MULTIPLICITY ADJUSTMENTS IN ADAPTIVE DESIGN

A Dissertation
Submitted to
the Temple University Graduate Board

---

in Partial Fulfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

---

by
Jingjing Chen
May, 2012

Examining Committee Members:

Sanat K. Sakar, Advisory Chair, Statistics
Damaraju Raghavarao, Statistics
Milton Parnes, Statistics
Jichun Xie, Statistics
Devan Mehrotra, External Member, Merck & Co., Inc.

ABSTRACT

MULTIPLICITY ADJUSTMENTS IN ADAPTIVE DESIGN

Jingjing Chen

DOCTOR OF PHILOSOPHY

Temple University, May, 2012

Professor Sanat K. Sarkar, Chair

There are a number of available statistical methods for adaptive designs, among which the combination method of Bauer and Köhne's (1994) is well known and widely used. In this work, we revisit the the Bauer-Köhne method in three ways: overall FWER control for single-hypothesis in a two-stage adaptive design, overall FWER control for two-hypothesis in a two-stage adaptive design, and overall FDR control for multiple-hypothesis in a two-stage adaptive design.

We first take the Bauer-Köhne method in a more direct manner to have more flexibility in the choice of the early rejection and acceptance boundaries as well as the second stage critical value based on the chosen combination function. Our goal is not to develop a new method, but focus primarily on developing a comprehensive understanding of two-stage designs. Rather than tieing up the early rejection and acceptance boundaries by considering the second stage critical value to be the same as that of the level α combination test, as done in the original Bauer-Köhne method, we allow the second-stage critical value to be determined

from prefixed early rejection and acceptance boundaries. An explicit formula is derived for the overall Type I error probability to determine the second stage critical value from these stopping boundaries not only for Fisher's combination function but also for other types of combination function. Tables of critical values corresponding to several different choices of early rejection and acceptance boundaries and these combination functions are presented. A dataset from a clinical study is used to apply the different methods based on directly computed second stage critical values from pre-fixed stopping boundaries and discuss the outcomes in relation to those produced by the original Bauer-Köhne method.

We then extend the Bauer-Köhne method to two-hypothesis setting and propose a stepwise-combination method for a two-stage adaptive design. In particular, we modify Holm's stepdown procedure (1979) and suggest a stepdown-combination method to control the overall FWER at a desired level α.

In many scientific studies requiring simultaneous testing of multiple null hypotheses, it is often necessary to carry out the multiple testing in two stages to decide which of the hypotheses can be rejected or accepted at the first stage and which should be followed up for further testing having combined their p-values from both stages. Unfortunately, no multiple testing procedure is available yet to perform this task meeting pre-specified boundaries on the first-stage p-values in terms of the false discovery rate (FDR) and maintaining a control over the overall FDR at a desired level. Our third goal in this work is to present two procedures, extending the classical Benjamini-Hochberg (BH) procedure and its adaptive version incor-

porating an estimate of the number of true null hypotheses from single-stage to a two-stage setting. These procedures are theoretically proved to control the overall FDR when the pairs of first- and second-stage p-values are independent and those corresponding to the null hypotheses are identically distributed as a pair (p_1, p_2) satisfying the p-clud property of Brannath, Posch and Bauer (2002, *Journal of the American Statistical Association*, 97, 236 -244). We consider two types of combination function, Fisher's and Simes', and present explicit formulas involving these functions towards carrying out the proposed procedures based on pre-determined critical values or through estimated FDR's. Simulations were carried to compare the proposed methods with class BH procedure using first stage data only and full data from both stages respectively. Our simulation studies indicate that the proposed procedures can have significant power improvement over the single-stage BH procedure based on the first stage data, at least under independence, and can continue to control the FDR under some dependence situations. Application of the proposed procedures to a real gene expression data set produces more discoveries compared to the single-stage BH procedure using the first stage data and full data as well.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to all those who have contributed to the work described in this thesis.

First and foremost, I am deeply indebted to my advisor, Dr. Sanat Sarkar, for his guidance and insight throughout this research. He helped me come up with this topic and guided me through the development. His stimulating suggestions, invaluable hints, and forever encouragement helped me in all the time of research for and writing of this thesis.

My sincere appreciation goes to Dr. Wenge Guo for numerous fruitful discussions and offering suggestions on the simulations for improvement. I also extend my thanks to all other faculty members in my committee, Dr. Damaraju Raghavarao, Dr. Milton Parnes, Dr. Jichun Xie, and Dr. Devan Mehrotra. I would like to thank them for all their support, interest, and editing remarks.

Special thanks go to my colleagues and friends, Gengqian Cai, Li He, Fang Liu, and Zijiang Yang, for being of great help in various ways in difficult times.

My heartfelt gratitude also goes to my supervisor at Octagon Research Solutions, Dr. Jeff Davidson, for his ultimate support and endless goodwill.

Last but not least, my gratitude is extended to my parents in China for their understanding, patience, motivation, and encouragement. Their consistent support has been unconditional all these years. They have cherished with me every great moment and supported me whenever I needed it. Their love enabled me to complete this work. I also would like to give my special thanks to my son Ryan for

being such a nice and sweet little boy during the past years of this research.

To my son:

Ryan

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Drug development in today's world has become increasingly costly and challenging. One white paper released by the US Food and Drug Administration (FDA) states that if the drug development processes do not become more efficient and effective, innovation may continue to stagnate and the biomedical revolution may fail to achieve its full potential (FDA, 2004). There are many ways that statistics and biometrics in general can contribute to improve the drug development cycle (Posch et al., 2005; Bretz et al., 2006, 2009). Considered as a potential mechanism for improving the development efficiency, adaptive design appears to evidence as one innovative statistical approach worthy of investigation. Adaptive designs are also seen favorably by regulatory agencies, if performed with care (CHMP, 2007; FDA, 2010).

Classical drug development consists of a sequence of independent trials. Adaptive design aims at interweaving these trials by combining them into one single

study conducted in two or more stages (Posch et al., 2005; Bretz et al., 2006, 2009). Thus, adaptive design is considered to have the potential to improve the performance of the trial (Bauer and Einfalt, 2006). Compared to other designs, the horizon of the adaptive design is clear because any design change is possible. The advantage of such a design is to facilitate the process of drug development by allocating resources more efficiently without lowering regulatory standards. Such an approach would provide flexibility in efficiently conducting clinical trials by reducing the decision-making time during drug development and saving cost through the combination of evidence across studies.

During the past decades, adaptive designs have received much attention in the literature, and there are numerous statistical methods developed that theoretically handle adaptive design, e.g. Fisher's combination test (Bauer and Köhne, 1994), the conditional error approach (Proschan and Hunsberger, 1995), bias-adjusted Proschan and Hunsberger method (Denn, 2000), the weighted statistic approach (Lehmacher and Wassmer, 1999; Cui et al., 1999), the "self-designing" and "variance spending" method (Fisher, 1998; Shen and Fisher, 1999), multistage adaptive design (Müller and Schäffer, 2001; Brannath et al., 2002), the likelihood ratio test approach (Li et al., 2002), the more recent work by Bartroff and Lai (2008), etc. Liu and Chi (2001) also gave a family of conditional error function in a different context. It has been addressed that these approaches are interrelated (Posch and Bauer, 1999; Wassmer, 2000; Bauer et al., 2001; Jennison and Turnbull, 2003, 2005). From the statistical point of view, essentially these methods can be categorized in-

to two major concepts: the combination test principle (Bauer and Köhne, 1994) and the conditional error principle (Proschan and Hunsberger, 1995). In fact, the conditional error function approach can be looked at in terms of combination tests and vice versa (Jennison and Turnbull, 2005). According to Bauer and Einfalt's review (2006) regarding the application of adaptive design, the most widely used methodology is the combination approach of Bauer and Köhne based on Fisher's combination test for independent p-values (Bauer and Köhne,1994), followed by the weighted inverse approach by Lehmacher and Wassmer (1999), and the conditional error function approach by Proschan and Hunsberger (1995).

Since the work of Bauer and Köhne (1994), the general combination principle has gained a lot of attention with regards of determining the early stopping boundaries in adaptive designs without compromising the overall FWER (i.e., Bauer and Röhmel,1995; Bauer and Kieser, 1999; Kieser et al., 1999; Hommel, 2001; Posch and Bauer, 2000; Brannath et al., 2002). In this work, we revisit the Bauer-Köhne method for a two-stage adaptive design with independent p-values. It is important to point out the following feature of this method. The early rejection and acceptance boundaries, α_L and α_U respectively, with the overall Type I error rate controlled at α, are tied up through the equation $\alpha = \alpha_L + c_\alpha(\ln \alpha_U - \ln \alpha_L)$, with c_α, the second stage critical value, chosen to be the same as that for the level α Fisher's combination test, that is, $c_\alpha = \exp\{-\frac{1}{2}\chi^2_{4;1-\alpha}\}$,where $\chi^2_{\nu;1-\alpha}$ denotes the $(1-\alpha)$-quantile of the χ^2distribution with ν degrees of freedom, and the restriction $c_\alpha \leq \alpha_L$. Thus, α_L and α_U are not being allowed to be pre-determined independently of each other

before proceeding to the second stage. For example, suppose that we like to control the overall Type I error at $\alpha = 0.05$ and desire to set the early rejection boundary at 0.010. The second stage critical value c_α will be 0.0087 and we are constrained to choose the early acceptance boundary as 0.9926. It means practically no futility is allowed under this design. On the contrary, if we like to have the early acceptance boundary to be 0.2 at Stage 1, then the corresponding early rejection boundary has to be 0.035, which is more than half of the overall Type I error rate α and considerably greater than the original early rejection boundary (0.010) that we desired.

The above idea in Bauer and Köhne of determining the early rejection and acceptance boundaries by pre-fixing the second stage critical value, a roundabout way of determining these boundaries, seems to defeat the main purpose of using adaptive design, which is to have the flexibility in the choice of the early stopping boundaries for both rejection and acceptance of the null hypothesis. Often these boundaries are pre-chosen, to meet some efficiency requirements, before the second stage critical value is determined; or one might want to have a general idea of how these boundaries can influence the second stage critical value before making a judicious choice of these boundaries as well as the second stage critical value. Thus, while considering Bauer and Köhne's general combination test principle in a two-stage adaptive design, it is often the case that the early rejection and acceptance boundaries are pre-fixed, and the second stage critical value is to be determined based on these boundaries subject to a control of the overall type I error probability.

This is the kind of situation we are considering in the first part of this work.

Similar to all other clinical trials, with the multi-stage data, multiplicity of inferences is definitely a concern for adaptive designs, specially when there is more than one hypothesis to be tested within each stage. For example, in a two-stage seamless phase II/III trial, several treatments are evaluated at the Stage 1, and one (or more) treatment (s) can be selected after the first stage at the interim, and then investigated further in the second stage. If multiplicity is not properly handled, unsubstantiated claims for the effectiveness of a drug may be made as a consequence of an inflated rate of false positive conclusions. Usually, the more objectives a clinical trial is set to achieve, the more complex the statistical methods are to safeguard the Type I error at a desired rate. In fact, there have been critical comments on the use of adaptive designs, such as protection of Type I error, preserving conditional Type I error, flexibility and credibility, etc. For instance, it is unclear whether validity of multiplicity adjustment still holds, how to interpret significance with respect to multiple responses, when is adjustment of multiplicity necessary, how should composite endpoints be handled statistically with respect to regulatory claims, how to best analyze important secondary endpoints after the primary endpoint is found to be positive, how to incorporate design efficiency consideration into entire drug development program, etc. (Koch, 2006). Thus, the second purpose of this research to extend Bauer and Köhne's general combination test principle to the statistical testing with two hypotheses and propose a stepwise-combination method for a two-stage adaptive design. In particular, we modify Holm's stepdown procedure (1979),

and suggest a stepdown-combination method to control the overall Type I error at a desired level α.

Furthermore, in the field of genomics, gene association or expression studies usually involve a large number of endpoints (i.e., genetic markers) and are often quite expensive. Multi-stage adaptive design with its feature of being cost effective and efficient, since it allows genes being screened in early stages and selected genes being further investigated in later stages using additional observations, has become more and more attractive in such genetic studies. To address the multiplicity concern in simultaneous testing of the hypotheses associated with the endpoints, controlling the FWER, the probability of at least one Type I error among all hypotheses, is a commonly applied concept. However, these studies are often exploratory, so controlling the false discovery rate (FDR), which is the expected proportion of Type I errors among all rejected hypotheses, is more appropriate than controlling the FWER (Weller et al., 1998; Benjamini and Hochberg, 1995; and Storey and Tibshirani, 2003). Moreover, with tens of thousands of hypotheses typically being tested in these studies, better power can be achieved in a multiple testing method under the FDR framework than under the more conservative FWER framework.

Construction of methods with the FDR control in the setting of a two-stage adaptive design allowing reduction in the number of tested hypotheses at the interim analysis does not seem to be a simple extension of standard FDR controlling methods in single-stage designs, like the BH (Benjamini and Hochberg, 1995) or methods related to it, from a single-stage to a two-stage design setting. Thus, our third goal

in this dissertation is to propose two BH type methods to control the FDR in a two-stage adaptive design with combination tests for multiple endpoints, one extending the original single-stage BH procedure, which we call the BH-TSADC Procedure (BH type procedure for two-stage adaptive design with combination tests), and the other extending an adaptive version of the single-stage BH procedure incorporating an estimate of the number of true null hypotheses, which we call the Plug-In BH-TSADC Procedure, from single-stage to a two-stage setting.

The rest of this dissertation is organized as follows. Chapter 2 briefly reviews the basic concepts and the current available statistical methods regarding data analysis in adaptive design. The general concepts of the multiple testing techniques and the closure principle are introduced in this chapter as well. In Chapter 3, we give an explicit formula for the overall Type I error probability in terms of early rejection and acceptance boundaries and the corresponding second stage critical value for each of Fisher's, Tippett's and Simes' combination functions for single hypothesis test in a two-stage combination test. Based on these formulas, we numerically compute the critical values for these combination functions having chosen some pairs of early rejection and acceptance boundaries and values of α, and present them in Tables. We then extend the Bauer-Köhne's combination approach to the two-hypothesis testing environment and suggest a stepwise-combination testing procedure to safeguard the overall Type I error in Chapter 4. We also propose two BH type procedures to control the FDR in a two-stage adaptive design with combination tests for multiple endpoints, extending the original BH method and its adaptive version incorporating

an estimate of the number of true null hypotheses from single-stage to a two-stage setting in Chapter 5. We attempt to use straightforward statistical ideas in a two-stage setting by prefixing the early rejection and early acceptance boundaries and then estimating the second stage critical values, while maintaining a strong control of the overall FDR. The proposed procedures are illustrated with simulation results and real data applications. Chapter 6 completes the dissertation with some final comments and a brief discussion on future research.

For simplicity and to avoid problems with conflicting directional decisions, we assume a two-stage adaptive design and one-sided hypothesis test throughout this work.

CHAPTER 2

LITERATURE REVIEW

In this chapter, we first introduce some basic concepts in adaptive design and multiple hypothesis testing techniques. We then review the current available statistical procedures regarding data analysis and the overall Type I error control in adaptive designs. We also investigate multiple testing techniques in adaptive designs while applying the closure principle.

2.1 Some Basics in Adaptive Design

A study design is called "adaptive" if the statistical methodology allows the modification of a design element (e.g. sample-size, randomization ratio, change or modification of endpoints, discontinuing treatment arms, etc.) at an interim analysis with full control of the Type I error (CHMP, 2007). Adaptive design uses accumulating data to decide on how to modify aspects of the study without undermining the validity and integrity of the trial. To maintain study validity means

providing correct statistical inference such as adjusted p-values, unbiased estimates and adjusted confidence intervals, etc., assuring consistency between different stages of the study, and minimizing operational bias. To maintain study integrity means providing convincing results to a broader scientific community, preplanning, based on intended adaptations, and maintaining the blind of interim analysis results.

Wald (1947) pioneered sequential analysis in 1947. Armitage (1957, 1975) first adopted it to the field of clinical trials. Pocock (1977) and O'Brien and Fleming (1979) introduced the group sequential test, which was considered more practical than the pure sequential test. Lan and DeMets (1983) proposed a more flexible approach with the alpha-spending function. On the basis of sequential analysis and group sequential analysis, adaptive design was initiated by Bauer (1989), who demonstrated the adaptive design was superior compared to the classic group sequential designs as adaptive design provided the potential for substantial data-driven re-design (Hellmich and Hommel, 2004). For example, the modification of adaptive randomization to achieve balance within strata, sample size re-estimation, early stopping due to efficacy or futility, dropping inferior treatment groups, and change of treatments, patient population, hypotheses (non-inferiority $\rightarrow$ superiority, superiority $\rightarrow$ non-inferiority) or the order of hypotheses. Using flexible designs implies that the statistical methods control the pre-specified Type I error, and the estimate of homogeneity of results from different stages are pre-planned (CHMP, 2007). In other words, adaptive design has become attractive, because it allows modifications to some aspects of the trial after its initiation without undermining the trial's

validity and integrity (Chang, 2005).

2.1.1 Adaptation Rules

The real merit of adaptive design is adaptations going beyond sample size modification. In general, the adaptation rules include but not limited to:

1. Randomization rules. It is desirable to randomize more patients to superior treatment groups, which can be achieved by increasing increasing the probability of assigning a patient to the treatment group when the evidence of responsive rate increases in a group (Chang, 2005; Rosenberger and Lachin, 2002).

2. Early stopping rules. It is desirable to stop trial when the efficacy or futility of the test drug becomes obvious during the trial.

3. Dropping loser rules. One can improve the efficiency of a trial by dropping some inferior groups during the trial.

4. Sample size adjustment rules. It is desirable to adjust the sample size according to the effect size of an ongoing trial.

2.1.2 Types of Adaptive Design

Based on the level of flexibility, adaptive design can be categorized into three classes: rigid, totally flexible, or partially flexible.

1. Rigid Adaptive Designs. The scope of possible adaptations and decisions are pre-specified up front in the protocol (PhRMA, 2006). The advantages of rigid

adaptive designs include that logistical problems such as changing treatments, patient eligibility, and accrual rates can be planned for in advance; there is no need to file protocol amendments; final analysis can be made to depend on sufficient statistics; sample space can be statistically identified. However, rigid adaptive designs are not able to respond to unexpected circumstances during a long-term trial, and information about progress of the trial is more easily inferred.

2. Totally Flexible Adaptive Designs. Unplanned design modifications can be made at unplanned interim analyses, which is also called "Self-designing trials" (Fisher, 1998). The advantage of totally flexible adaptive designs is the ultimate flexibility. However, ad hoc design modifications based on unblinded interim results can lead to loss of credibility, and it requires use of unfamiliar test statistics which can be a source of inefficiency (Jennison and Turnbull, 2003, 2006), and can lead to possible anomalous results (Burman and Sonesson, 2006). Also, point and interval estimation may be problematic.

3. Partially Flexible Adaptive Designs. Partially flexible adaptive design is a compromise. Let's take Bauer and Köhne's two-stage design (1994) as an example. The design and length of Stage 1 are fixed in advance. The design of Stage 2 is permitted to depend on Stage 1 results in an arbitrary and unplanned way. Final inference must be based on the p-values from the two stages according to a rule specified in advance. By applying the method recursively, multistage designs can be constructed.

Based on the adaptations employed, adaptive design in clinical trials can be categorized into but not limited to the following(Chow, 2008):

1. Adaptive Randomization Design. An adaptive randomization design allows modification of randomization schedules based on varied probabilities of treatment assignment in order to increase the probability of success. Although an adaptive randomization design could increase the probability of success, it may not be feasible for a large trial or a trial with a relatively long treatment duration because the randomization of a given subject depends on the response of the previous subject.

2. Group Sequential Design. A group sequential design allows for prematurely stopping a trial due to safety, futility, or efficacy based on interim analysis results. The stopping boundaries are obtained based on different boundary functions for the control of Type I error rate (Lan and DeMets, 1987; Wang and Tsiatis, 1987; Rosenberger et al., 2001; Jennison and Turnbull, 2000, 2005; Chow and Chang, 2006). The concept of two-stage adaptive design has led to the development of the adaptive group sequential design (Cui et al., 1999; Posch and Bauer, 1999; Lehmacher and Wassmer, 1999; Liu et al., 2002).

3. Sample Size Re-estimation Design. A sample size re-estimation design allows for sample size adjustment or re-estimation based on interim analysis results. It should be noted that the observed difference at interim based on a small number of subjects may not be statistically significant.

4. Drop-the-Losers Design. A drop-the-losers design allows dropping the inferior treatment groups or adding additional arms. A drop-the-losers design is useful in phase II clinical development especially when there are uncertainties regarding the dose levels (Bauer and Kieser,1999; Brannath et al., 2003; Sampson and Sill, 2005; Posch et al., 2005).

5. Adaptive Dose Finding Design. An adaptive dose finding design is often used to identify the minimum effective dose or the maximum tolerable dose for future clinical trials in early phase clinical development (Bauer and Röhmel, 1995; Whitehead, 1997; Zhang et al.,2006). A Bayesian approach is usually considered in this kind of study (O'Quigley et al., 1990; O'Quigley and Shen, 1996; Chang and Chow, 2005).

6. Biomarker Adaptive Design. A biomarker adaptive design allows for adaptations based on the response of biomarkers. It involves biomarker qualification and standard, optimal screening design, and model selection and validation.

7. Adaptive Treatment-Switching Design. An adaptive treatment-switching design allows the investigator to switch a patient's treatment from an initial assignment to an alternative treatment if there is evidence of lack of efficacy or safety of the initial treatment. However, a high percentage of subjects switching treatment due to disease progression could lead to change in hypotheses, especially in oncology clinical trials. In this case, estimation of survival could be a challenge and sample size adjustment for achieving a desired power may be necessary.

8. Adaptive Hypotheses Design. An adaptive hypotheses design allows modifications in hypotheses based on interim analysis results (Hommel, 2001). For example, the hypothesis may switch from superiority to non-inferiority, or switch between the primary endpoint and the secondary endpoints.

9. Adaptive Seamless Phase II/III Design. An adaptive seamless phase II/III trial design addresses objectives that are normally achieved through separate trials in phase IIb and phase III of clinical development within one single trial (see Figure 2.1). It is a two-stage design consisting of a learning stage (phase IIb) and a confirmatory stage (phase III). A typical approach is to power the study for the phase III confirmatory phase and obtain valuable information with certain assurance using confidence interval approach at the phase II learning stage. An adaptive seamless phase II/III design uses data from patients enrolled before and after the adaptation in the final analysis (Kelly et al., 2005; Maca et al., 2006). However, its validity and efficiency has been challenged (Tsiatis, 2003). Further, it is unclear how to perform a combined analysis if the study objectives are different at different phases (Chow et al., 2007) .

10. Multiple Adaptive Design. A multiple adaptive design combines any of the above adaptive designs. However, the statistical inference for a multiple-adaptation design is often difficult in practice.

In summary, adaptive design is all about flexibility and this flexility comes from careful statistical planning. Hence, adaptive design comes at a price of efficien-

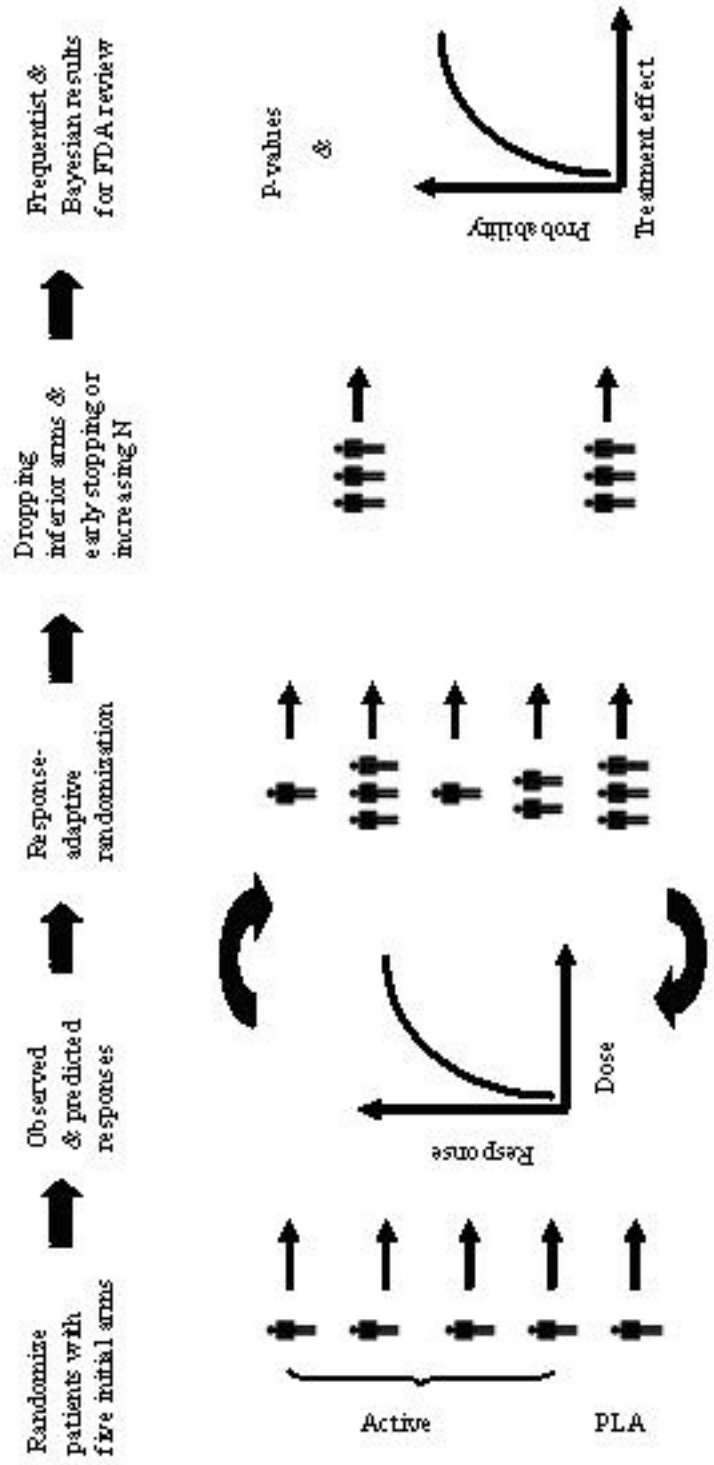

Figure 2.1: Phase II/III seamless trial design. (Chow and Chang, 2008).

cy, careful design evaluation, and scientific interpretation. During the past decades, adaptive designs have received much attention in the literature, and there are numerous statistical methods developed that theoretically handle adaptive design, e.g. Fisher's combination test (Bauer and Köhne, 1994), the conditional error approach (Proschan and Hunsberger, 1995), bias-adjusted Proschan and Hunsberger method (Denn, 2000), the weighted statistic approach (Lehmacher and Wassmer, 1999; Cui et al., 1999), the "self-designing" and "variance spending" method (Fisher, 1998; Shen and Fisher, 1999), multistage adaptive design (Müller and Schäffer, 2001; Brannath et al., 2002), the likelihood ratio test approach (Li et al., 2002), the more recent work by Bartroff and Lai (2008), etc. Liu and Chi (2001) also gave a family of conditional error function in a different context. It has been addressed that these approaches are interrelated (Posch and Bauer, 1999; Wassmer, 2000; Bauer et al., 2001; Jennison and Turnbull, 2003, 2005). In fact, the conditional error function approach can be looked at in terms of combination tests and vice versa (Jennison and Turnbull, 2005). More recently, adaptive designs have attracted interest to make drug development more efficient by interweaving a sequence of independent trials and combining them into one single study conducted in two or more stages (Posch et al., 2005; Bretz et al., 2006, 2009). Adaptive designs are also seen favorably by regulatory agencies, if performed with care (CHMP, 2007; FDA, 2010).

In fact, these methods can be categorized into two classes: combination test and conditional error function.

Combination test principle. The combination test principle uses stage-wise test

statistics which are combined according to a pre-defined combination function (Bauer, 1989; Bauer and Köhne, 1994).

Conditional error principle. The conditional error principle specifies that the conditional probability for a false rejection of the null hypothesis given that the previous stage is known. It states that any type of design modifications can be performed at any time of the trial as long as the conditional error of the new design does not exceed the conditional error of the pre-planned design (Proschan and Hunsberger, 1995; Müller and Schäfer, 2001, 2004).

2.2 Some Basics in Multiple Hypothesis Testing

Assume that we are testing m null hypotheses $H_1, ..., H_m$ simultaneously and denote by R the number of rejected hypotheses. Table 2.1 summarizes all possible outcomes in the frequentist setting (Benjamini and Hochberg, 1995). The specific m hypotheses are assumed to be known in advance. The number of true and false hypotheses m_0 and m_1 are unknown. R is an observable random variable, and S, T, U, and V are unobservable random variables. Specifically, R is the observed total number of rejections, A is the observed total number of acceptances, V is the number of false discoveries (Type I error), T is the number of false non-discoveries (Type II error), U is the number of correct acceptances and S is the number of correct rejections.

Table 2.1: Outcomes of simultaneously testing m hypotheses.

	Accepted	Rejected	Total
True Null	U	V	m_0
False Null	T	S	m_1
Total	A	R	m

2.2.1 Type-I Error Rates in Multiple Hypothesis Testing

The number of multiple testing procedures is fast growing. A fundamental issue of multiple testing is how to effectively control Type I error. There are a variety of measures of error rates in the multiple testing setting. The following listed error rates are the most commonly used (Hochberg and Tamhane, 1987).

Per-Family Error Rate (PFER)

The PFER is defined as the expected number of false rejections, i.e.,

$$PFER = E(V). \tag{2.1}$$

Per-Comparison Error Rate (PCER)

The PCER is defined as the expected proportion of false rejections, i.e.,

$$PCER = E(V)/m. \tag{2.2}$$

Familywise Error Rate (FWER)

The FWER is defined as the probability of at least one false rejection, i.e.,

$$FWER = Pr(V \geq 1). \tag{2.3}$$

The FWER has been the most widely used approach among these traditional error rates. Controlling the FWER is natural in the situation where even a single false rejection is a bad event (Gordon, 2007). However, in many applications, one might be willing to tolerate more than one false rejection. Thus, many researchers proposed alternative approaches to measure the error rates, such as the generalized FWER (Victor, 1982), the false discovery rate (Benjamini and Hochberg, 1995) and its generalization including the positive false discovery rate (Storey, 2003), the proportion of false positives (Fernando et al., 2004), the generalized false discovery rate (Sarkar, 2006, 2007), etc.

Generalized Familywise Error Rate (k-FWER)

This concept was introduced by Victor (1982) and reintroduced by Korn et al. (2004), Dudoit et al. (2004) and Lehmann and Romano (2005). Gordan (2007) gave the explicit formulas for the k-FWER. The k-FWER is defined as the probability of having at least k false rejections for a pre-specified integer k, i.e.,

$$k - FWER = Pr(V \geq k). \tag{2.4}$$

False Discovery Rate (FDR)

The FDR is defined as the expected proportion of false discoveries among all rejections. i.e.,

$$FDR = E(Q) = E(\frac{V}{R}|R > 0)Pr(R > 0). \tag{2.5}$$

where by definition

$$Q = \begin{cases} V/R, & \text{if } R > 0 \\ 0, & \text{if } R = 0. \end{cases}$$

Positive False Discovery Rate (pFDR).

The pFDR is defined as the expected proportion of false discoveries among all rejections given there is at least one rejection, i.e.,

$$pFDR = E(\frac{V}{R}|R > 0). \tag{2.6}$$

Proportion of False Positives (PFP)

The PFP is defined as the proportion of the expected false discoveries among all the expected rejection, i.e.,

$$PFP = \frac{E(V)}{E(R)}. \tag{2.7}$$

Generalized False Discovery Rate (k-FDR)

The k-FDR is defined as the expected proportion of k or more false discoveries among all rejections, where k is pre-specified, i.e.,

$$k - FDR = \begin{cases} V/R, & \text{if } V \geq K \\ 0, & \text{if } V < K. \end{cases} \tag{2.8}$$

From the definitions given above, it is easy to see that for a given multiple testing procedure, $PCER \leq FDR \leq FWER \leq PFER$. Under the complete null hypothesis, the PCER is the average of the $\alpha_i, i = 1, ..., m, PCER = (\alpha_1 + ... + \alpha_m)/m$. The PFER is the sum of $\alpha_i, PFER = \alpha_1 + ... + \alpha_m$. The FWER and the FDR are the functions not of α_i alone, but involves the joint distribution of the test statistics T_i.

2.2.2 Strong Control versus Weak Control

Strong Control refers to the control of Type I error rate under any configuration of true and false null hypotheses. In contrast, *Weak Control* refers to the control of the Type I error rate only when all the null hypotheses are assumed to be true, i.e., under the complete null hypothesis $H_0 = \bigcap_{i=1}^{m} H_i$ with $m_0 = m$. For the FWER, weak control means control of $Pr(V \geq 1 | H_0^C)$, whereas strong control means control of $\max_{\Lambda_0 \subseteq \{1,...,m\}} Pr(V \geq 1 | \bigcap_{j \in \Lambda_0} H_j)$. In general, controlling a rate in a weak sense is unsatisfactory as it is not realistic that all the null hypotheses are true.

2.2.3 Types of Multiple Testing Procedure

Multiple testing procedures can be classified into two classes: single-step procedure and stepwise procedure. Let $p_1, ..., p_m$ denote the p-values corresponding to the null hypotheses $H_1, ...H_m$, respectively. Sort these p-values so that $p_{(1)} \leq \cdots \leq p_{(m)}$. Let α denote the overall Type I error. Given a set of critical values $\alpha_1 \leq \cdots \leq \alpha_m$ (refer to Sections 2.2.4 and 2.2.5 for more details),

Single-step Procedure. All the hypotheses are tested in one single step. Usually, there is only one critical values for all the hypotheses, i.e. Bonferroni procedure and Sidák procedure.

- Bonferroni Procedure. The Bonferroni procedure is one of the first used procedures to control the FWER in a strong sense when conducting multiple tests. The Bonferroni procedure rejects $H_0 = \bigcap_{i=1}^{m} H_i$, if $p_i \leq \alpha/m, i = 1, 2, ..., m$. The Bonferroni Inequality ensures that

$$Pr\{\bigcup_{i=1}^{m}(p_i \leq \alpha/m)\} \leq \alpha, (0 \leq \alpha \leq 1).$$

  The Bonferroni procedure requires no distributional assumptions. The downside of the Bonferroni procedure is that it is conservative and lacks of power if numerous highly correlated tests are undertaken.

- Sidák procedure. The Sidák procedure rejects $H_0 = \bigcap_{i=1}^{m} H_i$, if $p_i \leq 1 - (1-\alpha)^{1/m}, i = 1, 2, ..., m$. The Sidák procedure gives slightly smaller adjusted p-values than Bonferroni. It guarantees the strict control of the FWER only when the comparisons are independent.

- Simes' Procedure. The Simes' procedure is a modification of Bonferroni procedure (1986). The Simes' procedure rejects $H_0 = \bigcap_{i=1}^{m} H_i$, if the ordered p-value $p_{(i)} \leq i\alpha/m, i = 1, 2, ..., m$. Simes' procedure controls the Type I error rate with independent test statistics.

Stepwise Procedure. The hypotheses are tested in more than one step and usually they are tested sequentially.

- Step-down Procedure (SDP). The SDP is based on the ordered p-values and controls the FWER. Start with the most significant p-value $p_{(1)}$. The goal is to find $j = \min\{1 \leq i \leq m : p_{(i)} > \alpha_i\}$, then reject the hypotheses: $H_{(1)}, ..., H_{(j-1)}$. The SDP was originally proposed by Miller (1966) and then was widely used in the multiple testing problems. For example, Holm's procedure (1979) is a step-down procedure.
- Step-up Procedure (SUP). The SUP is based on the ordered p-values as well. Start with the least significant p-value $p_{(m)}$. The goal is to find $j = \max\{1 \leq i \leq m : p_{(i)} \leq \alpha_i\}$, and reject the hypotheses $H_{(1)}, ..., H_{(j)}$. The SUP is uniformly more powerful than the step-down procedure. For example, Hochberg's procedure (1988) is a step-up procedure.
- Generalized Step-up-down Procedure (SUDP). The SUDP is a generalization of the SDP and SUP. It was proposed by Tamhane et al. (1998) When the objective is to reject a specified minimum number out of a family of n hypotheses, the SUDP is specially useful. Start with $p_{(r)}, 1 \leq r \leq m$. If $p_{(r)} > \alpha_r$, accept $H_{(r)}, ..., H_{(m)}$ and continues test-

ing the remaining hypotheses in a step-up manner using corresponding p-values. On the contrary, if $p_{(r)} \leq \alpha_r$, reject $H_{(1)}, ..., H_{(r)}$ and continues testing the remaining hypotheses in a step-down manner.

A stepwise procedure reduces to a single step procedure when the critical values are all the same. In general, the stepwise procedures are more powerful than the single step procedures, as the stepwise procedures learn about the true configuration of the parameters and use this information in the following steps of test.

2.2.4 Procedures Controlling FWER

Bonferroni Procedure

The Bonferroni procedure is one of the first used procedures to control the FWER in a strong sense when conducting multiple tests. The Bonferroni procedure rejects $H_0 = \bigcap_{i=1}^{m} H_i$, if $p_i \leq \alpha/m, i = 1, 2, ..., m$. The Bonferroni Inequality ensures that

$$Pr\{\bigcup_{i=1}^{m}(p_i \leq \alpha/m)\} \leq \alpha, (0 \leq \alpha \leq 1).$$

The Bonferroni procedure requires no distributional assumptions. The downside of the Bonferroni procedure is that it is conservative and lacks of power if numerous highly correlated tests are undertaken.

Holm's Procedure

Holm (1979) proposed a more powerful sequentially rejective Bonferroni procedure. While the Bonferroni procedure is a single-step procedure, Holm's Procedure is shortcut version of step-down procedure constructed with closure method.

Start with $p_{(1)}$, if $p_{(1)} > \alpha/m$, accept all hypotheses $H_i, i = 1, ..., m$. If $p_{(1)} \leq \alpha/m$, reject $H_{(1)}$ and go to $p_{(2)}$ to check if $p_{(2)} > \alpha/(m-1)$. If it is true, accept all the remaining hypotheses. Otherwise, reject $H_{(2)}$ and go to $p_{(3)}$ and so on. In summary, Holm's procedure procedure is to find,

$$j = \min\{1 \leq i \leq n : p_{(i)} > \alpha/(m-i+1)\}.$$

If the minimum exists, accept all H_i with $i \geq j$ and reject the rest.

Simes' Procedure

Simes (1986) proposed another modification of Bonferroni Procedure for the test of overall null hypothesis $H_0 = \bigcap_{i=1}^{m} H_i$. Simes' procedure rejects H_0 if $p_{(i)} \leq i\alpha/m$ for any $i = 1, ..., m$.

Simes proved that this test controls the Type I error rate with independent test statistics. Based on a simulation study, he also conjectured that his procedure conservatively controls Type I error rate for a large variety of distributions when the test statistics are correlated. Sarkar and Chang (1997) and Sarkar (1998, 2008a) analytically proved Simes' conjecture for random variables with common marginal and Multivariate Totally Positive of Order 2 (MTP2) property. When the overall hypothesis H_0 is rejected, Simes also suggested the following rule to make statements about

individual hypothesis: reject $H_{(1)}, ..., H_{(j)}$, where $j = \max\{i : p_{(i)} \leq i\alpha/m\}$. However, the suggested procedure can only weakly control FWER and cannot strongly control FWER even for independent test statistics. Hommel (1988) extended Simes' suggestion and obtained a procedure controlling FWER strongly.

Hommel's Procedure

Simes' test was proposed for testing overall hypothesis $H_0 = \bigcap_{i=1}^{m} H_i$. When H_0 has been rejected, the question remains which of the individual hypotheses H_i, $i = 1, ..., m$ should be rejected. Hommel (1988) employed the closure principle to extend Simes' procedure for making statements on individual hypotheses. Hommel's procedure finds

$$j = \max\{1 \leq i \leq m : p_{(m-i+k)} > k\alpha/i\}.$$

for $k = 1, ..., i$. If the maximum does not exist, reject all $H_i, i = 1, ..., m$, otherwise reject all H_i with $p_i \leq \alpha/j$. Hommel's procedure can be expressed as applying Simes' procedure to each subset of hypotheses. It follows that this procedure controls the FWER provided each of Simes' tests for subsets is a level α test.

Hommel's procedure is at least as powerful as Holm's procedure. The computations for testing the individual hypotheses are very simple and can be performed also for a large n.

Hochberg Procedure

Hochberg (1988) proposed another modification of Bonferroni Procedure for multiple testing. Hochberg's procedure is a step-up procedure in terms of test statistics. Start with $p_{(m)}$, if $p_{(m)} \leq \alpha$, reject all hypotheses $H_i, i = 1, ..., m$. If $p_{(m)} > \alpha$, accept $H_{(m)}$ and go to $p_{(m-1)}$ to check if $p_{(m-1)} \leq \alpha/2$. If it is true, reject all the remaining hypotheses. Otherwise, accept $H_{(m-1)}$ and go to $p_{(m-2)}$ and so on. In summary, Hochberg's procedure is to find

$$j = \max\{1 \leq i \leq m : p_{(i)} \leq \alpha/(m - i + 1)\}.$$

If the maximum exists, reject all H_i with $i \leq j$ and accept the rest.

Hochberg's procedure uses the same critical values as that in Holm's procedure, but it is more powerful than Holm's procedure. Generally, with the same set of critical values, the step-up procedure rejects more hypotheses than the step-down procedure.

2.2.5 Procedures Controlling FDR

Benjamini-Hochberg (BH) Procedure

Benjamini and Hochberg (1995) proposed False Discovery Rate (FDR) as an alternative error rate to control. This approach to multiple testing is philosophically different from the classical approaches. Benjamini and Hochberg (1995) proposed a step-up procedure with FDR controlling property, which is known as Benjamini-Hochberg (BH) Procedure. The BH Procedure finds j such that

$$j = \max\{1 \leq i \leq m : p_{(i)} < i\alpha/m\},$$

rejects $H_{(1)}, ..., H_{(k_0)}$ if j exists, otherwise retain all null hypotheses.

The BH procedure is a step-up procedure and uses Simes' critical values as in Simes' procedure. Benjamini and Hochberg (1995) proved that FDR can be controlled at $m_0\alpha/n$ by BH Procedure for independent test statistics, where m_0 is the number of true null hypothesis. Benjamini and Yekutieli (2001) showed that the BH procedure indeed conservatively controls the FDR if the joint distribution of the test statistics is Positive Regression Dependent on the subset of test statistics corresponding to the true null hypotheses. Sarkar (2002) strengthened the work of Benjamini and Yekutieli (2001) by proving that Simes' critical values can be adopted in a generalized step-up-down procedure as proposed in Tamhane, Liu, and Dunnett (1998), and the FDR can still be controlled under similar dependency. Since its introduction, the BH procedure has been accepted widely for multiple testing purposes, especially when the total number of simultaneously tested hypotheses is large.

Genovese and Wasserman (2002) show that for large number of hypotheses n and with an independence assumption, the BH procedure can be equivalent to a single step procedure with an appropriate p-value cutoff which is between α and α/n.

Adaptive BH Procedure of Benjamini & Hochberg

Benjamini and Hochberg (2000) introduced an adaptive procedure for the original BH procedure with independent statistics based on an estimate of m_0 using the so called the Lowest Slope (LSL) method. When $m_0 < m$, the BH procedure is

conservative. This adaptive procedure utilizes the data to estimate m_0 within the family as $\hat{m}_0$ and then uses the adjusted critical values $(i\alpha/\hat{m}_0)$ in the BH procedure.

When all the hypotheses are true and the test statistics are independent, the set of observed p-values $p_{(i)}$'s can be considered as a realization of an ordered sample from the uniform distribution over $[0, 1]$. The expected value of $p_{(i)}$ is thus $i/(m+1)$. The plot of $p_{(i)}$ versus i should exhibit linear relationship, along a line of slope $S = 1/(m+1)$ passing through the origin and the point $(m+1, 1)$.

When $m_0 < m$, the p-values corresponding to the false null hypotheses tend to be smaller than those corresponding to the true null hypotheses, so they concentrate on the left side of the plot. The relationship over the right side of the plot remains approximately linear, with slope $\beta = 1/(m_0+1)$. Using a suitable set of the largest p-values, fit a straight line through the point $(m+1, 1)$ with slope $\hat{\beta}$, and use it to estimate m_0 by $\hat{m}_0 = 1/\hat{\beta}$. Benjamini and Hochberg (2000) suggested estimating m_0 using the LSL method and their adaptive procedure as follows:

1. Apply the original BH procedure. If none is rejected, then accept all hypotheses and stop; otherwise continue.

2. Calculate the slopes $S_i = (1 - p_{(i)})/(m + 1 - i)$.

3. Starting with $i = 1$, proceed as long as $S_i \geq S_{i-1}$ and stop when the first time $S_j < S_{j-1}$. Let $\hat{m}_0 = \min\{m, 1/S_j + 1\}$.

4. Apply the BH procedure with $\alpha_i = i\alpha/\hat{m}_0$.

Although there is no proof that this procedure controls FDR, simulation

study shows that the adaptive method controls FDR.

Adaptive BH Method of Storey, Taylor and Siegmund

Storey, Taylor and Siegmund (2004) modified Storey's (2002) original estimate $\widehat{\text{FDR}}_\lambda(t)$ of FDR(t), when $0 < \lambda < 1$, to

$$\widehat{\text{FDR}}_\lambda^{STS}(t) = \begin{cases} \frac{m\hat{\pi}_0^{STS}(\lambda)t}{\max\{R(t),1\}} & \text{if } t \leq \lambda, \\ 1 & \text{if } t > \lambda, \end{cases} \tag{2.9}$$

with

$$\hat{\pi}_0^{STS}(\lambda) = \frac{m - R(\lambda) + 1}{m(1-\lambda)},$$

and suggested thresholding the p-values based on this new estimate as follows:

$$t_\alpha(\widehat{FDR}_\lambda^{STS}) = \sup\{0 \leq t \leq 1 : \widehat{FDR}_\lambda^{STS}(t) \leq \alpha\}.$$

The adaptive BH method corresponding to this new estimate, to be called the STS method, rejects $H_{(1)}, \ldots, H_{(r)}$ where

$$r = \max\left\{0 \leq i \leq m : p_{(i)} \leq \min(\frac{i\alpha}{\hat{m}_0^{STS}}, \lambda)\right\}, \tag{2.10}$$

with

$$\hat{m}_0^{STS}(\lambda) = \frac{m - R(\lambda) + 1}{1 - \lambda}.$$

The STS controls the FDR under independence of the p-values (Benjamini, Krieger and Yekutieli, 2006; Storey, Taylor and Siegmund, 2004; Sarkar,2004, 2008a), as well as under certain form of weak dependence asymptotically as $m \to \infty$ (Storey, Taylor and Siegmund, 2004).

Adaptive BH Method of Benjamini, Krieger and Yekutieli (2006)

Unlike Storey (2002) or Storey, Taylor and Siegmund (2004) where m_0 is estimated based on the number of significant p-values observed in a single-step test with an arbitrary critical value λ, Benjamini, Krieger and Yekutieli (2006) considered estimating m_0 from the BH method at level $\alpha/(1+\alpha)$. Their adaptive version of the BH method, to be called the BKY method, runs as follows:

1. Apply the BH method at level $q = \frac{\alpha}{1+\alpha}$. Let r_1 be the number of rejections. If $r_1 = 0$, accept all the null hypotheses and stop; if $r_1 = m$, reject all the null hypotheses and stop; otherwise continue to the next step.

2. Estimate m_0 as
$$\hat{m}_0^{BKY} = \frac{m - r_1}{1 - q} = (m - r_1)(1 + \alpha).$$

3. Apply the BH method with the critical values $\alpha_i = i\alpha/\hat{m}_0^{BKY}$, $i = 1, \ldots, m$.

As Benjamini, Krieger and Yekutieli (2006) have proved, the BKY method controls the FDR at α under independence of the p-values. While it is less powerful than the adaptive procedure proposed in Storey et al.(2004) when the p-values are independent, simulation studies have shown that, with the p-values generated from multivariate normals with common positive correlations, it can also control the FDR. Benjamini, Krieger and Yekutieli (2006) also extended the BKY method to a multiple-stage procedure (MST) by repeating the two-stage procedure as long as more hypotheses are rejected, which is stated as follows:

1. Let $r = \max\{i : \text{ for all } j \leq i, \text{ there exists } l \geq j \text{ so that } p_{(l)} \leq \alpha l/[m+1-j(1-\alpha)]\}$.

2. If such an r exists, reject $p_{(1)}, \ldots, p_{(r)}$; otherwise reject no hypotheses.

This multiple-stage procedure is a combination of step-up and step-down methods. They offered no analytical proof of its FDR control. Benjamini, Krieger and Yekutieli (2006) also mentioned that a multiple-stage step-down procedure (MSD) can be developed by choosing $l = j$ in MST. They provided numerical results showing that the MST method can also control the FDR, the theoretical justification of which is given later in Gavrilov, Benjamini and Sarkar (2009) to be reviewed in the following section.

Adaptive Method of Gavrilov, Benjamini and Sarkar (2009)

As mentioned above, Gavrilov, Benjamini and Sarkar (2009) reexamined the multiple-stage step-down procedure, the MSD method, mentioned in Benjamini, Krieger and Yekutieli (2006) and proved that this multiple-stage step-down procedure can control the FDR under the independence of the p-values. The following is the MSD method:

Find $k = \max\{1 \leq i \leq m : p_{(j)} \leq j\alpha/(m+1-j(1-\alpha)) \text{ for all } j = 1, \ldots, i\}$ and reject $H_{(1)}, \ldots, H_{(k)}$ if k exists; otherwise reject no hypotheses.

Although it has been referred to as a multiple-stage stepdown method by Benjamini, Krieger and Yekutieli (2006), it is actually, as Sarkar (2008a) argued, an adaptive version of the stepdown analog of the BH method considered in Sarkar

(2002). To see this, first note that, under the same setup involving the mixture model and a constant rejection threshold t for each p-value as in Storey (2002) or Storey, Taylor and Siegmund (2004), one can consider estimating m_0 based on the number of significant p-values compared to the t, rather than a different arbitrary constant λ. In other words, by considering the Storey, Taylor and Siegmund (2004) type estimate of $m_0 = m\pi_0$ with $\lambda = t$ and using this estimate in $\widehat{\mathrm{FDR}}_\lambda(t)$, Storey's original estimate of the FDR(t), one can develop the following alternative estimate of FDR(t):

$$\widehat{\mathrm{FDR}}^*(t) = \frac{[m - R(t) + 1]t}{(1-t)\max\{R(t), 1\}}.$$

A step-down method developed through this estimate, that is, the one that rejects $H_{(1)}, \ldots, H_{(r)}$ where

$$\begin{aligned} r &= \max\left\{1 \le i \le m : \widehat{\mathrm{FDR}}^*(p_{(j)}) \le \alpha \text{ for all } j = 1, \ldots, i\right\} \\ &= \max\left\{1 \le i \le m : \frac{p_{(j)}}{1 - p_{(j)}} \le \frac{j\alpha}{m - j + 1} \text{ for all } j = 1, \ldots, i\right\}, \end{aligned} \tag{2.11}$$

which is the same as the MSD, is an adaptive version of the step-down analog of the BH method.

Simulation studies were conducted to compare the above three FDR controlling adaptive procedures, the BKY, MSD and STS. The STS is the most powerful one when the test statistics are independent, with the MSD taking the second place, although sometimes the power is very close to that of the STS. Under dependence, the BKY method is the only one that seems to control the FDR. The MSD in this

case also appears to perform well and its control over the FDR does not break down by much from the desired level.

2.3 Data Analysis in Adaptive Design with Single Hypothesis Test

Assume that we have a one-sided null hypothesis H_0 on the difference θ in mean efficacy of two treatments, i.e., $H_0 : \theta \leq \delta$, for some given δ, to be tested against $H_a : \theta > \delta$. We consider a two-stage design with a single interim analysis and assume that the same null hypothesis H_0 is being tested against the same alternative throughout Stage 1 and Stage 2. Let p_i be the p-value of the test of H_0 at Stage $i = 1, 2$, $C(p_1, p_2)$ be a combination function, α_L and α_U be Stage 1 early rejection and acceptance boundaries, respectively, and c_α be the second-stage critical value. Then, a two-stage adaptive test is described as follows:

1. Define a test procedure for Stage 1, determining the stopping rules for the interim decision.

2. Conduct Stage 1 of the study, resulting in p_1.

3. Based on p_1, decide whether to stop at the interim (either reject or accept H) or to continue the study to the next stage.

4. If the study is continued, resulting in p_2.

2.3.1 p-value Combination Function Approach

Bauer and Köhne (1994) proposed a principle to combine p-values from separate stages for sequential adaptive tests. It assumes that when continuous test statistics are applied, under H_0 the p-values in a stochastically independent sample are generally uniformly distributed on $[0, 1]$. The resultant p-value in the later stage is stochastically independent of the previous one. Given H_0 is true, data-dependence does not change the independence and distribution of p-values. For the final analysis, a two-stage combination test (Bauer, 1989; Bauer and Köhne, 1994; Bauer and Kieser, 1999) is defined by a combination function $C(p_1, p_2)$ which is monotonically increasing in both arguments, early stopping boundaries α_L and α_U, and a critical value c_α. Then, with Stage 1 early rejection and acceptance boundaries α_L and α_U, respectively, and the second stage critical value c_α, all to be determined subject to a control of the overall Type I error at α, a two-stage adaptive test is described as follows: (1) Stop at the interim with a decision to reject H_0 if $p_1 \leq \alpha_L$, to accept H_0 if $p_1 > \alpha_U$, or to continue to Stage 2 if $\alpha_L < p_1 \leq \alpha_U$. (2) If continued to Stage 2, determine $C(p_1, p_2)$ combining p_2 with p_1 and reject H_0 if $C(p_1, p_2) \leq c_\alpha$, otherwise, accept it. The overall Type I error rat is given by

$$\alpha_L + \int_{\alpha_L}^{\alpha_U} \int_0^1 I_{C(p_1,p_2)\leq c_\alpha} I_{p_1\leq 1} I_{p_2\leq 1} dp_1 dp_2. \tag{2.12}$$

Fisher's inverse χ^2 Approach

Motivated by Fisher's combination test, Bauer and Köhne (1989,1994) proposed to combine the p-values from separate stages in a sequential adaptive design

by taking the product of these p-values, assuming of course that these p-values are generated from stochastically independent continuous test statistics and uniformly distributed on $[0,1]$ under H_0. Thus in the Bauer-Köhne method based on Fisher's combination function for a two-stage adaptive design, the early rejection and acceptance boundaries, α_L and α_U respectively, and the critical value c_α are all determined from the following single equation:

$$\begin{aligned} \alpha &= \alpha_L + \int_{\alpha_L}^{\alpha_U} \int_0^1 I_{p_1 p_2 \le c_\alpha} I_{p_1 \le 1} I_{p_2 \le 1} dp_1 dp_2, \\ &= \alpha_L + \int_{\alpha_L}^{\alpha_U} \int_0^{\min(c_\alpha/p_1, 1)} dp_2 dp_1. \end{aligned} \tag{2.13}$$

Bauer and Köhne assumed that $c_\alpha \le \alpha_L$, which simplifies (2.10) to

$$\alpha = \alpha_L + c_\alpha(\ln \alpha_U - \ln \alpha_L). \tag{2.14}$$

The second stage critical value c_α is chosen to be the same as that for the level α Fisher's combination test, that is

$$c_\alpha = \exp\{-\frac{1}{2}\chi^2_{4;1-\alpha}\},$$

where $\chi^2_{(\nu,1-\alpha)}$ denotes the $(1-\alpha)-$quantile of the χ^2 distribution with ν degrees of freedom. Thus, given α and α_U (or α_L), c_α and α_L (or α_U) can be derived from (2.11) with the additional restriction $c_\alpha \le \alpha_L$.

Alternatively, a more general expression (Bauer and Röhmel, 1995) is given by $c_{\alpha_2} = \exp[-\frac{1}{2}\chi^2_{(4,1-\alpha_2)}]$, where $\alpha_2 < \alpha$. Given α, α_U, and α_L, the value of c_{α_2} can be derived. So does the value of α_2.

This procedure can be easily generalized to three or more stages. Suppose we have $k (k \ge 2)$ stages, where k denotes the maximum number of stages. In an

adaptive design setting, $p_i, i = 1, ..., k$, is still $U[0, 1]$ distributed under H_0. As this distribution does not depend on other p's, it is also true unconditionally and thus $p_1, ..., p_k$ are statistically independent. Reject H_0 if

$$C(p_1, ..., p_k) = p_1 p_2 \cdots p_k \leq c_\alpha = \exp[-\frac{1}{2}\chi^2_{(2k,1-\alpha)}]. \tag{2.15}$$

Let's take the three-stage adaptive design as an example (Bauer and Köhne, 1994).

- If $p_1 \leq \alpha_L$, reject H_0 and stop at the Stage 1. If $p_1 > \alpha_U$, accept H_0 and stop at the Stage 1. If $\alpha_L < p_1 \leq \alpha_U$, the trial continues to the second stage.

- If Stage 2 is reached, stop at stage 2 with acceptance of H_0 if $p_2 > \alpha_U$. If $p_1 p_2 \leq c_{\alpha_2}$, stop at stage 2 with rejection of H_0, where c_{α_2} denotes $\exp[-\frac{1}{2}\chi^2_{(4,(1-\alpha_2))}]$.

- If stage 3 is reached, reject H_0 with $p_1 p_2 p_3 \leq d_\alpha = \exp[-\frac{1}{2}\chi^2_{(6,(1-\alpha))}]$. Choose $c_{\alpha_2} = d_\alpha/\alpha_U$, then no value of $p_3 > \alpha_U$ can lead to a rejection of H_0.

- The overall probability of the Type I error is given by

$$\alpha_L + \int_{\alpha_L}^{\alpha_U} \int_0^{d_\alpha/(\alpha_U p_1)} dp_2 dp_1 + \int_{\alpha_L}^{\alpha_U} \int_{d_\alpha/(\alpha_U p_1)}^{\alpha_U} \int_0^{d_\alpha/(p_1 p_2)} dp_3 dp_2 dp_1$$

$$= \alpha_L + \frac{d_\alpha}{\alpha_U}(\ln \alpha_U - \ln \alpha_L) + d_\alpha(2 \ln \alpha_U - \ln d_\alpha)(\ln \alpha_U - \ln \alpha_L) + \frac{d_\alpha}{2}(\ln \alpha^2 \alpha_U - \ln^2 \alpha_L).$$

Recursive Combination Test

The principle of the combination test approach is that test statistics are calculated separately from the disjoint subsamples of the different stages. The test

decision is derived from a predefined function that combines the test statistics into a single criterion after each stage. Brannath et al. (2002) generalized the combination test principle and introduced the method of recursive two-stage combination tests that allows the recursive calculation of an overall p-value and the construction of confidence intervals. They assume that the distribution of the p-values p_1 and p_2 under H_0 satisfies

$$Pr_{H_0}(p_1 \leq \alpha) \leq \alpha \text{ and } Pr_{H_0}(p_2 \leq \alpha | p_1) \leq \alpha$$

for all $0 \leq \alpha \leq 1$ and call this property of the distribution of the p-values "p-clud". This means that the distribution of p_1 and the conditional distribution of p_2 given p_1 are stochastically larger than or equal to the uniform distribution on [0,1]. If independent sample units are recruited at different stages and tests are applied that control the Type I error probability for any prechosen significance level α, then this will apply.

For a two-stage design,

$$(p_1, p_2) = \begin{cases} p_1, & \text{if } p_1 \leq \alpha_L \text{ or } p_1 > \alpha_U \\ \alpha_L + \int_{\alpha_L}^{\alpha_U} \int_0^1 I(C(x,y) \leq C(p_1,p_2) dy dx, & \text{otherwise} \end{cases}$$

For Fisher's combination test, the combined p-value $q(p_1, p_2)$ is given by

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

For a multiple-stage design, define the stopping boundaries $\alpha_{L,i}$ and $\alpha_{U,i}$ at the ith interim look. denote by t^* the final number of stages. The overall p-

value can be calculated by backward recursion: $q_{t^*} = p_{t*}$ and $q_{t-1} = q_{t-1}(p_{t-1}, q_t)$ for $t = t^*, ..., 2$. the p-value q_t summarizes the results of stage t and all of the proceeding stages. the computation may be summarized by

$$p = q_1(p_1, q_2(p_2, q_3(..., qt^* - 2(p_{t^*-2}, q_{t^*-1}(p_{t^*-1}, p_{t^*}))...)))$$

and reject H_0 if $p \leq \alpha$. If independent samples are drawn at every stage $t = 1, ..., t^*$ and conservative tests are used to compute the p_t, then the p-values p_t and p-clud, that is

$$Pr_{H_0}(p_t \leq \alpha | p_{t-1}, ..., p_1) \leq \alpha, \text{ for all } 0 \leq \alpha \leq 1. \tag{2.16}$$

Weighted Inverse Normal Approach

Mosteller and Bush (1954) first introduced weighted inverse normal method, which rejects H_0 if $\omega_1 Z_1 + ... + \omega_k Z_k > Z(\alpha)$, where ω_i is the arbitrary weights, and Z_i is the test statistics, $i = 1, ..., k$. Lehmacher and Wassmer (1999) adopted this weighted inverse normal idea to the two-stage adaptive design.

$$C(p_1, p_2) = 1 - \Phi[\omega_1 \Phi^{-1}(1 - p_1) + \omega_2 \Phi^{-1}(1 - p_2)]. \tag{2.17}$$

where $0 < \omega_i < 1, i = 1, 2$, are arbitrary weights subject to $\omega_1^2 + \omega_2^2 = 1$ and Φ denotes the standard normal cumulative distribution function (CDF). If the weights are properly chosen, this combination function is equal to a classical two-stage group sequential test (Bretz et al., 2006). This approach has been used by many researchers, such as Fisher (1998), Cui et al.(1999), etc.

Truncated Product Approach

Another alternative to Fisher's product is the truncated product method (Zaykin et al., 2002), where p-values within a certain range are used. The truncated product W_τ is defined as $W_\tau = \prod_{j=1}^{k} p_j^{I(p_j \leq \tau)}$, where $I(\cdot)$ is the indicator function. Since the p-values of the different stages are independent,

$$Pr(W_\tau \leq w) = \sum_{j=0}^{k} \binom{k}{j} (1-\tau)^{k-j} [w \sum_{j=0}^{j-1} \frac{(j \ln \tau - \ln w)^j}{j!} I(w \leq \tau^j) + \tau^j I(w > \tau^j)]. \tag{2.18}$$

holds for $w < 1$ under the overall null hypothesis.

2.3.2 Conditional Error Function Approach

The conditional error function approach is an alternative to the use of combination function, originally proposed by Proschan and Hunsberger (1995). Let $A(p_1)$ denote the conditional error function, which is the probability of rejecting H_0 in the final analysis given the first-stage p-value p_1 is known. The two-stage adaptive procedure rejects H_0 at the second stage if $p_2 \leq A(p_1)$, which controls the overall Type I error at the level of α.

$$A(p_1) = P_H(\text{reject H}|p_1) = \begin{cases} 1, & \text{if } p_1 \leq \alpha_L \\ 0, & \text{if } p_1 \geq \alpha_U \\ \max\{p_2 | C(p_1, p_2) \leq c_\alpha\}, & \text{if } p_1 \in (\alpha_L, \alpha_U). \end{cases} \tag{2.19}$$

Hence, the overall Type I error is given by,

$$\int_0^1 A(p_1)dp_1 = \alpha. \tag{2.20}$$

Essentially, Proschan and Hunsberger (1995) proposed the conditional error rate function is such

$$A(p_1) = \begin{cases} 0, & \text{if } p_1 \geq \alpha_U \\ 1, & \text{if } p_1 \leq 1 - \Phi(C_{PH}) \\ 1 - \Phi[\sqrt{C_{PH}^2 - (\Phi^{-1}(1-p_1))^2}], & \text{if } 1 - \Phi(C_{PH}) \leq p_1 \leq \alpha_U. \end{cases}$$

where $\Phi(\cdot)$ denotes the normal cumulative distribution function, Φ^{-1} is its inverse, and C_{PH} is determined by (2.19).

Liu and Chi (2001) also gave a family of conditional error function in a different context. In fact, the conditional power approach can be looked at in terms of combination tests and vice versa (Jennison and Turnbull, 2005).

If we adopt the conditional error function approach to *Fisher's Product Combination Test*, the condition error function is such

$$A(p_1) = \begin{cases} 0, & \text{if } p_1 \geq \alpha_U \\ 1, & \text{if } p_1 \leq \alpha_L \\ c_\alpha/p_1, & \text{if } \alpha_L < p_1 \leq \alpha_U. \end{cases}$$

where $c_\alpha = exp[-\frac{1}{2}\chi^2_{(4,(1-\alpha))}]$ as before. If we adopt it to *Weighted Inverse Normal Approach*, the conditional error function is such

$$A(p_1) = \Phi[\frac{\omega_1\Phi^{-1}(1-p_1) - z(\alpha)}{\omega_2}],$$

where $\omega_1^2 + \omega_2^2 = 1$ as before.

Bauer and Einfalt summarized it all in their review paper regarding the application of adaption (2006). The most widely used methodology is based on Fisher combination test for p-values proposed by Bauer and Köhne (1994), followed by the inverse normal combination function of Lehmacher and Wassmer (1999) and the conditional error function approach by Proschan and Hunsberger (1995). Apparently in application, the product of p-values to combine information from different stage of a trial is preferred. This may be because of the appealing simplicity of this criterion (Bauer and Einfalt, 2006).

2.4 Data Analysis in Adaptive Designs with Multiple Hypotheses

Due to multiple hypotheses, endpoints, treatment groups, or subgroups, multiplicity is present in almost all clinical trials. Assume we have k directional null hypotheses $H_i, i = 1, ..., k$. $H_i \in \mathcal{H}$, where $\mathcal{H}$ denotes a family of null hypotheses of interest. For example, the comparisons of k treatment arms with the control. It is widely acknowledged that controlling the Per-Comparison Error Rate (PCER) without reference to the corresponding family $\mathcal{H}$ is not sufficient. On the contrary, strong control of the FWER is desired (Hellmich and Hommel, 2004). The closure procedure is a general method to control the FWER in the strong sense (Peritz, 1970; Marcus et al., 1976). It considers all the possible interaction hypotheses constructed from the original hypotheses set $\mathcal{H}$. More specifically, the closure principle

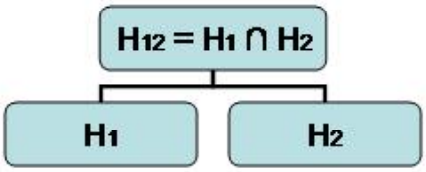

Figure 2.2: Closure principle for two null hypotheses H_1 and H_2 (Bretz et al., 2006).

is formally defined as:

1. Define a set of hypotheses $H_i, i = 1, ..., k$.

2. Construct all possible m intersection hypotheses $H_I = \bigcap_{i \in I} H_i, I \subseteq 1, ..., k, m \geq k$.

3. Find a local level-α' for each of the m hypotheses.

4. If all hypotheses H_i are rejected at the local level α', we conclude that H_i is rejected at the FWER α.

For example, if we have two hypotheses $H_i, i = 1, 2$, such as two treatment arms compared to one control arm. The resulting closed set of hypotheses $\mathcal{H} = \{H_1, H_2, H_{12}\}$, where $m = 3$. H_1 is rejected at the FWER α if both H_1 and H_{12} are rejected at the local level α' (see Figure 2.2).

One key advantage of the closure principle is that the error properties is not affected by the presence of interim looks as long as each $H \in \mathcal{H}$ is decided by a prefixed local level α test (Hellmich and Hommel, 2004). Hence, adaptive treatment selection relies on the application of the closure principle (Marcus et al., 1976)

together with combination tests. To apply the closure principle, level α tests have to be defined for all individual and intersection hypotheses $H_s = \bigcap_{i \in \mathcal{S}} H_i, \mathcal{S} \subseteq T_1$. To reject the elementary null hypothesis $H_j, j \in T_1$, at multiple level α, for all subsets $S \subseteq T_1$ that contain j the intersection hypotheses H_s have to be rejected at level α.

However, Hellmich (2001) pointed out that it may be problematic if the design of the experiment is modified in consequence of the interim results. Unless some adaptive testing method is used, any change of the prefixed test statistic is not covered by the closed testing principle.

2.4.1 The Closure Principle in Adaptive Design

In an adaptive design setting, to apply the closure principle, construct all intersection hypotheses and test each resulting hypothesis with a suitable combination test (Hommel, 1997, 2001; Bauer and Kieser, 1999; Kieser et al., 1999). A null hypothesis H_i is rejected if all hypotheses implying H_i are rejected as well. For the aforementioned example with two hypotheses H_1 and H_2, let's now consider to test them adaptively using a two-stage adaptive design. Similarly, the hypotheses set is defined as $\mathcal{H} = \{H_1, H_2, H_{12}\}$ according to the closure principle. Let $p_{i,j}$ denote the p-value for hypothesis H_j, where $j \in \{1, 2, 12\}$ denotes each individual hypothesis to be tested and $i = 1, 2$ denotes the testing stages (see Figure 2.3). Let $C(p_{1,j}, p_{2,j})$ be the combination function from each stage, where $j \in \{1, 2, 12\}$ and $i = 1, 2$. Following the closure principle, H_1 is rejected at the FWER α, if H_1 and H_{12} are both rejected at the local level α'. If the combination test approach is applied, in order to

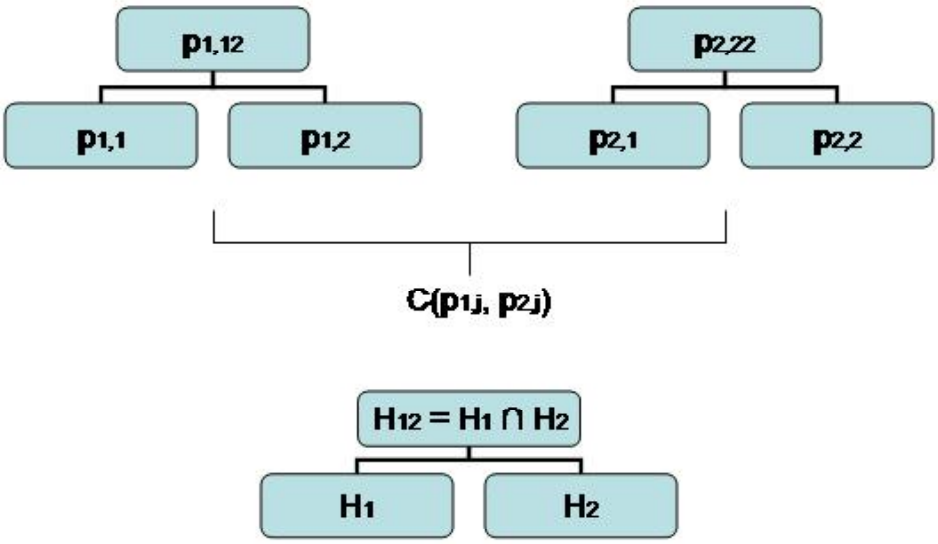

Figure 2.3: Closure principle for testing adaptively $n = 2$ null hypotheses H_1 and H_2. (Bretz et al., 2006).

reject H_1, we need $C(p_{1,1}, p_{2,1}) \leq c$ and $C(p_{1,12}, p_{2,12}) \leq c$. If the conditional error function approach is applied, we need $p_{2,1} \leq A(p_{1,1})$ and $p_{2,12} \leq A(p_{1,12})$.

2.4.2 Multiple Testing Techniques in Adaptive Design - FWER

In general, within an adaptive design frame, the null hypotheses may be dropped, the new null hypotheses may be included, or the order of null hypotheses may be modified. The research results show that when using an adaptive design (Hellmich and Hommel, 2004),

1. The efficacy decision relies on the combination test with the p-values from the separate stages. The multiple inference on the null hypotheses at the separate stages can be achieved by a prefixed closure procedure.

2. The treatment arms may be terminated for efficacy or safety assessment, or new hypotheses may be included at the interim analysis, the control of error rate won't corrupt. The inclusion and exclusion of null hypotheses at interim analysis are just reverse strategies.

3. The order of a fixed sequence of hypotheses may be altered, reflecting a corresponding shift in interest or importance. The rearrangement of the order of null hypotheses is a special case of an adaptive choice of test statistics to gain power.

Many researchers investigated the closure principle with an adaptive design (Marcus et al., 1976; Bauer and Budde, 1994; Rom et al., 1994; Tamhane et al., 1996). The common method is to utilize combination tests and the closure principle. Several applications have been described (Bauer and Röhmel, 1995; Kieser et al., 1999; Bauer and Kieser, 1999; Lehmacher et al., 2000, Kropf et al., 2000). A similar strategy has been applied by Kieser el al. (1999) for inference on multiple endpoints and by Bauer and Kieser (1999) for multiple comparison with a common control. Hellmich (2001) discussed the problem of pairwise comparisons between multiple treatments. In particular, Bretz et al. (2006) investigated the probability to reject correctly at least one of the hypotheses at the final analysis. He also compared adaptive Dunnett, adaptive hierarchical Dunnett, single stage Dunnett and single stage Bonferroni methods.

- Adaptive Dunnett Method. Adaptive Dunnett is an adaptive combination test using many-to-one Dunnett (1955) for the intersection hypotheses at each

stage, and combining the stagewise p-values with the inverse normal method with equal weights.

- Adaptive Hierarchical Method. Adaptive hierarchical is an adaptive combination test using the many-to-one Dunnett test for Stage 1 intersection hypothesis. Based on the interim results, the most promising treatment is chosen and a fixed sequence test procedure (Westfall and Krishen, 2001) starting with the selected treatment is applied for Stage 2 intersection hypothesis. Furthermore, Brannath et al. (2007) discussed the step-down Dunnett approach.

- Single Stage Dunnett Method. Single stage Dunnett uses the Dunnett adjustment in the final analysis no matter whether Stage 2 is conducted with one or more treatments.

- Single Stage Bonferroni Method. Similar to single stage Dunneet, single stage Bonferroni procedure uses the Bonferroni adjustment irrespective of whether stage 2 is conducted with one or more treatments. Obviously, this procedure is uniformly less powerful than other aforementioned procedures.

However, there are drawbacks. Directional errors are not always controlled by the closure principle (Hellmich and Hommel, 2004). Westfall et al. (1999) pointed out that special care should be taken where directional inference is present in closure procedures. Hellmich (2001) pointed out that it may be problematic if the design of the experiment is modified in consequence of the interim results. Unless some adaptive testing method is used, any change of the prefixed test statistic is not

covered by the closed testing principle.

In short, not many publications have given the explicit formula for determining the early stopping boundaries, while applying the combination test with the closure principle to control the FWER in a strong sense.

2.4.3 Multiple Testing Techniques in Adaptive Design - FDR

In the field of gene expression or gene associated studies, a large number of hypotheses are often investigated. Conventional single-stage design may lack power due to low sample size for individual hypothesis. Multi-stage adaptive design has been considered in the literature under both the FWER and FDR frameworks.

Extending single-stage design, there are two types of two-stage designs that have gained attention.

- Type 1: The total number of observations (across stages and hypotheses) is random. Stage-wise sample size for each hypotheses are preplanned. Only a limited number of hypotheses for which the first stage data showed promising effects will continue to Stage 2. This approach has been discussed under both FWER and FDR frameworks, i.e., Miller et al., 2001; Satagopan and Elston, 2003; Benjamini and Yekutieli, 2005.

- Type 2: The total number of observations is fixed and the sample size for Stage 2 is random. A certain fraction of these observations is spent in Stage 1. The remaining observations are then distributed among the hypotheses selected for Stage 2. But this approach neither controls the FWER nor the FDR.

FDR Control Method of Zehetmayer et al. (2005, 2008)

In a single-stage design, an estimator of the FDR is given by (Storey et al., 2004)

$$\widehat{FDR_\lambda}(\gamma) = \frac{\hat{\pi}_0 \gamma m_1}{\max\left(\#\{p_i < \gamma\}, 1\right)}, \tag{2.21}$$

where λ is a constant chosen a priori and $\#\{p_i < \gamma\}$ denotes the number of p-values exceeding λ and $\hat{\pi}_0 = \#\{p_i > \lambda\}/[(1-\lambda)m_1]$.

Zehetmayer et al. (2005) extended the two-stage designs to control the FDR where promising hypotheses are selected using a constant rejection threshold for each p-value at the first stage and an estimation based approach to controlling the FDR asymptotically (as the number of hypotheses goes to infinity, i.e., Storey, 2002; Storey et al., 2004) was taken at the second stage to test the selected hypotheses using more observations. The ultimate goal in that paper has been to determine asymptotically optimal values of the first-stage threshold and the fraction of observations to be spent at the first stage, given FDR level, the number of hypotheses to be selected at the first stage, and the total number of observations, based on maximizing power under the setting of multiple testing of normal means.

$$\widehat{FDR_\lambda}(\gamma) = \frac{\hat{\pi}_0 m_1 \gamma(\gamma_2)}{\max\left(\#\{p_i^{(1)} \leq \gamma_1, p_i < \gamma_2\}, 1\right)}, \tag{2.22}$$

where γ as a function of γ_2.

Zehetmayer et al. (2008) have extended this work from two-stage to multi-stage adaptive designs under both FDR and FWER frameworks, and provided useful insights into the power performance of optimized multi-stage adaptive designs with

respect to the number of stages, and into the power difference between optimized integrated design and optimized pilot design.

FDR Control Method of Victor and Hommel (2007)

Construction of methods with the FDR control in the setting of a two-stage adaptive design allowing reduction in the number of tested hypotheses at the interim analysis has been discussed in Victor and Hommel (2007) who focused on controlling the FDR in terms of a generalized global p-values for a two-stage adaptive design permitting a flexible decision for stopping at the interim analysis. The term "global p-value" refers to a final p-value which combines the p-values from all stages of the adaptive design into one single p-value. It should not be confused with global p-values in multiple testing.

The workflow of Victor and Hommel's procedure using the explorative Simes procedure where the two-stage adaptive design is determined by a family of global rejection regions is as follows:

1. Conduct the first stage.

2. Compute the first stage p-values. All hypotheses with first stage p-values greater than α_0 are stopped at the interim analysis without rejection of the corresponding hypotheses.

3. Calculate the worst case global p-value for each remaining hypothesis. Look at the Simes' boundary attained by the worst case global p-values. Hypotheses corresponding to worst case global p-values below this boundary can already

be rejected. Decide whether there are other hypotheses that may be stopped but not yet rejected.

4. Decide which of the hypotheses whose investigation is continued in the second stage should be considered in the sample size reassessment. Calculate the necessary Simes' boundary for these hypotheses (according to the order of their first stage p-values). It should be mentioned that this calculated Simes' boundary for each hypothesis is only a hypothetical one and depends on the global p-values of other hypotheses. Calculate the necessary second stage p-values for reaching this boundary. Perform sample size considerations for each of these hypotheses using the information gathered in the first stage. Choose an adequate sample size for the second stage.

5. Perform the second stage for all remaining hypotheses. Calculate the second stage (if applicable) and the global p-values for all hypotheses. Use the explorative Simes' procedure on all global p-values to decide upon the rejection of each hypothesis.

Victor and Hommel also considerd a special case of global rejection regions defined by using the Bauer-Köhne combination function where the global p-value is defined as:

$$q(p_1,p_2) = \begin{cases} \alpha/m + p_1p_2(\ln(\alpha_0) - \ln(\alpha/m)), & \text{if } \alpha/m < p_1 < \alpha_0 \bigwedge p_1p_2 < \alpha/m \\ p_1p_2 + p_1p_2(\ln(\alpha_0) - \ln(p_1p_2)), & \text{if } p_1 \leq \alpha_0 \bigwedge p_1p_2 \geq \alpha/m \\ p_1, & \text{if } p_1 < \alpha/m \bigvee p_1 > \alpha_0. \end{cases} \quad (2.23)$$

For $\alpha/m < p_1 \leq \alpha_0$, the worst case global p-value is $q(p_1, 1) = p_1 + p_1(\ln(\alpha_0) - \ln(p_1))$, and p_1 in all other situations.

In summary, controlling the FDR does not seem to be as simple as they should be in extension of single-stage design to two-stage design. Moreover, the existing methods do not appear to be a natural extension of standard FDR controlling methods, like the BH (Benjamini and Hochberg, 1995) or methods related to it, from a single-stage to a two-stage design setting.

2.5 Correlated Test Statistics

So far, we have reviewed statistical methodologies using an adaptive design, assuming that the data for each stage come from different units, and the p-values for each stage are independent. However in reality, the test statistics of each stage may be dependent, and p-values may not be uniformly distributed on $[0, 1]$. Hommel, Lindig, and Faldum (2005) discussed whether combination tests which were developed for independent p-values are robust enough to be used in dependent situations. They also proposed a modified Simes test for two-stage adaptive designs with correlated test statistics.

The rejection region for the original Simes test (Hochberg and Hommel, 1998; Simes, 1986) is $\{p_1 \leq \alpha/2\} \bigcup \{\max(p_1, p_2) \leq \alpha\} \bigcup \{p_2 \leq \alpha/2\}$. The level of this test is α for independent p-values (Simes, 1986). Samuel-Cahn (1996) showed the level α is controlled for non-negative correlation from a bivariate normal distribution. Sarkar (1998) and Sarkar and Chang (1997) proved the control of level α

for positively dependent test statistics.

Assuming that p_1 and p_2 are from a bivariate normal distribution with correlation ρ,

$$\begin{pmatrix} \Phi^{-1}(p_1) \\ \Phi^{-1}(p_2) \end{pmatrix} \sim N_2(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}),$$

where $\Phi(\cdot)$ is the distribution function of the univariate standard distribution. The rejection region for Hommel's proposed modified Simes test is defined by $\{p_1 \leq \alpha_{1L}\} \bigcup \{\alpha_{1L} < p_1 \leq \alpha_{1U} \text{ and } p_2 \leq \alpha_2\}$, where $0 \leq \alpha_{1L} < \alpha_{1U} \leq 1$ and $0 < \alpha_2 < 1$.

When $\rho = 0$, that is independent and uniformly distributed p-values,

$$\alpha := \alpha_{1L} + (\alpha_{1U} - \alpha_{1L}) \cdot \alpha_2. \tag{2.24}$$

When $\rho \neq 0$,

$$\alpha \leq \alpha_{1L} + \min(\alpha_{1U} - \alpha_{1L}, \alpha_2). \tag{2.25}$$

Hommel et al. (2005) also argued in the case when no correlation is present, the modified Simes' test seems to be a reasonable combination test as well.

CHAPTER 3

OVERALL FWER CONTROL FOR SINGLE HYPOTHESIS IN TWO-STAGE COMBINATION TEST

In this Chapter, we give an explicit formula for the overall Type I error probability in terms of early rejection and acceptance boundaries and the corresponding second stage critical value for each of Fisher's, Tippett's and Simes' combination functions. Based on these formulas, we numerically compute the critical values for these combination functions having chosen some pairs of early rejection and acceptance boundaries and values of α, and present them in Tables 3.1-3.3. Comparison of power shows that the loss in power is small when early stopping

occurs. We also apply the different methods based on the directly computed second stage critical values given pre-fixed stopping boundaries to data from a clinical study and discuss the outcomes in relation to those produced by Bauer and Köhne's original method.

3.1 Motivation

Assume that we have a one-sided null hypothesis H_0 on the difference θ in mean efficacy of two treatments, i.e., $H_0 : \theta \leq \delta$, for some given δ, to be tested against $H_a : \theta > \delta$. We consider a two-stage design with a single interim analysis and assume that the same null hypothesis H_0 is being tested against the same alternative throughout Stage 1 and Stage 2. Let p_i be the p-value of the test of H_0 at Stage $i = 1, 2$, and $C(p_1, p_2)$ be a combination function. Then, with Stage 1 early rejection and acceptance boundaries α_L and α_U, respectively, and the second-stage critical value c_α, all to be determined subject to a control of the overall Type I error at α, a two-stage adaptive test is described as follows: (1) Stop at the interim with a decision to reject H_0 if $p_1 \leq \alpha_L$, to accept H_0 if $p_1 > \alpha_U$, or to continue to Stage 2 if $\alpha_L < p_1 \leq \alpha_U$. (2) If continued to Stage 2, determine $C(p_1, p_2)$ combining p_2 with p_1 and reject H_0 if $C(p_1, p_2) \leq c_\alpha$, otherwise, accept it. The overall Type I error rate is given by

$$\alpha_L + \int_{\alpha_L}^{\alpha_U} \int_0^1 I(C(p_1, p_2) \leq c_\alpha) I(p_2 \leq 1) I(p_1 \leq 1) dp_2 dp_1. \tag{3.1}$$

Motivated by Fisher's combination test, Bauer and Köhne proposed to combine the p-values from separate stages in a sequential adaptive design by taking

the product of these p-values, assuming of course that these p-values are generated from stochastically independent continuous test statistics and uniformly distributed on $[0,1]$ under H_0. Thus, in the Bauer-Köhne method based on Fisher's combination function for a two-stage adaptive design, the early rejection and acceptance boundaries, α_L and α_U respectively, and the final critical value c_α are all determined from the following single equation:

$$\begin{aligned} \alpha &= \alpha_L + \int_{\alpha_L}^{\alpha_U} \int_0^1 I(p_1 p_2 \le c_\alpha) I(p_2 \le 1) I(p_1 \le 1) dp_2 dp_1, \\ &= \alpha_L + \int_{\alpha_L}^{\alpha_U} \int_0^{\min(c_\alpha/p_1, 1)} dp_2 dp_1. \end{aligned} \tag{3.2}$$

Bauer and Köhne assumed that $c_\alpha \le \alpha_L$, which simplifies (3.2) to

$$\alpha = \alpha_L + c_\alpha(\ln \alpha_U - \ln \alpha_L), \tag{3.3}$$

and chose c_α to be the same as that for the level α Fisher's combination test, that is, $c_\alpha = \exp\{-\frac{1}{2}\chi^2_{4;1-\alpha}\}$, where $\chi^2_{\nu;1-\alpha}$ is the $(1-\alpha)$-quantile of the χ^2 distribution with ν degrees of freedom. Thus, given α and α_U (or α_L), c_α and α_L (or α_U) can be derived from (3.3) with the additional restriction $c_\alpha \le \alpha_L$.

As noted in the introduction, often it is necessary to solve (3.2) directly for c_α given pre-chosen $0 \le \alpha_L < \alpha_U \le 1$. What we will do in the next section is to remove the restriction of Bauer and Köhne's method and provide explicit formulas for the equations needed to be solved for c_α given $0 \le \alpha_L < \alpha_U \le 1$. Clearly, this can be done, not only for Fisher's combination function but also for other combination functions like Tippett's and Simes'.

In the next section, we consider two-stage adaptive designs based on Fish-

er's, Tippett's, and Simes' combination functions. For each of these combination functions, we first derive an explicit formula for the overall Type I error rate as a function of the second stage critical value c_α given early rejection and acceptance boundaries α_L and α_U, respectively. We then use this formula to numerically compute the values of c_α subject to a control of the overall Type I error probability at α for different choices of the early stopping boundaries and present them in tables for different value of α. Sidak's combination test is often used in application. However, since it is equivalent to Tippett's, we will present the results for this combination function directly from those for Tippett's, in case one wishes to use it.

3.2 Fisher's Combination Function

As described in Chapter 2, Fisher's (1932) combination function is defined as

$$C_{Fisher}(p_1, p_2) = p_1 p_2. \tag{3.4}$$

with the corresponding overall Type I error rate as a function of $0 \leq \alpha_L < \alpha_U \leq 1$ and c_α being equal to

$$\begin{aligned}
\text{Type I error} &= \alpha_L + \int_{\alpha_L}^{\alpha_U} \int_0^1 I(C_{Fisher}(p_1, p_2) \leq c_\alpha) I(p_2 \leq 1) I(p_1 \leq 1) dp_2 dp_1 \\
&= \alpha_L + \int_{\alpha_L}^{\alpha_U} \int_0^{\min(c_\alpha/p_1, 1)} dp_2 dp_1 \\
&= \alpha_L + \int_{\alpha_L}^{\alpha_U} \min\left(1, \frac{c_\alpha}{p_1}\right) dp_1 \\
&= \begin{cases} \alpha_L + c_\alpha(\ln \alpha_U - \ln \alpha_L), & \text{if } c_\alpha \leq \alpha_L \\ c_\alpha(1 + \ln \alpha_U - \ln c_\alpha), & \text{if } \alpha_L < c_\alpha < \alpha_U \\ \alpha_U, & \text{if } c_\alpha \geq \alpha_U. \end{cases}
\end{aligned} \tag{3.5}$$

This is a more general expression for the overall Type I error rate than what Bauer and Köhne (1994) originally considered for Fisher's combination function. Although Brannath et al. (2002) generalized the Bauer and Köhne method, it is in a different context. The above equation provides a more complete picture of how c_α can be determined from $0 \leq \alpha_L < \alpha_U \leq 1$ and vice-versa subject to a control of the overall Type I error rate at α (see Figure 3.1). The equation determining c_α and satisfying $c_\alpha \leq \alpha_L$ is $\alpha_L + c_\alpha (\ln \alpha_U - \ln \alpha_L) = \alpha$, same as in Bauer and Köhne , although they have considered finding α_L (and hence α_L) by prefixing c_α. To determine c_α under the condition $\alpha_L < c_\alpha < \alpha_U$, the equation to be used is $c_\alpha (1 + \ln \alpha_U - \ln c_\alpha) = \alpha$, which interestingly does not depend on α_L. One can set the early acceptance boundary at α and can control the overall Type I error rate at α by choosing α_L and c_α arbitrarily subject to $\alpha_L < \alpha < c_\alpha$. But, it would be undesirable to do so, as it boils down to not continuing the trial to the second stage. The above expression also covers the situation considered in Bauer and Röhmel

(1995) who chose $c_{\alpha_2} = \exp\{-\frac{1}{2}\chi^2_{4;1-\alpha_2}\}$ with an $\alpha_2 < \alpha$.

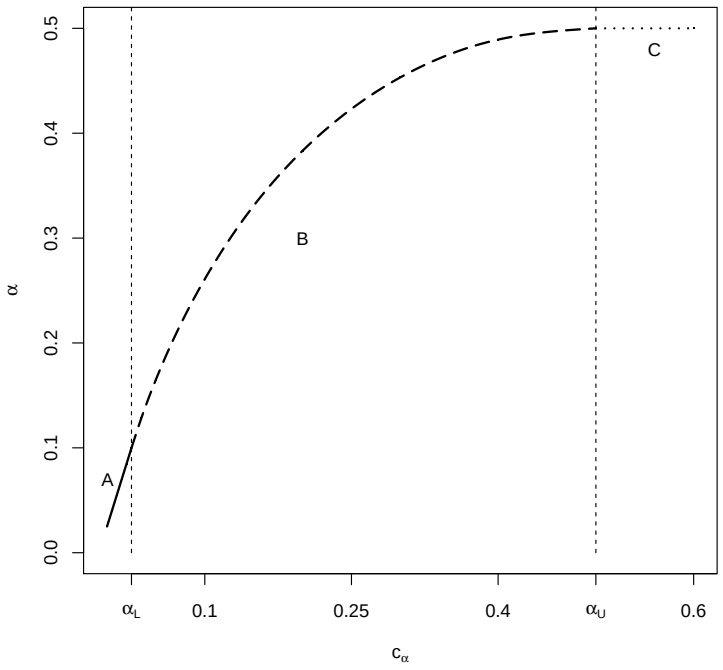

Figure 3.1: Plot for Fisher's combination function in $c_\alpha - \alpha$ plane ($\alpha_L = 0.025, \alpha_U = 0.5$). A: $\alpha = \alpha_L + c_\alpha(\ln \alpha_U - \ln \alpha_L)$; B: $\alpha = c_\alpha(1 + \ln \alpha_U - \ln c_\alpha)$; C: $\alpha = \alpha_U$.

Tables 3.1 and 3.2 present the values of c_α for some choices of pre-fixed early stopping boundaries α_L and α_U and $\alpha = 0.010, 0.025, 0.050$, and 0.100, in a two-stage adaptive design based on Fisher's combination function. The bold numbers in the table are the ones that Bauer and Köhne presented in their paper.

3.3 Tippett's or Sidak's Combination Function

Tippett's combination test (Tippett, 1931), also known as the "MinP" test, is based on following combination function:

$$C_{Tippett}(p_1, p_2) = 2\min(p_1, p_2). \tag{3.6}$$

For a two-stage adaptive design based on this combination function, the overall Type I error rate is given by

$$\begin{aligned}
\text{Type I error} &= \alpha_L + \int_{\alpha_L}^{\alpha_U}\int_0^1 I(C_{Tippett}(p_1,p_2)\le c_\alpha) I(p_2\le 1) I(p_1\le 1) dp_2 dp_1 \\
&= \alpha_L + \int_{\alpha_L}^{\min(\frac{1}{2}c_\alpha,\alpha_U)}\int_{p_1}^1 dp_2 dp_1 + \int_{\alpha_L}^{\alpha_U}\int_0^{\min(\frac{1}{2}c_\alpha,p_1)} dp_2 dp_1 \\
&= \alpha_L + \int_{\alpha_L}^{\min(\frac{1}{2}c_\alpha,\alpha_U)} (1-p_1) dp_1 + \int_{\alpha_L}^{\alpha_U} \min(\frac{1}{2}c_\alpha, p_1) dp_1 \\
&= \begin{cases} \alpha_L + \frac{c_\alpha}{2}(\alpha_U - \alpha_L), & \text{if } \frac{1}{2}c_\alpha \le \alpha_L \\ (1+\alpha_U)\frac{c_\alpha}{2} - \frac{1}{4}{c_\alpha}^2, & \text{if } \alpha_L < \frac{1}{2}c_\alpha < \alpha_U \\ \alpha_U, & \text{if } \frac{1}{2}c_\alpha \ge \alpha_U. \end{cases}
\end{aligned} \tag{3.7}$$

The expression (3.7) provides a complete picture of how c_α depends on the early rejection and acceptance boundaries and vice-versa in a two-stage adaptive design with Tippett's combination function controlling the overall Type I error rate at α (see Figure 3.2). It is very similar to that for Fisher's combination function. Just like what Bauer and Köhne (1994) did for Fisher's combination function, we may consider the equation $\alpha_L + \frac{c_\alpha}{2}(\alpha_U - \alpha_L) = \alpha$, with c_α chosen to be equal to $2\left(1-\sqrt{1-\alpha}\right)$, the critical value of the MinP test, and determine α_L and α_U subject to $\frac{1}{2}c_\alpha \le \alpha_L$. However, instead of prefixing c_α, we will consider determining it from α_L and α_U. Tables 3.3 and 3.4 present values of c_α for some choices of α_L and α_U

under the conditions $\frac{c_\alpha}{2} \leq \alpha_L$ and $\alpha_L < \frac{c_\alpha}{2} < \alpha_U$ for a two-stage adaptive design based on Tippett's combination function for $\alpha = 0.010, 0.025, 0.050$, and 0.100. Again, one can set the early acceptance boundary at α and can control the overall Type I error rate at α by choosing α_L and c_α arbitrarily subject to $\alpha_L < \frac{1}{2}\alpha < c_\alpha$, but as before it would be the same as not continuing the trial to the second stage.

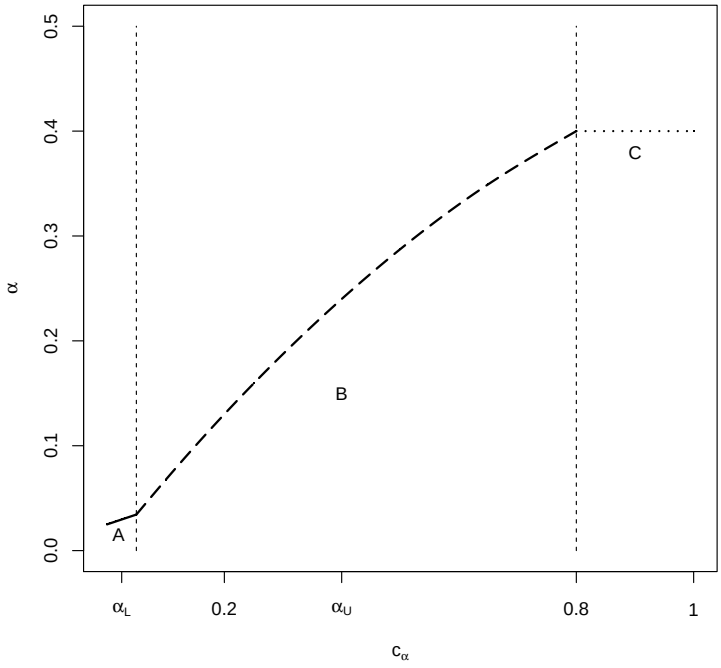

Figure 3.2: Plot for Tippett's combination function in $c_\alpha - \alpha$ plane ($\alpha_L = 0.025, \alpha_U = 0.5$). A: $\alpha = \alpha_L + \frac{c_\alpha}{2}(\alpha_U - \alpha_L)$; B: $\alpha = (1 + \alpha_U)\frac{c_\alpha}{2} - \frac{1}{4}{c_\alpha}^2$; C: $\alpha = \alpha_U$.

Sidak's combination test (1967) is based on the following combination function:

$$C_{Sidak}(p_1, p_2) = 1 - [1 - \min(p_1, p_2)]^2 \tag{3.8}$$

Since for any constant $0 < c < 1$,

$$\{C_{Sidak}(p_1, p_2) \leq c\} \equiv \{C_{Tippett}(p_1, p_2) \leq 2[1 - (1-c)^{\frac{1}{2}}]\},$$

a two-stage adaptive design based on this combination function is equivalent to that based on Tippett's, where the

$$\text{Type I error} = \alpha_L + \int_{\alpha_L}^{\alpha_U} \int_0^1 I(C_{Sidak}(p_1, p_2) \leq c_\alpha) I(p_2 \leq 1) I(p_1 \leq 1) dp_2 dp_1, \quad (3.9)$$

is the same as that given in (3.7) with c_α replaced by $2[1 - \sqrt{1 - c_\alpha}]$. Thus, for Sidak's combination function, we have

$$\text{Type I error} = \begin{cases} \alpha_U - (\alpha_U - \alpha_L) d_\alpha, & \text{if } d_\alpha \geq 1 - \alpha_L \\ \alpha_U(1 - d_\alpha) + d_\alpha(1 - d_\alpha), & \text{if } 1 - \alpha_U < d_\alpha < 1 - \alpha_L \\ \alpha_U, & \text{if } d_\alpha \leq 1 - \alpha_U. \end{cases} \quad (3.10)$$

where $d_\alpha = \sqrt{1 - c_\alpha}$.

3.4 Simes' Combination Function

Let $p_{(1)} \leq p_{(2)}$ be the ordered versions of p_1 and p_2. Then, Simes' combination test (1986) based on these p-values uses the following combination function:

$$C_{Simes}(p_1, p_2) = 2 \min\{p_{(1)}, \frac{p_{(2)}}{2}\}. \quad (3.11)$$

With each $p_i \sim U(0, 1)$, it is distributed as $U(0, 1)$ under independence of p_1 and p_2, as we assume in this article; that is, rejecting H_0 if $C_{Simes}(p_1, p_2) \leq \alpha$ provides a α-level test. In fact, it is still valid α-level test under positive dependence (Sarkar and Chang, 1997; Sarkar, 1998).

For any fixed $0 < c < 1$, we have

$$\begin{aligned}&\{C_{Simes}(p_1, p_2) \le c\}\\ &\equiv \{p_1 \le \tfrac{1}{2}c, 0 \le p_2 \le 1\} \cup \{\tfrac{1}{2}c < p_1 \le c, p_2 \le c\} \cup \{p_1 > c, p_2 \le \tfrac{1}{2}c\},\end{aligned} \tag{3.12}$$

from which one can obtain the overall Type I error probability for a two-stage adaptive design based on Simes' combination function as follows:

$$\begin{aligned}\text{Type I error} &= \alpha_L + \int_{\alpha_L}^{\alpha_U}\int_0^1 I(C_{Simes}(p_1,p_2) \le c_\alpha) I(p_2 \le 1) I(p_1 \le 1) dp_2 dp_1\\
&= \alpha_L + \int_{\alpha_L}^{\min(\alpha_U, \frac{c_\alpha}{2})}\int_0^1 dp_2 dp_1 + \int_{\max(\alpha_L, \frac{c_\alpha}{2})}^{\min(\alpha_U, c_\alpha)}\int_0^{c_\alpha} dp_2 dp_1\\
&+ \int_{\max(c_\alpha, \alpha_L)}^{\alpha_U}\int_0^{\frac{c_\alpha}{2}} dp_2 dp_1\\
&= \alpha_L + \int_{\alpha_L}^{\min(\alpha_U, \frac{c_\alpha}{2})} dp_1 + c_\alpha \int_{\max(\alpha_L, \frac{c_\alpha}{2})}^{\min(\alpha_U, c_\alpha)} dp_1 + \frac{1}{2}c_\alpha \int_{\max(c_\alpha, \alpha_L)}^{\alpha_U} dp_1\\
&= \begin{cases}
\alpha_L + \frac{1}{2}c_\alpha(\alpha_U - \alpha_L), & \text{if } c_\alpha \le \alpha_L\\
\alpha_L + c_\alpha(\frac{1}{2}\alpha_U - \alpha_L) + \frac{1}{2}{c_\alpha}^2, & \text{if } \alpha_L < c_\alpha \le \min(2\alpha_L, \alpha_U)\\
\alpha_L + c_\alpha(\alpha_U - \alpha_L), & \text{if } \alpha_U < c_\alpha \le 2\alpha_L\\
\frac{1}{2}c_\alpha(1 + \alpha_U), & \text{if } 2\alpha_L < c_\alpha \le \alpha_U\\
\frac{1}{2}c_\alpha(1 + 2\alpha_U) - \frac{1}{2}{c_\alpha}^2, & \text{if } \max(2\alpha_L, \alpha_U) \le c_\alpha \le 2\alpha_U\\
\alpha_U, & \text{if } c_\alpha \ge 2\alpha_U
\end{cases}\end{aligned} \tag{3.13}$$

This provides a complete picture of how c_α can be determined from early rejection and acceptance boundaries, α_L and α_U respectively, or vice-versa in a two-stage adaptive design based on Simes' combination function and controlling the overall Type I error rate at α (see Figure 3.3). Tables 3.5, 3.6, and 3.7 present the values of c_α for some choices of α_L and α_U under the conditions $2\alpha_L < c_\alpha \le \alpha_U$, $\alpha_L < c_\alpha \le \min(2\alpha_L, \alpha_U)$ and $c_\alpha \le \alpha_L$ for a two-stage adaptive design with

the overall significance level $\alpha = 0.010, 0.025, 0.050$, and 0.100 based on Simes' combination function.

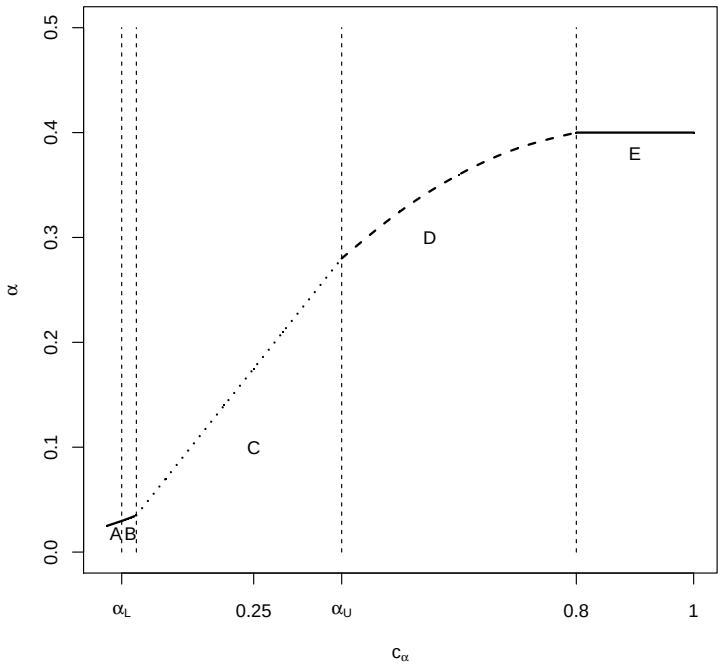

Figure 3.3: Plot for Simes' combination function in $c_\alpha - \alpha$ plane ($\alpha_L = 0.025, \alpha_U = 0.5$). A: $\alpha = \alpha_L + \frac{1}{2}c_\alpha(\alpha_U - \alpha_L)$; B: $\alpha = \alpha_L + c_\alpha(\frac{1}{2}\alpha_U - \alpha_L) + \frac{1}{2}{c_\alpha}^2$; C: $\alpha = \frac{1}{2}c_\alpha(1 + \alpha_U)$; D: $\alpha = \frac{1}{2}c_\alpha(1 + 2\alpha_U) - \frac{1}{2}{c_\alpha}^2$; E: $\alpha = \alpha_U$.

3.5 Power Analysis

The maximum sample size for such an adaptive procedure coincides with a non-sequential combination test. Thus in terms of power analysis, this test procedure suffers a loss of power if early stopping occurs. For instance in a two-stage adaptive design compared with the classic non-sequential combination test, the loss

in power is given by

$$Pr[\{C(p_1,p_2) \leq c_\alpha\} \bigcap \{p_1 \geq \alpha_U\}] - Pr[\{C(p_1,p_2) \geq c_\alpha\} \bigcap \{c_\alpha \leq p_1 \leq \alpha_L\}] \quad (3.14)$$

The first term of this expression refers to the situation where the classical combination test would reject H_0, but the condition $p_1 \geq \alpha_U$ leads to an early acceptance at Stage 1. The second term refers to the gain in power from early rejection in an adaptive design under the condition $p_1 \leq \alpha_L$ which would not end up with a rejection with the classic non-sequential combination test (Bauer and Köhne, 1994). More specifically, Bauer and Köhne (1994) stated that the loss in power for Fisher's combination test is less or equal to

$$Pr(p_2 \leq \frac{c_\alpha}{\alpha_U}) Pr(p_1 \geq \alpha_U) - Pr(p_2 \geq \alpha_U) Pr(\frac{c_\alpha}{\alpha_U} \leq p_1 \leq \alpha_L).$$

The loss in power was evaluated by 1000 simulation runs to test a hypothesis of $\mu = 0$ versus $\mu = \delta$ with known variance $\sigma^2 = 1$ for a two-stage combination test with the overall type I error controlled at $\alpha = 0.05$. Figure 3.4 shows the loss in power for a classic two-stage non-sequential combination test and an adaptive design with early stopping based on different combination functions. The loss in power due to early stopping is small for all three tests.

Figure 3.5 compares power properties of different combination functions with pre-fixed early stopping boundaries at $\alpha_L = 0.025$ and $\alpha_U = 0.5$ and overall type I error controlled at $\alpha = 0.05$. With the same pre-fixed early stopping boundaries, Fisher's combination function is slightly more powerful than the other two tests.

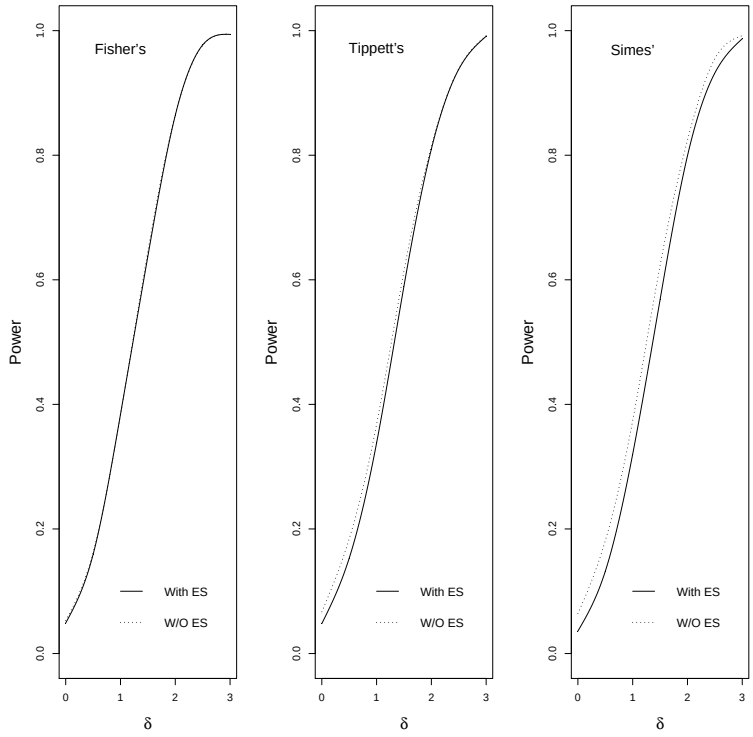

Figure 3.4: Power of two-stage combination test with early stopping (With ES) and without early stopping (W/O ES) for Fisher's, Tippett's and Simes' combination functions ($\alpha = 0.05, \alpha_L = 0.025, \alpha_U = 0.5$).

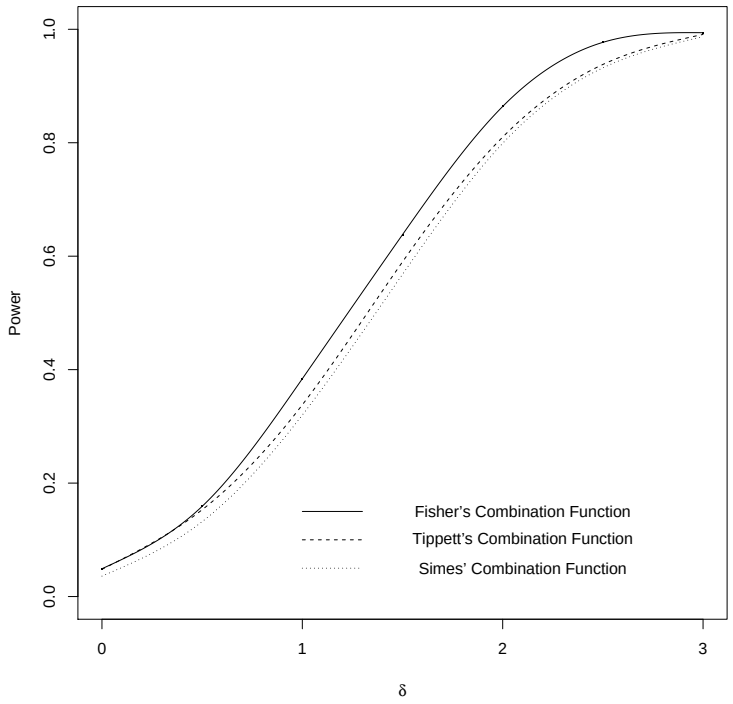

Figure 3.5: Power of two-stage combination test for Fisher's, Tippett's and Simes' combination functions ($\alpha = 0.05, \alpha_L = 0.025, \alpha_U = 0.5$).

3.6 Example

We use the data from a clinical study on patients with acne papulopustulosa to illustrate an application of the formula described above. It was a randomized, placebo-controlled, double-blind study, comparing the effect of treatment under a combination of 1% choloramphenicol (CAS 56-75-7) and 0.5% pale ulfonated shale oil versus the placebo (Fluhr, 1998). The original data was collected from a traditional one stage design with 24 patients assigned in the combination therapy and 26 patients in the placebo group. A two-sided t-test ($p = 0.0008$) showed that the combination therapy significantly reduced bacteria as compared to placebo. Lehmacher and Wassmer (1999) applied this data to a three-stage adaptive Pocock's design (1977) with an overall Type I error rate $\alpha = 0.01$. There were 24 patients (12 per group) in Stage 1. The resulting one-sided p-value p_1 was 0.0070 ($t = 2.672$), which led the trial to Stage 2 with next 12 patients (6 per group). The resulting Stage 2 one-sided p-value p_2 turned out to be 0.0468 ($t = 1.853$), yielding a significant result in treatment arm. The study then stopped at Stage 2 for efficacy.

For illustration purpose, we intended to use this data and applied the two-stage adaptive designs with the same overall significance level $\alpha = 0.01$ based on the different combination functions considered in this article. We consulted Tables 3.1-3.3 to select appropriate pairs of early stopping boundaries and the corresponding second stage critical values for these designs before applying them to the data. To have a meaningful comparison of these designs, we will keep the early acceptance boundaries same for all these designs. These values and the conclusions are

summarized in Table 3.4.

3.6.1 Fisher's Combination Function

Tables 3.1 and 3.2 offer a variety of possible values for the pair (α_L, α_U) and the corresponding c_α for this combination function, each providing a control of the overall Type I error rate at $\alpha = 0.01$. Interestingly, since the first stage p-value is $p_1 = 0.0070$ and the combined p-value based on this combination function in the second stage is $C(p_1, p_2) = p_1 p_2 = 0.0070 \times 0.0468 = 0.0003$, we would have the same conclusion as Fluhr (1998) and Lehmacher and Wassermer (1999), that is, the superiority of the combination treatment over the placebo would be demonstrated, no matter what values are chosen for these quantities from this table. Nevertheless, as said above, we choose $\alpha_U = 0.4$ for this as well as for the other two designs. The corresponding values of α_L and c_α, satisfying the constraint $\alpha_L < c_\alpha$, are 0.0035 and 0.0014, respectively. Under the constraint $\alpha_L < c_\alpha < \alpha_U$, the corresponding value of c_α is 0.0015, and since α_L can be any value between 0 and 0.0015, we choose in particular $\alpha_L = 0.0010$.

3.6.2 Tippett's Combination Function

Tables 3.3 and 3.4 present some possible values of α_L, α_U, and c_α for this combination function. With $\alpha_U = 0.4$, we can set α_L at any value in the interval $(0, 0.0072)$, with the corresponding $c_\alpha = 0.0144$, under the restriction $\alpha_L < \frac{1}{2} c_\alpha < \alpha_U$. Since the first stage p-value is 0.0070, the trial would stop at Stage 1 for efficacy if we select $\alpha_L = 0.0072$, and there would be no need to continue to the second stage.

Similarly, when $\frac{1}{2}c_\alpha \leq \alpha_L$, α_L could be set at 0.0075 or even greater with the same α_U at 0.4, and again the trial would stop at Stage 1 for efficacy.

3.6.3 Simes' Combination Function

Tables 3.5, 3.6, and 3.7 show some possible pairs of early stopping boundaries and the corresponding second stage critical value when Simes' combination function is used. When $2\alpha_L < c_\alpha \leq \alpha_U$ and $\alpha_U = 0.4$, α_L can be set at 0.0071, and the trial would stop early at Stage 1 for efficacy with Stage 1 p-value $p_1 = 0.0070$. Similarly, for situation $\alpha_L < c_\alpha \leq \min(2\alpha_L, \alpha_U)$ or $c_\alpha \leq \alpha_L$, we can also find pairs of early stopping boundaries allowing the trial to stop at Stage 1 with the superiority of the combination treatment over placebo demonstrated (see Table 3.8).

In short, in this example the application of the proposed formula for overall Type I error rate led to the same statistical conclusion as the original global t-test but with only partial patients, which implied our results were unbiased with respect to data-driven adaptation of the design. Moreover, by applying different combination functions and early stopping boundaries, the trial could be stopped early at Stage 1 due to efficacy, suggesting that with good planning, a two-stage adaptive design can reduce lead time and save cost in clinical development.

3.7 Discussion

Inevitably, there is no superior strategy to determine early stopping boundaries or the critical value in two-stage combination function approach, even though

the Bauer-Köhne method is very commonly used. While choosing the early stopping boundaries and the second stage critical value subject to a control of the overall type I error rate at α, Bauer and Köhne restricted the second stage critical value to that of the level α test based on the chosen combination function, and thus gave only one choice for the early rejection boundary with a prefixed acceptance boundary. As this may be restrictive in some instances, we have decided to look at the formula for the overall type I error rate more explicitly as a function of the early stopping boundaries and the second stage critical value before giving a general idea through numerical calculations of how the second stage critical value can be chosen from pre-fixed early stopping boundaries. This offers increased flexibility in designing an adaptive design where the early stopping boundaries are chosen upfront. The formulas and the related calculations have been given not only for Fisher's combination function, the one Bauer and Köhne originally considered, but also for other commonly used combination functions, Tippett's and Simes', which were not available in the literature, as far as we know. Since recursive application of two-stage adaptive design is valid (Bauer and Köhne, 1994), conceptually, one can extend the present results to a three stage design, although the final expressions in that case might be more involved to work with. In addition, the choice of c_α can be done based on optimal power calculation, which will be further investigated.

We believe this paper offers a good understanding of a two-stage adaptive design from the point of view of choosing proper early stopping boundaries and the second stage critical value. Of course, there is an arbitrariness in these choices,

because the control of the overall Type I error rate at the desired level is the only criterion used while choosing these quantities. This arbitrariness can be removed by bringing in other considerations like power, which we will address in a different communication.

Table 3.1: The critical value c_α while prefixing the early stopping boundaries α_L and α_U under the condition $c_\alpha \leq \alpha_L$, based on Fisher's Combination Function.

$\alpha = 0.010$					α_L			
α_U	0.0015	0.0020	0.0027	0.0031	0.0035	0.0040	0.0045	0.0050
0.1	-	-	0.0020	0.0020	0.0019	0.0019	0.0018	0.0017
0.2	-	0.0017	0.0017	0.0017	0.0016	0.0015	0.0014	0.0014
0.3	-	0.0016	0.0015	0.0015	0.0015	0.0014	**0.00131**	0.0012
0.4	-	0.0015	0.0015	0.0014	0.0014	**0.00131**	0.0012	0.0011
0.5	0.0015	0.0014	0.0014	0.0014	**0.00131**	0.0012	0.0012	0.0011
0.6	0.0014	0.0014	0.0014	**0.00131**	0.0013	0.0012	0.0011	0.0010
0.7	0.0014	0.0014	**0.00131**	0.0013	0.0012	0.0011	0.0011	0.0010
0.8	0.0014	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0010
0.9	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0010	0.0010
1.0	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0010	0.0009
$\alpha = 0.025$					α_L			
α_U	0.0045	0.0050	0.0060	0.0080	0.0090	0.0100	0.0115	0.0150
0.1	-	-	-	0.0067	0.0066	0.0065	0.0062	0.0053
0.2	-	-	0.0054	0.0053	0.0052	0.0050	0.0047	0.0039
0.3	-	0.0049	0.0049	0.0047	0.0046	0.0044	0.0041	0.0033
0.4	-	0.0046	0.0045	0.0043	0.0042	0.0041	**0.0038**	0.0030
0.5	0.0044	0.0043	0.0043	0.0041	0.0040	0.0038	0.0036	0.0029
0.6	0.0042	0.0042	0.0041	0.0039	**0.0038**	0.0037	0.0034	0.0027
0.7	0.0041	0.0040	0.0040	**0.0038**	0.0037	0.0035	0.0033	0.0026
0.8	0.0040	0.0039	0.0039	0.0038	0.0036	0.0035	0.0032	0.0025
0.9	0.0039	0.0039	0.0038	0.0036	0.0035	0.0033	0.0031	0.0024
1.0	0.0038	0.0038	0.0037	0.0035	0.0034	0.0033	0.0030	0.0024
$\alpha = 0.050$					α_L			
α_U	0.0100	0.0150	0.0200	0.0250	0.0299	0.0350	0.0400	0.0450
0.1	-	-	0.0186	0.0180	0.0166	0.0143	0.0109	0.0063
0.2	-	0.0135	0.0130	0.0120	0.0106	0.0086	0.0062	0.0034
0.3	-	0.0117	0.0111	0.0101	**0.0087**	0.0070	0.0050	0.0026
0.4	-	0.0107	0.0100	0.0090	0.0077	0.0062	0.0043	0.0023
0.5	-	0.0100	0.0093	0.0083	0.0071	0.0056	0.0040	0.0021
0.6	0.0098	0.0095	0.0088	0.0079	0.0067	0.0053	0.0037	0.0019
0.7	0.0094	0.0091	0.0084	0.0075	0.0064	0.0050	0.0035	0.0018
0.8	0.0091	0.0088	0.0081	0.0072	0.0061	0.0048	0.0033	0.0017
0.9	0.0089	0.0085	0.0079	0.0070	0.0059	0.0046	0.0032	0.0017
1.0	0.0087	0.0083	0.0077	0.0068	0.0057	0.0045	0.0031	0.0016
$\alpha = 0.100$					α_L			
α_U	0.0250	0.0300	0.0400	0.0500	0.0600	0.0700	0.0800	0.0900
0.1	-	-	-	-	-	-	-	-
0.2	-	-	0.0373	0.0361	0.0332	0.0286	0.0218	0.0125
0.3	-	-	0.0298	0.0279	0.0249	0.0206	0.0151	0.0083
0.4	-	0.0270	0.0261	0.0240	0.0211	0.0172	0.0124	0.0067
0.5	0.0250	0.0249	0.0238	0.0217	0.0189	0.0153	0.0109	0.0058
0.6	0.0236	0.0234	0.0222	0.0201	0.0174	0.0140	0.0099	0.0053
0.7	0.0225	0.0222	0.0210	0.0189	0.0163	0.0130	0.0092	0.0049
0.8	0.0216	0.0213	0.0200	0.0180	0.0154	0.0123	0.0087	0.0046
0.9	0.0209	0.0206	0.0193	0.0173	0.0148	0.0117	0.0083	0.0043
1.0	0.0203	0.0200	0.0186	0.0167	0.0142	0.0113	0.0079	0.0042

Table 3.2: The critical value c_α while prefixing the early stopping boundaries α_L and α_U under the condition $\alpha_L < c_\alpha < \alpha_U$, based on Fisher's Combination Function.

	$\alpha_L < c_\alpha < \alpha_U$			
	α			
α_U	0.0100	0.0250	0.0500	0.1000
0.1	-	0.0068	0.0187	-
	-	$(0 < \alpha_L < 0.0068)$	$(0 < \alpha_L < 0.0187)$	-
0.2	0.0017	-	0.0135	0.0373
	$(0 < \alpha_L < 0.0017)$	-	$(0 < \alpha_L < 0.0135)$	$(0 < \alpha_L < 0.0373)$
0.3	-	-	-	0.0304
	-	-	-	$(0 < \alpha_L < 0.0304)$
0.4	0.0015	0.0046	-	0.0271
	$(0 < \alpha_L < 0.0015)$	$(0 < \alpha_L < 0.0046)$	-	$(0 < \alpha_L < 0.0271)$
0.5	-	-	-	-
0.6	0.0014	0.0042	0.0098	0.0236
	$(0 < \alpha_L < 0.0014)$	$(0 < \alpha_L < 0.0042)$	$(0 < \alpha_L < 0.0098)$	$(0 < \alpha_L < 0.0236)$
0.7	-	-	0.0094	-
	-	-	$(0 < \alpha_L < 0.0094)$	-
0.8	-	-	-	-
0.9	0.0013	-	-	0.0210
	$(0 < \alpha_L < 0.0013)$	-	-	$(0 < \alpha_L < 0.0210)$
1.0	0.0013	-	-	0.0205
	$(0 < \alpha_L < 0.0013)$	-	-	$(0 < \alpha_L < 0.0205)$

Table 3.3: The critical value c_α while prefixing the early stopping boundaries α_L and α_U under the condition $\alpha_L < \frac{c_\alpha}{2} < \alpha_U$, based on Tippett's Combination Function.

	$\alpha_L < \frac{c_\alpha}{2} < \alpha_U$			
	α			
α_U	0.0100	0.0250	0.0500	0.1000
0.1	0.0183	0.0464	0.0950	0.2000
	$(0 < \alpha_L < 0.0092)$	$(0 < \alpha_L < 0.0232)$	$(0 < \alpha_L < 0.0475)$	$(0 < \alpha_L < 0.1000)$
0.2	0.0168	0.0424	0.0864	0.1802
	$(0 < \alpha_L < 0.0084)$	$(0 < \alpha_L < 0.0212)$	$(0 < \alpha_L < 0.0432)$	$(0 < \alpha_L < 0.0901)$
0.3	0.0155	0.0390	0.0793	0.1642
	$(0 < \alpha_L < 0.0077)$	$(0 < \alpha_L < 0.0195)$	$(0 < \alpha_L < 0.0397)$	$(0 < \alpha_L < 0.0821)$
0.4	0.0144	0.0362	0.0734	0.1510
	$(0 < \alpha_L < 0.0072)$	$(0 < \alpha_L < 0.0181)$	$(0 < \alpha_L < 0.0367)$	$(0 < \alpha_L < 0.0755)$
0.5	0.0134	0.0337	0.0682	0.1399
	$(0 < \alpha_L < 0.0067)$	$(0 < \alpha_L < 0.0169)$	$(0 < \alpha_L < 0.0341)$	$(0 < \alpha_L < 0.0700)$
0.6	0.0125	0.0316	0.0638	0.1303
	$(0 < \alpha_L < 0.0063)$	$(0 < \alpha_L < 0.0158)$	$(0 < \alpha_L < 0.0319)$	$(0 < \alpha_L < 0.0652)$
0.7	0.0118	0.0297	0.0599	0.1220
	$(0 < \alpha_L < 0.0059)$	$(0 < \alpha_L < 0.0148)$	$(0 < \alpha_L < 0.0299)$	$(0 < \alpha_L < 0.0610)$
0.8	0.0111	0.0280	0.0564	0.1148
	$(0 < \alpha_L < 0.0056)$	$(0 < \alpha_L < 0.0140)$	$(0 < \alpha_L < 0.0282)$	$(0 < \alpha_L < 0.0574)$
0.9	0.0106	0.0265	0.0534	0.1084
	$(0 < \alpha_L < 0.0053)$	$(0 < \alpha_L < 0.0133)$	$(0 < \alpha_L < 0.0267)$	$(0 < \alpha_L < 0.0542)$
1.0	0.0100	0.0252	0.0506	0.1026
	$(0 < \alpha_L < 0.0050)$	$(0 < \alpha_L < 0.0126)$	$(0 < \alpha_L < 0.0253)$	$(0 < \alpha_L < 0.0513)$

Table 3.4: The critical value c_α while prefixing the early stopping boundaries α_L and α_U under the condition $\frac{1}{2}c_\alpha \leq \alpha_L$, based on Tippett's Combination Function.

	$\frac{1}{2}c_\alpha \leq \alpha_L$							
$\alpha = 0.010$	α_L							
α_U	0.0060	0.0065	0.0070	0.0075	0.0080	0.0085	0.0090	0.0095
0.1	-	-	-	-	-	-	-	-
0.2	-	-	-	-	-	0.0157	0.0105	0.0052
0.3	-	-	-	-	0.0137	0.0103	0.0069	0.0034
0.4	-	-	-	0.0127	0.0102	0.0077	0.0051	0.0026
0.5	-	-	0.0122	0.0102	0.0081	0.0061	0.0041	0.0020
0.6	-	0.0118	0.0101	0.0084	0.0068	0.0051	0.0034	0.0017
0.7	0.0115	0.0101	0.0087	0.0072	0.0058	0.0043	0.0029	0.0014
0.8	0.0101	0.0088	0.0076	0.0063	0.0051	0.0038	0.0025	0.0013
0.9	0.0089	0.0078	0.0067	0.0056	0.0045	0.0034	0.0022	0.0011
1.0	0.0080	0.0070	0.0060	0.0050	0.0040	0.0030	0.0020	0.0010
$\alpha = 0.025$	α_L							
α_U	0.0150	0.0160	0.0170	0.0180	0.0190	0.0200	0.0220	0.0240
0.1	-	-	-	-	-	-	-	0.0263
0.2	-	-	-	-	-	-	0.0337	0.0114
0.3	-	-	-	-	-	0.0357	0.0216	0.0072
0.4	-	-	-	-	0.0315	0.0263	0.0159	0.0053
0.5	-	-	0.0331	0.0290	0.0249	0.0208	0.0126	0.0042
0.6	-	0.0308	0.0274	0.0241	0.0207	0.0172	0.0104	0.0035
0.7	0.0292	0.0263	0.0234	0.0205	0.0176	0.0147	0.0088	0.0030
0.8	0.0255	0.0230	0.0204	0.0179	0.0154	0.0128	0.0077	0.0026
0.9	0.0226	0.0204	0.0181	0.0159	0.0136	0.0114	0.0068	0.0023
1.0	0.0203	0.0183	0.0163	0.0143	0.0122	0.0102	0.0061	0.0020
$\alpha = 0.050$	α_L							
α_U	0.0300	0.0320	0.0350	0.0380	0.0400	0.0420	0.0450	0.0480
0.1	-	-	-	-	-	-	-	0.0769
0.2	-	-	-	-	-	-	0.0645	0.0263
0.3	-	-	-	-	0.0769	0.0620	0.0392	0.0159
0.4	-	-	-	0.0663	0.0556	0.0447	0.0282	0.0114
0.5	-	-	0.0645	0.0519	0.0435	0.0349	0.0220	0.0088
0.6	-	0.0634	0.0531	0.0427	0.0357	0.0287	0.0180	0.0072
0.7	0.0597	0.0539	0.0451	0.0363	0.0303	0.0243	0.0153	0.0061
0.8	0.0519	0.0469	0.0392	0.0315	0.0263	0.0211	0.0132	0.0053
0.9	0.0460	0.0415	0.0347	0.0278	0.0233	0.0186	0.0117	0.0047
1.0	0.0412	0.0372	0.0311	0.0249	0.0208	0.0167	0.0105	0.0042
$\alpha = 0.100$	α_L							
α_U	0.0700	0.0750	0.0800	0.0850	0.0875	0.0900	0.0925	0.0950
0.1	-	-	-	-	-	-	-	-
0.2	-	-	-	-	-	-	0.1395	0.0952
0.3	-	-	-	0.1395	0.1176	0.0952	0.0723	0.0488
0.4	-	-	0.1250	0.0952	0.0800	0.0645	0.0488	0.0328
0.5	0.1395	0.1176	0.0952	0.0723	0.0606	0.0488	0.0368	0.0247
0.6	0.1132	0.0952	0.0769	0.0583	0.0488	0.0392	0.0296	0.0198
0.7	0.0952	0.0800	0.0645	0.0488	0.0408	0.0328	0.0247	0.0165
0.8	0.0822	0.0690	0.0556	0.0420	0.0351	0.0282	0.0212	0.0142
0.9	0.0723	0.0606	0.0488	0.0368	0.0308	0.0247	0.0186	0.0124
1.0	0.0645	0.0541	0.0435	0.0328	0.0274	0.0220	0.0165	0.0110

Table 3.5: The critical value c_α while prefixing the early stopping boundaries α_L and α_U under the conditions $2\alpha_L < c_\alpha \leq \alpha_U$, based on Simes' Combination Function.

	$2\alpha_L < c_\alpha \leq \alpha_U$			
	α			
α_U	0.0100	0.0250	0.0500	0.1000
0.1	0.0182	0.0455	0.0909	0.1818
	$(0 < \alpha_L \leq 0.0091)$	$(0 < \alpha_L \leq 0.0227)$	$(0 < \alpha_L \leq 0.0455)$	$(0 < \alpha_L \leq 0.0909)$
0.2	0.0167	0.0417	0.0833	0.1667
	$(0 < \alpha_L \leq 0.0083)$	$(0 < \alpha_L \leq 0.0208)$	$(0 < \alpha_L \leq 0.0417)$	$(0 < \alpha_L \leq 0.0833)$
0.3	0.0154	0.0385	0.0769	0.1538
	$(0 < \alpha_L \leq 0.0077)$	$(0 < \alpha_L \leq 0.0192)$	$(0 < \alpha_L \leq 0.0385)$	$(0 < \alpha_L \leq 0.0769)$
0.4	0.0143	0.0357	0.0714	0.1429
	$(0 < \alpha_L \leq 0.0071)$	$(0 < \alpha_L \leq 0.0179)$	$(0 < \alpha_L \leq 0.0357)$	$(0 < \alpha_L \leq 0.0714)$
0.5	0.0133	0.0333	0.0667	0.1333
	$(0 < \alpha_L \leq 0.0067)$	$(0 < \alpha_L \leq 0.0167)$	$(0 < \alpha_L \leq 0.0333)$	$(0 < \alpha_L \leq 0.0667)$
0.6	0.0125	0.0313	0.0625	0.1250
	$(0 < \alpha_L \leq 0.0063)$	$(0 < \alpha_L \leq 0.0156)$	$(0 < \alpha_L \leq 0.0313)$	$(0 < \alpha_L \leq 0.0625)$
0.7	0.0118	0.0294	0.0588	0.1176
	$(0 < \alpha_L \leq 0.0059)$	$(0 < \alpha_L \leq 0.0147)$	$(0 < \alpha_L \leq 0.0294)$	$(0 < \alpha_L \leq 0.0588)$
0.8	0.0111	0.0278	0.0556	0.1111
	$(0 < \alpha_L \leq 0.0056)$	$(0 < \alpha_L \leq 0.0139)$	$(0 < \alpha_L \leq 0.0278)$	$(0 < \alpha_L \leq 0.0556)$
0.9	0.0105	0.0263	0.0526	0.1053
	$(0 < \alpha_L \leq 0.0053)$	$(0 < \alpha_L \leq 0.0132)$	$(0 < \alpha_L \leq 0.0263)$	$(0 < \alpha_L \leq 0.0526)$
1.0	0.0100	0.0250	0.0500	0.1000
	$(0 < \alpha_L \leq 0.0050)$	$(0 < \alpha_L \leq 0.0125)$	$(0 < \alpha_L \leq 0.0250)$	$(0 < \alpha_L \leq 0.0500)$

Table 3.6: The critical value c_α while prefixing the early stopping boundaries α_L and α_U under the conditions $\alpha_L < c_\alpha \leq \min(2\alpha_L, \alpha_U)$, based on Simes' Combination Function.

	$\alpha_L < c_\alpha \leq \min(2\alpha_L, \alpha_U)$							
$\alpha = 0.010$	α_L							
α_U	0.0065	0.0068	0.0070	0.0073	0.0075	0.0080	0.0085	0.0090
0.1	-	-	-	-	-	-	-	-
0.2	-	-	-	-	-	-	0.0151	0.0104
0.3	-	-	-	-	-	0.0134	0.0102	-
0.4	-	-	-	0.0135	0.0126	0.0101	-	-
0.5	-	0.0128	0.0120	0.0109	0.0101	0.0081	-	-
0.6	0.0117	0.0107	0.0101	0.0091	0.0084	-	-	-
0.7	0.0100	0.0092	0.0086	0.0078	-	-	-	-
0.8	0.0088	0.0081	0.0076	-	-	-	-	-
0.9	0.0078	0.0072	-	-	-	-	-	-
1.0	0.0070	-	-	-	-	-	-	-
$\alpha = 0.025$	α_L							
α_U	0.0170	0.0175	0.0180	0.0185	0.0190	0.0195	0.0200	0.0225
0.1	-	-	-	-	-	-	-	-
0.2	-	-	-	-	-	-	-	0.0274
0.3	-	-	-	-	-	0.0185	0.0170	-
0.4	-	-	0.0351	0.0328	0.0306	0.0283	0.0259	-
0.5	0.0321	0.0303	0.0284	0.0266	0.0247	0.0227	0.0208	-
0.6	0.0270	0.0254	0.0238	0.0222	0.0206	-	-	-
0.7	0.0232	0.0218	0.0205	0.0191	-	-	-	-
0.8	0.0203	0.0191	-	-	-	-	-	-
0.9	0.0181	-	-	-	-	-	-	-
1.0	-	-	-	-	-	-	-	-
$\alpha = 0.050$	α_L							
α_U	0.0300	0.0325	0.0350	0.0375	0.0400	0.0425	0.0450	0.0475
0.1	-	-	-	-	-	-	-	0.0683
0.2	-	-	-	-	-	0.0778	0.0591	-
0.3	-	-	-	-	0.0692	0.0555	-	-
0.4	-	-	-	0.0642	0.0535	-	-	-
0.5	-	-	0.0611	0.0524	0.0432	-	-	-
0.6	-	0.0589	0.0516	0.0439	-	-	-	-
0.7	0.0574	0.0510	0.0445	0.0377	-	-	-	-
0.8	0.0506	0.0449	0.0390	-	-	-	-	-
0.9	0.0452	0.0400	-	-	-	-	-	-
1.0	0.0408	0.0360	-	-	-	-	-	-
$\alpha = 0.100$	α_L							
α_U	0.0650	0.0700	0.0725	0.0750	0.0775	0.0800	0.0850	0.0900
0.1	-	-	-	-	-	-	-	-
0.2	-	-	-	-	-	-	0.1589	0.1318
0.3	-	-	-	-	0.1517	0.1418	0.1200	0.0936
0.4	-	-	0.1394	0.1312	0.1225	0.1132	0.0929	-
0.5	-	0.1240	0.1166	0.1089	0.1009	0.0925	-	-
0.6	0.1189	0.1060	0.0992	0.0922	0.0849	-	-	-
0.7	0.1039	0.0920	0.0858	0.0794	-	-	-	-
0.8	0.0919	0.0810	0.0753	-	-	-	-	-
0.9	0.0821	0.0721	-	-	-	-	-	-
1.0	0.0741	-	-	-	-	-	-	-

Table 3.7: The critical value c_α while prefixing the early stopping boundaries α_L and α_U under the conditions $c_\alpha \leq \alpha_L$, based on Simes' Combination Function.

	$c_\alpha \leq \alpha_L$							
$\alpha = 0.010$	α_L							
α_U	0.0070	0.0075	0.0080	0.0085	0.0087	0.0090	0.0093	0.0095
0.1	-	-	-	-	-	-	-	-
0.2	-	-	-	-	-	-	0.0073	0.0052
0.3	-	-	-	-	0.0089	0.0069	0.0048	0.0034
0.4	-	-	-	0.0076	0.0066	0.0051	0.0036	0.0026
0.5	-	-	-	0.0061	0.0053	0.0041	0.0029	0.0020
0.6	-	-	0.0068	0.0051	0.0044	0.0034	0.0024	0.0017
0.7	-	0.0072	0.0058	0.0043	0.0038	0.0029	0.0020	0.0014
0.8	-	0.0063	0.0051	0.0038	0.0033	0.0025	0.0018	0.0013
0.9	0.0067	0.0056	0.0045	0.0034	0.0029	0.0022	0.0016	0.0011
1.0	0.0060	0.0050	0.0040	0.0030	0.0026	0.0020	0.0014	0.0010
$\alpha = 0.025$	α_L							
α_U	0.0200	0.0210	0.0215	0.0220	0.0225	0.0230	0.0235	0.0240
0.1	-	-	-	-	-	-	-	-
0.2	-	-	-	-	-	0.0226	0.0170	0.0114
0.3	-	-	-	0.0216	0.0180	0.0144	0.0108	0.0072
0.4	-	-	0.0185	0.0159	0.0132	0.0106	0.0080	0.0053
0.5	-	0.0167	0.0146	0.0126	0.0105	0.0084	0.0063	0.0042
0.6	0.0172	0.0138	0.0121	0.0104	0.0087	0.0069	0.0052	0.0035
0.7	0.0147	0.0118	0.0103	0.0088	0.0074	0.0059	0.0044	0.0030
0.8	0.0128	0.0103	0.0090	0.0077	0.0064	0.0051	0.0039	0.0026
0.9	0.0114	0.0091	0.0080	0.0068	0.0057	0.0046	0.0034	0.0023
1.0	0.0102	0.0082	0.0072	0.0061	0.0051	0.0041	0.0031	0.0020
$\alpha = 0.050$	α_L							
α_U	0.0350	0.0400	0.0410	0.0420	0.0430	0.0450	0.0470	0.0480
0.1	-	-	-	-	-	-	-	-
0.2	-	-	-	-	-	-	0.0392	0.0263
0.3	-	-	-	-	-	0.0392	0.0237	0.0159
0.4	-	-	-	-	0.0392	0.0282	0.0170	0.0114
0.5	-	-	0.0392	0.0349	0.0306	0.0220	0.0132	0.0088
0.6	-	0.0357	0.0322	0.0287	0.0251	0.0180	0.0108	0.0072
0.7	-	0.0303	0.0273	0.0243	0.0213	0.0153	0.0092	0.0061
0.8	-	0.0263	0.0237	0.0211	0.0185	0.0132	0.0080	0.0053
0.9	0.0347	0.0233	0.0210	0.0186	0.0163	0.0117	0.0070	0.0047
1.0	0.0311	0.0208	0.0188	0.0167	0.0146	0.0105	0.0063	0.0042
$\alpha = 0.100$	α_L							
α_U	0.0750	0.0800	0.0085	0.0900	0.0920	0.0950	0.0970	0.0990
0.1	-	-	-	-	-	-	-	-
0.2	-	-	-	-	-	-	0.0583	0.0198
0.3	-	-	-	-	0.0769	0.0488	0.0296	0.0100
0.4	-	-	-	0.0645	0.0519	0.0328	0.0198	0.0066
0.5	-	-	0.0723	0.0488	0.0392	0.0247	0.0149	0.0050
0.6	-	0.0769	0.0583	0.0392	0.0315	0.0198	0.0119	0.0040
0.7	-	0.0645	0.0488	0.0328	0.0263	0.0165	0.0100	0.0033
0.8	0.0690	0.0556	0.0420	0.0282	0.0226	0.0142	0.0085	0.0029
0.9	0.0606	0.0488	0.0368	0.0247	0.0198	0.0124	0.0075	0.0025
1.0	0.0541	0.0435	0.0328	0.0220	0.0176	0.0110	0.0066	0.0022

Table 3.8: Sample pair of Stage 1 early stopping boundaries and corresponding second stage critical value for two-stage adaptive design using Fisher's, Tippett's, and Simes' combination functions.

	α_L	α_U	c_α	Results
Fisher's				
$c_\alpha \leq \alpha_L$	0.0035	0.4	0.0014	Continue to Stage 2 and reject H_0
$\alpha_L < c_\alpha < \alpha_U$	0.0010	0.4	0.0015	Continue to Stage 2 and reject H_0
Tippett's				
$\alpha_L < \frac{1}{2}c_\alpha < \alpha_U$	0.0072	0.4	0.0144	Reject H_0 and stop at Stage 1
$\frac{1}{2}c_\alpha < \alpha_L$	0.0075	0.4	0.0127	Reject H_0 and stop at Stage 1
Simes'				
$2\alpha_L < c_\alpha \leq \alpha_U$	0.0071	0.4	0.0143	Reject H_0 and stop at Stage 1
$\alpha_L < c_\alpha \leq \min(2\alpha_L, \alpha_U)$	0.0073	0.4	0.0135	Reject H_0 and stop at Stage 1
$c_\alpha \leq \alpha_L$	0.0085	0.4	0.0076	Reject H_0 and stop at Stage 1

CHAPTER 4

STEPDOWN-COMBINATION APPROACH FOR TWO HYPOTHESES IN TWO-STAGE COMBINATION TEST

It is very likely that multiple study objectives need to be achieved or multiple hypotheses need to be tested in practice. In this chapter, we extend the Bauer and Köhne's (1994) combination test approach to the multiple-hypothesis situations, and consider a specific stepwise-combination procedure to control the overall FWER strongly.

4.1 Notations

Assume we are testing an endpoint with two responses. In other words, we are testing two null hypotheses, H_1 and H_2, simultaneously. Let p_{ij} denote the individual p-value at each stage, where $i = 1, 2$ refers to the study stage and $j = 1, 2$ refers to the testing response or hypothesis (see Table 4.1). For example, p_{11} is the individual p-value for Response 1 at Stage 1 and p_{21} is the p-value for Response 1 at Stage 2. If either response or both continue to Stage 2, define a combination function $C^j(p_{1j}, p_{2j})$ for the two-stage combined p-value and define c_{α_j} as the second stage critical value for the combination test. For example, following the Bauer and Köhne's combination test principle, let $p_1 = C^1(p_{11}, p_{21}) = p_{11}p_{21}$ denote the combined p-value for Response 1 and $p_2 = C^2(p_{12}, p_{22}) = p_{12}p_{22}$ for Response 2. Let α_L and α_U denote the early rejection and early acceptance bounds for Stage 1. A two-stage-two-hypothesis adaptive test procedure is described as follows:

1. Define a test procedure for Stage 1, determining the stopping rules α_L and α_U for the interim decision.

2. Conduct Stage 1 of the study, resulting in p_{11} for Response 1 and p_{12} for Response 2.

3. Based on p_{11} and p_{12}, decide whether to stop at the interim (stop either Response 1 or Response 2 or both), or proceed the study to Stage 2.

    - Stop at the interim with a decision to reject H_1 if $p_{11} \leq \alpha_L$, to accept H_1

if $p_{11} > \alpha_U$, or to continue the investigation on Response 1 to Stage 2 if $\alpha_L < p_{11} \leq \alpha_U$ and result in Stage 2 p-value for Response 1, p_{21}. The two-stage combined p-value for Response 1 is $p_1 = C^1(p_{11}, p_{21})$.

- Stop at the interim with a decision to reject H_2 if $p_{12} \leq \alpha_L$, to accept H_2 if $p_{12} > \alpha_U$, or to continue the investigation on Response 1 to Stage 2 if $\alpha_L < p_{12} \leq \alpha_U$ and result in Stage 2 p-value for Response 2, p_{22}. The two-stage combined p-value for Response 2 is $p_2 = C^2(p_{12}, p_{22})$.

The possible scenarios for a two-hypothesis-two-stage adaptive design are:

- Both Responses stop at Stage 1.
  - $p_{11} \leq \alpha_L, p_{12} \leq \alpha_L$.
  - $p_{11} \leq \alpha_L, p_{12} > \alpha_U$.
  - $p_{11} > \alpha_U, p_{12} \leq \alpha_L$.
  - $p_{11} > \alpha_U, p_{12} > \alpha_U$.
- Response 1 stops at Stage 1, and Response 2 is continued to Stage 2.
  - $p_{11} \leq \alpha_L, \alpha_L \leq p_{12} \leq \alpha_U, p_2 = C^2(p_{12}, p_{22})$.
  - $p_{11} > \alpha_U, \alpha_L \leq p_{12} \leq \alpha_U, p_2 = C^2(p_{12}, p_{22})$.
- Response 2 stops at Stage 1, and Response 1 is continued to Stage 2.
  - $p_{12} \leq \alpha_L, \alpha_L \leq p_{11} \leq \alpha_U, p_1 = C^1(p_{11}, p_{21})$.
  - $p_{12} > \alpha_U, \alpha_L \leq p_{11} \leq \alpha_U, p_1 = C^1(p_{11}, p_{21})$.

- Both Responses continue to Stage 2, which implies $\alpha_L \leq p_{11} \leq \alpha_U$ and $\alpha_L \leq p_{12} \leq \alpha_U$.

  – $p_1 = C^1(p_{11}, p_{21}), p_2 = C^2(p_{12}, p_{22})$.

  The study $FWER$ is given by

$$\begin{aligned}
\text{FWER} &= Pr(p_{11} \leq \alpha_L, p_{12} \leq \alpha_L) + Pr(p_{11} \leq \alpha_L, p_{12} > \alpha_U) \\
&\quad + Pr(p_{11} > \alpha_U, p_{12} \leq \alpha_L) \\
&\quad + Pr(p_{11} \leq \alpha_L, \alpha_L \leq p_{12} \leq \alpha_U, p_2 \leq c_{\alpha_2}) \\
&\quad + Pr(p_{11} \leq \alpha_L, \alpha_L \leq p_{12} \leq \alpha_U, p_2 > c_{\alpha_2}) \\
&\quad + Pr(p_{11} > \alpha_U, \alpha_L \leq p_{12} \leq \alpha_U, p_2 > c_{\alpha_2}) \\
&\quad + Pr(p_{12} \leq \alpha_L, \alpha_L \leq p_{11} \leq \alpha_U, p_1 \leq c_{\alpha_1}) \\
&\quad + Pr(p_{12} \leq \alpha_L, \alpha_L \leq p_{11} \leq \alpha_U, p_1 > c_{\alpha_1}) \\
&\quad + Pr(p_{12} > \alpha_U, \alpha_L \leq p_{11} \leq \alpha_U, p_1 \leq c_{\alpha_1}) \\
&\quad + Pr(p_1 \leq c_{\alpha_1}, p_2 \leq c_{\alpha_2} | \alpha_{1L} \leq p_{11} \leq \alpha_U, \alpha_{2L} \leq p_{12} \leq \alpha_U) \\
&\quad + Pr(p_1 \leq c_{\alpha_1}, p_2 > c_{\alpha_2} | \alpha_L \geq p_{11} \leq \alpha_U, \alpha_L \leq p_{12} \leq \alpha_U) \\
&\quad + Pr(p_1 > c_{\alpha_1}, p_2 \leq c_{\alpha_2} | \alpha_L \leq p_{11} \leq \alpha_U, \alpha_L \leq p_{12} \leq \alpha_U).
\end{aligned}$$

Obviously, when both Response 1 and Response 2 are continued to Stage 2, multiplicity adjustment is necessary to analyze the combined p-values p_1 and p_2 for the overall control of the Type I error. To simplify the calculation, we only discuss how to determine the early stopping rules under the condition that both Responses 1 and 2 proceed to Stage 2 in this work. We then investigate whether

these resulting stopping cutoff boundaries can be applied to the situation where either response or both responses have an early stop at Stage 1. Bonferroni and the stepdown procedures along with the combination test are applied, separately.

4.2 Bonferroni-Combination Procedure

Bonferroni procedure is one of the most popular multiplicity adjustment procedures. It controls the FWER in a strong sense. Let's consider a situation that both Responses 1 and 2 proceed to Stage 2. Apply Bonferroni procedure as a natural extension to the Bauer-Köhne combination function method. We first conduct the combination test to the two-stage p-values of Response j, where $j = 1, 2$. We then perform the Bonferroni procedure to test the null hypotheses H_j based on the resulting combined p-values. Reject H_j for Response j, if the combined p-value $p_j \leq \alpha_j$, where $\alpha_j = \frac{\alpha}{2}$. The Type I error rate for Response j is given by

$$\alpha_L + \int_{\alpha_L}^{\alpha_U} \int_0^1 I(C^j(p_{1j}, p_{2j}) \leq c_{\alpha_j}) I(p_{2j} \leq 1) I(p_{1j} \leq 1) dp_{2j} dp_{1j} = \alpha_j. \qquad (4.1)$$

4.2.1 Apply to the Bauer-Köhne Combination Function Method

Following the Bauer-Köhne combination function method, reject H_j for Response j if $p_j = p_{1j}p_{2j} \leq c_{\alpha_j}$, where $c_{\alpha_j} = \exp(-\frac{1}{2}\chi^2_{4;1-\alpha_j})$ and $\chi^2_{4;1-\alpha_j}$ is the $(1-\alpha_j)$-quantile of the central χ^2-distribution with 4 degrees of freedom. Equation (4.1) for Response j can be written as

$$\alpha_L + \int_{\alpha_L}^{\alpha_U} \int_0^{c_{\alpha_j}/p_{1j}} dp_{2j} p_{1j} = \alpha_L + c_{\alpha_j}[\ln \alpha_U - \ln \alpha_L] = \alpha_j. \qquad (4.2)$$

As mentioned in earlier chapters, the Bauer-Köhne combination function method adds the restriction that $c_{\alpha_j} \leq \alpha_L$. Thus, α_L lies in the interval $[c_{\alpha_j}, \alpha]$. If the Stage 1 p-value for Response j $p_{1j} \leq c_{\alpha_j}$, one could stop the study at the interim for Response j with rejection of H_j, because the condition $0 \leq p_{2j} \leq 1$ guarantees that the combination test must reject H_j. The critical value c_{α_j} obtained based on both responses continued to Stage 2 can also be applied to the situation that one or both responses stop at Stage 1. For example, if $\alpha = 0.05$, we have $\alpha_1 = \alpha_2 = \frac{\alpha}{2} = 0.025, c_{\alpha_1} = c_{\alpha_2} = \exp(-\frac{1}{2}\chi^2_{4;1-\frac{\alpha}{2}}) = 0.0038$. Then for a pre-fixed early acceptance boundary $\alpha_U = 0.5$, we can easily get the early rejection boundary $\alpha_L = 0.0102$. More specifically, if $p_{1j} \leq 0.0102$ or $p_{1j} > 0.5$, stop the study at Stage 1 with rejection or acceptance of H_j for Response j respectively. If $0.0102 \leq p_{1j} < 0.5$, Response j is continued to Stage 2. If the resulting combined p-value $p_j = p_{1j}p_{2j} < c_{\alpha_j} = 0.0038$, reject H_j for Response j at Stage 2.

4.3 Stepdown-Combination Procedure

It is well known that Bonferroni procedure is too conservative. Let's consider a more powerful approach, stepdown testing procedure (see Figure 4.1). Rather than choosing the same early rejection and acceptance bounds α_L and α_U for both responses, we use different stopping rules for Response 1 and Response 2 in this procedure. Let α_{1L} and α_{1U} denote the early rejection and acceptance bounds for Response 1, and α_{2L} and α_{2U} for Response 2. Define the combined p-values for Response 1 and Response 2 are $p_1 = C^1(p_{11}, p_{21})$ and $p_2 = C^2(p_{12}, p_{22})$. Similar to

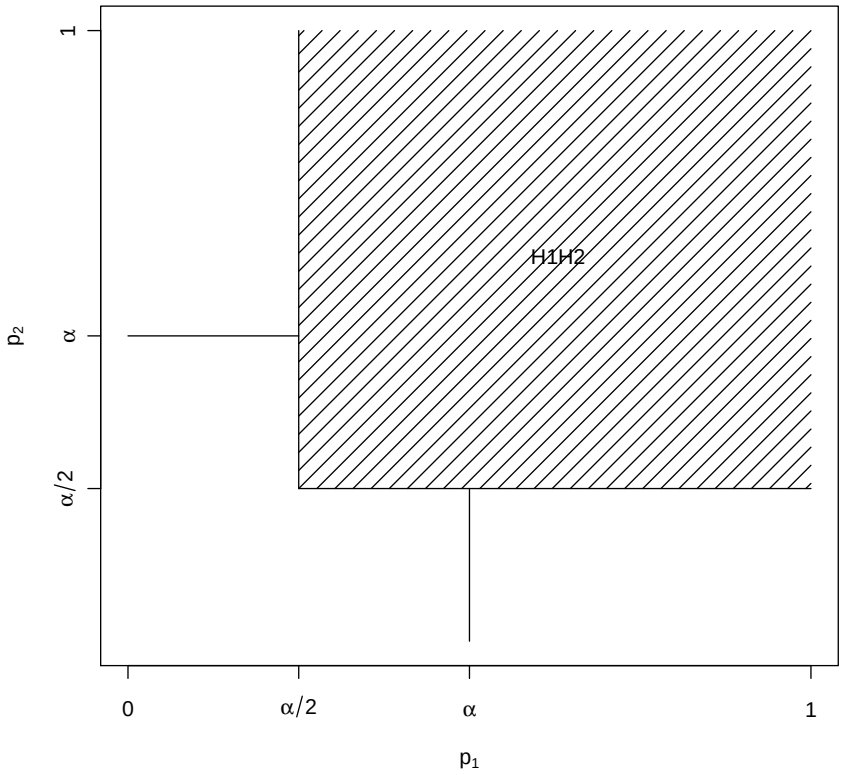

Figure 4.1: Holm's step-down procedure with two testing hypotheses.

the previous discussion, in order to simplify the calculation we first investigate the situation that both responses continue to Stage 2 which implies $\alpha_{1L} < p_{11} \leq \alpha_{1U}$ and $\alpha_{2L} < p_{12} \leq \alpha_{2U}$, and then evaluate whether the obtained second stage critical value can be applied to other situations where either response or both have an early stop at Stage 1. The proposed stepdown-combination approach is described as follows:

1. Obtain the combined two-stage p-values $p_1 = C^1(p_{11}, p_{21})$ and $p_2 = C^2(p_{12}, p_{22})$ via combination function for each response respectively.

2. Perform the stepdown testing procedure on the combined p-values p_1 and p_2, while controlling the FWER at a desired level α. First, order the combined p-values p_1 and p_2 so that $p_{(1)} \leq p_{(2)}$. Let $c_{(\alpha_1)}$ and $c_{(\alpha_2)}$ denote the ordered second stage critical values corresponding to $p_{(1)}$ and $p_{(2)}$. Then start with the most significant p-value $p_{(1)}$, and proceed to $p_{(2)}$ as needed.

   - If $p_{(1)} > c_{(\alpha_1)}$, accept H_1 and H_2 for both responses.
   - If $p_{(1)} \leq c_{(\alpha_1)}$, reject $H_{(1)}$ and go to the next step, looking into $p_{(2)}$. If $p_{(2)} > c_{(\alpha_2)}$, accept $H_{(2)}$. Otherwise, reject $H_{(2)}$.

The corresponding rejection region is $\{p_{(1)} \leq c_{(\alpha_1)} \text{ and } p_{(2)} > c_{(\alpha_2)}\} \bigcup \{p_{(1)} \leq c_{(\alpha_1)} \text{ and } p_{(2)} \leq c_{(\alpha_2)}\}$. The study FWER is given by

$$\begin{aligned} & Pr(p_{(1)} \leq c_{(\alpha_1)}, p_{(2)} > c_{(\alpha_2)}) + Pr(p_{(1)} \leq c_{(\alpha_1)}, p_{(2)} \leq c_{(\alpha_2)}) \\ = \quad & Pr(p_{(1)} \leq c_{(\alpha_1)}) = \alpha, \end{aligned} \tag{4.3}$$

which is equivalent to

$$\begin{aligned} & Pr(p_1 \leq c_{\alpha_1}, p_2 > c_{\alpha_2}) + Pr(p_1 \leq c_{\alpha_1}, p_2 \leq c_{\alpha_2}) \\ = & \int_0^{c_{\alpha_1}} \int_{c_{\alpha_2}}^1 f(p_1) f(p_2) dp_2 dp_1 + \int_0^{c_{\alpha_1}} \int_0^{c_{\alpha_2}} f(p_1) f(p_2) dp_2 dp_1 \\ = & \frac{\alpha}{2}, \end{aligned} \tag{4.4}$$

where $f(p_1)$ and $f(p_2)$ are the distribution functions for the combined p-values p_1 and p_2, respectively. More specifically, for a stepdown procedure we have

$$\begin{cases} Pr(p_1 \leq c_{\alpha_1}, p_2 > c_{\alpha_2}) + Pr(p_1 \leq c_{\alpha_1}, p_2 \leq c_{\alpha_2}) = \frac{\alpha}{2} \\ Pr(p_1 > c_{\alpha_1}, p_2 \leq c_{\alpha_2}) + Pr(p_1 \leq c_{\alpha_1}, p_2 \leq c_{\alpha_2}) = \alpha \end{cases} \tag{4.5}$$

Assuming the observed p_{11}, p_{21}, p_{12}, and p_{22} follow stochastically independent uniform distribution $[0, 1]$ under H_0, the Type I error probability for Response 1 and Response 2 is given by

$$\begin{cases} \text{Response 1: } \alpha_{1L} + \int_{\alpha_{1L}}^{\alpha_{1U}} \int_0^1 I(C^1(p_{11}, p_{21}) I(p_{21} \leq 1) I(p_{11} \leq 1) dp_{21} p_{11} = \alpha_1 \\ \text{Response 2: } \alpha_{2L} + \int_{\alpha_{2L}}^{\alpha_{2U}} \int_0^1 I(C^2(p_{12}, p_{22}) I(p_{22} \leq 1) I(p_{12} \leq 1) dp_{22} p_{12} = \alpha_2 \end{cases} \tag{4.6}$$

Mathematically, $c_{\alpha_1}, c_{\alpha_2}, \alpha_{1L}$ and α_{2L} can be obtained with prefixed α_{1U} and α_{2U} by equating equations (4.5) and (4.6).

This procedure can be easily applied to the proposed general formula for the overall Type I error rate for different combination functions.

4.3.1 Apply to the Bauer-Köhne Method

Assuming both responses are continued to Stage 2 and following the Bauer and Köhne's combination function approach (1994) with a restriction $c_\alpha \leq pha_L$, we first obtain the Stage 1 and Stage 2 combined p-values $p_1 = p_{11} p_{21}$ and $p_2 =$

$p_{12}p_{22}$ for either response respectively. Note that the combined p-values p_1 and p_2 do not follow the uniform distribution on $[0,1]$ any more. Instead, $-2\ln p_1$ and $-2\ln p_2$ follow a χ^2-distribution with 4 degrees of freedom. It is easy to obtain the distribution functions of p_1 and p_2, which are $f(p_1) = -\frac{1}{2}\ln p_1$ and $f(p_2) = -\frac{1}{2}\ln p_2$ where $0 < p_1, p_2 < 1$. Plug the distribution functions $f(p_1)$ and $f(p_1)$ into the equation (3.22). We obtain an expression for the second stage critical values c_{α_1} and c_{α_2}.

$$\begin{cases} \frac{1}{2}c_{\alpha_1}(1-\ln c_{\alpha_1}) = \alpha \\ \frac{1}{4}c_{\alpha_2}(1-\ln c_{\alpha_2}) = \alpha \end{cases} \tag{4.7}$$

For a specific α value, the numeric solutions for c_{α_1} and c_{α_2} can be obtained (see Table 4.2). Table 4.2 lists the numeric solutions for c_{α_1} and the corresponding Type I error α_1 as well as the numeric solutions for c_{α_2} and α_2 at the overall Type I error $\alpha = 0.010, 0.020, 0.025, 0.050, 0.10$.

Assuming the observed p_{11}, p_{21}, p_{12}, and p_{22} follow stochastically independent uniform distribution $[0,1]$ under H_0, the Type I error rate for either response (equation 4.6) can be written as

$$\begin{cases} \text{Response 1: } \alpha_{1L} + c_{\alpha_1}[\ln\alpha_{1U} - \ln\alpha_{1L}] = \alpha_1 \\ \text{Response 2: } \alpha_{2L} + c_{\alpha_2}[\ln\alpha_{2U} - \ln\alpha_{2L}] = \alpha_2 \end{cases} \tag{4.8}$$

Based on the numeric results for c_{α_1} and c_{α_2} obtained from equation (4.8), the early rejection bounds α_{1L} and α_{2L} can be determined by equating equation (4.8) with the prefixed early acceptance boundaries α_{1U} and α_{2U} for Responses 1 and 2 (see Table 4.3). Table 4.3 gives the numeric solutions for the early rejection boundary α_L

with the prefixed early acceptance boundary α_U for different choices of the overall Type I error $\alpha = 0.010, 0.020, 0.025, 0.050, 0.100$.

For example, with the overall Type I error $\alpha = 0.050$ and the prefixed early acceptance boundary α_{1U} at 0.5, the early rejection boundary for Response 1 is $\alpha_{1L} = 0.0548$. In other words, the minimum requirement for continuation to Stage 2 of Response 1 is $0.0548 < p_{11} \leq 0.5$. If $p_{11} > 0.5$, accept H_1 and H_2 and stop the trial at Stage 1. If $p_{11} \leq 0.0548$, reject H_{11} and stop the analysis for response 1 at Stage 1. If $0.0548 < p_{11} \leq 0.5$, continue Response 1 to Stage 2 and look at the combined two-stage p-value $p_1 = p_{11}p_{21}$. If $p_1 \leq c_{\alpha_1}$ where $c_{\alpha_1} = 0.02045$, then reject H_1 at Stage 2. Figure 4.2 shows the rejection region of Response 1 in the $p_{11} - p_{21}$ plane ($\alpha = 0.050, \alpha_{1U} = 0.5, \alpha_{1L} = 0.0548$). Figure 4.3 shows the rejection region of Response 2 in the $p_{12} - p_{22}$ plane ($\alpha = 0.050, \alpha_{2U} = 0.5, \alpha_{2L} = 0.1937$). The area below the hyperbola $p_{2j} = c_{\alpha_j}/p_{1j}$ is the rejection region for Fisher's combination test. The rectangular region to the right of α_{jU} corresponds to early stopping with the "lack of efficacy" decision. The area to the left of α_{jL} corresponds to early rejection of H_{1j}. Hence, the solid curve confines the rejection region of the proposed procedure.

Now consider the situation that either or both responses stop at Stage 1. For Response (1), if $p_{(11)} > \alpha_{1U}$ or $p_{(11)} \leq \alpha_{1L}$, stop the investigation on Response (1) at Stage 1 and accept or reject $H_{(1)}$ for Responses (1) accordingly. For Response (2), if $p_{(12)} > \alpha_{2U}$ or $p_{(12)} \leq \alpha_{2L}$, stop the investigation on Response (2) at Stage 1 and accept or reject $H_{(2)}$ for Response (2) accordingly. Similar to the previous

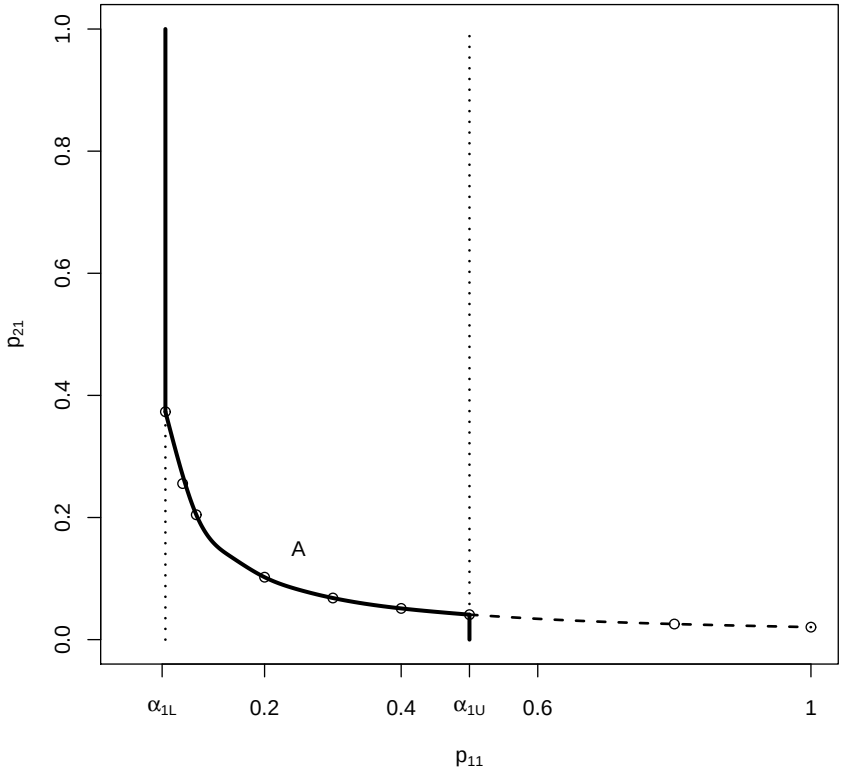

Figure 4.2: Plot for the rejection region of Response 1 in the $p_{11} - p_{21}$ plane ($\alpha = 0.050, \alpha_{1U} = 0.5, \alpha_{1L} = 0.0548$). A represents $p_{11}p_{21} = c_{\alpha_1}$.

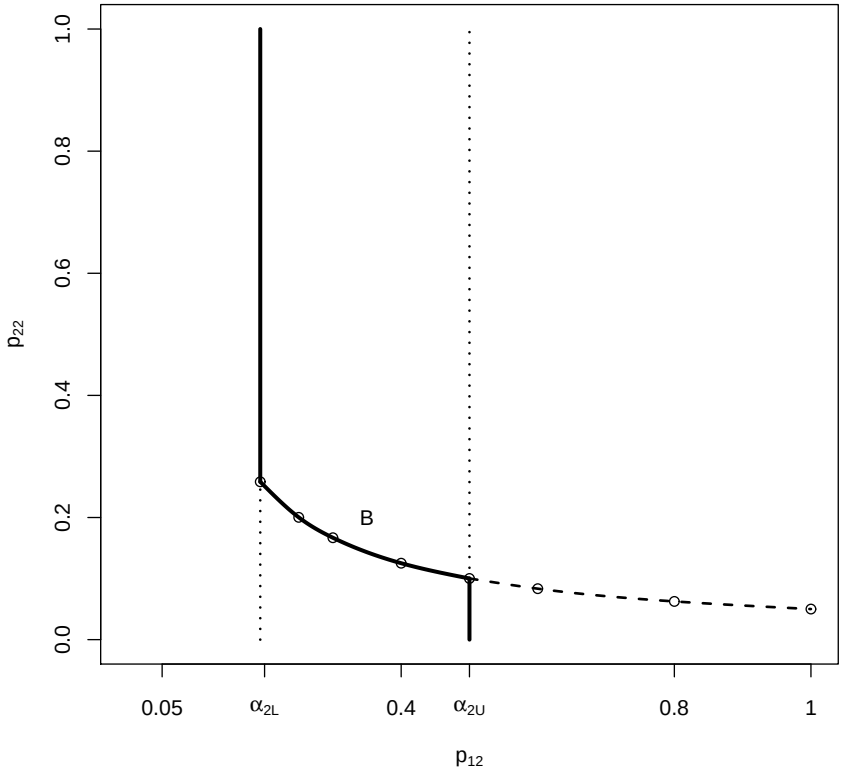

Figure 4.3: Plot for the rejection region of Response 2 in the $p_{12} - p_{22}$ plane ($\alpha = 0.050, \alpha_{2U} = 0.5, \alpha_{2L} = 0.1937$). B represents $p_{12}p_{22} = c_{\alpha_2}$.

discussion, the condition $0 \leq p_{(21)} \leq 1$ guarantees that the combination test must reject $H_{(1)}$, if $p_{(11)} \leq c_{(\alpha_1)}$. We can easily get the same conclusion for $p_{(12)}$ and $p_{(22)}$. Hence, the critical value $c_{(\alpha_j)}$ and early rejection bound α_{jL} based on both responses continued to Stage 2 can be applied to the situation that either or both responses stop at Stage 1 as well.

Similarly, based on the numeric results for c_{α_1} and c_{α_2} obtained from equation (3.21), the early rejection bounds α_{1L} and α_{2L} can be determined by equating equation (3.22) with the prefixed early acceptance boundaries α_{1U} and α_{2U} for the other two situations as well.

4.4 Example

We use data from a clinical study on colorectal cancer patients as an example to illustrate an application of the stepdown-combination approach described above. It was a group sequential study to examine two systemic chemotherapy regimens for metastatic colorectal carcinoma, MOF-Strep versus MTX-FU (Pocock et al.,1987). There were two endpoints of interest, tumor response after 2 months' treatment and patient survival time. The original trial aimed to enroll (at most) five groups of 17 patients per arm. At the first interim analysis, there were six tumor responses on MOF-Strep and one on MTX-FU, yielding an uncorrected $\chi^2 = 4.50$ ($p_{11} = 0.034$). For survival time, the log-rank test yielded $\chi^2 = 2.11$ ($p_{12} = 0.15$) with 18 deaths all together, 7 on MOF-Strep and 11 on MTX-FU.

We intended to use this data and apply the two-stage-two-hypotheses de-

sign with an overall Type error rate at 0.05. We consider both are primary endpoints and assume these two endpoints are independent. To have a meaningful comparison, we keep the early acceptance boundary the same at 0.5. For illustration purpose, we only discuss the Bauer-Köhne method for Fisher's combination function below. The results and conclusions are summarized in Table 4.4.

Bonferroni-Combination Approach

The early rejection boundary α_L can be chosen at 0.0102, providing a control of the Type I error rate at 0.025 for either response. It is obvious that either tumor response ($p_{11} = 0.034$) or survival time ($p_{12} = 0.15$) did not qualify for an early stop at Stage 1 and would continue to Stage 2 with a prefixed early acceptance boundary $\alpha_U = 0.5$.

4.4.1 Stepdown-Combination Approach

Table 4.4 offers the numeric solutions for c_{α_1}, $c_{\alpha_2}, \alpha_{1L}$, and α_{2L}. With the overall Type I error at 0.05, c_{α_1} is 0.02045 and c_{α_2} is 0.05007. With the early acceptance boundaries $\alpha_{1U} = 0.5$ and $\alpha_{2U} = 0.5$, we can set the early rejection boundary α_{1L} at 0.0548 for tumor response and 0.1937 for survival time. Since the first stage p-value for tumor response p_{11} is 0.034, the trial would stop the investigation on tumor response at Stage 1 for efficacy. Similarly, with a first stage p-value for survival time of $p_{12} = 0.15$, the trial would stop the investigation on survival time at Stage 1 as well.

In short, in this example the application of the stepdown-combination ap-

Table 4.1: Two-stage adaptive design with a two-response endpoint.

	Response 1	Response 2
Stage 1	p_{11}	p_{12}
	(H_1)	(H_2)
Stage 2	p_{21}	p_{22}
	(H_1)	(H_2)
Combined p-value	$p_1 = C^1(p_{11}, p_{21})$	$p_2 = C^2(p_{12}, p_{22})$

Table 4.2: The critical values c_{α_1} and c_{α_2} for Response 1 and Response 2, based on the Bauer and Köhne's combination approach (1994) and the proposed Stepdown-Combination procedure.

	α				
	0.010	0.020	0.025	0.050	0.100
c_{α_1}	0.00293	0.00665	0.0087	0.02045	0.05007
$\chi^2_{4(1-\alpha_1)}$	11.6655	10.0263	9.4889	7.7795	5.9887
α_1	0.02002	0.03999	0.04998	0.099998	0.199994
c_{α_2}	0.00665	0.01547	0.02045	0.05007	0.13234
$\chi^2_{4(1-\alpha_2)}$	10.0263	8.3377	7.7795	5.9887	4.0448
α_2	0.03999	0.07996	0.099998	0.199994	0.39998

proach led to a possible early stopping due to efficacy at Stage 1, suggesting that with good planning a two-stage-two-hypothesis adaptive design can reduce lead time and save cost in clinical development.

Table 4.3: The critical values, c_{α_1} and c_{α_2}, and the early rejection and acceptance bounds $\alpha_{1L}, \alpha_{1U}, \alpha_{2L}, \alpha_{2U}$, based on Fisher's Combination Function and proposed Stepdown-Combination approach.

	α	0.010	0.020	0.025	0.050	0.100
α_{1U}	c_{α_1}	0.00293	0.00665	0.0087	0.02045	0.05007
				α_{1L}		
0.3		0.0101	0.0229	0.0299	0.0703	0.1722
0.4		0.0089	0.0201	0.0263	0.0618	0.1513
0.5		0.0078	0.0178	0.0233	0.0548	0.1341
0.6		0.0070	0.0158	0.0207	0.0486	0.1190
0.7		0.0061	0.0140	0.0183	0.0429	0.1050
0.8		0.0053	0.0121	0.0159	0.0373	0.0913
0.9		0.0045	0.0102	0.0133	0.0313	0.0767
1.0		0.0031	0.0070	0.0092	0.0211	0.0510
α_{2U}	c_{α_2}	0.00665	0.01547	0.02045	0.05007	0.13234
				α_{2L}		
0.3		0.0229	0.0703	0.0922	0.1972	0.4019
0.4		0.0201	0.0682	0.0901	0.1952	0.4000
0.5		0.0178	0.0666	0.0885	0.1937	0.3985
0.6		0.0258	0.0652	0.0872	0.1924	0.3972
0.7		0.0140	0.0641	0.0861	0.1914	0.3962
0.8		0.0122	0.0631	0.0851	0.1905	0.3953
0.9		0.0102	0.0622	0.0843	0.1896	0.3945
1.0		0.0070	0.0614	0.0835	0.1889	0.3938

Table 4.4: Sample pairs of Stage 1 early stopping boundaries and corresponding second stage critical value for two-stage-two-hypothesis adaptive design following Bonferonni-combination and the proposed Stepdown-Combination approach ($\alpha = 0.05$).

	Bonferonni-Combination	Stepdown-Combination
Response 1	$\alpha_{1L} = 0.0102, \alpha_{1U} = 0.5$	$\alpha_{2L} = 0.0548, \alpha_{2U} = 0.5$
Response 2	$\alpha_{2L} = 0.0102, \alpha_{1U} = 0.5$	$\alpha_{2L} = 0.1937, \alpha_{2U} = 0.5$
Results	Continue to Stage 2	Reject H_1 and H_2 and stop at Stage 1

CHAPTER 5

OVERALL FDR CONTROL FOR MULTIPLE HYPOTHESES IN TWO-STAGE COMBINATION TEST

In this chapter, we focus on our proposed BH type procedures to control the FDR in a two-stage adaptive design with multiple endpoints using combination tests. We provide theoretical proofs of the FDR control of these methods, assuming that the p-values are independent across the hypotheses and p-clud dependence (Brannath et al., 2002) across the stages. We present numerical evidence from

simulation studies that the FDR is well controlled under independence and this control can continue to hold in some commonly occurring dependence situations. We show that our proposed methods take advantage of more information in the data, and thus are more effective, flexible, and powerful.

5.1 Motivation

Gene association or expression studies usually involve a large number of endpoints (i.e., genetic markers) are often expensive. Multi-stage adaptive design can be cost effective and efficient. In a multi-stage design, genes are screened in early stages and selected genes are further investigated in later stages using additional observations. For multi-stage adaptive designs, as for single stage designs, multiplicity in simultaneous tests is an important issue. To address the multiplicity concern in simultaneous testing of the hypotheses associated with the endpoints, controlling the familywise error rate (FWER), the probability of at least one type I error among all hypotheses, is a commonly applied concept. However, these studies are often exploratory, so controlling the false discovery rate (FDR), which is the expected proportion of type I errors among all rejected hypotheses, is more appropriate than controlling the FWER (Weller et al., 1998; Benjamini and Hochberg, 1995; and Storey and Tibshirani, 2003). Moreover, with large number of hypotheses typically being tested in these studies, better power can be achieved in a multiple testing method under the FDR framework than under the more conservative FWER framework.

Adaptive designs with multiple endpoints have been considered in the literature under both the FWER and FDR frameworks. Miller et al. (2001) suggested using a two-stage design in gene experiments, and proposed using the Bonferroni method to control the FWER in testing the hypotheses selected at the first stage, although only the second stage observations are used for this method. This was later improved by Satagopan and Elston (2003) by incorporating the first stage data through group sequential schemes in the final Bonferroni test. Zehetmayer et al. (2005) considered a two-stage adaptive design where promising hypotheses are selected using a constant rejection threshold for each p-value at the first stage and an estimation based approach to controlling the FDR asymptotically (as the number of hypotheses goes to infinity) was taken (Storey, 2002; Storey, Taylor and Siegmund, 2004) at the second stage to test the selected hypotheses using more observations. Zehetmayer et al. (2008) have extended this work from two-stage to multi-stage adaptive designs under both FDR and FWER frameworks, and provided useful insights into the power performance of optimized multi-stage adaptive designs with respect to the number of stages, and into the power difference between optimized integrated design and optimized pilot design. Posch et al. (2009) showed that a data-dependent sample size increase for all the hypotheses simultaneously in a multi-stage adaptive design has no effect on the asymptotic (as the number of hypotheses goes to infinity) control of the FDR if the hypotheses to be rejected are determined only by the test at the final interim analysis, under all scenarios except the global null hypothesis when all the null hypotheses are true.

Construction of methods with FWER or FDR control in the setting of a two-stage adaptive design allowing reduction in the number of tested hypotheses at the interim analysis has been discussed, as a separate issue from sample size adaptations, in Bauer and Kieser (1999) and Kieser, Bauer and Lehmacher (1999) , who presented methods with the FWER control, and in Victor and Hommel (2007) who focused on controlling the FDR in terms of a generalized global p-values. We revisit this issue in the present paper, but focusing primarily on the FDR control in a non-asymptotic setting (with the number of hypothesis not being infinitely large).

Our motivation behind this paper lies in the fact that the theory presented so far (see, for instance, Victor and Hommel, 2007) towards developing an FDR controlling procedure in the setting of a two-stage adaptive design with combination tests does not seem to be as simple as one would hope for. Moreover, it does not allow setting boundaries on the first stage p-values in terms of FDR and operate in a manner that would be a natural extension of standard single-stage FDR controlling methods, like the BH (Benjamini and Hochberg, 1995) or methods related to it, from a single-stage to a two-stage design setting. So, we consider the following to be our main problem in this paper:

> To construct an FDR controlling procedure for simultaneous testing of the null hypotheses associated with multiple endpoints in the setting of a two-stage adaptive design where the hypotheses are sequentially screened at the first stage as rejected or accepted based on pre-specified boundaries on their p-values in terms of the FDR, and the null hypotheses that are left out at the first stage are again sequentially tested at the second stage having combined their p-values from the two stages through a combination function.

We propose two FDR controlling procedures, one extending the original

single-stage BH procedure, which we call the BH-TSADC Procedure (BH type procedure for two-stage adaptive design with combination tests), and the other extending an adaptive version of the single-stage BH procedure incorporating an estimate of the number of true null hypotheses, which we call the Plug-In BH-TSADC Procedure, from single-stage to a two-stage setting. Let (p_{1i}, p_{2i}) be the pair of first- and second-stage p-values corresponding to the ith null hypothesis. We provide a theoretical proof of the FDR control of the proposed procedures under the assumption that the (p_{1i}, p_{2i})'s are independent and those corresponding to the true null hypotheses are identically distributed as (p_1, p_2) satisfying the p-clud property (Brannath et al., 2002), and some standard assumption on the combination function. We consider two special types of combination function, Fisher's and Simes', which are often used in multiple testing applications, and present explicit formulas for probabilities involving them that would be useful to carry out the proposed procedures at the second stage either using critical values that can be determined before observing the p-values or based on estimated FDR's that can be obtained after observing the p-values.

We carried out extensive simulations to see how well the proposed procedures control the FDR and perform in terms of power under independence, and how selections of different early stopping boundaries affect the FDR and power. In order to provide a relatively fair comparison to the classic BH procedure under independence, we compared our proposed procedure to the BH method based on p-values from the first stage data and from full data across two stages respectively. Simu-

lations were also performed to evaluated whether or not the proposed procedures can continue to control the FDR under the different types of (positive) dependence among the underlying test statistics we consider, such as equal, clumpy and auto-regressive of order one [AR(1)] dependence. Our simulation studies indicate that between the two proposed procedures, the BH-TSADC seems to be the better choice in terms of controlling the FDR and power improvement over the single-stage BH procedure when π_0, the proportion of true nulls, is large. If π_0 is not large, the Plug-In BH-TSADC procedure is better, but it might lose the FDR control when the p-values exhibit equal or AR(1) type dependence with a large equal- or auto-correlation. Cost efficiency of such a two-stage design with the overall FDR control was discussed via simulation studies as well.

We applied the proposed procedures to reanalyze the data on multiple myeloma considered before by Zehetmayer et al. (2008), of course, for a different purpose. The data consist of a set of 12625 gene expression measurements for each of 36 patients with bone lytic lesions and 36 patients in a control group without such lesions. We considered this data in a two-stage framework, with the first 18 subjects per group for Stage 1 and the next 18 per group for Stage 2. With some pre-chosen early rejection and acceptance boundaries, these procedures produce significantly more discoveries than single-stage BH procedure based on the first stage data at the same FDR level.

5.2 Controlling the FDR in a Single-Stage Design

Suppose that there are m endpoints and the corresponding null hypotheses H_i, $i = 1, \ldots, m$, are to be simultaneously tested based on their respective p-values p_i, $i = 1, \ldots, m$, obtained in a single-stage design. The FDR of a multiple testing method that rejects R and falsely rejects V null hypotheses is $E(\text{FDP})$, where FDP $= V/\max\{R, 1\}$ is the false discovery proportion. Multiple testing is often carried out using a stepwise procedure defined in terms of $p_{(1)} \leq \cdots \leq p_{(m)}$, the ordered p-values. With $H_{(i)}$ the null hypothesis corresponding to $p_{(i)}$, a stepup procedure with critical values $\gamma_1 \leq \cdots \leq \gamma_m$ rejects $H_{(i)}$ for all $i \leq k = \max\{j : p_{(j)} \leq \gamma_j\}$, provided the maximum exists; otherwise, it accepts all null hypotheses. A stepdown procedure, on the other hand, with these same critical values rejects $H_{(i)}$ for all $i \leq k = \max\{j : p_{(i)} \leq \gamma_i \text{ for all } i \leq j\}$, provided the maximum exists, otherwise, accepts all null hypotheses. The following are formulas for the FDR's of a stepup or single-step procedure (when the critical values are same in a stepup procedure) and a stepdown procedure in a single-stage design, which can guide us in developing stepwise procedures controlling the FDR in a two-stage design. We will use the notation FDR_1 for the FDR of a procedure in a single-stage design.

Result 1. (Sarkar, 2008b). Consider a stepup or stepdown method for testing m null hypotheses based on their p-values p_i, $i = 1, \ldots, m$, and critical values $\gamma_1 \leq \cdots \leq \gamma_m$ in a single-stage design. The FDR of this method is given by

$$\text{FDR}_1 \leq \sum_{i \in J_0} E\left[\frac{I\left(p_i \leq \gamma_{R_{m-1}^{(-i)}(\gamma_2, \ldots, \gamma_m)+1}\right)}{R_{m-1}^{(-i)}(\gamma_2, \ldots, \gamma_m) + 1}\right],$$

with equality holding in the case of stepup method, where I is the indicator function, J_0 is the set of indices of the true null hypotheses, and $R_{m-1}^{(-i)}(\gamma_2, \ldots, \gamma_m)$ is the number of rejections in testing the $m-1$ null hypotheses other than H_i based on their p-values and using the same type of stepwise method with the critical values $\gamma_2 \leq \cdots \leq \gamma_m$.

With p_i having the cdf $F(u)$ when H_i is true, the FDR of a stepup or stepdown method with the thresholds γ_i, $i = 1, \ldots, m$, under independence of the p-values, satisfies the following:

$$\mathrm{FDR}_1 \leq \sum_{i \in J_0} E\left(\frac{F(\gamma_{R_{m-1}^{(-i)}(\gamma_2, \ldots, \gamma_m)+1})}{R_{m-1}^{(-i)}(\gamma_2, \ldots, \gamma_m) + 1} \right).$$

When F is the cdf of $U(0,1)$ and these thresholds are chosen as $\gamma_i = i\alpha/m$, $i = 1, \ldots, m$, the FDR equals $\pi_0 \alpha$ for the stepup and is less than or equal to $\pi_0 \alpha$ for the stepdown method, where π_0 is the proportion of true nulls, and hence the FDR is controlled at α. This stepup method is the so called BH method (Benjamini and Hochberg, 1995), the most commonly used FDR controlling procedure in a single-stage deign. The FDR is bounded above by $\pi_0 \alpha$ for the BH as well as its stepdown analog under certain type of positive dependence condition among the p-values (Benjamini and Yekutieli, 2001; Sarkar, 2002, 2008b).

The idea of improving the FDR control of the BH method by plugging into it a suitable estimate $\hat{\pi}_0$ of π_0, that is, by considering the modified p-values $\hat{\pi}_0 p_i$, rather than the original p-values, in the BH method, was introduced by Benjamini and Hochberg (2000), which was later brought into the estimation based approach to controlling the FDR by Storey (2002). A number of such plugged-in versions of

the BH method with proven and improved FDR control mostly under independence have been put forward based on different methods of estimating π_0 (for instance, Benjamini, Krieger, and Yekutieli, 2006; Blanchard and Roquain, 2009; Gavrilov, Benjamini and Sarkar, 2009; Sarkar, 2008b; and Storey, Taylor and Siegmund, 2004).

5.3 Controlling the FDR in a Two-Stage Adaptive Design

Now suppose that the m null hypotheses H_i, $i = 1, \ldots, m$, are to be simultaneously tested in a two-stage adaptive design setting. When testing a single hypothesis, say H_i, the theory of two-stage combination test can be described as follows: Given p_{1i}, the p-value available for H_i at the first stage, and two constants $\lambda < \lambda'$, make an early decision regarding the hypothesis by rejecting it if $p_{1i} \leq \lambda$, accepting it if $p_{1i} > \lambda'$, and continuing to test it at the second stage if $\lambda < p_{1i} \leq \lambda'$. At the second stage, combine p_{1i} with the additional p-value p_{2i} available for H_i using a combination function $C(p_{1i}, p_{2i})$ and reject H_i if $C(p_{1i}, p_{2i}) \leq \gamma$, for some constant γ. The constants λ, λ' and γ are determined subject to a control of the type I error rate by the test.

For simultaneous testing, we consider a natural extension of this theory from single to multiple testing. More specifically, given the first-stage p-value p_{1i} corresponding to H_i for $i = 1, \ldots, m$, we first determine two thresholds $0 \leq \hat{\lambda} < \hat{\lambda}' \leq 1$, stochastic or non-stochastic, and make an early decision regarding the hypotheses at this stage by rejecting H_i if $p_{1i} \leq \hat{\lambda}$, accepting H_i if $p_{1i} > \hat{\lambda}'$, and

continuing to test H_i at the second stage if $\hat{\lambda} < p_{1i} \leq \hat{\lambda}'$. At the second stage, we use the additional p-value p_{2i} available for a follow-up hypothesis H_i and combine it with p_{1i} using the combination function $C(p_{1i}, p_{2i})$. The final decision is taken on the follow-up hypotheses at the second stage by determining another threshold $\hat{\gamma}$, again stochastic or non-stochastic, and by rejecting the follow-up hypothesis H_i if $C(p_{1i}, p_{2i}) \leq \hat{\gamma}$. Both first-stage and second-stage thresholds are to be determined in such a way that the overall FDR is controlled at the desired level α.

Let $p_{1(1)} \leq \cdots \leq p_{1(m)}$ be the ordered versions of the first-stage p-values, with $H_{(i)}$ being the null hypotheses corresponding to $p_{1(i)}$, $i = 1, \ldots, m$, and $q_i = C(p_{1i}, p_{2i})$. We describe in the following a general multiple testing procedure based on the above theory, before proposing our FDR controlling procedures that will be of this type.

A General Stepwise Procedure.

1. For two non-decreasing sequences of constants $\lambda_1 \leq \cdots \leq \lambda_m$ and $\lambda'_1 \leq \cdots \leq \lambda'_m$, with $\lambda_i < \lambda'_i$ for all $i = 1, \ldots, m$, and the first-stage p-values p_{1i}, $i = 1, \ldots, m$, define two thresholds as follows: $R_1 = \max\{1 \leq i \leq m : p_{1(j)} \leq \lambda_j \text{ for all } j \leq i\}$ and $S_1 = \max\{1 \leq i \leq m : p_{1(i)} \leq \lambda'_i\}$, where $0 \leq R_1 \leq S_1 \leq m$ and R_1 or S_1 equals zero if the corresponding maximum does not exist. Reject $H_{(i)}$ for all $i \leq R_1$, accept $H_{(i)}$ for all $i > S_1$, and continue testing $H_{(i)}$ at the second stage for all i such that $R_1 < i \leq S_1$.

2. At the second stage, consider $q_{(i)}$, $i = 1, \ldots, S_1 - R_1$, the ordered versions of the combined p-values $q_i = C(p_{1i}, p_{2i})$, $i = 1, \ldots, S_1 - R_1$, for the follow-up null

hypotheses, and find $R_2(R_1, S_1) = \max\{1 \le i \le S_1 - R_1 : q_{(i)} \le \gamma_{R_1+i}\}$, given another non-decreasing sequence of constants $\gamma_{r_1+1}(r_1, s_1) \le \cdots \le \gamma_{s_1}(r_1, s_1)$, for every fixed $r_1 < s_1$. Reject the follow-up null hypothesis $H_{(i)}$ corresponding to $q_{(i)}$ for all $i \le R_2$ if this maximum exists, otherwise, reject none of the follow-up null hypotheses.

REMARK 1. We should point out that the above two-stage procedure screens out the null hypotheses at the first stage by accepting those with relatively large p-values through a stepup procedure and by rejecting those with relatively small p-values through a stepdown procedure. At the second stage, it applies a stepup procedure to the combined p-values. Conceptually, one could have used any type of multiple testing procedure to screen out the null hypotheses at the fisrt stage and to test the follow-up null hypotheses at the second stage. However, the particular types of stepwise procedure we have chosen at the two stages provide flexibility in terms of developing a formula for the FDR and eventually determining explicitly the thresholds we need to control the FDR at the desired level.

Let V_1 and V_2 denote the total numbers of falsely rejected among all the R_1 null hypotheses rejected at the first stage and the R_2 follow-up null hypotheses rejected at the second stage, respectively, in the above procedure. Then, the overall FDR in this two-stage procedure is given by

$$\mathrm{FDR}_{12} = E\left[\frac{V_1 + V_2}{\max\{R_1 + R_2, 1\}}\right].$$

The following theorem will guide us in determining the first- and second-

stage thresholds in the above procedure that will provide a control of FDR_{12} at the desired level. This is one of the procedures that we will propose in this work. Before stating the theorem, we need to define some notations.

Let $R_1^{(-i)}$ be defined as R_1 in terms of the $m-1$ first-stage p-values $\{p_{11},\ldots,p_{1m}\}\setminus\{p_{1i}\}$ and the sequence of constants $\lambda_2 \le \cdots \le \lambda_m$, $\tilde{R}_1^{(-i)}$ and $S_1^{(-i)}$ be defined as R_1 and S_1, respectively, in terms of $\{p_{11},\ldots,p_{1m}\}\setminus\{p_{1i}\}$ and the two sequences of constants $\lambda_1 \le \cdots \le \lambda_{m-1}$ and $\lambda'_2 \le \cdots \le \lambda'_m$, and $R_2^{(-i)}$ be defined as R_2 with R_1 replaced by $\tilde{R}_1^{(-i)}$ and S_1 replaced by $S_1^{(-i)}+1$ and noting the number of rejected follow-up null hypotheses based on all the combined p-values except the q_i and the critical values other than the first one; that is,

$$\begin{aligned} R_2^{(-i)} &\equiv R_2^{(-i)}(\tilde{R}_1^{(-i)}, S_1^{(-i)}+1) \\ &= \max\{1 \le j \le S_1^{(-i)} - \tilde{R}_1^{(-i)} : q_{(j)}^{(-i)} \le \gamma_{\tilde{R}_1^{(-i)}+j+1}(\tilde{R}_1^{(-i)}, S_1^{(-i)}+1)\}, \end{aligned}$$

where $q_{(j)}^{(-i)}$'s are the ordered versions of the combined p-values for the follow-up null hypotheses except the q_i.

Theorem 1. The FDR of the above general multiple testing procedure satisfies the following inequality

$$\begin{aligned} \mathrm{FDR}_{12} \le & \sum_{i\in J_0} E\left[\frac{I(p_{1i} \le \lambda_{R_1^{(-i)}+1})}{R_1^{(-i)}+1}\right] + \\ & \sum_{i\in J_0} E\left[\frac{I(\lambda_{\tilde{R}_1^{(-i)}+1} < p_{1i} \le \lambda'_{S_1^{(-i)}+1},\ q_i \le \gamma_{\tilde{R}_1^{(-i)}+R_2^{(-i)}+1, S_1^{(-i)}+1})}{\tilde{R}_1^{(-i)}+R_2^{(-i)}+1}\right]. \end{aligned}$$

Proof of Theorem 1.

$$\mathrm{FDR}_{12} = E\left[\frac{V_1+V_2}{\max\{R_1+R_2,1\}}\right] \le E\left[\frac{V_1}{\max\{R_1,1\}}\right] + E\left[\frac{V_2}{\max\{R_1+R_2,1\}}\right].$$

Now,

$$E\left[\frac{V_1}{\max\{R_1,1\}}\right] = \sum_{i\in J_0} E\left[\frac{I(p_{1i}\le\lambda_{R_1})}{\max\{R_1,1\}}\right] = \sum_{i\in J_0} E\left[\frac{I(p_{1i}\le\lambda_{R_1})}{\max\{R_1,1\}}\right]$$
$$\le \sum_{i\in J_0} E\left[\frac{I(p_{1i}\le\lambda_{R_1^{(-i)}+1})}{R_1^{(-i)}+1}\right];$$

(as shown in Sarkar, 2008; see also Result 1). And,

$$E\left[\frac{V_2}{\max\{R_1+R_2,1\}}\right]$$
$$= \sum_{i\in J_0} E\left[\frac{I(\lambda_{R_1+1}<p_{1i}\le\lambda'_{S_1},\ q_i\le\gamma_{R_1+R_2,S_1}, S_1>R_1, R_2>0)}{R_1+R_2}\right]. \quad (5.1)$$

Writing R_2 more explicitly in terms of R_1 and S_1, we see that the expression in (5.1) is equal to

$$\sum_{i\in J_0}\sum_{s_1=1}^{m}\sum_{r_1=0}^{s_1-1}\sum_{r_2=1}^{s_1-r_1}$$
$$E\left[\frac{I(\lambda_{r_1+1}<p_{1i}\le\lambda'_{s_1},\ q_i\le\gamma_{r_1+r_2,s_1}, R_1=r_1, S_1=s_1, R_2(r_1,s_1)=r_2)}{r_1+r_2}\right]$$
$$= \sum_{i\in J_0}\sum_{s_1=1}^{m}\sum_{r_1=0}^{s_1-1}\sum_{r_2=1}^{s_1-r_1}$$
$$E\left[\frac{I(\lambda_{r_1+1}<p_{1i}\le\lambda'_{s_1},\ q_i\le\gamma_{r_1+r_2,s_1}\tilde{R}_1^{(-i)}=r_1, S_1^{(-i)}=s_1-1, R_2^{(-i)}(r_1,s_1)=r_2-1)}{r_1+r_2}\right]$$
$$= \sum_{i\in J_0}\sum_{s_1=0}^{m-1}\sum_{r_1=0}^{s_1}\sum_{r_2=0}^{s_1-r_1}$$
$$E\left[\frac{I(\lambda_{r_1+1}<p_{1i}\le\lambda'_{s_1+1},\ q_i\le\gamma_{r_1+r_2+1,s_1+1}, \tilde{R}_1^{(-i)}=r_1, S_1^{(-i)}=s_1, R_2^{(-i)}(r_1,s_1+1)=r_2)}{r_1+r_2+1}\right]$$
$$= \sum_{i\in J_0} E\left[\frac{I(\lambda_{\tilde{R}_1^{(-i)}+1}<p_{1i}\le\lambda'_{S_1^{(-i)}+1},\ q_i\le\gamma_{\tilde{R}_1^{(-i)}+R_2^{(-i)}+1,S_1^{(-i)}+1})}{\tilde{R}_1^{(-i)}+R_2^{(-i)}+1}\right].$$

Thus, the theorem is proved.

5.3.1 BH Type Procedures

We are now ready to propose our FDR controlling multiple testing procedures in a two-stage adaptive design setting with combination function. Before that, let us state some assumptions we need.

ASSUMPTION 1. The combination function $C(p_1, p_2)$ is non-decreasing in both arguments.

ASSUMPTION 2. The pairs (p_{1i}, p_{2i}), $i = 1, \ldots, m$, are independently distributed and the pairs corresponding the null hypotheses are identically distributed as (p_1, p_2) with a joint distribution that satisfies the 'p-clud' property (Brannath et al., 2002), that is,

$$\Pr(p_1 \leq u) \leq u \text{ and } \Pr(p_2 \leq u \mid p_1) \leq u \text{ for all } 0 \leq u \leq 1.$$

Let us define

$$H(c; t, t') = \int_t^{t'} \int_0^1 I(C(u_1, u_2) \leq c) du_2 du_1.$$

Definition 1. (BH-TSADC Procedure).

1. Given the level α at which the overall FDR is to be controlled, three sequences of constants $\lambda_i = i\lambda/m$, $i = 1, \ldots, m$, $\lambda'_i = i\lambda'/m$, $i = 1, \ldots, m$, for some prefixed $\lambda < \alpha < \lambda'$, and $\gamma_{r_1+1,s_1} \leq \cdots \leq \gamma_{s_1,s_1}$, satisfying

$$H(\gamma_{r_1+i,s_1}; \lambda_{r_1}, \lambda'_{s_1}) = \frac{(r_1 + i)(\alpha - \lambda)}{m},$$

$i = 1, \ldots, s_1 - r_1$, for every fixed $1 \leq r_1 < s_1 \leq m$, find $R_1 = \max\{1 \leq i \leq$

$m : p_{1(j)} \leq \lambda_j$ for all $j \leq i\}$ and $S_1 = \max\{1 \leq i \leq m : p_{1(i)} \leq \lambda'_i\}$, with R_1 or S_1 being equal to zero if the corresponding maximum does not exist.

2. Reject $H_{(i)}$ for $i \leq R_1$; accept $H_{(i)}$ for $i > S_1$; and continue testing $H_{(i)}$ for $R_1 < i \leq S_1$ making use of the additional p-values p_{2i}'s available for all such follow-up hypotheses at the second stage.

3. At the second stage, consider the combined p-values $q_i = C(p_{1i}, p_{2i})$ for the follow-up null hypotheses. Let $q_{(i)}$, $i = 1, \ldots, S_1 - R_1$, be their ordered versions. Reject $H_{(i)}$ [the null hypothesis corresponding to $q_{(i)}$] for all $i \leq R_2(R_1, S_1) = \max\{1 \leq j \leq S_1 - R_1 : q_{(j)} \leq \gamma_{R_1+j,S_1}\}$, provided this maximum exists, otherwise, reject none of the follow-up null hypotheses.

PROPOSITION 1. Let π_0 be the proportion of true null hypotheses. Then, the FDR of the BH-TSADC method is less than or equal to $\pi_0\alpha$, and hence controlled at α, if Assumptions 1 and 2 hold.

PROOF OF PROPOSITION 1.

$$\begin{aligned}
\mathrm{FDR}_{12} &\leq \sum_{i \in J_0} E\left[\frac{Pr_H(p_1 \leq \lambda_{R_1^{(-i)}+1})}{R_1^{(-i)}+1}\right] + \\
&\qquad \sum_{i \in J_0} E\left[\frac{Pr_H(\lambda_{\tilde{R}_1^{(-i)}+1} < p_1 \leq \lambda'_{S_1^{(-i)}+1},\ C(p_1,p_2) \leq \gamma_{\tilde{R}_1^{(-i)}+R_2^{(-i)}+1,S_1^{(-i)}+1})}{\tilde{R}_1^{(-i)}+R_2^{(-i)}+1}\right] \\
&\leq \sum_{i \in J_0} E\left[\frac{\lambda_{R_1^{(-i)}+1}}{R_1^{(-i)}+1}\right] + \\
&\qquad \sum_{i \in J_0} E\left[\frac{Pr(\lambda_{\tilde{R}_1^{(-i)}+1} < u_1 \leq \lambda'_{S_1^{(-i)}+1},\ C(u_1,u_2) \leq \gamma_{\tilde{R}_1^{(-i)}+R_2^{(-i)}+1,S_1^{(-i)}+1})}{\tilde{R}_1^{(-i)}+R_2^{(-i)}+1}\right].
\end{aligned} \tag{5.2}$$

The first sum in (5.2) is less than or equal to $\pi_0\lambda$, since $\lambda_{R_1^{(-i)}+1} = [R_1^{(-i)}+1]\lambda/m$, and the second sum is less than or equal to $\pi_0(\alpha-\lambda)$, since the probability in the numerator in this sum is equal to

$$= \frac{H(\gamma_{\tilde{R}_1{}^{(-i)}+R_2^{(-i)}+1, S_1^{(-i)}+1}; \lambda_{\tilde{R}_1{}^{(-i)}+1}, \lambda'_{S_1^{(-i)}+1})}{\left[\tilde{R}_1^{(-i)}+1+R_2^{(-i)}\right](\alpha-\lambda)}{m}.$$

Thus, the proposition is proved.

The BH-TSADC procedure can be implemented alternatively, and often more conveniently, in terms of some FDR estimates at both stages. With $R^{(1)}(t) = \#\{i : p_{1i} \le t)$ and $R^{(2)}(c;t,t') = \#\{i : t < p_{1i} \le t', C(p_{1i}, p_{2i}) \le c\}$, let us define

$$\widehat{\mathrm{FDR}}_1(t) = \begin{cases} \frac{mt}{R^{(1)}(t)} & \text{if } R^{(1)}(t) > 0 \\ 0 & \text{if } R^{(1)}(t) = 0, \end{cases}$$

$$\text{and } \widehat{\mathrm{FDR}}_{2|1}(c;t,t') = \begin{cases} \frac{mH(c;t,t')}{R^{(1)}(t)+R^{(2)}(c;t,t')} & \text{if } R^{(2)}(c;t,t') > 0 \\ 0 & \text{if } R^{(2)}(c;t,t') = 0, \end{cases}$$

Then, we have the following:

The BH-TSADC procedure: An alternative definition. Reject $H_{(i)}$ for all $i \le R_1 = \max\{1 \le k \le m : \widehat{\mathrm{FDR}}_1(p_{1(j)}) \le \lambda \text{ for all } j \le k\}$; accept $H_{(i)}$ for all $i > S_1 = \max\{1 \le k \le m : \widehat{\mathrm{FDR}}_1(p_{1(k)}) \le \lambda'\}$; continue to test $H_{(i)}$ at the second stage for all i such that $R_1 < i \le S_1$. Reject $H_{(i)}$, the follow-up null hypothesis corresponding to $q_{(i)}$, at the second stage for all $i \le R_2(R_1, S_1) = \max\{1 \le k \le S_1 - R_1 : \widehat{\mathrm{FDR}}_{2|1}(q_{(k)}; R_1\lambda/m, S_1\lambda'/m) \le \alpha - \lambda\}$.

REMARK 2. The BH-TSADC procedure is an extension of the BH procedure, from a method of controlling the FDR in a single-stage design to that in a two-stage adaptive design with combination tests. When $\lambda = 0$ and $\lambda' = 1$, that is, when we have a single-stage design based on the combined p-values, this method reduces to the usual BH method. Notice that $\widehat{\text{FDR}}_1(t)$ is a conservative estimate of the FDR of the single-step test with the rejection $p_i \leq t$ for each H_i. So, the BH-TSADC procedure screens out those null hypotheses as being rejected (or accepted) at the first stage the estimated FDR's at whose p-values are all less than or equal to λ (or greater than λ').

Clearly, the BH-TSADC procedure can potentially be improved in terms of having a tighter control over its FDR at α by plugging a suitable estimate of π_0 into it while choosing the second-stage thresholds, similar to what is done for the BH method in a single-stage design. As said in Section 2, there are different ways of estimating π_0, each of which has been shown to provide the ultimate control of the FDR, of course when the p-values are independent, by the resulting plugged-in version of the single-stage BH method (see, e.g., Sarkar, 2008). However, we will consider the following estimate of π_0, which is of the type considered in Storey, Taylor and Siegmund (2004) and seems natural in the context of the present adaptive design setting where $m - S_1$ of the null hypotheses are accepted as being true at the first stage:

$$\hat{\pi}_0 = \frac{m - S_1 + 1}{m(1 - \lambda')}.$$

The following theorem gives a modified version of the the BH-TSADC procedure

using this estimate.

Definition 2. (Plug-In BH-TSADC Procedure).

Consider the BH-TSADC procedure with the early decision thresholds R_1 and S_1 based on the sequences of constants $\lambda_i = i\lambda/m$, $i = 1, \ldots, m$, and $\lambda'_i = i\lambda'/m$, $i = 1, \ldots, m$, given $0 \leq \lambda < \lambda' \leq 1$, and the second-stage critical values $\gamma^*_{R_1+i,S_1}$, $i = 1, \ldots, S_1 - R_1$, given by the equations

$$H(\gamma^*_{r_1+i,s_1}; \lambda_{r_1}, \lambda'_{s_1}) = \frac{(r_1 + i)(\alpha - \lambda)}{m\hat{\pi}_0}, \tag{5.3}$$

for $i = 1, \ldots, s_1 - r_1$.

Proposition 2. The FDR of the Plug-In BH-TSADC method is less than or equal to α if Assumptions 1 and 2 hold.

Proof of Proposition 2. This can be proved as in Proposition 1. More specifically, first note that the FDR here, which we call the FDR^*_{12}, satisfies the following:

$$\begin{aligned} FDR^*_{12} \quad \leq \quad & \sum_{i \in J_0} E\left[\frac{I(p_{1i} \leq \lambda_{R_1^{(-i)}+1})}{R_1^{(-i)} + 1}\right] + \\ & \sum_{i \in J_0} E\left[\frac{I(\lambda_{\tilde{R}_1^{(-i)}+1} \leq p_{1i} \leq \lambda'_{S_1^{(-i)}+1}, q_i \leq \gamma^*_{\tilde{R}_1^{(-i)}+R_2^{*(-i)}+1, S_1^{(-i)}+1})}{\tilde{R}_1^{(-i)} + R_2^{*(-i)} + 1}\right], \end{aligned} \tag{5.4}$$

where

$$\begin{aligned} R_2^{*(-i)} &\equiv R_2^{*(-i)}(\tilde{R}_1^{(-i)}, S_1^{(-i)} + 1) \\ &= \max\{1 \le j \le S_1^{(-i)} - \tilde{R}_1^{(-i)} : q_{(j)}^{(-i)} \le \gamma^*_{\tilde{R}_1^{(-i)}+j+1, S_1^{(-i)}+1}\}, \end{aligned}$$

with $q_{(j)}^{(-i)}$ being the ordered versions of the combined p-values except the q_i. As in Proposition 1, the first sum in (5.4) is less than or equal to $\pi_0\lambda$. Before working with the second sum, first note that the γ^* satisfying Eqn. (5.3), that is, the following equation

$$H(\gamma^*_{r_1+i,s_1}; \lambda_{r_1}, \lambda'_{s_1}) = \frac{(r_1+i)(\alpha-\lambda)(1-\lambda')}{m - S_1 + 1},$$

is less than or equal to the γ^{**} satisfying

$$H(\gamma^{**}_{r_1+i,s_1}; \lambda_{r_1}, \lambda'_{s_1}) = \frac{(r_1+i)(\alpha-\lambda)(1-\lambda')}{m - S_1^{(-j)}},$$

for any fixed $j = 1, \ldots, m$. So, the second sum in (5.4) is less than or equal to

$$\begin{aligned} & \sum_{i \in J_0} E\left[\frac{I(\lambda_{\tilde{R}_1^{(-i)}+1} \le p_{1i} \le \lambda'_{S_1^{(-i)}+1}, q_i \le \gamma^{**}_{\tilde{R}_1^{(-i)}+R_2^{*(-i)}+1, S_1^{(-i)}+1})}{\tilde{R}_1^{(-i)} + R_2^{*(-i)} + 1}\right] \\ = & \sum_{i \in J_0} E\left[\frac{H(\gamma^{**}_{\tilde{R}_1^{(-i)}+R_2^{*(-i)}+1, S_1^{(-i)}+1}; \lambda_{\tilde{R}_1^{(-i)}+1}, \lambda'_{S_1^{(-i)}+1})}{\tilde{R}_1^{(-i)} + R_2^{*(-i)} + 1}\right] \\ = & (\alpha-\lambda) \sum_{i \in J_0} E\left[\frac{1-\lambda'}{m - S_1^{(-i)}}\right] \le \alpha - \lambda, \end{aligned}$$

since $\sum_{i \in J_0} E\left[\frac{1-\lambda'}{m-S_1^{(-i)}}\right] \le 1$; see, for instance, Sarkar (2008, p. 151). Hence, $\text{FDR}_{12}^* \le \pi_0\lambda + \alpha - \lambda \le \alpha$, which proves the proposition.

As in the BH-TSADC procedure, the Plug-In BH-TSADC procedure can

also be described alternatively using estimated FDR's at both stages. Let

$$\widehat{\mathrm{FDR}}^{*}_{2|1}(c;t,t') \;=\; \begin{cases} \frac{m\hat{\pi}_0 H(c;t,t')}{R^{(1)}(t)+R^{(2)}(c;t,t')} & \text{if } R^{(2)}(c;t,t')>0 \\ 0 & \text{if } R^{(2)}(c;t,t')=0, \end{cases}$$

Then, we have the following:

The Plug-In BH-TSADC procedure: An alternative definition. At the first stage, decide the null hypotheses to be rejected, accepted, or continued to be tested at the second stage based on $\widehat{\mathrm{FDR}}_1$, as in (the alternative description of) the BH-TSADC procedure. At the second stage, reject $H_{(i)}$, the follow-up null hypothesis corresponding to $q_{(i)}$, for all $i \le R_2^*(R_1,S_1) = \max\{1 \le k \le S_1 - R_1 : \widehat{\mathrm{FDR}}^{*}_{2|1}(q_{(k)}; R_1\lambda/m, S_1\lambda'/m) \le \alpha - \lambda\}$.

5.3.2 Two Special Combination Functions

We now present explicit formulas of $H(c;t,t')$ for two special combination functions - Fisher's and Simes' - often used in multiple testing applications. The Simes' combination function emphasizes on the smaller p-value, but the Simes' combined p-value can never be smaller than $\min(p_1,p_2)$. Fisher's combination function allows several small p-values to reinforce one another to produce a more powerful test than min-P based method.

Fisher's combination function: $C(p_1, p_2) = p_1 p_2$.

$$\begin{aligned} H_{Fisher}(c;t,t') &= \int_t^{t'} \int_0^1 I(C(u_1,u_2) \le c)\, du_2 du_1 \\ &= \begin{cases} c \ln\left(\frac{t'}{t}\right) & \text{if} \quad c < t \\ c - t + c \ln\left(\frac{t'}{c}\right) & \text{if} \quad t \le c < t' \\ t' - t & \text{if} \quad c \ge t' \ , \end{cases} \end{aligned} \tag{5.5}$$

for $c \in (0,1)$.

Simes' combination function: $C(p_1, p_2) = \min\{2\min(p_1, p_2), \max(p_1, p_2)\}$.

$$\begin{aligned} H_{Simes}(c;t,t') &= \int_t^{t'} \int_0^1 I(C(u_1,u_2) \le c) du_2 du_1 \\ &= \begin{cases} \frac{c}{2}(t'-t) & \text{if} \quad c \le t \\ c(\frac{t'}{2} - t) + \frac{c^2}{2} & \text{if} \quad t < c \le \min(2t, t') \\ c(t'-t) & \text{if} \quad t' < c \le 2t \\ \frac{c}{2}(1+t') - t & \text{if} \quad 2t < c \le t' \\ \frac{c}{2}(1+2t') - \frac{c^2}{2} - t & \text{if} \quad \max(2\mathrm{t}, \mathrm{t}') \le \mathrm{c} \le 2\mathrm{t}' \\ t' - t & \text{if} \quad c \ge 2t', \end{cases} \end{aligned}$$

for $c \in (0,1)$.

See also Brannath et al. (2002) for the formula (5.5). These formulas can be used to determine the critical values γ_i's before observing the combined p-values or to estimate the FDR after observing the combined p-values at the second stage in the BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions. Of course, for large values of m, it is numerically more challenging to determine the γ_i's than estimating the FDR at the second stage,

and so in that case we would recommend using the alternative versions of these procedures.

5.4 Simulation Studies

This section presents the results of simulation studies we conducted to investigate the following three questions related to the proposed procedures:

Q1. How well do the proposed BH-TSADC and Plug-In BH-TSADC procedures perform under independence compared to the single-stage BH procedure in terms of FDR control and power?

Q2. Can the proposed BH-TSADC and Plug-In BH-TSADC procedures continue to control the FDR for dependent p-values?

Q3. How well do the proposed BH-TSADC and Plug-In BH-TSADC procedures perform in terms of cost-saving?

5.4.1 Under Independence

To investigate Q_1, (i) we generated two independent sets of m uncorrelated random variables $Z_i \sim N(\mu_i, 1)$, $i = 1, \ldots, m$, one for Stage 1 and the other for Stage 2, having set $m\pi_0$ of these μ_i's at zero and the rest at 2, (ii) tested $H_i : \mu_i = 0$ against $K_i : \mu_i > 0$, simultaneously for $i = 1, \ldots, m$, by applying the (alternative versions of) BH-TSADC and Plug-In BH-TSADC procedures at level α with both Fisher's and Simes' combination functions, $\lambda = 0.025$ and $\lambda' = 0.5$ to the generated data for both stages and the level α BH procedure to the data for the first stage and full

data from both stages respectively, and (iii) noted the false discovery proportion and the proportion of false nulls that are rejected. We repeated steps (i)-(iii) 1000 times and averaged out the above proportions over these 1000 runs to obtain the final simulated values of FDR and average power (the expected proportion of false nulls that are rejected) for each of these procedures.

In some sense, there does not exist a fair comparison of our two-stage method with the single-stage BH method, because the data requirement for these two methods is different. In the single-stage BH method, the information of all markers needs to be available for the subjects in the single-stage design. If there is no information of some markers for one subject, then the subject cannot be used in the study. However, in our proposed two-stage methods, the information of all markers is only required for the subjects in the first stage, whereas, for the subjects in the second stage, only the information of the markers selected to the second stage is required.

In order to perform a relatively fair numerical comparison between our suggested two-stage method and the single-stage BH method in terms of the FDR control and power, we firstly apply the single-stage BH method to the data from the first stage in our simulation studies. Secondly, we apply the single-stage BH method to the full data from both stages. The simulated FDR's and average powers for these four procedures have been graphically displayed in Figures 5.1 and 5.2. Figure 5.1 compares the proposed BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes combination functions with those of the BH procedures

for different values of π_0, $\alpha = 0.05$, and $m = 10, 100$, and 1000, in terms of the simulated FDR, while Figure 5.2 does the same in terms of the simulated average power. The simulation results indicate that our proposed two-stage method has significantly power improvement over the single-stage BH method using the first stage data only while controlling the overall FDR.

Different Early Stopping Boundaries

Our proposed two-stage BH-type methods control the overall FDR in all circumstance under independence. However, the power of such a two-stage design relies on the choice of early stopping boundaries. Figures 5.3 and 5.5 show the simulated FDR versus π_0 for early rejection boundary $\lambda = 0.005, 0.010$, and 0.025 and early acceptance boundary $\lambda' = 0.5$. Figures 5.4 and 5.6 show the corresponding simulated average power versus π_0. Figures 5.7 to 5.10 do the same but with early rejection boundary $\lambda = 0.025$ and early acceptance boundary $\lambda' = 0.5, 0.8, 0.9$ respectively. The simulation results show that a relative smaller early rejection boundary tends to lead to greater power when the early acceptance boundary is fixed. Similarly, a relatively larger early acceptance boundary tends to lead to greater power when the early rejection boundary is fixed

Exponentially Decreasing Effect Sizes

To examine the performance of proposed procedures in a more complicated genetic mode, we explored a model with exponentially decreasing effect sizes. (i) We generated two independent sets of $m = 1000$ uncorrelated random variables $Z_i \sim$

$N(\mu_i, 1), i = 1, ..., m$, one for Stage 1 and the other for Stage 2. We set $m\pi_0$ of these μ_i's at zero and the rest with equally spaced exponentially decreasing effect sizes at $1.5 \times (2^2, 2^1, 2^{0.5}, 2^0)$. (ii) We tested $H_i : \mu_i = 0$ against $H_i : \mu_i > 0$, simultaneously for $i = 1, ..., m$, by applying the (alternative versions of) BH-TSADC and Plug-In BH-TSADC procedures at level α with both Fisher's and Simes' combination functions to the generated data for both stages and the level α BH procedure to the data for the first stage and full data from both stages respectively. For the two-stage design, the early acceptance boundary λ' was set at 0.5 and the early rejection boundary λ was set at 0.005, 0.010, and 0.025, respectively. (iii) The false discovery proportion and the proportion of false nulls that are rejected. We repeated steps (i)-(iii) 1000 times and average out the above proportions over these 1000 runs to obtain the final simulated values of FDR and average power (the expected proportion of false nulls that are rejected) for each of these procedures.

Figures 5.11 and 5.12 show that in the setting with the exponentially decreasing effect sizes at $1.5 \times (2^2, 2^1, 2^{0.5}, 2^0)$, the power differences between our suggested procedures and the BH procedure applied to the first stage data and full data from both stages is decreasing compared to that in the setting with the constant effect size at 2.

5.4.2 Under Dependence

In our simulation study to investigate Q_2, we considered three different scenarios for dependent p-values. In particular, we generated two independent sets of $m = 100$ correlated normal random variables $Z_i \sim N(\mu_i, 1)$, $i = 1, \ldots, m$, one for

Stage 1 and the other for Stage 2, with $m\pi_0$ of the μ_i's being equal to 0 and the rest being equal to 2, and a correlation matrix exhibiting one of three different types of dependence - equal, clumpy and AR(1) dependence. In other words, the Z_i's were assumed to have a common, non-negative correlation ρ in case of equal dependence, were broken up into ten independent groups with 10 of the Z_i's within each group having a common, non-negative correlation ρ in case of clumpy dependence, and were assumed to have correlations $\rho_{ij} = \text{Cor}(Z_i, Z_j)$ of the form $\rho_{ij} = \rho^{|i-j|}$ for all $i \neq j = 1, \ldots, m$, and some non-negative ρ in case of AR(1) dependence. We then applied the (alternative versions of) the BH-TSADC and Plug-In BH-TSADC procedures at level $\alpha = 0.05$ with both Fisher's and Simes combination functions, $\lambda = 0.025$, and $\lambda' = 0.5$ to these data sets. These two steps were repeated 1000 times before obtaining the simulated FDR's and average powers for these procedures, as in our study related to Q_1.

Figures 5.13 to 5.15 graphically display the simulated FDR's of these procedures for different values of π_0 and types of dependent p-values considered.

As seen from Figures 5.1 to 5.12, the proposed procedures with Fisher's combination function seem to have a slight edge over the corresponding ones with Simes' combination function in terms of FDR control and power. Between these two procedures, whether it's based on Fisher's or Simes' combination function, the BH-TSADC appears to be the better choice when π_0 is large, which is often the case in practice. It controls the FDR not only under independence, which is theoretically known, but also the FDR control seems to be maintained even under different types

of positive dependence. Also, it provides a better power improvement, at least in the independence case, over the single-stage BH procedure. If, however, π_0 is not large, the Plug-In BH-TSADC procedure provides a better control of the FDR and its power improvement, again at least in the independence case, over the single-stage BH procedure seems more significant than the BH-TSDADC procedure; of course, it may lose the FDR control when the p-values exhibit equal or AR(1) type dependence with a moderately large equal- or auto-correlation.

5.4.3 Cost Saving

In real application, one of the main advantages of our suggested two-stage method is cost savings, compared to the single-stage BH method. For example, in a genome-wide association study, because of high cost genotyping hundreds of thousands of markers on thousands of subjects, many investigations have used a two-stage design, in which a proportion of the available samples are genotyped on a large number of markers in the first stage, and a small proportion of these markers are selected and then followed up by genotyping them on the remaining samples in the second stage.

Let's take the genome-wide association study as an example. Suppose the unit cost of genotyping one marker for each patient is c, then when applying the single-stage BH method to the full data for both stages, the total cost for genotyping all markers for each patient is $m \times N \times c$, where N is the total number of patients assigned to stage 1 and 2, and m is the total number of markers for each patient. On the other hand, when applying our two-stage methods, suppose s is the total

number of rejected and accepted hypotheses in the first stage, then the total cost is $f \times N \times m \times c + (1-f) \times N \times (m-s) \times c$, where f and $(1-f)$ are the ratios of the N patients assigned to stage 1 and 2 and s is a function of f denoted by $s(f)$. Thus, the total cost saving for our method is $(1-f) \times N \times s(f) \times c$, so the proportion of cost saving of the two-stage method relative to the BH method is

$$\frac{(1-f) \times N \times s(f) \times c}{m \times N \times c} = \frac{(1-f) \times s(f)}{m}.$$

In our simulation study to investigate Q_3, we considered four different scenarios for sample allocation ratio $f = 0.25, 0.50, 0.75, 1.00$. In particular, we generated three datasets of $m = 100, 1000, 5000$ independent normal random variables $Z_i \sim N(\mu_i, 1/(2 \times f))$, $i = 1, \ldots, m$, for Stage 1, with $m\pi_0$ of the μ_i's being equal to 0 and the rest being equal to 2. The p-values for Stage 2 do not matter in terms of evaluating cost saving. We averaged out the total number of early stoppings $s(f)$ at the first stage from 1000 simulation runs. Figure 5.16 displays proportional cost saving versus π_0 for $m = 100, 1000, 5000$ with $\lambda = 0.025$ and $\lambda' = 0.5$ by sample allocation rate $f = 0.25, 0.50, 0.75, 1.00$ across two stages.

5.5 A Real Data Application

To illustrate how the proposed procedures can be implemented in practice, we reanalyzed a dataset taken from an experiment by Tian et al. (2003) and post-processed by Jeffery et al. (2006). Zehetmayer et al. (2008) considered this data for a different purpose. In this data set, multiple myeloma samples were generated with Affymetrix Human U95A chips, each consisting $12,625$ probe sets. The samples

were split into two groups based on the presence or absence of focal lesions of bone.

The original dataset contains gene expression measurements of 36 patients without and 137 patients with bone lytic lesions, However, in our reanalysis, we used the gene expression measurements of 36 patients with bone lytic lesions and a control group of the same sample size without such lesions. We considered this data in a two-stage framework, with the first 18 subjects per group for Stage 1 and the next 18 subjects per group for Stage 2. We prefixed the Stage 1 early rejection boundary λ at 0.025 and the early acceptance boundary λ' at 0.5, and applied the proposed (alternatives versions of) BH-TSADC and plug-in BH-TSADC procedures at the overall FDR level 0.05.

In particular, we considered all $m = 12,625$ probe set gene expression measurements for the first stage data of 36 patients (18 patients per group) and the full data of 72 patients (36 patients per group) across two stages respectively, and analyzed them based on a stepdown procedure with the critical values $\lambda_i = i0.025/m$, $i = 1, \ldots, m$, and a stepup procedure with the critical values $\lambda'_i = i0.5/m$, $i = 1, \ldots, m$, using the corresponding p-values generated from one-sided t-tests. We noted the probe sets that were rejected by the stepdown procedure and those that were accepted by the stepup procedure. With these numbers being r_1 and $m - s_1$, respectively, we took the probe sets that were neither rejected by the stepdown procedure nor accepted by the stepup procedure, that is, the probe sets with the first-stage p-values more than $r_1\lambda/m$ but less than or equal to $s_1\lambda'/m$, for further analysis using estimated FDR based on their first-stage and second-stage p-values

combined through Fisher's and Simes' combination functions, as described in the alternative versions of the BH-TSADC and plug-in BH-TSADC procedures.

The results of this analysis are reported in Table 5.1. As seen from this table, the BH-TSADC procedure with Fisher's combination function and its plug-in version produce 144 and 93 discoveries, respectively; whereas, these numbers are 40 and 32, respectively, for the Simes' combination function. These numbers are significantly larger than 18, the number of discoveries made by the single-stage BH procedure using the first stage data. When the full data across two stages are used, single-stage BH procedure results in more rejections which is consistent with the simulation results.

Table 5.1: The results of two-stage combination tests with Fisher's and Simes' combination functions, $\lambda = 0.025$, $\lambda' = 0.5$, and $\alpha = 0.05$ of 12625 probe sets in the Affymetrix Human U95A Chips data taken from Tian et al. (2003).

		Fisher's		Simes'		BH	
			Plug-in		Plug-in	Single-Stage	
		BH-TSADC	BH-TSADC	BH-TSADC	BH-TSADC	Stage 1 Data	Full Data
Stage 1	Reject	4	4	4	4	18	417
	Accept	10520	10520	10520	10520	12607	12108
Stage 2	Reject	140	89	36	28	NA	NA
	Accept	1961	2012	2065	2073	NA	NA
Total	Reject	144	93	40	32	18	417

To examine the effect of early stopping boundaries, we also explored different pairs of early rejection and early acceptance boundaries, i.e. $\lambda = 0.005, 0.010, 0.025$

Table 5.2: The total number of rejections of two-stage combination tests with Fisher's and Simes' combination functions, different $\lambda = 0.005, 0.010, 0.015$ and $\lambda' = 0.5, 0.8, 0.9$, and $\alpha = 0.025$ of 12625 probe sets in the Affymetrix Human U95A Chips data taken from Tian et al. (2003).

	Fisher's		Simes'		BH	
	BH-TSADC	Plug-in BH-TSADC	BH-TSADC	Plug-in BH-TSADC	Stage 1 Data	Full Data
$\lambda = 0.005$						
$\lambda' = 0.5$	84	58	33	17	2	127
$\lambda' = 0.8$	97	35	42	17	2	127
$\lambda' = 0.9$	106	34	54	18	2	127
$\lambda = 0.010$						
$\lambda' = 0.5$	74	41	24	13	2	127
$\lambda' = 0.8$	81	31	30	16	2	127
$\lambda' = 0.9$	90	31	37	18	2	127
$\lambda = 0.015$						
$\lambda' = 0.5$	56	31	17	12	2	127
$\lambda' = 0.8$	63	29	23	15	2	127
$\lambda' = 0.9$	69	27	30	18	2	127

and $\lambda' = 0.5, 0.8, 0.9$ (see Table 5.2).

5.6 Discussion

Our main goal in this article has been to construct a two-stage multiple testing procedure that allows making early decisions on the null hypotheses in terms of rejection, acceptance or continuation to the second stage for further testing with more observations and eventually controls the FDR. Such two-stage formulation of multiple testing is of practical importance in many statistical investigations; nev-

ertheless, generalizations of the classical BH type methods from single-stage to the present two-stage setting, which seem to be the most natural procedures to consider, have not been put forward until the present work. We have been able to construct two such generalizations with proven FDR control, provided simulation results and practical examples of their improved power performances compared to the corresponding single-stage BH type methods under independence. We also have presented numerical evidence that they can maintain a control over the FDR even under some dependence situations.

It is important to emphasize that the theory behind the developments of our proposed two-stage FDR controlling methods has been driven by the idea of setting the early decision boundaries $\lambda < \lambda'$ on the (estimated) FDR at the first-stage p-values, rather than on these p-values themselves. In other words, we flag those null hypotheses for rejection (or acceptance) at the first stage at whose p-values the estimated FDR's are all less than or equal to λ (or greater than λ') before proceeding to the second stage; see Remark 2. This, we would argue, is often practical and meaningful when we are testing multiple hypotheses in two-stages in an FDR framework.

Brannath et al. (2002) have defined a global p-value $\tilde{p}(p_1, p_2)$ for testing a single hypothesis in a two-stage adaptive design with combination function $C(p_1, p_2)$. With the boundaries $\lambda < \lambda'$ set on each p_{1i}, the global p-value for each H_i is defined

by

$$\tilde{p}_i \equiv \tilde{p}(p_{1i}, p_{2i}) = \begin{cases} p_{1i} & \text{if} \quad p_{1i} \leq \lambda \text{ or } p_{1i} > \lambda' \\ \lambda + H(C(p_{1i}, p_{2i}); \lambda, \lambda') & \text{if} \quad \lambda < p_{1i} \leq \lambda' \ . \end{cases}$$

They have shown that each $\tilde{p}_i$ is stochastically larger than or equal to $U(0,1)$ when (p_{1i}, p_{2i}) satisfies the p-clud property, and the equality holds when p_{1i} and p_{2i} are independently distributed as $U(0,1)$. So, one may consider the BH method based on the $\tilde{p}_i$'s. This would control the overall FDR under the assumptions considered in the paper, maybe under some positive dependence conditions as well. However, it does not set the early decision boundaries on the FDR.

We proposed our FDR controlling procedures in this paper considering a non-asymptotic setting. However, one may consider developing procedures that would asymptotically control the FDR by taking the following approach towards finding the first- and second-stage thresholds subject to the early boundaries $\lambda < \lambda'$ and the final boundary α on the FDR. Given two constants $t < t'$, make an early decision regarding H_i by rejecting it if $p_{1i} \leq t$, accepting it if $p_{1i} > t'$, and continuing to test it at the second stage if $t < p_{1i} \leq t'$. At the second stage, reject H_i if $C(p_{1i}, p_{2i}) \leq c$. Storey's (2002) estimate of the first-stage FDR is given by

$$\widehat{\text{FDR}}_1^*(t) = \begin{cases} \frac{m\hat{\pi}_0 t}{R^{(1)}(t)} & \text{if} \quad R^{(1)}(t) > 0 \\ 0 & \text{if} \quad R^{(1)}(t) = 0 \ , \end{cases}$$

for some estimate $\hat{\pi}_0$ of π_0. Similarly, the overall FDR can be estimated as follows:

$$\widehat{\text{FDR}}_{12}^*(c,t,t') = \begin{cases} \frac{m\hat{\pi}_0[t+H(c;t,t')]}{R^{(1)}(t)+R^{(2)}(c;t,t')} & \text{if} \quad R^{(1)}(t) + R^{(2)}(c;t,t') > 0 \\ 0 & \text{if} \quad R^{(1)}(t) + R^{(2)}(c;t,t') = 0 \end{cases}$$

Let

$$\begin{aligned} \hat{t}_{\lambda} &= \sup\{t:\ \widehat{\mathrm{FDR}}_1(t) \le \lambda \text{ for all } t' \le t\}, \\ \hat{t}_{\lambda'} &= \inf\{t:\ \widehat{\mathrm{FDR}}_1(t) > \lambda' \text{ for all } t' > t\}, \\ \text{and} \qquad \hat{c}_{\alpha}(\lambda, \lambda') &= \sup\{c:\ \widehat{\mathrm{FDR}}_{12}(c, \hat{t}_{\lambda}, \hat{t}_{\lambda'}) \le \alpha\}. \end{aligned}$$

Then, reject H_i if $p_{1i} \le \hat{t}_{\lambda}$ or if $\hat{t}_{\lambda} < p_{1i} \le \hat{t}_{\lambda'}$ and $C(p_{1i}, p_{2i}) \le \hat{c}_{\alpha}(\lambda, \lambda')$. This may control the overall FDR asymptotically under the weak dependence condition and the consistency property of $\hat{\pi}_0$ (as in Storey, Taylor and Siegmund, 2004).

There are a number of other important issues related to the present problem which we have not touched in this paper but hope to address in different communications. There are other combination functions, such as Fisher's weighted product (Fisher 1932) and weighted inverse normal (Mosteller and Bush, 1954), performances of which would be worth investigating. Consideration of conditional error function (Proschan and Hunsberger, 1995) while defining a two-stage design before constructing FDR controlling methods is another important issue. Now that we know how to test multiple hypotheses in a two-stage design subject to first-stage boundaries on and the overall control of the FDR, we should be able to address issues relate to sample size determinations.

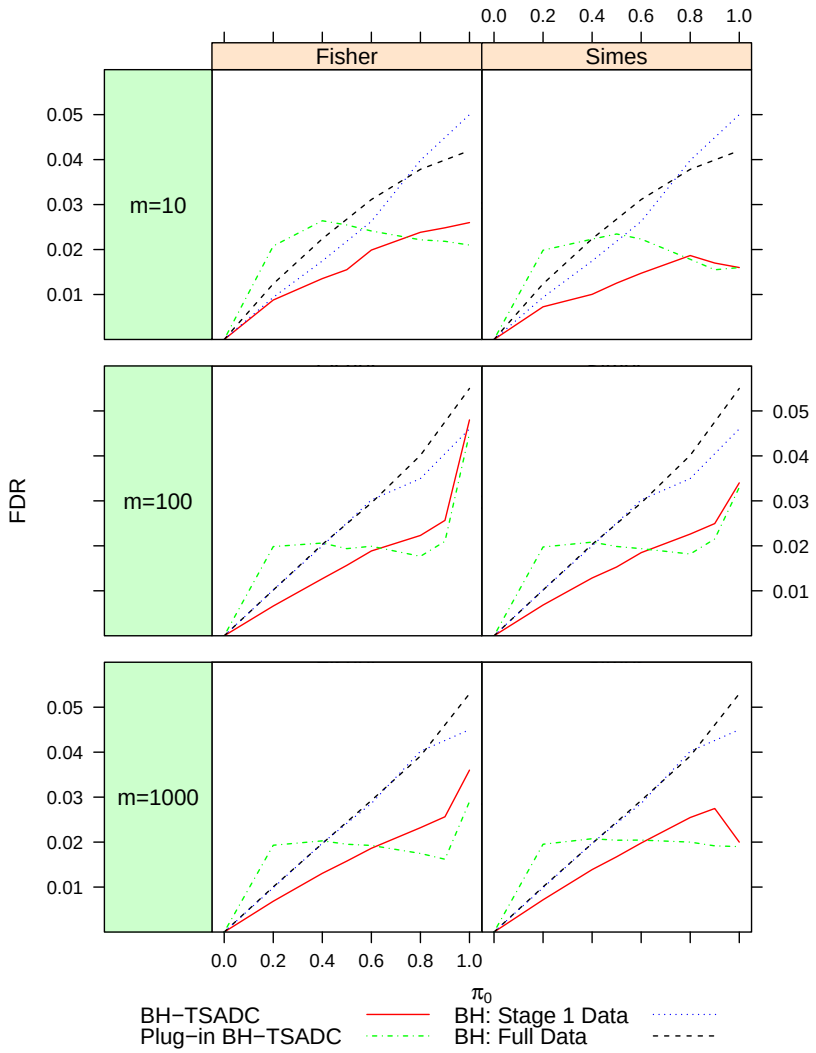

Figure 5.1: Comparison of simulated FDR's of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions, $\lambda = 0.025$, and $\lambda' = 0.5$, with simulated FDR of single-stage BH procedure based on full data from both stages, at $\alpha = 0.05$. [BH-TSADC: solid line, Plug-In BH-TSADC: dotdash line, BH Stage 1 Data: dotted line, and BH Full Data: dash line.]

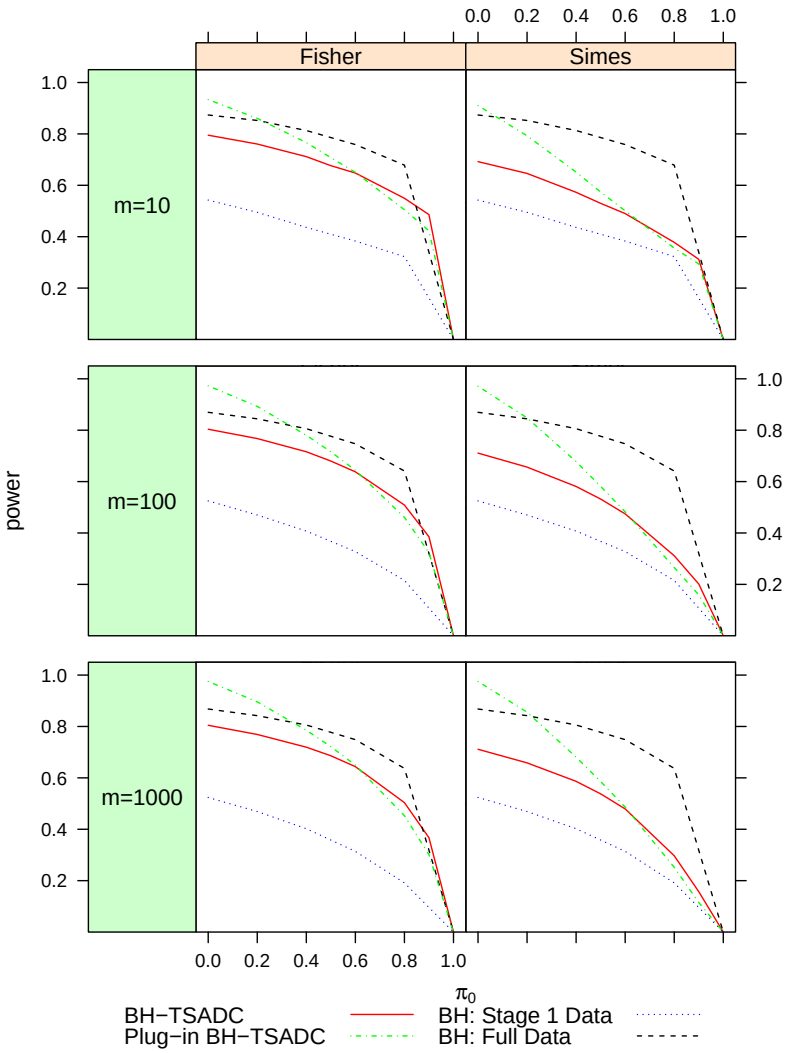

Figure 5.2: Comparison of simulated average powers of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions, $\lambda = 0.025$, and $\lambda' = 0.5$, with simulated FDR of single-stage BH procedure based on the first stage data and full data from both stages, at $\alpha = 0.05$. [BH-TSADC: solid line, Plug-In BH-TSADC: dotdash line, BH Stage 1 Data: dotted line, and BH Full Data: dash line.]

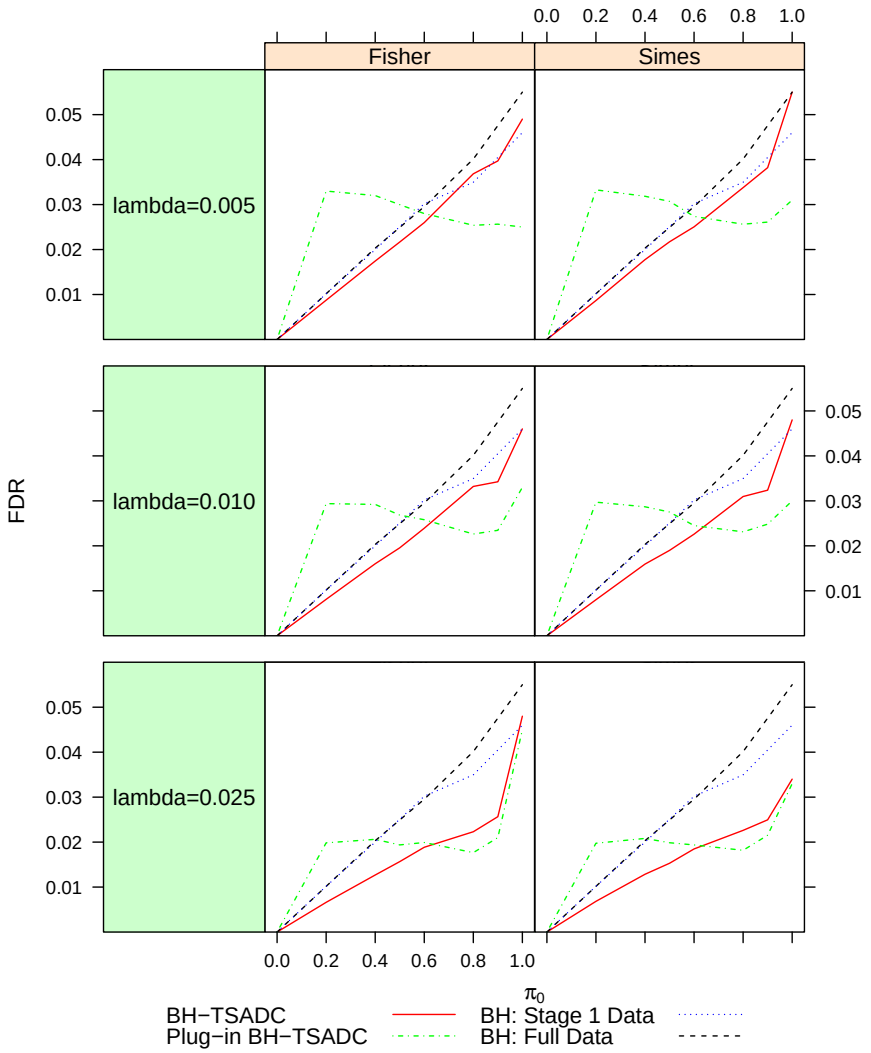

Figure 5.3: Comparison of simulated FDR's of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions for $m = 100$, $\lambda = 0.005, 0.010, 0.025$, and $\lambda' = 0.5$, with simulated FDR of single-stage BH procedure based on the first stage data and full data from both stages, at $\alpha = 0.05$. [BH-TSADC: solid line, Plug-In BH-TSADC: dotdash line, BH Stage 1 Data: dotted line, and BH Full Data: dash line.]

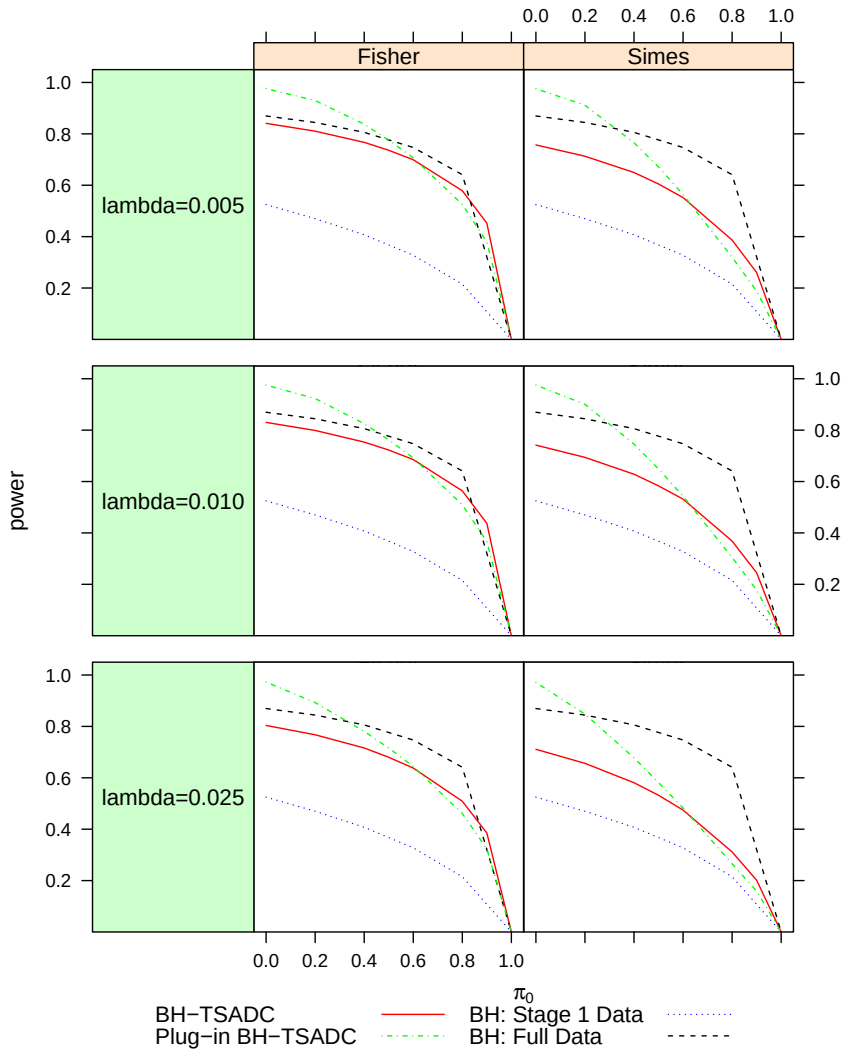

Figure 5.4: Comparison of simulated average powers of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions for $m = 100$, $\lambda = 0.005, 0.010, 0.025$, and $\lambda' = 0.5$, with the simulated average power of single-stage BH procedure based on the first stage data and full data from both stages, at $\alpha = 0.05$. [BH-TSADC: solid line, Plug-In BH-TSADC: dotdash line, BH Stage 1 Data: dotted line, and BH Full Data: dash line.]

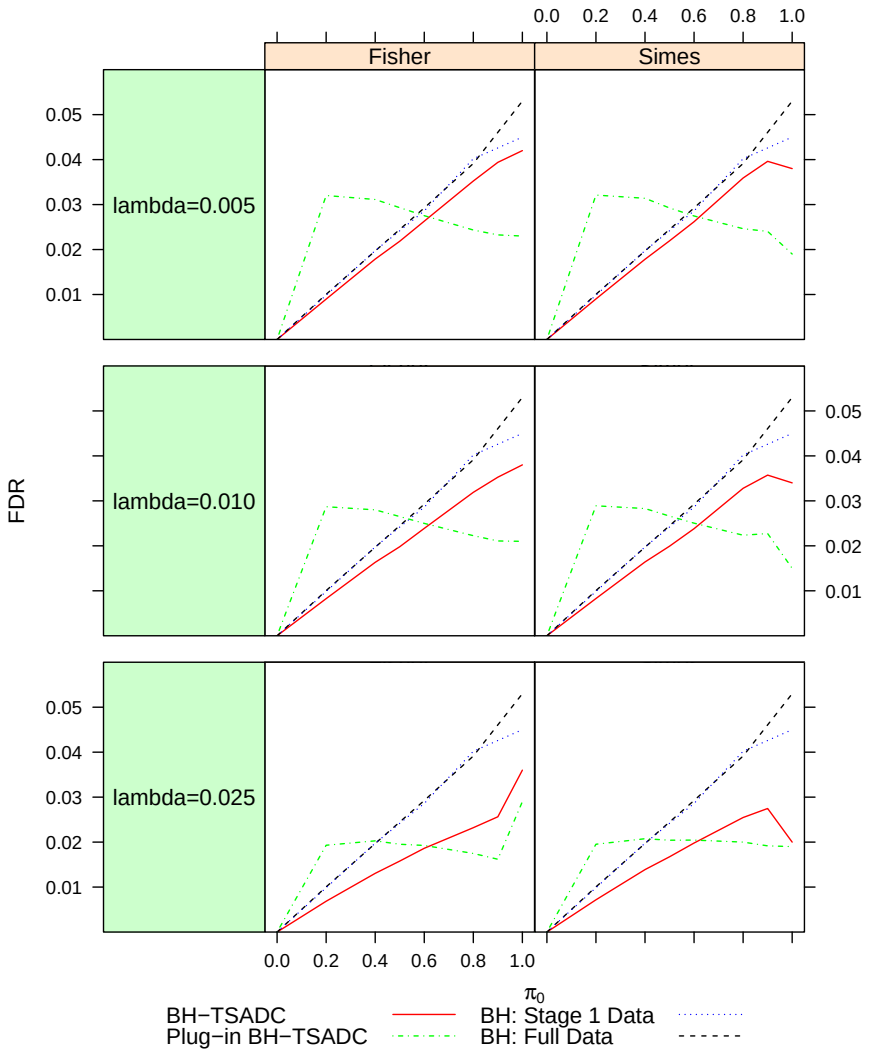

Figure 5.5: Comparison of simulated FDR's of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions for $m = 1000$, $\lambda = 0.005, 0.010, 0.025$, and $\lambda' = 0.5$, with simulated FDR of single-stage BH procedure based on the first stage data and full data from both stages, at $\alpha = 0.05$. [BH-TSADC: solid line, Plug-In BH-TSADC: dotdash line, BH Stage 1 Data: dotted line, and BH Full Data: dash line.]

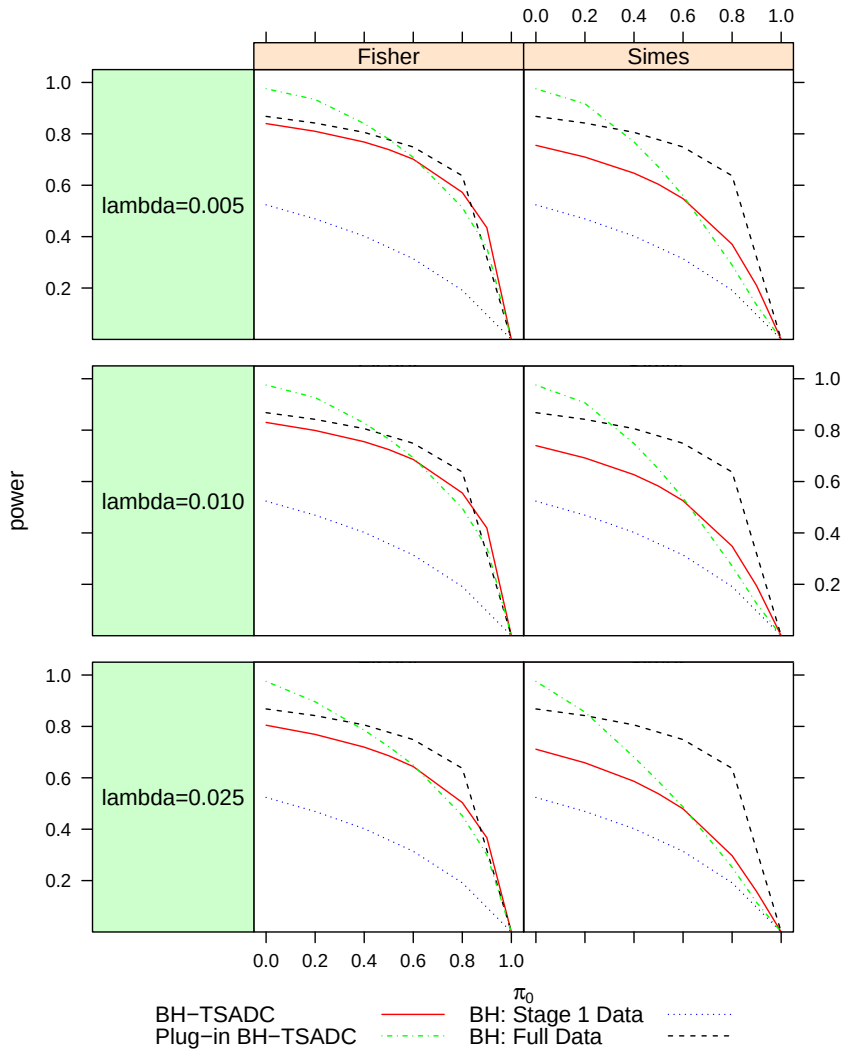

Figure 5.6: Comparison of simulated average powers of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions for $m = 1000$, $\lambda = 0.005, 0.010, 0.025$, and $\lambda' = 0.5$, with the simulated average power of single-stage BH procedure based on the first stage data and full data from both stages, at $\alpha = 0.05$. [BH-TSADC: solid line, Plug-In BH-TSADC: dotdash line, BH Stage 1 Data: dotted line, and BH Full Data: dash line.]

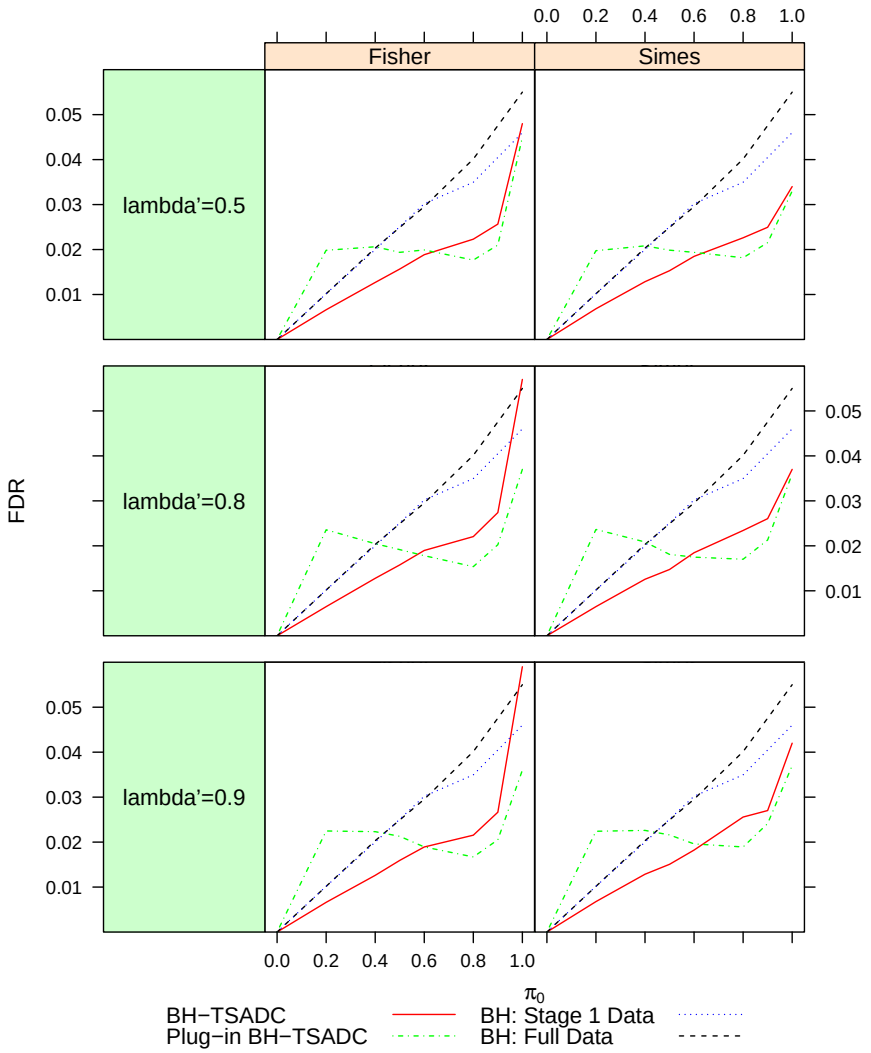

Figure 5.7: Comparison of simulated FDR's of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions for $m = 100$, $\lambda = 0.025$, and $\lambda' = 0.5, 0.8, 0.9$, with simulated FDR of single-stage BH procedure based on the first stage data and full data from both stages, at $\alpha = 0.05$. [BH-TSADC: solid line, Plug-In BH-TSADC: dotdash line, BH Stage 1 Data: dotted line, and BH Full Data: dash line.]

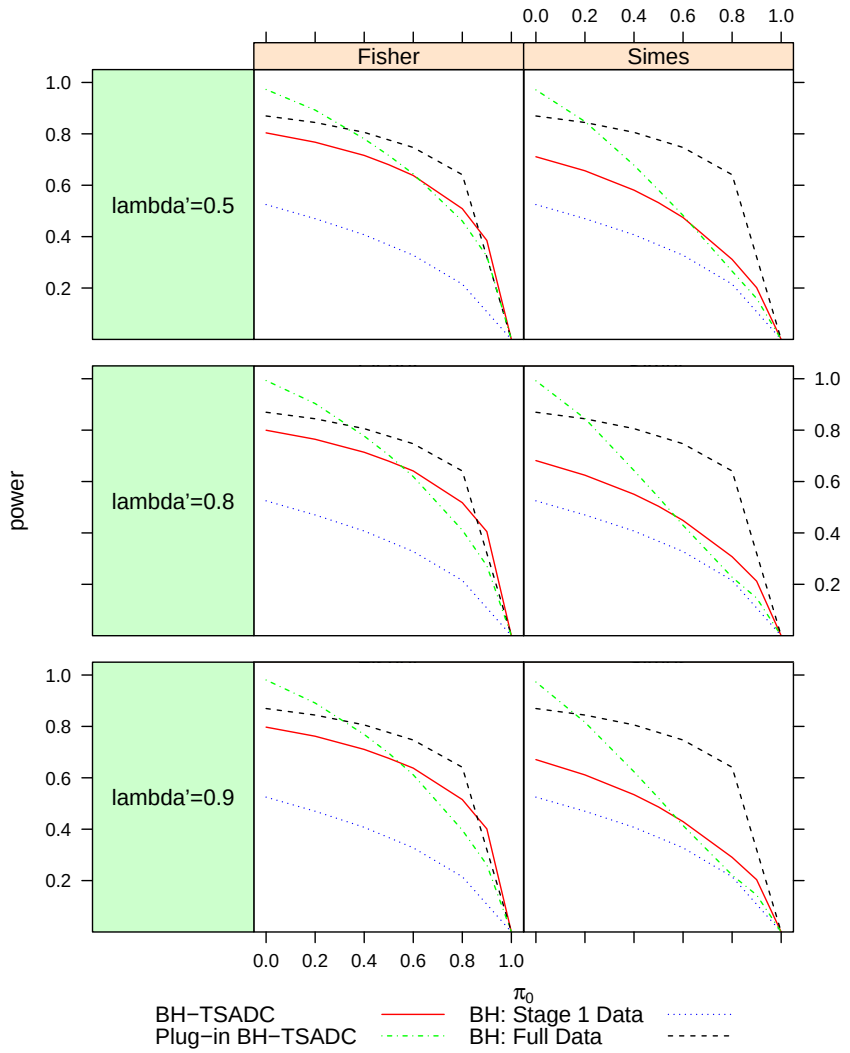

Figure 5.8: Comparison of simulated average powers of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions for $m = 100$, $\lambda = 0.025$, and $\lambda' = 0.5, 0.8, 0.9$, with the simulated average power of single-stage BH procedure based on the first stage data and full data from both stages, at $\alpha = 0.05$. [BH-TSADC: solid line, Plug-In BH-TSADC: dotdash line, BH Stage 1 Data: dotted line, and BH Full Data: dash line.]

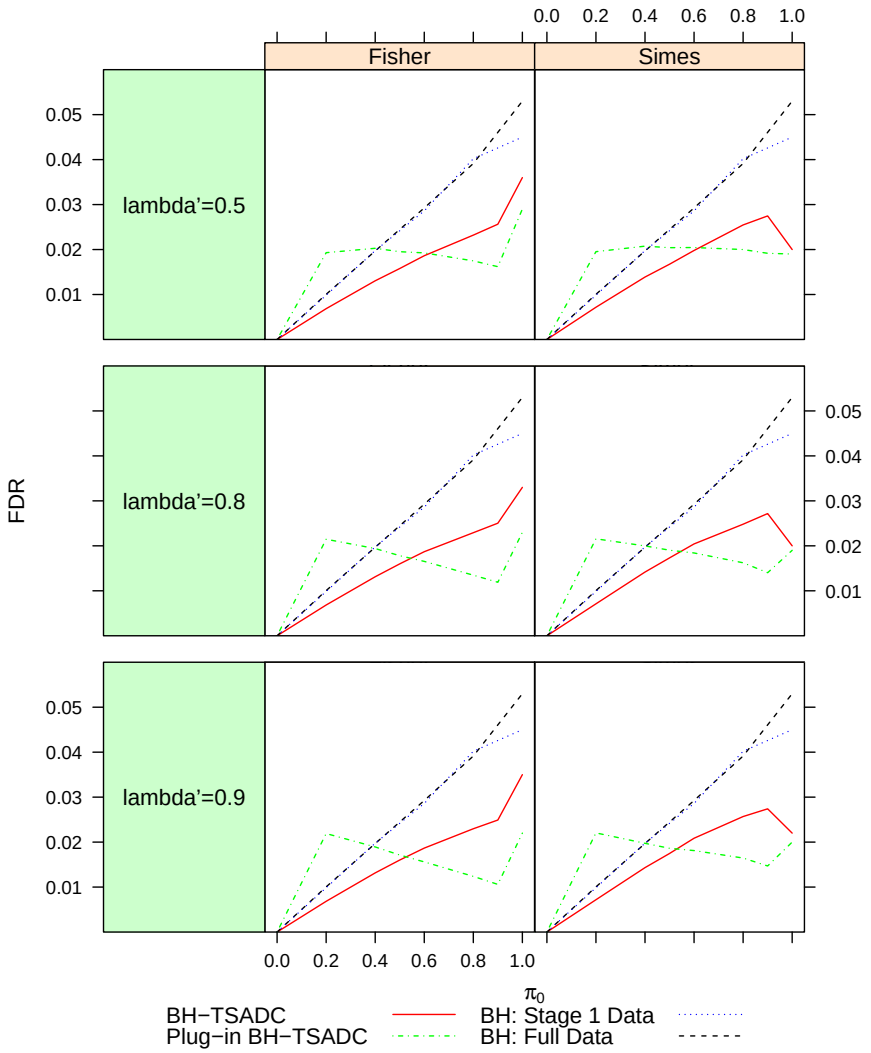

Figure 5.9: Comparison of simulated FDR's of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions for $m = 1000$, $\lambda = 0.025$, and $\lambda' = 0.5, 0.8, 0.9$, with simulated FDR of single-stage BH procedure based on the first stage data and full data from both stages, at $\alpha = 0.05$. [BH-TSADC: solid line, Plug-In BH-TSADC: dotdash line, BH Stage 1 Data: dotted line, and BH Full Data: dash line.]

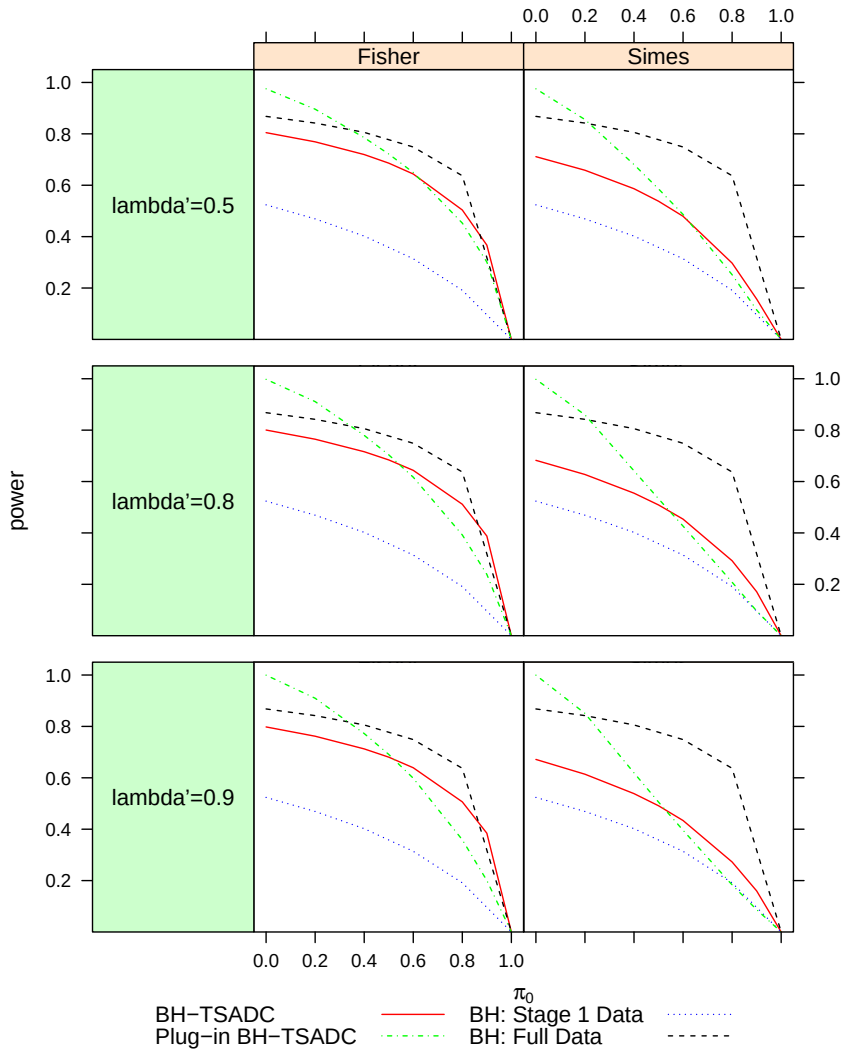

Figure 5.10: Comparison of simulated average powers of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions for $m = 1000$, $\lambda = 0.025$, and $\lambda' = 0.5, 0.8, 0.9$, with the simulated average power of single-stage BH procedure based on the first stage data and full data from both stages, at $\alpha = 0.05$. [BH-TSADC: solid line, Plug-In BH-TSADC: dotdash line, BH Stage 1 Data: dotted line, and BH Full Data: dash line.]

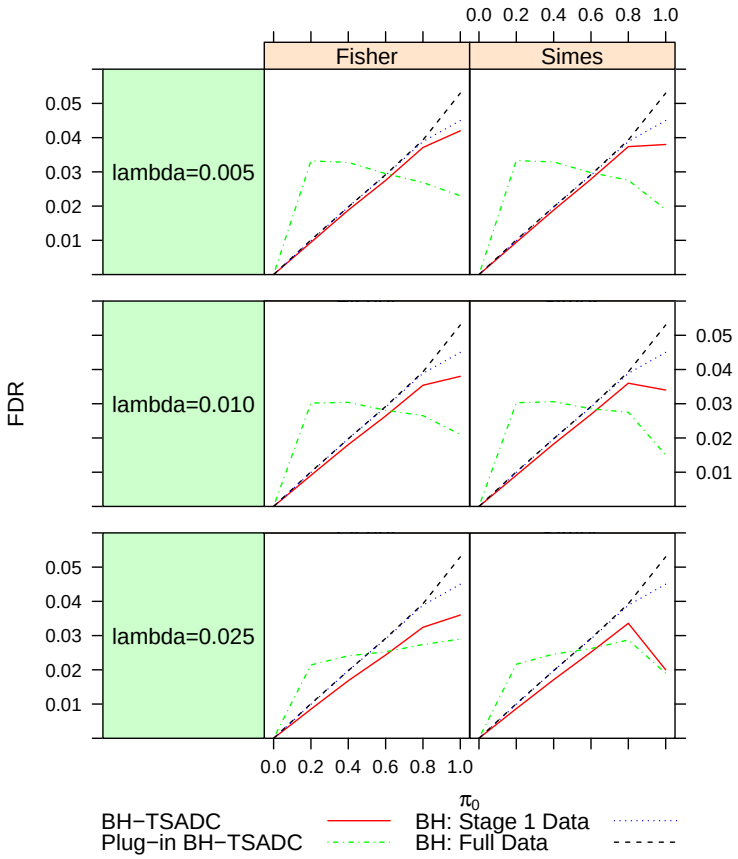

Figure 5.11: Comparison of simulated FDR's of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions for $m = 1000$ with equally spaced exponential decreasing effect sizes $1.5 \times (2^2, 2^1, 2^{0.5}, 2^0)$, $\lambda = 0.005, 0.010, 0.025$, and $\lambda' = 0.5$, with simulated FDR of single-stage BH procedure based on the first stage data and full data from both stages, at $\alpha = 0.05$. [BH-TSADC: solid line, Plug-In BH-TSADC: dotdash line, BH Stage 1 Data: dotted line, and BH Full Data: dash line.]

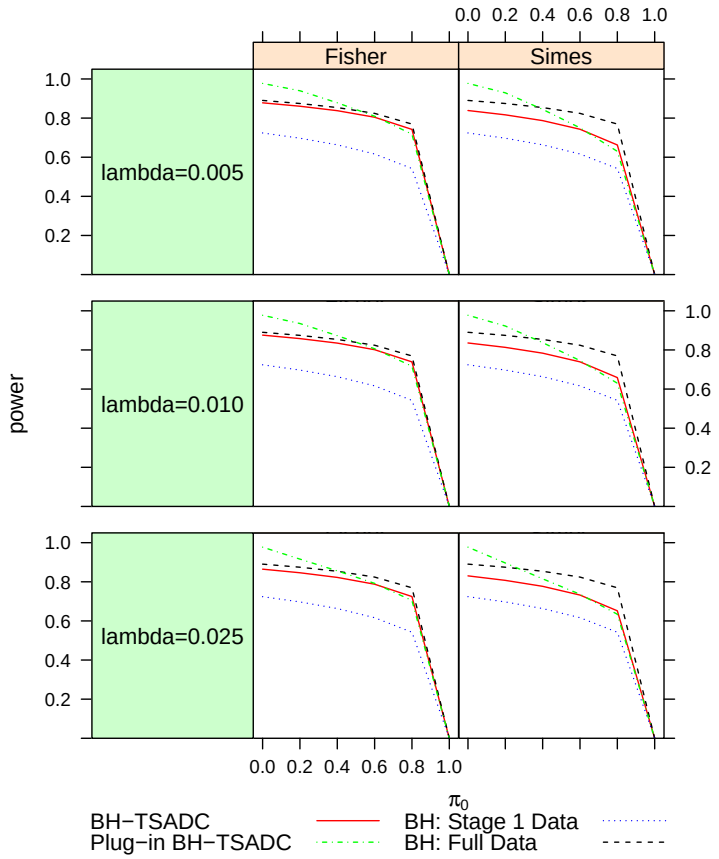

Figure 5.12: Comparison of simulated average powers of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions for $m = 1000$ with equally spaced exponential decreasing effect sizes $1.5 \times (2^2, 2^1, 2^{0.5}, 2^0)$, $\lambda = 0.005, 0.010, 0.025$, and $\lambda' = 0.5$, with the simulated average power of single-stage BH procedure based on the first stage data and full data from both stages, at $\alpha = 0.05$. [BH-TSADC: solid line, Plug-In BH-TSADC: dotdash line, BH Stage 1 Data: dotted line, and BH Full Data: dash line.]

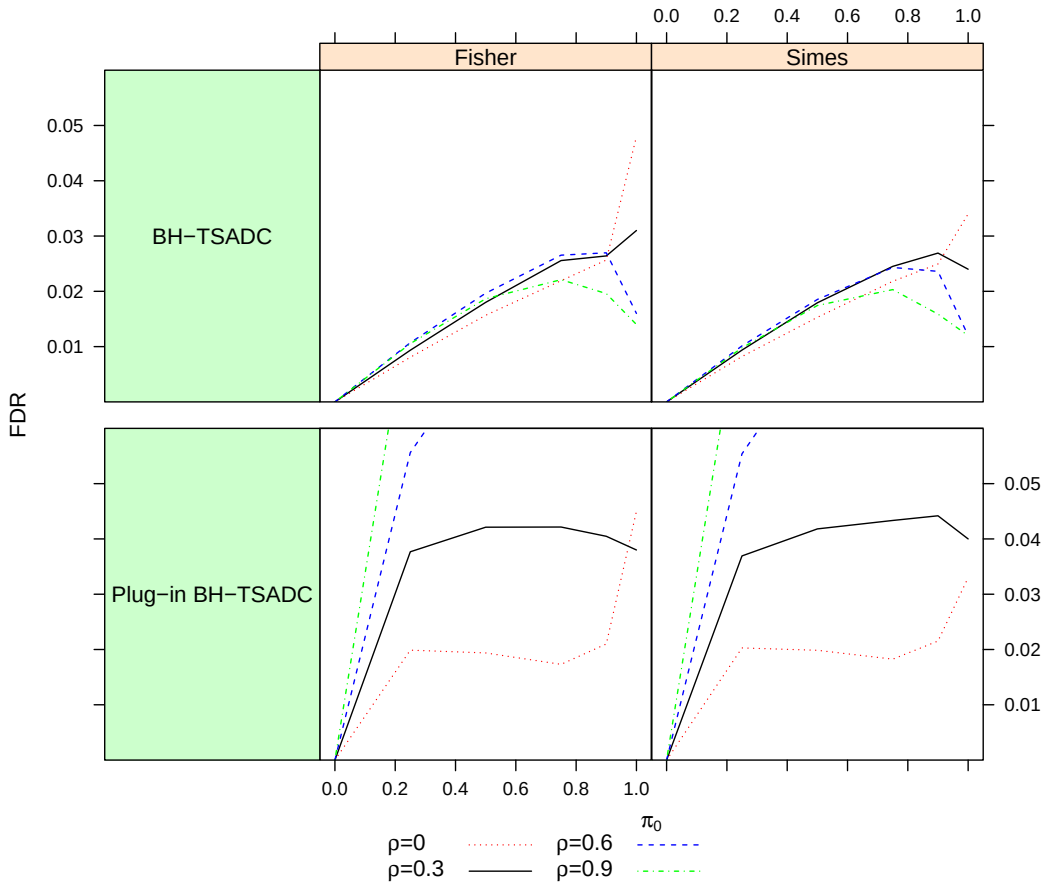

Figure 5.13: Comparison of simulated FDR's of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions under equal dependence with $\lambda = 0.025$, $\lambda' = 0.5$, $m = 100$, and $\alpha = 0.05$. [Dotted line: $\rho = 0$; solid line: $\rho = 0.3$; dash line: $\rho = 0.6$; dotdash line: $\rho = 0.9$.]

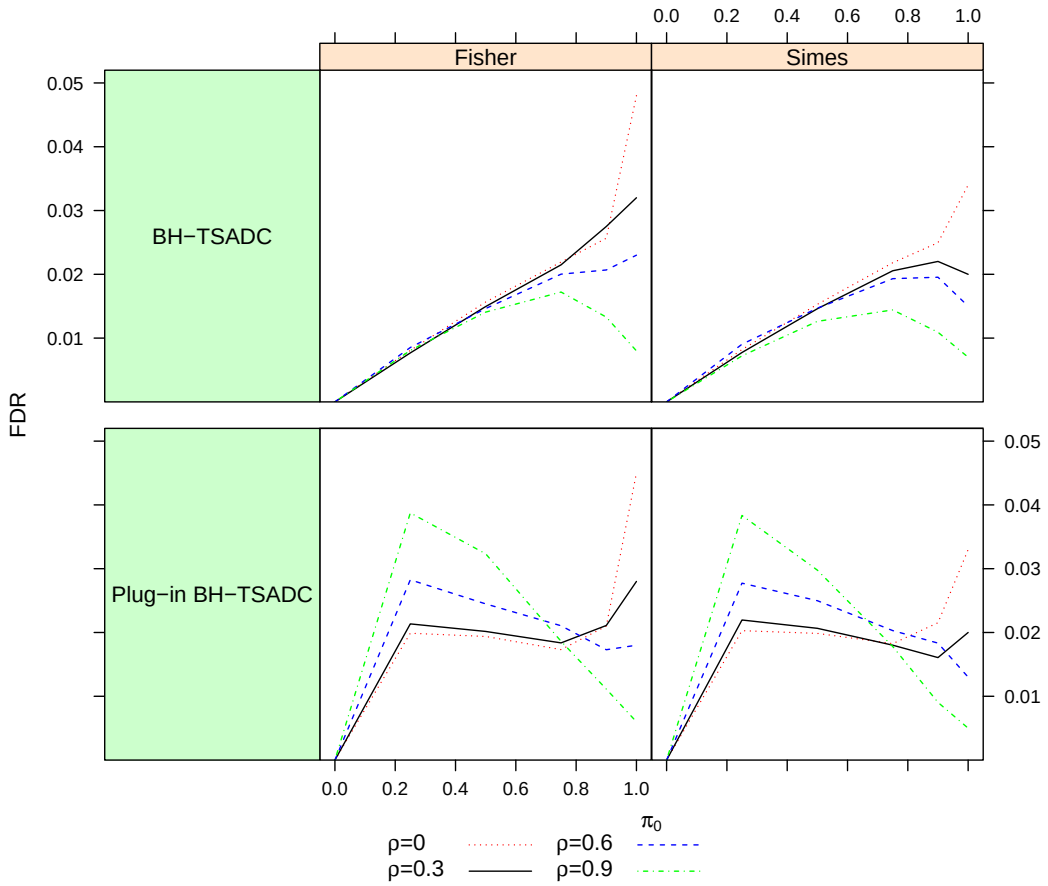

Figure 5.14: Comparison of simulated FDR's of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions under clumpy dependence with $\lambda = 0.025$, $\lambda' = 0.5$, $m = 100$, and $\alpha = 0.05$. [Dotted line: $\rho = 0$; solid line: $\rho = 0.3$; dash line: $\rho = 0.6$; dotdash line: $\rho = 0.9$.]

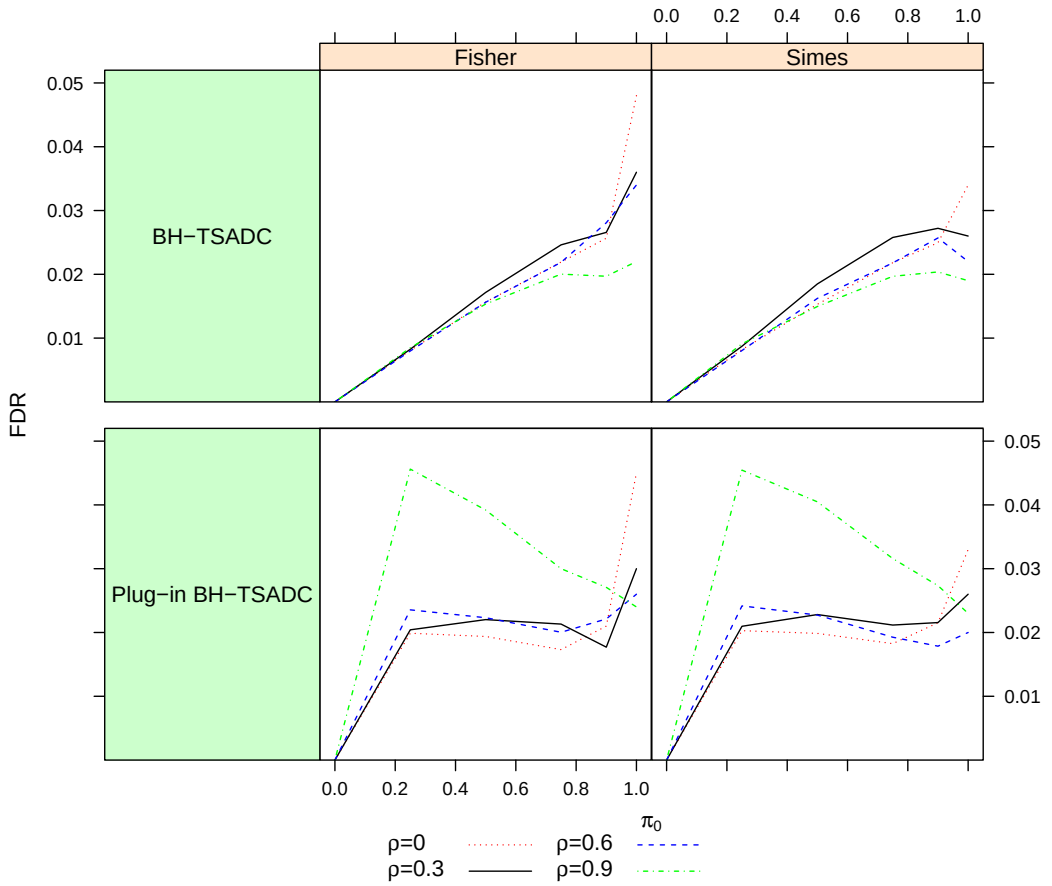

Figure 5.15: Comparison of simulated FDR's of BH-TSADC and Plug-In BH-TSADC procedures with Fisher's and Simes' combination functions under AR(1) dependence with $\lambda = 0.025$, $\lambda' = 0.5$, $m = 100$, and $\alpha = 0.05$. [Dotted line: $\rho = 0$; solid line: $\rho = 0.3$; dash line: $\rho = 0.6$; dotdash line: $\rho = 0.9$.]

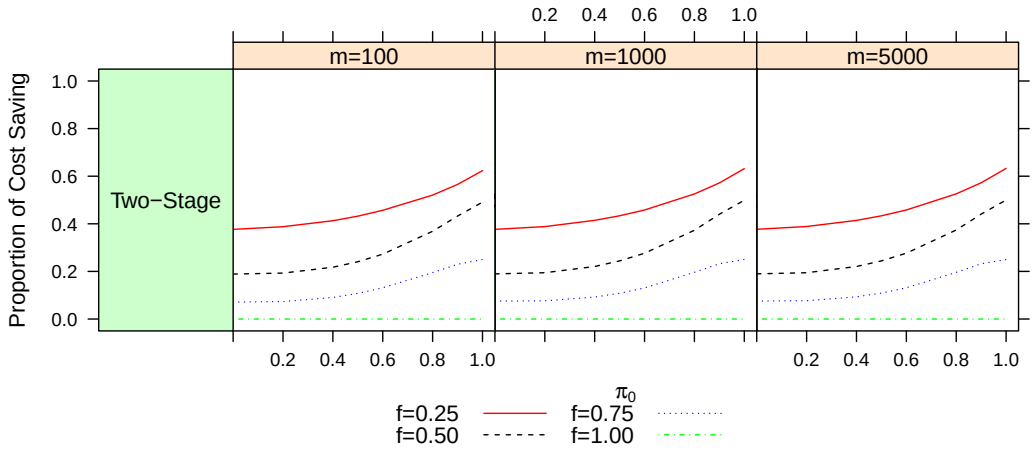

Figure 5.16: Plot of proportional cost saving versus π_0 for number of hypotheses $m = 100, 1000, 5000$ by sample allocation rate f across two stages, $\lambda = 0.025$ and $\lambda' = 0.5$ at $\alpha = 0.05$. [Solid line: $f = 0.25$; dash line: $f = 0.50$; dotted line: $f = 0.75$; dotdash line: $f = 1.00$.]

CHAPTER 6

SUMMARY AND FUTURE RESEARCH

Adaptive designs offer the opportunity of mid-course design change in clinical trials, where additional flexibility comes from careful statistical planning. They are known as adaptive, sequential, flexible, self-designing, multi-stage, dynamic, response-driven, smart, and novel. At any stage, the data may be analyzed and subsequent stages can be redesigned taking into account all available data. The statistical methodology for adaptive designs has developed quickly over the past decades.

6.1 Overall FWER Control for Single-Hypothesis in Two-Stage Combination Test

Inevitably, there is no superior strategy to determine early stopping boundaries or the critical value in two-stage adaptive designs, even though the Bauer-Köhne method is widely used. We believe our proposed method offers a comprehensive understanding of a two-stage adaptive design in terms of choosing proper early stopping boundaries and the second stage critical value. Of course, there is an arbitrariness in these choices, because the control of the overall Type I error rate at the desired level is the only criterion used while choosing these quantities. This arbitrariness can be removed by bringing in other considerations like power. There are other combination functions, such as Fisher's weighted product (Fisher, 1932) and weighted inverse normal (Mosteller and Bush, 1954), performance of which would be worth investigating as well.

6.2 Overall FWER Control for Two-Hypothesis Test in Two-Stage Combination Test

As not many current publications investigated the construction of adaptive rules while multiple hypotheses are present, the proposed stepdown-combination approach fits well in the gap, which takes advantage of combination test and closure testing principle. In theory, this method can be generalized to multiple stages since recursive application of two-stage adaptive design is valid (Bauer and Köhne, 1994).

However, it is mathematically difficult to find numerical solutions for the critical values. A further step in this aspect is worth investigating. Moreover, we can also seek to apply other multiple testing techniques such as a stepup procedure with the combination test.

6.3 Overall FDR Control for Multiple-Hypothesis in Two-Stage Combination Test

The third goal of this work was to construct a two-stage testing procedure that allows early stopping as well as controls the overall FDR. Similar to the existing procedures that control the overall FWER in literatures, we formulated the overall FDR and FDR estimate for a two-stage adaptive design, where the early stopping boundaries are prefixed and second stage critical values can be determined. When the early rejection boundary is set at 0 and early acceptance boundary is set at 1, no early stopping for efficacy or futility will be allowed and all hypotheses will continue to Stage 2. The test is equivalent to a single-stage test.

When test statistics are independent, we demonstrated in a simulation study for a two-stage adaptive design using the proposed BH-TSADC and plug-in BH-TSADC procedures with Fisher's and Simes' combination functions, where the FDR is well controlled with finite number of hypotheses m and increased power compared to a classic single-stage BH procedure. The results were emphasized by a real data application from microarray experiment. In summary, the proposed BH-TSADC procedures successfully screened the hypotheses at Stage 1, controlling the

overall FDR under independence, and leading to substantial power increase, hence increase the effectiveness of the test.

We proposed our FDR controlling procedures in this paper considering a non-asymptotic setting. However, one may consider developing procedures that would asymptotically control the FDR by taking an approach towards finding the first- and second-stage thresholds subject to the early boundaries $\lambda < \lambda'$ and the final boundary α on the FDR.

There are a number of other important issues related to the present problem which we have not touched in this work but hope to address in different communications. There are other combination functions, such as Fisher's weighted product (Fisher, 1932) and weighted inverse normal (Mosteller and Bush, 1954), performances of which would be worth investigating. Consideration of conditional error function (Proschan and Hunsberger, 1995) while defining a two-stage design before constructing FDR controlling methods is another important issue. Now that we know how to test multiple hypotheses in a two-stage design subject to first-stage boundaries on and the overall control of the FDR, we should be able to address issues relate to sample size determinations.

6.4 FDR Based Sample Size Re-Estimation Method in Large Scale Multiple Testing

An area of future of research is an FDR based sample size re-estimation method in large scale multiple testing. Sample size calculation is critical for design-

ing a study. It would be a waste of resources if the sample size is not adequate to draw reliable statistical inferences. When testing a single hypothesis, the determination of sample size rests upon the idea of maintaining control over the type I error rate at a specified level, in addition to achieving a desired power in the test to detect the true effect size. However, when there are multiple hypotheses to test simultaneously, this traditional idea of calculating sample size is not directly applicable without adjusting it to take into account the multiplicity of tests. There are two popular ways in which this multiplicity can be taken into account when it comes to controlling an overall measure of type I errors, either by considering the familywise error rate (FWER), which is the probability of at least one type I error, or by considering the false discovery rate (FDR), which is the expected proportion of type I errors (i.e., false discoveries) among all rejections (i.e., discoveries). However, in genetic studies, where the number of hypotheses is excessively large, the FDR is known to be a more reasonable and powerful measure of false discoveries than the FWER (Storey and Tibshirani, 2003). So, it is important to obtain a sample size calculation strategy while testing multiple hypotheses in an FDR framework prior to undertaking a genetic experiment.

A number of papers have been written in the past decade dealing with sample size calculation with FDR control in genetic experiments (Pawitan et al., 2005; Jung, 2005; Pounds and Cheng, 2005; Liu and Hwang, 2007; Tong and Zhao, 2008; Hu et al., 2005; Tibshirani, 2006). However, all these papers have focused on single-stage design, with such calculation depending on preliminary assumptions

or estimates about unknown design parameters.

Sample size calculation in the framework of a single-stage design, whether it is for single or multiple testing, often suffers from the uncertainty or limited information about the effect sizes, causing such calculation questionable at the planning stage. One way out of this is to carry out this calculation in two stages, with the data in the first stage providing information about the effect sizes that can be used to make a correction at the second stage to a single-stage sample size calculation method that one would have used based on all the data. Such sample size adjustment strategy has been quite popular in the context of single hypothesis testing, particularly in pharmaceutical research (Li et al., 2002), but not yet been adapted to multiple hypothesis testing.

A two-stage method of sample size calculation for large-scale multiple testing in an FDR framework may solve the problem. One possible solution is to extend Proschan and Hunsbarger (1995) from single to multiple testing, which uses the idea of maximizing the conditional power given the first-stage data to determine the final sample size. More specifically, we plan to consider the marginal false discovery rate (mFDR), which is an asymptotically equivalent (when the number of hypotheses is infinitely large) form of the usual notion of the FDR due to Benjamini and Hochberg (1995), and provide a method of re-adjusting the sample size for each hypothesis subject to controlling the mFDR at a specified level and achieving a desired average power of detecting a targeted set of alternatives with given effect sizes conditional on the first stage results, which we call the conditional average power. However, this

problem is more difficult than one might expect, owing to multiplicity of simultaneous tests as well as interim looks of multi-stage designs. It is also computationally challenging to find numeric solutions.

In summary, although statistical methodology has been developed to allow for different types of adaptive designs, these methods should not be used to replace the careful planning of a clinical trial. Before starting the trial, an efficient design must be detailed in the protocol and then adaptive design methodology provides a valuable tool for reasonable design change, data analysis, or making statistical inference.

REFERENCES

Armitage, P. (1957). Restricted sequential procedure. *Biometrika*, *44*, 9-56.

Armitage, P. (1975). *Sequential medical trials* (2nd ed.). New York: John Willey and Sons.

Bartroff, J., & Lai, T. (2008). Efficient adaptive designs with mid-course sample size adjustment in clinical trials. *Statistics in Medicine*, *27*, 1593-1611.

Bauer, P. (1989). Multistage testing with adaptive designs (with discussion). *Biometrie Inform. Med. Biol*, *20*, 130-148.

Bauer, P., Brannath, W., & Posch, M. (2001). Flexible two-stage designs: an overview. *Methods Inf. Med.*, *40*(2), 117-121.

Bauer, P., & Budde, M. (1994). Multiple testing for detecting efficient dose steps. *Biometrical Journal*, *36*, 3-15.

Bauer, P., & Einfalt, J. (2006). Application of adaptive designs - a review. *Biometrical Journal*, *48*(4), 493-506.

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

Bauer, P., & Köhne, K. (1994). Evaluation of experiments with adaptive interim

analysis. *Biometrics*, *50*, 1029-1041.

Bauer, P., & Röhmel, J. (1995). An adaptive method for establishing a dose-response relationship. *Statistics in Medicine*, *14*, 1595-1607.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*, 289-300.

Benjamini, Y., & Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, *25*, 6083.

Benjamini, Y., Krieger, A., & Yekutieli, D. (2006). Adaptive linear step-up false discovery rate controlling procedures. *Biometrika*, *93*(3), 491-507.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, *29*, 1165-1188.

Blanchard, G., & Roquain, E. (2009). Adaptive fdr control under independence and dependence. *Journal of Machine Learning Research*, *10*, 2837-2871.

Brannath, W., Bauer, P., Maurer, W., & Posch, M. (2003). Sequential tests for non-inferiority and superiority. *Biometrics*, *59*(106), 14.

Brannath, W., Koenig, F., & Bauer, P. (2007). Multipplicity and flexibility in clinical trials. *Pharmaceutical Statistics*, *6*, 205-216.

Brannath, W., Posch, M., & Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association*, *97*(457), 236-244.

Bretz, F., Koenig, F., Brannath, W., Glimm, E., & Posch, M. (2009). Adaptive

designs for confirmatory clinical trials. *Statistics in Medicine*, *28*, 1181-1217.

Bretz, F., Schmidli, H., König, F., Racine, A., & Maurer, W. (2006). Confirmatory seamless phase ii/iii clinical trials with hypotheses selection at interim: general concepts. *Biometrical Journal*, *48*(4), 623-634.

Burman, C., & Sonesson, C. (2006). Are flexible designs sound? *Biometrics*, *62*(3), 664-669.

Chang, M. (2005, May). *Adaptive design for clinical trials.* (Invited paper for International Symposium on Applied Stochastic Models and Data Analysis, France)

CHMP. (2007, October). *Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design.* EMEA CHMP/EWP/2459/02. (European Medicines Agency)

Chow, S., & Chang, M. (2006). *Adaptive design methods in clinical trials.* New York: Chapman and Hall/CRC Press, Taylor and Francis.

Chow, S., & Chang, M. (2008). Adaptive design methods in clinical trials a review. *Orphanet Journal of Rare Disease*, *3(11)*.

Chow, S., Lu, Q., & Tse, S. (2007). Statistical analysis for two-stage adaptive design with different study points. *Journal of Biopharmaceutical Statistics*, *17*, 1163-1176.

Cui, L., Huang, H., & Wang, S. (1999). Modification of sample size in group sequential clinical trials. *Biometrics*, *55*, 853-857.

Denn, J. (2000). Estimation following extension of a study on the basis of conditional

power. *Journal of Biopharmaceutical Statistics*, *10*(2), 131-144.

Dudoit, S., Laan, M. van der, & Pollard, K. (2004). Multiple testing. part i. single-step procedures for control of general type i error rates. *Statistical Appl. Genetics Molecular Biol.*, *3*(1). (http://www.bepress.com/sagmb/vol13/iss1/art13)

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

FDA. (2004, March). *Innovation/stagnation: challenge and opportunity on the critical path to new medical products.* Available at http://www.fda.gov/oc/initiatives/criticalpath/whitepapter.html.

FDA. (2010, February 25). *Draft guidance for industry: adaptive design clinical trials for drugs and biologics.* Available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf, for public comment. (US Food and Drug Administration)

Fernando, R., Nettleton, D., Southey, B., Dekkers, J., & Rothschild, M. (2004). Controlling the proportion of false positives in multiple dependent tests. *Genetics*, *166*, 611-619.

Fisher, L. (1998). Self-designing clinical trials. *Statistics in Medicine*, *17*, 1551-1562.

Fisher, R. (1932). *Statistical methods for research workers* (4th ed.). London:Oliver and Boyd.

Fluhr, J. (1998). Antibacterial and sebosuppressive efficacy of a combination of chloramphenicol and pale sulfonated shale oil. *ArzneimittelForschung/Drug Research*, *48(I)*, 188-196.

Gavrilov, Y., Benjamini, Y., & Sarkar, S. (2009). An adaptive step-down procedure with proven fdr control under independence. *Annals of Statistics*, *37*(2), 619-629.

Genovese, C., & Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society, Series B*, *64*(499-517).

Gordon, A. (2007). Explicit formulas for generalized family-wise error rates and unimprovable step-down multiple testing procedures. *Journal of Statistical Planning and Inference*, *137*, 3497-9512.

Hellmich, M. (2001). Monitoring clinical trials with multiple arms. *Biometrics*, *57*, 892-898.

Hellmich, M., & Hommel., G. (2004). Mmultiple testing in adaptive designs - a review. *Recent developments in multiple comparison procedures institute of mathematical statistics, Lecture Notes - Monograph Series*, *47*, 33-47.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, *75*, 800-802.

Hochberg, Y., & Hommel, G. (1998). Multiple hypotheses, simes' test of in kotz,s., c.b.read and d.l.banks eds. *Encyclopedia of Statistical Sciences*, *2*, 418-422.

Hochberg, Y., & Tamhane, A. (1987). *Multiple comparison procedure*. John Wiley

and Sons.

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

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a midified test procedure. *Biometrika*, *75*(383-386).

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

Hommel, G., Lindig, V., & Faldum, A. (2005). Two-stage adaptive designs with correlated test statistics. *Journal of biopharmaceutical statistics*, *15*, 613-623.

Jennison, C., & Turnbull, B. (2000). *Group sequential methods with applications to clinical trials*. New York: Chapman and Hall/CRC Press.

Jennison, C., & Turnbull, B. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine*, *22*, 971-993.

Jennison, C., & Turnbull, B. (2005). Meta-analyses and adaptive group sequential designs in the clinical development process. *Journal of Biopharmaceutical Statistics*, *15*, 537-558.

Jennison, C., & Turnbull, B. (2006). Adaptive and nonadaptive group sequential tests. *Biometrika*, *93*(1), 1-21.

Kelly, P., Stallard, N., & Todd, S. (2005). An adaptive group sequential design for phase ii/iii clinical trials that select a single treatment from several. *Journal of Biopharmaceutical Statistics*, *15*, 641-658.

Kieser, M., Bauer, P., & Lehmacher., W. (1999). Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biometrical Journal*, *41*, 261-277.

Koch, A. (2006). Confirmatory clinical trials with an adaptive design. *Biometrical Journal*, *48*(574-585).

Korn, E., Tronendle, J., McShane, L., & Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*, *124*, 379-398.

Kropf, S., Hommel, G., Schmidt, U., Brickwedel, J., & Jepsen, M. (2000). Multiple comparisons of treatments with stable multivariate tests in a two-stage adaptive design, includng a test for non-inferiority. *Biometrical Journal*, *42*, 951-965.

Lan, K., & DeMets, D. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, *70*, 659-663.

Lan, K., & DeMets, D. (1987). Group sequential procedures: calendar versus information time. *Statistics in Medicine*, *8*, 1191-1198.

Lehmacher, W., Kieser, M., & Hothorn, L. (2000). Sequential and multiple testing for dose-response analysis. *Drug Inf. J.*, *34*, 591-597.

Lehmacher, W., & Wassmer, G. (1999). Adaptive sample size calculation in group sequential trials. *Biometrics*, *55*, 1286-1290.

Lehmann, E., & Romano, J. (2005). Generalizations of the familywise error rate. *Annals of Statistics*, *33*, 1138-1154.

Li, G., Shih, W., Xie, T., & Lu, J. (2002). A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics*, *3*, 277-287.

Liu, Q., & Chi, G. (2001). On sample size and inference for two-stage adaptive designs. *Biometrics*, *57*(1), 172-177.

Liu, Q., Proschan, M., & Wassmer, G. (2002). A unified theory of two-stage adaptive designs. *Journal of American Statistical Association*, *97*, 1034-1041.

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*, *40*, 463-474.

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

Miller, R., Galecki, A., & Shmookler-Reis, R. (2001). Interpretation, design, and analysis of gene array expression experiments. *J Gerontol A-Biol*, *56*, B52-B57.

Mosteller, F., & Bush, R. (1954). Selected quantitative techniques. *Handbook of Social Psychology*, *1*.

Müller, H., & Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*, *57*, 886-891.

Müller, H., & Schäfer, H. (2004). A general statistical principle for changing a design any time druing the course of a trial. *Statistics in Medicine*, *23*, 2497-2508.

O'Brien, P., & Fleming, T. (1979). A multiple testing procedure for clinical trials.

Biometrics, *35*, 549-556.

O'Quigley, J., Pepe, M., & Fisher, L. (1990). Continual reassessment method: A practical design for phase i clinical trial in cancer. *Biometrics*, *46*, 33-48.

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

Peritz, E. (1970). *A note on multiple comparisons.* (Unpublished manuscript)

PhRMA. (2006). *Full white paper.* (DIJ40, 421-484)

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

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

Posch, M., & Bauer, P. (1999). Adaptive two-stage designs and the conditional error function. *Biometrical Journal*, *41*, 689-696.

Posch, M., & Bauer, P. (2000). Interim analysis and sample size reassessment. *Biometrics*, *56*, 1170-1176.

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*, *24*, 36973714.

Posch, M., Zehetmayer, S., & Bauer, P. (2009). Hunting for significance with the false discovery rate. *Journal of the American Statistical Association*, *104*(486), 832-840.

Proschan, M., & Hunsberger, S. (1995). Designed extension of studies based on

conditional power. *Biometrics*, *51*, 1315-1324.

Rom, D., Costello, R., & Connell, L. (1994). On closed test procedures for dose-response analysis. *Statistics in Medicine*, *13*, 1583-1596.

Rosenberger, W., & Lachin, J. (2002). *Randomization in clinical trials.* New York: John Willey and Sons.

Sampson, A., & Sill, M. (2005). Drop-the-loser design: normal case (with discussions). *Biometrical Journal*, *47*, 257-281.

Samuel-Cahn, E. (1996). Is the simes improved bonferrorni procecdure conservative? *Biometrika*, *83*(928-933).

Santagopan, J., & Elston, R. (2003). Optimal two-stage genotyping in population-based association studies. *Genetic Epidemiology*, *25*, 149-157.

Sarkar, S. (1998). Some probability inequalities for ordered $mtp2$ random variables: a proof of the simes conjecture. *Annals of Statistics*, *26*(494-504).

Sarkar, S. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Annals of Statistics*, *0*, 239-257.

Sarkar, S. (2006). False discovery and false nondiscovery rates in single-step multiple testing procedures. *Annals of Statistics*, *34*, 394-415.

Sarkar, S. (2007). Stepup procedures controlling generalized $fwer$ and generalized fdr. *Annals of Statistics*, *35*(6), 2405-2420.

Sarkar, S. (2008a). On the simes inequality and its generalization. *IMS Collections Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, *1*, 231-242.

Sarkar, S. (2008b). On methods controlling the false discovery rate. *Sankhya: The Indian Journal of Statistics*, *70-A*(part 2), 135-168.

Sarkar, S., & Chang, C. (1997). The simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association*, *92*(1601-1608).

Shen, Y., & Fisher, L. (1999). Statistical inference for self-designming clinical trials with a one-sided hypothesis. *Biometrics*, *55*, 190-197.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, *62*, 626-633.

Simes, R. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika*, *73*(751-754).

Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B*, *64*, 479-498.

Storey, J. (2003). The positive false discovery rate: a bayesian interpretation and the q-value. *Annals of Statistics*, *31*, 2013-2035.

Storey, J., Taylor, J., & Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B*, *66*, 187-205.

Storey, J., & Tibshirani, R. (2003). Statistical significance in genomewide studies. *Proceedings of the National Academy of Science USA*, *100*, 9440-9445.

Tamhane, A., Hochberg, Y., & Dunnett, C. (1996). Multiple test procedures for dose finding. *Biometrics*, *52*, 21-37.

Tamhane, A., Liu, W., & Dunnett, C. (1998). A generalized step-up-down multiple test procedure. *The Canadian Journal of Statistics*, *26*(2), 353-363.

Tippett, L. (1931). *The method of statistics*. Williams and Northgate, London.

Tsiatis, A., & Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, *90*, 367-378.

Victor, A., & Hommel, G. (2007). Combining adaptive design with control of the false discovery rate - a generalized definition for a global p-value. *Biometrical Journal*, *49*, 94-106.

Victor, N. (1982). Explorary data analysis and clinical research. *Edthods Inform. Med.*, *21*, 53-54.

Wald, A. (1947). *Sequential analysis*. New York: John Wiley and sons.

Wang, S., & Tsiatis, A. (1987). Approximately optimal one-parameter boundaries for a sequential trials. *Biometrics*, *43(1)*, 193-200.

Wassmer, G. (2000). Basic concepts of group sequential and adaptive group sequential test procedures. *Statistical Papers*, *41*, 253-279.

Weller, J., Song, J., Heyen, D., Lewin, H., & Ron, M. (1998). A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics*, *150*, 1699-1706.

Westfall, P., & Krishen, A. (2001). Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *Journal of Statistical Planning and Inference*,

99, 25-40.

Westfall, P., Tobias, R., D. Rom, R. W., & Hochberg, Y. (1999). *Multiple comparisons and multiple tests using the sassystem.* Books by Users. Cary: SAS®Institute Inc.

Whitehead, J. (1997). Bayesian decision procedures with application to dose-finding studies. *International Journal of Pharmaceutical Medicine*, *11*, 201-208.

Zaykin, D., Zhivotovsky, L., Westfall, P., & Weir, B. (2002). Original article truncated product method for combining p-values. *Genetic Epidemiology*, *22*(2), 170 - 185.

Zehetmayer, S., Bauer, P., & Posch, M. (2005). Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics*, *21*, 3771-3777.

Zehetmayer, S., Bauer, P., & Posch, M. (2008). Optimized multi-stage designs controlling the false discovery or the family-wise error rate. *Statistics in Medicine*, *27*, 4145-4160.

Zhang, W., Sargent, D., & Mandrekar, S. (2006). An adaptive dose-finding design incorporating both toxicity and efficacy. *Statistics in Medicine*, *25*, 2365-2383.