.Soc.Sc in Economics/Statistics, Second Honor Division I, 1992

RESEARCH PROFESSIONAL EXPERIENCE

- **Biostatistician**, Prometrika, Cambridge, MA, May - Dec 2013
  Perform clinical trial data analysis and statistical programming in SAS. Produce or validate project deliverables such as tables, figures, and listings according to regulatory standards. Participate in writing statistical section of study protocols, Statistical Analysis Plans (SAP), and other study-related documents. Propose statistical analyses and corresponding table shells after reviewing protocols, SAPs, and clinical report forms (CRF).
- **Biostatistician Intern**, Prometrika, Cambridge, MA, Sep 2012 - Apr 2013
  Same as above.
- **Research Assistant**, Biostatistics Department, Boston University School of Public Health, Sep 2011 - Aug 2012
  Under the supervision of Biostatistics Professor (Adrienne Cupples, Ph.D.), conduct QC, meta-analyses, and other statistical analyses for (1) the GIANT WAIST project which was dedicated to finding genetic association for adiposity-related traits and (2) GWAS data from the Framingham Heart Study. Knowledge of statistical genetic applications such as METAL, PLINK 1.07 and LocusZoom.
- **IRB Intern**, Boston University Institutional Review Board, Jan - Feb 2010
  Under the direction and supervision of IRB Chair (James Feldman, M.D.) and Vice-Chair (Louis Vachon, M.D.), review several assigned research protocols and consent forms, particularly on the statistical analysis section and presented findings.
- **Biostatistics Intern**, Harvard Clinical Research Institute, Boston, MA, Jan - Jun 2010
  Under the direction and supervision of senior biostatistics consultant (Joseph Massaro, Ph.D.), conduct statistical analyses for assigned clinical trials, performed SAS programming, summarization and interpretation of results.
- **Research Assistant**, Family Medicine, Boston Medical Center, Boston, MA, Feb - Sep 2010
  Conduct statistical analyses on data generated by the Project RED, a series of randomized controlled trials, dedicated to studying the effect of a hospitalization discharge plan on hospital re-admission outcomes.

PUBLICATIONS

1. **Wu. J.**, Menon, S., Doros, G., Barker, K., Chang, M. (2015) "Bayesian hierarchical bias model for establishing biosimilarity using composite endpoint." (To be submitted to Journal of Biopharmaceutical Statistics)
2. **Wu, J.**, Menon, S., Chang, M. (2014). "An adaptive staggered dose design for a normal endpoint." (In press for the Journal of Biopharmaceutical Statistics)
3. **Wu, J.**, Menon, S., Chang, M. "Fitting the dose - adaptive staggered dose design." In Clinical and Statistical Considerations of Personalized Medicine, Carini, C., Menon, S., Chang, M. (ed.). Chapman & Hall/CRC Press, 2014.
4. Stallwood, C.G., Shergill, K.S., **Wu, J.**, et al. (2014). "Phenotypic manifestations of inflammatory bowel disease in patients of Haitian and Cape Verdean Descent." (Under review at Journal of Crohn's and Colitis)
5. Shungin, D., Winkler, T.W., Croteau-Chonka, D.C., Ferreira, T., Locke, A.E., Magi, R., Strawbridge, R.J., Pers, T.H., Fischer, K., Justice, A.E., Workalemahu, T., **Wu, J.**, et al. (2013). "New genetic loci link adipocyte and insulin biology to body fat distribution." (Under review by Nature)

6. **Wu, J.**, Gupta, M., Gerstenfeld, L. "Bayesian approach to modeling factorial time-course microarray data with an application to bone aging." (To be submitted to Annal of Applied Statistics)
7. Lo, L., **Wu, J.** "Chinese translated IEP: Do they do more harm than good?" July, 2010 AAPI Nexus.
8. Foley, SM, Raphael, R, Adolphe, M, **Wu, J.**, Tamene, B, Leung, C, Yusuf, A. "Four Boston Area Community Perspectives with Research: Haitian, Chinese, Ethiopian, and Asian/Pacific Islander perspectives on knowledge creation." Dec, 2010, Journal of American Academy of Pediatrics

TEACHING EXPERIENCE

- **Instructor**, Department of Biostatistics, Boston University School of Public Health, Fall 2013 & Spring 2014
  BS723 Introduction of Statistical Computing. Give lectures on introductory SAS and biostatistics. Manage graders in grading homework assignments and coordinate examinations. Meet with students. Class size of about 20.
- **Teaching Assistant**, Department of Biostatistics, Boston University School of Public Health, Spring 2012
  BS852 Statistical Methods in Epidemiology. Assist Biostatistics Professor (Paola Sebastiani, Ph.D.) in teaching. Class size was 30. Responsible for grading homework assignments and final exam, holding office hour, and answering studentsÕ questions.

HONORS & AWARDS

- **Winner of Student Paper Competition**, Department of Biostatistics, Boston University School of Public Health, Spring 2014
  "An adaptive staggered dose design for a normal endpoint"
- **Mu Sigma Rho**, Boston Chapter, Inducted since May 4, 2010
- **Dean's List**, Boston University, 2009 - 2011
- **Interdisciplinary Training in Biostatistics Program**, Department of Biostatistics, Boston University School of Public Health, Sep 2009 - May 2011

SKILLS

- Major statistical programming: SAS and R (proficient in both PC and UNIX environments)
- Other statistical and scientific softwares: LaTeX, WinBUGS, nQuery, STATA, SPSS, and METAL
- Language proficiency: English, Cantonese, and Mandarin

NON-RESEARCH PROFESSIONAL EXPERIENCE & CREDENTIALS

- **Licensed Mental Health Counselor (LMHC)**, Commonwealth of Massachusetts, Sep 2004 - Dec 2010
- **National Certified Counselor (NCC)**, National Board for Certified Counselors, Jun 2005 - Aug 2010

- **Family Services Director**, Boston Chinatown Neighborhood Center, Boston, MA, Apr 2008 - Jan 2010
  Under the direction of the Agency Director, I was responsible for developing and implementing a center-based program of Family Services. I coordinated clinical case management for families and children with special needs, provided assessment, counseling, and brief therapy, and organized educational activities for adults and children. In addition, I provided support to research initiatives related to family health, organized and conducted trainings for staff in areas of health, disabilities, and other family issues, and provided consultation to all agency staff. My administrative duties included supervision of two program part time staff, interns, and volunteers, and writing grant proposal and reports. In 3 years, the program had reached out to more than 300 families, and was featured in an article in the Boston Globe in 2008.
- **Family Services Coordinator**, Boston Chinatown Neighborhood Center, Boston, MA, Sep 2005 - Mar 2008
  Same as above.
- **Mental Health Clinician**, South Cove Community Health Center, Boston, MA, Jun 2002 - Aug 2005
  Provided clinical mental health services including individual, group, and family therapy. Conducted intakes and assessments, furnished diagnostic, treatment, and outcome reports, and provided multi-cultural counseling, mainly to Asian population but also to other cultural groups.

REFERENCES

- **Mark Chang, Ph.D.**, Vice President, Biometrics, AMAG Pharmaceuticals, Inc., Waltham, MA
- **Sandeep Menon, Ph.D.**, Executive Director and Head of Biotx R&D, Biostatistics, Pfizer Inc., Cambridge, MA
- **Mayetri Gupta, Ph.D.** Instructor, University of Glasgow, Scotland, UK
- **Adrienne Cupples, Ph.D.**, Professor Emeritus, Boston University, MA
- **Varsha Ghosh**, Program Director, Public Network Program, Harvard University, Cambridge, MA