ON GROUP-SEQUENTIAL MULTIPLE TESTING CONTROLLING FAMILYWISE ERROR RATE

A Dissertation
Submitted to
the Temple University Graduate Board

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

by
Yiyong Fu
July, 2015

Examining Committee Members:

Sanat K. Sarkar, Advisory Chair, Statistics
Xu Han, Statistics
Zhigen Zhao, Statistics
Cheng Yong Tang, Statistics
Dror M. Rom, External Reader, Prosoft Clinical

ABSTRACT

ON GROUP-SEQUENTIAL MULTIPLE TESTING CONTROLLING FAMILYWISE ERROR RATE

Yiyong Fu

DOCTOR OF PHILOSOPHY

Temple University, July, 2015

Professor Sanat K. Sarkar, Chair

In modern scientific research, the importance of multiplicity adjustment has gained wide recognition. Without it, there will be too many spurious results and reproducibility becomes an issue; with it, if overtly conservative, discoveries will be made more difficult. Since, in the current literature on repeated testing of correlated multiple hypotheses, Bonferroni-based methods are still the main vehicle carrying the bulk of multiplicity adjustment, there is room for power improvement by suitably utilizing both hypothesis-wise and analysis-wise dependencies. This research will contribute to the development of a natural group-sequential extension of some classical stepwise multiple testing procedures, such as Dunnett's stepdown and Hochberg's stepup procedures. It is shown that, under some commonly seen dependency structure of test statistics, the proposed group-sequential procedures strongly control the familywise error rate (FWER) while being more powerful than the recently developed class of group-sequential Bonferroni-Holm's procedures. Particularly in this research, a convexity property is discovered for the distribution of the maxima of pairwise null P-values with the underlying test statistics having distributions such as bivariate normal, t, Gamma, F, or Archimedean copulas. Such property renders itself for an immediate use in improving Holm's procedure by incorporating pairwise dependencies of P-values. The improved Holm's procedure, as all stepdown multiple testing procedures, can also be naturally extended to group-sequential setting.

ACKNOWLEDGEMENTS

It was a privilege to have been able to work under Professor Sanat K. Sarkar, to whom I am deeply indebted for his research guidance and inspiration. And, I've benefited enormously in the past 8 years from his rigorous teaching and his cutting-edge research expertise in the field of mathematical statistics.

And, I wish to extend my heartfelt thanks to Professor Xu Han, Professor Zhigen Zhao, and Dr. Dror Rom at Prosoft Clinical for their valuable comments and suggestions.

I also wish to thank Professor Ajit C. Tamhane and one of his graduate students Mr. Jiangtao Gou, at Northwestern University, for having provided me with some comments on a weighted version of Dunnett & Tamhane's step-down procedure as presented in Chapter 4.

To my Mom

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Multiplicity Problems

Multiplicity of inferences are comprehensively discussed in the introductory chapters of two excellent books *Multiple Comparison and Multiple Tests Using the SAS System*, Westfall et al. (1999), and *Multiple Comparisons Using R*, Bretz et al. (2011). The two introduction chapters are complementary to each other on multiplicity problems, with the latter one focused on the problems and their implication in clinical trials.

It is sometimes on the news that a company's Phase-III registration trial failed in meeting the primary endpoint. That is to say, the large-scale confirmatory trial legally required for launching a medicinal product failed to replicate the effect size as previously observed. Sometimes the gap is very large. For example, it is not unusual to see that the effect size attained from a Phase-III trial is less than half of that concluded from several Phase-II trials. Why does a large-scale confirmatory trial fail to replicate those earlier results? The problem is complex. We are only seeking for answers from the perspective of statistical science. Statistically, the likely culprit is selection bias or selection effect, which is one of the main sources of multiplicity of inference. There are sometimes no clean-cut as to a Phase-II trial being aimed for learning or for confirmatory. Early phase clinicians often have some leeway as to what to do under unanticipated or unspecified circumstances, where selection bias can kick-in in subtle and unwitting ways, resulting in an inflated false positive (the Type I error) rate. Specific examples of selection bias are given in the introduction chapter of Bretz et al. (2011).

One of ways to reduce the Type-I error rate (and the huge cost associated with it) is to systematically adopt appropriate FWER-controlling procedures in the Phase-II trials prior to conducting registration Phase-III trials. For example, Fisher's protected least significance method is fairly popular and widely used in Phase II trials, where multiple dose-and-formulations are compared. But that method controls the FWER weakly. It is easier to yield a significant p-value. On the other hand, by adhering to some existing strong FWER-controlling procedures in Phase-II trials, discovery will be made more difficult. This dilemma calls for improving some existing FWER-controlling procedures, if it is possible. We find occasions where there is room for improvement by fully utilizing hypothesis-wise dependencies or by appropriately estimating the proportion of the null hypotheses, so that the power of the procedures will be optimized or nearly so. This dissertation is focused on incorporating hypothesis-wise correlations, and not touching on the adaptiveness of the procedures, which is another frontier undergoing rapid development in recent years. Clinical trials can be fixed-sample designed or group-sequential designed (GSD), with the latter gaining popularity in multi-year multi-center large-scale later-phase clinical trials. We see there is room for power improvement of the recently proposed group-sequential multiple testing procedures with multiple endpoints (Maurer & Bretz, 2013; Ye et al. 2012).

1.2 The Crossroad of Multiple Testing and Group-sequential Testing

The subject studied stands where the classical multiple testing procedures intersect with the classical group-sequential methodologies. In the field of multiple testing (for fixed-sample setting), the class of FDR-controlling procedures is undergoing rapid development to suit some study objectives of modern day large-scale multiple testings in the field of genomics, etc. Within the class of FWER-controlling procedures, recent years witnessed some development of adaptive procedures. However, in this dissertation, we exclusively study the class of classical multiple testing procedures, such as Holm's stepdown procedure, Dunnett's stepdown procedure, and Hochberg's stepup procedure, the Weighted Bonferroni procedure, etc. The field of group-sequential testing methodologies (for single-hypothesis) also saw rapid development in recent years. Adaptive trials, in which earlier interim results can be used for re-designing the trials, such as sample size re-estimation, have been proposed. Our method, however, is based on the original, the classical group-sequential methodology, which means no interim results will be used for re-designing. We assume that information fractions (not necessary equally spaced) are pre-specified. Our exclusive focus is to adjust the critical boundary appropriately to reflect the hypothesis-wise dependencies.

That is to say, this research is focused on how to "best" perform the classical FWER-controlling multiple testing in the group-sequential setting where the clinical objective is to claim efficacy on at least one of the endpoints. In other words, we are focused on improving the type of group-sequential trial that is designed to reject at least one of the hypotheses (endpoints), with the more rejections being the more desirable. In contrast, the type of group-sequential trial that is designed to reject all of the hypotheses is out of the scope of our current research. Two cohorts of researchers have made some significant contributions in this joint area. Tang et al. 1999 (probably first) proposed the use of closed testing principle in group-sequential setting. They proposed a group-sequential procedure testing multiple primary endpoints. Their procedure is constructed based on the closed testing principle and controls the FWER strongly at a pre-specified level. But their procedure is not based on individual marginal P-values. It is based on a combined test statistic, which is "centered linear combination test statistic" as they called. It is often cumbersome to calculate the test statistic. And the procedure does not yield a short-cut of the closed testing procedure. In the most recent years (2012, 2013), the work of Maurer et al., Bretz et al., Ye et al. etc, formulated a class of group-sequential weighted Bonferroni multiple testing procedure, which is also constructed per the closed testing principle. This recent class is conveniently based on individual P-values and is sequentially rejective for each of the analysis time points, requiring at most $\max(n, J)$ steps of the testing, where n is the number of hypotheses, and J is the number of analysis time points.

Hypothesis-wise correlation is utilized in some ways in the proposed group-sequential procedure of Tang et al (1999), but the procedure lacks of convenience in obtaining critical values, and it could take too many steps (tests) to complete, as many as $J(2^n - 1)$ tests. And, there are other drawbacks. The group-sequential weighted Bonferroni procedure is convenient for use, since an α-level critical boundary is immediately available by invoking R function *gsDesign*, given pre-specified error spending function and information fractions. But the procedure does not take into account hypothesis-wise correlations, thus is highly conservative under highly positive dependencies. Based on the class of group-sequential weighted Bonferroni procedure, we propose two newer group-sequential procedures that utilize the hypothesis-wise dependencies. Our proposed procedures maintain strong and tighter control of the FWER. The increase in computational burden is either none or modest.

1.3 Chapter Organization and Disclosure

Chapter 2 is a review of the literatures in the areas of multiple testings and group-sequential testings. The chapter is focused on the classical FWER-controlling multiple testing procedures and the classical (non-adaptive) group-sequential methodologies.

Chapter 3 presents some new results obtained in my research toward improving one of the classical stepdown procedures, the Holm's procedure, for multiple testing by incorporating pairwise correlations in some commonly used model settings.

Chapter 4 proposes a group-sequential parametric stepdown procedure.

Chapter 5 proposes a group-sequential stepup procedure. An extension of stepup multiple testing procedures to the group-sequential setting does not yet exist in the literature. Some researchers have been working on this topic, and research results will probably soon appear in publication.

Chapter 6 lists a few further topics for future research.

CHAPTER 2

LITERATURE REVIEW

2.1 Group-Sequential Design (GSD)

2.1.1 *An early milestone in the development of sequential design*

Modern theory of sequential analysis largely stems from the work of Wald (1947), who devised a sequential probability ratio test (SPRT). Suppose the observations are sequentially taken from a known form of distribution but with unknown parameter θ. Then, for testing the simple null $H_0 : \theta = \theta_0$ against the simple alternative $H_1 : \theta = \theta_1$, the SPRT works as follows: Likelihood ratio are successively calculated based on the sequentially accumulated observations, as long as the calculated likelihood ratio stays within interval (a, b), where a, b are dictated by a pre-specified α (the Type-I error rate) and β (the Type-II error rate), and is given as $a = log\frac{1-\beta}{\alpha}$ and $b = log\frac{\beta}{1-\alpha}$.

The SPRT has theoretical optimality in that, among all tests with error probability not exceeding α, β, the SPRT attains the smallest possible average sample number (ASN), provided either H_0 or H_i is true. However the ASN can be very large if θ deviates much from θ_0 and θ_1. The SPRT applies well for industry quality control in which the experimental unit (piece of manufactured product) is sampled one by one, and continuous assessment of quality (e.g., proportion of defect products) is made.

2.1.2 *Group-sequential design of experiment*

Armitage (1957) and others pioneered the use of sequential design method in comparative clinical trial, where assessment is feasible after group(s) of experimental units (such as patients) are observed. In a group sequential design, data inspection is conducted following subgroup(s) of observations, as opposed to following the completion of the overall group of observations. The trial can stop earlier based on the accumulated observations. Suppose there are K sequential groups of observations. Data analyses conducted at $i = 1, \ldots, K-1$ are called interim analyses, and the analysis at the final stage K is called final analysis. Multiple inspections (looks, or analyses) of the data need to be taken into account in controlling the experimental-wise Type-I error rate. Multiple inspections during the accumulating of data is called *repeated significance testing* (Armitage 1969). Assume the null hypothesis is true, if a level-α significance test is repeated a few times during the accumulating of data, the probability of obtaining a significant result (rejection of the null) will exceed the nominal level α. A suitable GSD makes appropriate adjustment for the multiplicity issues arising from multiple looks, and controls the experimental-wise Type-I error rate.

2.1.3 *Fundamental calculations for group-sequential design*

Same as that of a fixed-sample design, a GSD is designed to control experimental-wise Type-I error rate, and to attain a specific power $(1-\beta)$ for detecting an assumed effect size θ_1. Once α, β, and θ_1 values are set, the problem is to choose a set of critical regions and to solve for the planned sample size N.

Let's consider the normal model setting, that is, the observation (response) is normally distributed, with unknown mean θ and known variance σ^2. We are to test the null hypothesis $H : \theta = \theta_0$ against the alternative $K : \theta \neq \theta_0$.

Let $n_i, i = 1, \ldots, K$, denote the planned stage-wise sample sizes for the K stages (sequences) of observations. Total sample size $N = \sum_i^K n_i$. Let $\bar{x}_{k,cum}$ denote the cumulative sample mean up to stage k, and let Z_i denote the standardized cumulative sample mean $\bar{x}_{i,cum}$ at stage i, $i = 1, \ldots, K$. Consider a GSD that has no early acceptance of the null hypothesis. Such a GSD consists of specifying the continuation regions C_i, $i = 1, \ldots, K-1$, at the interim analyses and the acceptance region C_K at the final analyses. If at any interim analysis i, if $Z_i \notin C_i$, then the trial stops and the null hypothesis is rejected. If the trial continues to the final analysis K, and if $Z_K \in C_K$ then the null hypothesis is retained, and if $Z_K \notin C_K$, then the null hypothesis is rejected. Experimental-wise Type I error α is controlled by meeting the following condition:

$$\mathbb{P}_{H_0}\left(\bigcap_{i=1}^{K}\{Z_i \in C_i\}\right) = 1 - \alpha$$

Under H_0, the probability of exiting the boundary at stage k is denoted as

$$\pi_0(k) = \mathbb{P}_{H_0}\left(\bigcap_{j=1}^{k-1}\{Z_j \in C_j\} \cap \{Z_k \notin C_k\}\right)$$

Type I error rate α is the sum of the probability of exiting boundaries over stage $k = 1$ to K, that is, $\alpha = \sum_{k=1}^K \pi_0(k)$. For a given test, defined by a set of stage-wise critical boundaries, power at stage k, denoted as $\pi_1(k)$, is calculated as

$$\pi_1(k) = \mathbb{P}_{H_1}\left(\bigcap_{j=1}^{k-1}\{Z_j \in C_j\} \cap \{Z_k \notin C_k\}\right)$$

The power of the test, $1 - \beta$, is the sum of $\pi_1(k)$ over $k = 1$ to K. Note that power at $\theta \neq 0$ is a function of sample size(s). Given an assumed effect size $\theta \neq 0$, the power of the test is attained at a pre-specified value by adjusting the maximum sample size N, and hence the stage-wise sample sizes, because of a fixed ratio of allocation of sample size among the stages.

Once a GSD has been defined in terms of a set of stage-wise critical boundaries and stage-wise sample sizes, ASN is calculated as:

$$ASN = n_1 + \sum_{i=2}^{K} n_i \mathbb{P}_\theta \left(\bigcap_{j=1}^{i-1} \{Z_i \in C_i\} \right)$$

Note that the maximum sample size, $N_{max} = \sum n_i$, is always bigger than sample size of the fixed-sample design for the same value of α and β. Note that N_{max} is independent of effect size. It can be seen that ASN under the null H_0 is bigger than ASN under the alternative H_1, since the probability of continuing to the next stage is larger under H_0 than under H_1.

The following is a perspective on the sample space. Consider a K-stage GSD that has no early acceptance of the null H_0. Let's define a random variable τ standing for the stage at which rejection of the null is made:
$\tau = k$ if a rejection of H_0 occurs at stage k;
$\tau = 0$ if no rejection is made.
Then the total sample space is $\Omega = \{0, 1, 2, \ldots, K\}$. It can been seen that Ω is partitioned into $k + 1$ exclusive sub-spaces per the values of τ. The subset $\{1, 2, \ldots, K\}$ constitutes the critical region, and the subset $\{0\}$ is the acceptance region.

2.1.4 *The Wang & Tsiatis (1987) class of group-sequential design*

There are many choices of $C_i, i = 1, \ldots, K$, each leading to a valid GSD. Wang & Tsiatis (1987) made some generalization out of a few commonly used GSDs. Classical GSDs such as that of Pocock (1977) and that of O'Brien & Fleming (1979) are members of the Wang & Tsiatis class, which is a class of symmetric boundaries in the following form:

$$C_i = (-u_i, u_i), \quad u_i = c(k, \alpha, \Delta) i^{\Delta - 0.5}$$

where $c(k, \alpha, \Delta)$ is chosen to control the Type-I error rate. The decision rule is: H_0 is rejected when $|Z_i| > u_i$ at any $i = 1, \ldots, K$; otherwise, H_0 is retained.

- Set $\Delta = 0.5$, the boundary values are all equal, leading to the Pocock (PO) design.
- Set $\Delta = 0$, the boundary values decrease from earlier stages to later stages. This is the O'Brien & Fleming (OBF) design.
- Set $\Delta = 0.25$, the boundary values, at early stage, are not as large as that of OBF, and at later stage, are not as small as the of OBF. This type of design was originally proposed by Wang & Tsiatis (1987).

2.1.5 *The error spending function approach*

A further development or generalization was added by Lan & DeMets (1983), who proposed the error spending function approach to address the problem of unequal spacing in interim looks. The (cumulative) Type-I error rate function α^* is a non-decreasing function of information fraction t, $t \in [0, 1]$, $\alpha^*(0) = 0$ and $\alpha^*(1) = \alpha$. Several choices of error spending functions were proposed. One of them is $\alpha^*(t_i) = \alpha \ln(1 + (e-1)t_i)$. For two-sided tests, critical boundaries for the first interim analysis is given as $u_1 = \Phi^{-1}(1 - \alpha^*(t_i)/2)$, while the critical values for the remaining analyses are computed successively through:

$$\mathbb{P}_{H_0}\left(\bigcap_{j=1}^{i-1} \{|Z_j| < u_j\} \cap \{|Z_i| > u_i\}\right) = \alpha^*(t_i) - \alpha^*(t_{i-1})$$

Note that, per the classical group sequential designs, the timing of analysis $j+1$ can be dependent on logistics such as patient recruitment, but cannot be dependent on interim data (outcome) of analysis j or earlier. In practise, the interim data is typically blinded to sponsors and investigators. The Type-I error rate will be inflated if the timing for the next analysis is based on the earlier analysis results.

2.1.6 *Some concepts and results*

A useful concept introduced by Jennison & Turnbull (2000) is information level $\mathcal{I}$ in the following context. Let $X_{Ai} \sim N(\mu_A, \sigma_A^2)$, $i = 1, 2, \ldots$, for subjects allocated to treatment A, and let $X_{Bi} \sim N(\mu_B, \sigma_B^2)$, $i = 1, 2, \ldots$, for subjects allocated to treatment B. Let n_{Ak} and n_{Bk} respectively denote the cumulative number of observations on treatment A and B at the kth analysis. The natural estimate of $\mu_A - \mu_B$ is

$$\bar{X}_A^{(k)} - \bar{X}_B^{(k)} \quad \sim \quad N(\mu_A - \mu_B, \quad \sigma_A^2/n_{Ak} + \sigma_B^2/n_{Bk})$$

The *information* (in essence, same as Fisher's Information) for $\mu_A - \mu_B$ is defined as $\mathcal{I}_k = (\sigma_A^2/n_{Ak} + \sigma_B^2/n_{Bk})^{-1}$, which is the reciprocal of the estimate's variance, and $\mathcal{I}_k$ is used in expressing the standardized statistic (cumulative sample mean) at analysis k:

$$Z_k = (\bar{X}_A^{(k)} - \bar{X}_B^{(k)})\sqrt{\mathcal{I}_k}$$

The vector $(Z_1, \ldots, Z_K)$ is multivariate normal since each Z_k is a linear combination of the independent normal variates X_{Ai} and X_{Bi}, $i = 1, 2, \ldots$, and marginally, $Z_k \sim N(\theta\sqrt{\mathcal{I}_k}, \quad 1)$, $\quad k = 1, \ldots, K$ and $\theta = \mu_A - \mu_B$.

Suppose a group sequential study with up to K analyses yields the sequence of test statistics $\{Z_1, \ldots, Z_K\}$, of which the canonical joint distribution is:
(i) $(Z_1, \ldots, Z_K)$ is multivariate normal,
(ii) $E(Z_k) = \theta\sqrt{\mathcal{I}_k}$,
(iii) $Cov(Z_{k1}, Z_{k2}) = \sqrt{(\mathcal{I}_{k1}/\mathcal{I}_{k2})}, \quad 1 \leq k1 \leq k2 \leq K$

A useful identity:

Let $\bar{x}_k$ denote the sample mean (stage-wise) at stage k, and let $\bar{x}_{k,cum}$ denote the cumulative sample mean up to stage k. We have

$$\bar{x}_{k,cum} = \frac{\sum_{i=1}^{k}(n_i \bar{x}_i)}{\sum_{i=1}^{k} n_i}$$

Let Y_k denote the standardized sample mean $\bar{x}_k$ at stage k, and Z_k denote the standardized cumulative sample mean $\bar{x}_{k,cum}$ at stage k, $k = 1, \ldots, K$. We have $Z_k = \sum_{l=1}^{k} \sqrt{w_l} Y_l$, where $w_l = \frac{n_l}{N}$.

Results related to Brownian motion are useful. It has potentials of simplifying the calculations in elegant ways.

2.1.7 *Comparing with fixed-sample design*

The advantage or disadvantage of a group-sequential designed trial relative to a fixed-sample designed trial depends on the specific class of GSD. In confirmatory clinical trial, the most commonly used is the Wang & Tsiatis class, which, by design, has no early acceptance of the null hypothesis. The Wang & Tsiatis class has advantage over the fixed-sample design only if the actual effect size θ is close to or exceeds the assumed value used for design (which is usually the parameter value postulated by the alternative hypothesis). In practical settings, if the drug development team has little doubt over the assumption of a clinically meaningful effect size that was observed and concluded from earlier studies, the team would prefer a group-sequential designed confirmatory trial. The use of GSD would yield the following benefits over the fixed-sample design: (i) Strong signal (such efficacy in clinical trials) is likely to be detected and claimed earlier. (ii) The ASN is smaller than the sample size required by the fixed-sample design.

Also, for logistical consideration, a GSD is particularly ideal, in fact commonly used, for large-scale multi-years confirmatory trials. Also for the consideration of safety monitoring, a GSD is often preferred in situations where clinicians have some safety concerns over the experimental therapy. In such a situation, safety events of special interest can be examined at the first interim analysis. Usually statistical rules for stopping for efficacy and rules for stopping for safety are independent.

However, if the experimenters have some doubt over the magnitude of the assumed effect size, a fixed-sample design would be preferred. Otherwise, using a GSD would likely end up with a larger sample size. And, in the situation where an unbiased and more precise estimate of the effect size is of paramount importance, a fixed-sample design would be more appropriate. This is because the observed effect size tends to be biased upward if a group-sequential designed trial stops early for efficacy.

2.2 Testing Multiple Hypotheses in Group-Sequential Setting

2.2.1 *Testing of subset intersection hypothesis*

The global testing problem can be generically formulated: testing $H_0 : \cap_{i=1}^{m} H_{0i}$ against $H_1 : \cup_{i=1}^{m} H_{1i}$, where $H_{0i} : \theta_i = 0$ for all $i = 1, \ldots, m$, and $H_{1i} : \theta_i \neq 0$, for some $i = 1, \ldots, m$.

Bonferroni test

Bonferroni method is used across all hypotheses and analysis time points. This is the most conservative test.

A group-sequential Hotelling test

Suppose a group-sequential study with a maximum of K analyses yields a sequence of summary statistics $Y_k, k = 1, 2, \ldots, K$, where Y_k is a column vector of length p (for example, p is the number of endpoints). Consider the testing of the global null hypothesis H_0, with each element of Y_k having a mean zero, against a general alternative. Assume that the underlying data are multivariate normal and that the covariance of Y_k is completely known, so that standardized statistics $Z_k = (Z_{1k}, \ldots, Z_{pk})^T$ can be created. Let the correlation matrix of Z_k be denoted by Σ_k. The test statistics used is $Z_k^T \Sigma_k^{-1} Z_k$

for each $k = 1, \ldots, K$, which has a χ_p^2 distribution under H_0. Jennison & Turnbull (1991, 1997) showed that the sequence of $\{Z_k^T \Sigma_k^{-1} Z_k; k = 1, 2, \ldots, K\}$ is Markov, and they showed how boundary values based on such statistics can be calculated, giving the tests with a pre-specified Type-I error rate α.

A group-sequential O'Brien test

O'Brien (1984) proposed a test of the null hypothesis that a multivariate distribution has mean zero against alternatives under which means of all elements have the same sign. Adaptation of O'Brien (1984) to a group sequential setting has been made by Tang, Gnecco & Geller (1989a) and Tang, Geller & Pocock (1993).

A few other tests that are based on global statistics

Tang et al. (1993) proposes a group sequential test based on the approximate likelihood ratio (AlR) statistics of Tang, Gnecco & Geller (1989b). Nonparametric tests are of value when data depart substantially from normality. Su & Lachin (1992) describe a group sequential test based on a summary rank statistic and Lin (1991) proposes use of a summary statistics based on ranks in a group sequential test for multivariate survival data.

2.2.2 *Procedures for testing multiple hypotheses*

Although rejecting the global null hypothesis would be evidence for the efficacy of the experimental treatment, rejecting single-endpoint hypotheses would make the evidence more specific and compelling. Common to all the MCPs introduced in this section, the trial will continue unless all individual hypotheses are rejected or the final analysis is completed. This is particularly suitable for clinical trials with multiple endpoints, with the clinical objective of claiming 'success' on at least one of the endpoints. The following procedures belong to Union Intersection (UI) test applied in group-sequential setting.

Group-sequential unweighted Bonferroni-based procedures

The job for group sequential (multi-stage) testing of multiple hypotheses is how to appropriately adjust for multiplicity that comes from two sources. Conservative adjustment based on Bonferroni inequality is commonly used. Consider the problem of testing a family of m hypotheses in the setting of a J-stage GSD, with the goal of controlling the familywise Type-I error rate (FWER) at a level α in the strong sense, where the FWER is referred to the probability of making at least one false rejections across all the J analyses and all the m hypotheses, and "in the strong sense" means that the FWER is controlled under any truth/falsity configurations of the m null hypotheses. Per the unweighted Bonferroni method, α is equally divided among the m hypotheses, with α/m allocated to each hypothesis. Then a standard group sequential methodology, such as O'Brien-Fleming is applied to each of the α/m-level tests. This procedure is too conservative if the hypothesis-wise correlation is high.

Group-sequential weighted Bonferroni-based closed test procedures (Maurer & Bretz 2013)

Based on the class of group-sequential unweighted Bonferroni-based procedure, Maurer & Bretz (2013) proposed a weighted version. In our views, they made two important contributions to group-sequential applications of Bonferroni-based tests as the followings. (1) Weighting of individual hypotheses is incorporated, thus providing some flexibility for application. For example, one of the multiple hypotheses is allocated a higher weight to reflect its relative clinical importance. (2) α-reshuffling (α-propagation by some other authors) is incorporated, thus improving power. Note that the term α-reshuffling means that, in the stepdown procedure, upon the rejection of one hypothesis, say H_i, the significance level initially assigned to H_i is transferred to some or all of those unrejected hypotheses, and those unrejected hypotheses will be re-tested at higher significance levels.

Since a group-sequential procedure can be viewed as an extension of its fixed-sample counterpart, it is helpful to review the developmental history of its fixed-sample counterpart. Built on some earlier developments, Bretz et al. (2009, 2011) refined a generalized class of sequentially rejective weighted Bonferroni MTPs via graphical approaches. The generalized class of MTPs dissociates the underlying weighting strategy from the stepdown procedure, thus providing a large degree of flexibility for applications. With their systematic treatment, many seemingly different MTPs, such as fixed sequence, fallback, gatekeeping procedures can be placed into one framework due to their common features.

Consider all nonempty hypotheses $H_J = \cap_{i \in J} H_i, \;\; J \subseteq I = \{1, \ldots, m\}$. For each $J \subseteq I$, define a collection of weights $\{w_i(J) : i \in J, \;\; 0 \leq w_i \leq 1\}$. Reject H_J if $p_i \leq \alpha_i(J) = w_i \alpha$ for at least one $i \in J$. H_i is rejected if all H_J is rejected with $i \in J \subseteq I$ at their corresponding α-level test. A weighting strategy is devised to make monotone condition hold, thus consonance is ensured. The monotone condition is: $w_i(J) \leq w_i(J')$ for all $J' \subseteq J \subseteq I$ and $i \in J'$, where intersection hypothesis H_J implying $H_{J'}$, that is, $H_J \subseteq H_{J'}$. Consonance means that rejection of an intersection hypothesis H_J implies rejection of at least one elementary null hypothesis $H_i, \;\; i \in J$, which in turn allows one to perform the rejective sequential procedure at most m steps.

Ye et al. (2013) proposed a group-sequential weighted Holm's procedure for testing multiple primary endpoints. This can be viewed as a member of the class of group-sequential weighted Bonferroni-based closed testing procedure.

Group-sequential parametric-based closed test procedure (Tang & Geller 1999)

Consider the problem of testing a family of m hypotheses in the setting of a J-stage GSD. Tang & Geller(1999) would first of all choose a standard sequential method for a α-level test as if they were to test a single hypothesis in a group sequential trial. In their example (1999) of the chronic respira-

tory disease, it is the Pocock method that they chose. It is not explicitly mentioned in their paper (1999), but we are able to reconstruct the boundary values using the Pocock method. Essentially, error rate (spending level) is divided among analyses, and such division is made according to a standard univariate sequential method, such as the Pocock method or the O'Brien-Flemming method. Tang & Geller (1999) proposed a procedure that strongly controls the FWER based on the closed testing principle. Specifically, H_i is rejected if all $H_{0,J}$ with $i \in J$ are rejected, where $J \subseteq I = \{1, 2, \ldots, m\}$. At each analysis time point t, a normally distributed "centered linear combination" test statistics $Z_{J,t}$ is used for testing of the null intersection hypothesis $H_{0,J}$, and $H_{0,J}$ is rejected for larger $Z_{J,t}$. Note that $Z_{J,t}$ is a scalar. Consider a group-sequential trial with g analyses at approximately equally-spaced time. Let $\{c_{J,t}, t = 1, 2, \ldots, g\}$ be a one-sided group-sequential boundary for testing $H_{0,J}$ at level α. i.e. $P_{H_{0,J}}\{Z_{J,t} > c_{J,t}; t = 1, 2, \ldots, g\} \leq \alpha$.
The procedure proposed by Tang & Geller (1999) is outlined by the following steps:

Step 1: Conduct interim analysis to test $H_{0,I}$, based on the group-sequential boundary $\{c_{I,t}, t = 1, 2, \ldots, g\}$.

Step 2: When $H_{0,I}$ is rejected at t^*, apply the closed testing procedure to test all the other hypotheses $H_{0,J}$, using Z_{J,t^*}, with c_{J,t^*} as critical boundary.

Step3: If any hypothesis is not rejected, continue the trial to the next stage, in which the closed testing procedure is repeated, with previously rejected hypothesis automatically rejected without retesting.

Step 4: Reiterate Step 3 until all hypotheses are rejected or the last stage is reached.

It is seen that the drawback of the procedure is that there is no short-cut in the closed testing procedure, resulting in potentially too many intersection tests (as many as $2^m - 1$) to be performed.

Group-sequential hierarchical multiple testing procedure

Glimm et al. (2009) and Tamhane et al. (2010) independently considered a 2-stage trial with a primary and a secondary endpoint where the secondary endpoint is tested only if the primary endpoint is significant. The problem is to determine the boundary for the secondary endpoint to maintain strong control of FWER. Test statistics are assumed to be bivariate normal with correlation coefficient ρ. Compared with a α-level secondary boundary obtainable by applying the closure testing principle, the secondary boundary can be made more liberal (having more secondary power) by exploiting the correlation between the primary and the secondary hypotheses.

2.2.3 *Testing of intersection union (IU) hypothesis*

All of the multiple testing procedures aforementioned in *Sec 2.2.2* belong to the union intersection (UI) test. One particular type of multi-endpoints testing problem is the intersection union (IU) test. The testing problem is formulated:

$$H' = \bigcup_{i \in M} H_i \text{ against } K' = \bigcap_{i \in M} K_i$$

The IU test rejects H' at the level α if *all* elementary hypothesis H_i is rejected by their locally α-level test. In the following cases, IU test will be needed to address the clinical research problems.

Case 1: Suppose A and B are 2 existing therapies for treating a particular disease. Investigator is interested in the combination of therapy A and B. The combination therapy will be approved for marketing only if it is both better than A and better than B. The test statistics for the two component hypotheses maybe correlated.

Case 2: An experimental therapy will be approved if it is both efficacious in a primary endpoint and safe in a safety endpoint of major concern. It is also known that the two endpoints are correlated. For example, it is suspected that an experimental hormone therapy for post-menopausal women has a greater benefit of reducing the incidence of breast cancer (primary endpoint) by as much as 40% in a 5-year period, but with some moderate risk of increasing the incidence of venous thrombus (safety endpoint) by about 10%. The indication for breast cancer prevention will be approved only if both the reduction in breast cancer incidence in a 5-year period is over 30% and the increase in venous thrombus incidence is below 10%.

Jennison et al. (1993) developed a method for such testing problem in group-sequential setting as described in the following. Suppose there are p correlated response variables (endpoints), $j = 1, \ldots, p$, associated with the new experimental therapy. After a suitable translation of the parameter θ_j, the set of values of $\theta = \{\theta_1, \ldots, \theta_p\}$ for which new treatment is acceptable (I informally call it 'good outcome') is represented in the form: $\Omega_A = \{\theta : \theta_j > 0 \quad \text{for all} \quad j = 1, \ldots, p\}$. Let Ω_R denote complement of Ω_A in Ω, i.e., $\Omega_R = \Omega \backslash \Omega_A$. After the above setups, it is to test:

$$H_0 : \theta \in \Omega_R \quad \text{(informally the 'bad' region) against}\, H_1 : \theta \in \Omega_A \quad \text{(the 'good' region).}$$

Let $\pi(\theta)$ denote the probability of deciding in favor of the new therapy for a given value of θ. The Type-I error is defined as: the maximum probability of accepting the new therapy when $\theta \in \Omega_R$, and the size of the test is set to α. That is to have the following constraints:

$$\sup_{\theta \in \Omega_R} \{\pi(\theta)\} \leq \alpha$$

Corresponding to H_j, univariate test statistics Z_j is defined. For each endpoint j, a complete set of critical values $(a_{jk},\, b_{jk})$ can be calculated for a univariate one-sided group sequential test. It turns out that if each univariate test is defined to achieve the Type-I error rate α, the constraints above will be met, regardless of the correlation matrix Σ_k. But the power function and ASN depends on Σ_k.

For each endpoint, at analysis k, critical values $(a_{jk},\, b_{jk})$ form a good region $\mathcal{A}_k$ (supporting the new therapy) and a bad region $\mathcal{R}_k$ (opposing the new therapy), and an indifference (continuation) region $\mathcal{C}_k$. New therapy is accepted (which means H_0 is rejected) at stage k if *every* of the observed univariate test statistics Z_{jk} is $\mathcal{A}_k$ at k. Once H_0 is rejected, the trial will stop. New therapy is rejected (H_0 is accepted) at stage k if *any* univariate statistics is $\mathcal{R}_k$ at k. And the trial will stop once H_0 is accepted.

In their paper, an example of $p = 2$ is elaborated. The 2 endpoints are correlated and an covariance structure is estimated, and multivariate calculation is involved in the calculation of power. The continuation region is L-shaped.

One would think of the partition principle in connection with the intersection union (IU) test, and the closure principle in connection with the union intersection (UI) test.

2.3 Multiple Testing Procedures Controlling Familywise Error Rate (FWER)

2.3.1 *Fundamental concepts and principles*

Familywise error rate (FWER)

It is not the same as the *experiment-wise error rate* that was coined by Tukey (1953), though sometimes they are used exchangeably. In statistics literature, the FWER is often implicitly referred to the Type-I error rate. A rigorous definition of the FWER is given by Hochberg & Tamhane (1987) as the followings: Let $\mathbb{F}$ denote a family of inference and let $\mathbb{P}$ denote an MCP for this family. Let $\mathbf{M}(\mathbb{F}, \mathbb{P})$ be the random number of wrong inferences. $\text{FWER}(\mathbb{F}, \mathbb{P}) = Pr\{\mathbf{M}(\mathbb{F}, \mathbb{P}) > 0\}$.

Familywise power

Unlike that of testing single hypothesis, there are a few different concepts of power in the settings of multiple hypotheses testing.

(i) Familywise complete power (FWCP) = Pr(reject *all* H_i that are false).

(ii) Familywise minimal power = Pr(reject *at least one* H_i that is false)

(iii) Proportional power = average *proportion* of false H_i that are rejected.

(iv) Individual power = Pr(reject a *particular* H_i that is false). This is the same as the power used for single hypothesis testing.

Generalised familywise error rate (k-FWER)

It is defined as the probability of having at least k false rejections. $k-FWER = Pr(V > k)$, where V is the number of true nulls that are falsely rejected.

Multiple comparison procedure (MCP) and multiple testing procedure (MTP)

Hochberg & Tamhane (1987) clarify the concept of an MCP. To them, multiple comparison is to assess each comparison (pre-specified or selected by data-snooping) separately by a suitable procedure (a hypothesis test or confidence estimate) at a level deemed appropriate for that single inference. The procedure needs to account for *multiplicity* or selection effect (Tukey 1977). Some other authors (e.g., Lehmann et al. in their book 2005) make distinction of multiple test and multiple comparison. Depending on the context, an MCP can mean a multiple testing method or a simultaneous confidence interval method, or both. Hsu (in his book 1996) takes the narrower definition of MCP, which is concerned with estimating simultaneous confidence interval.

Strength of multiple comparison inference

Different MCPs may address different inferential objectives. The following types of inference are ordered from the strongest to the weakest, per the classification first given by Hsu (1996):

(i) Confidence interval based method. For example, estimating simultaneous confidence interval of a parameter (such as treatment effect) of interest.

(ii) Confidence direction method. For example, assessing the inequalities involving parameters of interest (such as the mean is less for one group than for another).

(iii) Testing-based method. For example, yes/no decision concerning a family of hypotheses of interest. The decision is made on each member hypothesis.

(iv) Tests of homogeneity. For example, testing of the global intersection hypothesis. With the rejection of the global null, little can be inferred regarding the individual hypotheses.

Single-step versus stepwise method

An MCP uses single-step method if all the hypotheses in the family are tested in one step, usually based on only one critical value. In such single-step procedure, one test does not depend on any other tests. Specifically critical value of one test will not be affected by critical values of other tests. Whereas in a stepwise procedure, there exists dependency among the tests. A simultaneous testing procedure is a special class of single-step test procedure for the hierarchical families of hypotheses. It is characterized by a collection of test statistics $Z_i, i \in I$, and a common critical constant ξ such that the procedure rejects H_i if $Z_i > \xi, i \in I$. Under simultaneous testing procedure, for any choice of the critical constant ξ, monotone $\Leftrightarrow$ coherence (Gabriel 1969). Stepwise procedures are categorized into stepdown procedures and step-up procedures. Both procedures assume a sequence of hypotheses, $H_1, \ldots, H_m$ corresponding to the observed p-values in non-decreasing order.

Stepdown procedure starts with testing the hypothesis H_1 corresponding to the most significant (smallest) p-value. Only when H_1 is rejected, the 2nd test (testing the null hypothesis corresponding to 2nd smallest p-value) will be performed. Such steps are repeated until non-rejection occurs, say at H_i, and then $H_1, \ldots, H_{i-1}$ are rejected. Stepup procedures start with testing H_m corresponding to the highest p-value, and step up (in the t-values) through the sequence while retaining the null hypotheses. The procedure stops at the first rejection, say at H_i, and then $H_1, \ldots, H_i$ are all rejected. A stepwise procedure can be seen as a shortcut of a closed testing procedure.

Closure principle

A general method for constructing stepdown procedures was proposed by Marcus et al. (1976). let $\{H_i, i = 1, 2, \ldots, m\}$ be a finite family of hypotheses. Form the closure of this family by taking all non-empty intersection hypotheses $H_P = \cap_{i \in P} H_i$ for $P \subseteq \{1, 2, \ldots, m\}$. If an α-level test of each hypothesis H_P is available, then the closed testing procedure rejects any H_P if and only if every H_Q is rejected by its associated α-level test for all $Q \supseteq P$. The closed testing procedure stated above strongly controls the FWER at level α. A rigorous proof is also given in the book (Hochberg and Tamhane, 1987). There is a total of $2^m - 1$ intersection hypotheses in the closed family, it is desirable to make each of the $2^m - 1$ tests the most powerful. A closed testing procedure is generally required to possess a property of logical consistency called *coherence*. Suppose $H_j \subseteq H_i$. That is, H_i is a component of H_j. If H_j is not rejected then H_i is not rejected; if H_i is rejected, then H_j is rejected. An MCP that satisfies this requirement is called *coherent*. Another desirable (though not required) property for an MCP to have is *consonance*, which is described as the following. If H_j is rejected, at least one of its components, such as H_i, is also rejected. *Consonance* implies that if an intersection hypothesis is rejected, at least one elementary hypothesis is rejected.

Partitioning principle

The partitioning principle was formally introduced by Finner & Strassburger (2002) based on some earlier development of the partitioning ideas. When applying the closure testing procedure, it is difficult to obtain somewhat more informative result such as simultaneous confidence interval. With the application of the partitioning method, one can derive some more powerful MCPs and obtain simultaneous confidence intervals as well. In the general case of m hypotheses $H_1, \ldots, H_m$, the partitioning principle can be implemented as the follows:

(i) Choose an appropriate partition $\{\Theta_l : l \in L\}$ of the parameter space Θ for some index set L.

(ii) Test each Θ_l with an α-level test.

(iii) Reject the null hypothesis H_i if all Θ_l having $\Theta_l \cap H_i \neq \varnothing$ are rejected.

(iv) The union of all retained Θ_l constitute a confidence set for θ at level 1-α.

FWER control: Weakly or Strongly

An MCP controls the FWER weakly if the error control is achieved under the sole configuration of all true nulls, whereas an MCP controls the FWER strongly if the error control is made under all possible configurations of true and false nulls. Controlling the FWER weakly is often not satisfactory because the configuration of all true nulls is not realistic.

2.3.2 *Procedures that control FWER weakly*

(i) Fisher's protected least significance difference (LSD) test.

The LSD test is often used for placebo-controlled clinical trials investigating multiple dosage and/or formulation groups versus one placebo group. Its popularity may have origin in that it is relatively easier to claim a statistically significant result. The LSD test performs multiple t-tests each at level-α

only if the preliminary F-test is significant at α. Assuming one of the mean differences (say, between the highest dosage group and the placebo) is very high, while all other differences are zeros or trivial, the F-test will certainly reject the global nulls, so the subsequent multiple t-tests get conducted at level-α. And obviously, in this case, this procedure inflates the FWER.

(ii) Simes' test (1986).

Simes' procedure rejects $H_0 : \cap_{i=1}^{n} H_i$ if $p_{(i)} < i\alpha/m$ for some $i = 1, \ldots, m$. Simes proved that under independence of the null p-values, the procedure controls the Type-I error rate exactly at the pre-specified level α. Based on simulation, Simes conjectured that the test controls the Type-I error rate when the test statistics are dependent with some specific multivariate distributions. Sarkar & Chang (1997) and Sarkar (1998) provided a theoretical proof for the class of distributions characterized by the MTP_2 property.

In a speculative sense, Simes (1986) suggested a stepup procedure for making inferences for the individual hypotheses: Reject $H_{(1)}, \ldots, H_{(k)}$, with $k = max\{j : p_{(j)} \leq j\alpha/m\}$. However, Hommel (1988) showed that the procedure does not strongly control the FWER even under independent test statistics.

2.3.3 *Procedures that control FWER strongly*

(i) Classical Bonferroni procedure

(ii) Sidak procedure

(iii) Holm's procedure (1979)

Holm's procedure rejects $H_{(1)}, \ldots, H_{(k)}$ where $k = \max\{j \in \{1, \ldots, m\} : p_{(i)} < \frac{\alpha}{(n-i+1)} \text{ for all } i = 1, \ldots, j\}$. If k does not exist, then make no rejection. It's a stepdown procedure that strongly controls the FWER under any dependency structures.

(iv) Hommel's procedure (1988)

It can be seen as a two-step procedure. First find $\hat{n_0} = \max\{k : p_{(m-k+1)} >$

$i\alpha/k, \forall i = 1, \ldots, k\}$. Then reject all H_i's with $p_i \leq \alpha/\hat{n_0}$. If $\hat{n_0}$ does not exist, then reject all H_i. Note that the first step is to estimate the number of true nulls based on Simes' critical values. The second step is to apply classical Bonferroni procedure that splits α equally for the true nulls. Hommel's procedure is proved to be uniformly more powerful than Holm's procedure in the case where Simes' test controls the global type-I error rate. Hommel's procedure maybe the first adaptive procedure.

(v) Hochberg's procedure (1988)

Hochberg's procedure rejects $H_{(1)}, \ldots, H_{(k)}$ where $k = max\{i \in \{1, \ldots, m\} : p_{(i)} < \alpha/(n-i+1)\}$. If k does not exist, then make no rejection. It's a stepup procedure that strongly controls the FWER under independence or positive dependence of the null p-values.

(vi) Rom's procedure (1990)

Hochberg's procedure is conservative in that it controls the FWER at a level slightly lower than the nominal α. Rom devised another stepup procedure similar to Hochberg's procedure, but with slightly more power by making FWER $= \alpha$ exactly under independence of p-values. Rom calculated critical points $c_1, \ldots, c_k$ using a recursive formula (Rom 1990). The decision rule is similar to that of any stepup procedures: if $p_{(k)} \leq c_k$ then all hypotheses are rejected; otherwise, $H_{(k)}$ is retained. Then compare $p_{(k-1)}$ with c_{k-1}. However, the power improvement over Hochberg's procedure is not large (about 2%).

(vii) A hybrid Hochberg-Hommel's stepup procedure (Gou et al., 2014).

This is a new class of procedures that was recently developed from recent efforts in improving Hommel's procedure. Romano et al. (2011) show that a non-consonant procedure can be improved by a more powerful consonant procedure if power is defined as the minimal power. The original Hommel's procedure is non-consonant, thus it can be improved by a consonant procedure. The resulting Hochberg-Hommel's procedure also maintains a simpler step-up structure similar to Hochberg's procedure, in additional to being more powerful than Hommel's. Power advantage is proved in the paper for the case of independency among test statistics, and for the cases of dependency, both positive and negative, power advantage is suggested by simulation results. Again, as that of Hochberg's and Hommel's, this class of hybrid step-up procedures has the convenience of basing on marginal p-values.

(viii) Seneta-Chen's procedure (2005)

This procedure is based on the 2nd order Bonferroni approximation. It involves evaluating some bivariate probability in determining the critical values of the procedure. It is a stepdown procedure as Holm's, but it sharpens critical bounds over Holm's procedure for all values of correlation. And power improvement is considerable for moderate or large correlations.

(ix) Dunnett's single-step procedure, Dunnett's stepdown procedure (1991) and Dunnett's stepup procedure (1992).

These are parametric-based, exact, α-exhaustive procedures.

2.3.4 *Dependencies among null P-values*

P-value is treated as a random variable, since it is a function of random variable (test statistics). One particular dependency structure of the P-values has been studied. It is called positive dependence through stochastic ordering (PDS) condition in some earlier literature (Block et al. 1985). The PDS is same as the concept of positive regression dependence on subset (PRDS) of the null P-values in some recent literature (Benjamini et al 2001; Sarkar 2002). The condition is:

$$\mathbf{E}\{\phi(P_1, \ldots, P_n) \mid P_i = u\} \;\uparrow\; u \in (0,1),$$

for any (coordinatewise) non-decreasing function ϕ of $P_1, \ldots, P_n$, where P_i refers to each of these P-values in case of the PDS and to the null P-values in case of the PRDS.

A weaker condition (Finner et al. 2007; Sarkar 2008) is:

$$\mathbf{E}\{\phi(P_1, \ldots, P_n) \mid P_i \leq u\} \;\uparrow\; u \in (0,1),$$

for any (coordinatewise) non-decreasing function ϕ.

The strong or weaker condition is satisfied by a number of multivariate distributions, such as multivariate normal test statistics with positive correlation, and multivariate t and F that arise from many multiple testing situations.

2.3.5 *Recent development of adaptive FWER-controlling procedures*

FWER-controlling MCPs are often conservative by a factor which is the unknown proportion of true null hypothesis. Conservativeness in these procedures can be reduced and power can be improved if they can be adapted to the data in the sense of estimating the proportion of true null hypotheses π_0 and incorporating that estimate into the procedure.

Benjamini & Hochberg (1990) first presented adaptive procedures for controlling FWER. But it was not proved that their adaptive procedures control FWER.

Guo (2009) proposed a simplified version of Benjamini & Hochberg's adaptive Bonferroni and Holm's procedures, in which π_0 is estimated using the estimator of Storey et al. (2004). He proved that the adaptive Bonferroni procedure controls the FWER in finite samples while the adaptive Holm's procedure controls the FWER approximately.

Finner & Gontscharuk (2009) proposed an adaptive Bonferroni procedure and an adaptive Sidak's procedure using a slight variant of the Storey's estimator.

Recently, Sarkar et al. (2012) proposed a class of adaptive Bonferroni procedures with proven FWER controls under some distributional settings. In particular, they proposed a different adaptive Holm's procedure and its stepup analogue, referred to as an adaptive Hochberg's procedure. The newer adaptive Holm's procedure and the adaptive Hochberg's procedure are seen numerically often outperform those adaptive Holm's procedures that were previously proposed. And, these newer adaptive procedures are proved to control the

FWER asymptotically.

2.4 Multiple Testing Procedures Controlling False Discovery Rate (FDR)

Benjamini and Hochberg (1995) formally introduced the notion of the false discovery rate (FDR) as an overall measure of Type I errors while testing multiple hypotheses. They developed a procedure (the BH procedure, 1995) controlling the FDR. The FDR is suitable for large-scale hypothesis testing, which arises, for instance, in genomic study, whereas the FWER is applicable for small-scale hypothesis testing, which is encountered in confirmatory clinical trials.

The FDR is defined as the expected proportion of false rejections (Type-I errors) among all rejections. $FDR = E(Q) = E(\frac{V}{R}|R > 0)Pr(R > 0)$, where V stands for the number of false rejections (Type-I errors) and R stands for the total number of rejections. Note that $Q = V/R$ if $R > 0$, and $Q = 0$ if $R = 0$.

The pFDR (positive false discovery rate) is defined as the expected proportion of false rejections among all rejections given there is at least one rejection. $pFDR = E(Q) = E(\frac{V}{R}|R > 0)$.

The false non-discovery rate (also termed false negative rate) (Genovese et al. 2002; Sarkar 2004), is defined as $FNR = E\{\frac{T}{A}I(A > 0)\}$, where T stands for the number of false non-rejections (Type-II errors), and A stands for the total number of non-rejections.

The pairwise FDR was introduced and defined by Sarkar (2008).

2.4.1 *Two approaches for constructing FDR-controlling procedures*

The fixed level approach: An MCP determines a fixed rejection region for a given level of significance. Representative of this approach is the first FDR-controlling procedure, the BH procedure (1995).

The fixed rejection region approach: the FDR is suitably estimated for a pre-specified fixed rejection region. Representative of this approach is Storey's procedure (2002).

2.4.2 *Single-step and stepwise FDR-controlling procedures*

Analogous to that of FWER-controlling procedures, FDR-controlling procedures can be performed using single-step procedures and multi-step procedures.

(i) A stepup procedure (method). Let $P_{(1)} \leq \cdots \leq P_{(n)}$ be the ordered P-values, and $H_{(1)}, \ldots H_{(n)}$ be the null hypotheses corresponding to these P-values. Then given a non-decreasing set of critical constants $0 < t_1 \leq \cdots \leq t_n < 1$, a stepup method rejects the set $\{H_{(i)}, i < i^*_{SU}\}$ and accepts the remaining, where $i^*_{SU} = \max\{1 \leq i \leq n : P_{(i)} \leq t_i\}$.

(ii) A stepdown procedure (method). A stepdown method rejects the set $\{H_{(i)}, i < i^*_{SD}\}$ and accepts the remaining, where $i^*_{SD} = \max\{1 \leq i \leq n : P_{(j)} \leq t_j, \forall j \leq i\}$.

(iii) A single-step procedure (method). In the above (i) or (ii), if the critical constants $t_i = t$, $\forall i = n$, a stepwise procedure reduces to a single-step procedure.

2.4.3 *FDR-controlling procedures*

Non-adaptive procedures

(i) The BH procedure (Benjamini & Hochberg, 1995). It is the first FDR-controlling procedure developed in the field. It is the stepup procedure as defined above, using the critical values $t_i = i\alpha/n, i = 1, \ldots, n$. Based on the same critical values, its stepdown analogue was developed by Sarkar (2002). When the P-values are independent, the FDR is controlled exactly at $n_0\alpha/n$ for the stepup procedure, and at $\leq n_0\alpha/n$ for the stepdown analogue, where n_0 is the number of true nulls; When the P-values are of PRDS, the FDR is controlled at $\leq n_0\alpha/n$ for both stepup and stepdown procedures.

(ii) The BY procedure (Benjamini & Yekutieli, 2001). It controls the FDR strongly under *any* dependency structure of P-values. It is a stepup procedure using the critical values $\alpha_i = i\alpha/(n\sum_{j=1}^{n} 1/j), i = 1, \ldots, n$. Another stepup procedure (Sarkar, 2008) that works under any dependency structure was developed using the critical values $\alpha_i = i(i+1)\alpha/2n^2, i = 1, \ldots, n$. The BY procedure and Sarkar's procedure are both too conservative because it is seen that the critical values tend to 0 as n increases.

(iii) Some stepdown procedures that strongly control the FDR under a variety of dependency structure are given in: Benjamini & Liu (1999), Romano & Shaikh (2006), Gavrilov-Benjamini-Sarkar (2009), and Blanchard & Roquain (2008).

Some adaptive procedures

FDR-controlling procedures can be improved in terms of power, by incorporating into it an appropriate estimate of the number n_0 or proportion π_0 of the true nulls. Once an appropriate estimate $\hat{n_0}$ is obtained, it is incorporated in the critical values by replacing n by this estimate. There are a few established procedures, such as (i) Adaptive BH procedure (2006), (ii) Storey's procedure (2004), (iii) Adaptive stepdown procedure Gavrilov-Benjamini-Sarkar (2008), that controls the FDR under independence of the P-values. However, it is still an open problem as to whether it controls the FDR when the independence condition does not hold.

CHAPTER 3

IMPROVING HOLM'S PROCEDURE BY INCORPORATING PAIRWISE DEPENDENCIES

3.1 Summary

Seneta and Chen (2005) tightened the familywise error rate (FWER) control of Holm's procedure by sharpening its critical values utilizing pairwise dependencies of the p-values. We propose further sharpening of these critical values when the distribution functions of the pairwise maximums of null p-values are convex, a property shown to hold in many standard applications of Holm's method. The newer critical values are uniformly larger providing tighter familywise error rate control than Seneta-Chen's, significantly so under high pairwise positive dependencies, as numerically seen. They are further improved under exchangeable null p-values with modest additional computation.

3.2 Introduction

Controlling the familywise error rate, the probability of falsely rejecting at least one true null hypothesis, is commonly undertaken when testing multiple hypotheses. Among several familywise error rate controlling methods available in the literature, Holm's (1979) is one of the most popular. Seneta and Chen (2005) attempted to improve it in situations where pairwise dependencies among the p-values can be quantified. They applied Kounias (1968) inequality that gives tighter upper bound for the distribution function of the minimum of a set of null p-values than Bonferroni's while modifying the Holm's critical values. The modification tightens the familywise error rate control of Holm's method, and can even be more powerful than Hochberg's original step-up method as they numerically showed for some multiple testing problems associated with normally distributed test statistics with known correlations.

We propose two different versions of improved Holm's step-down method, depending on whether the null p-values are exchangeable or not, for p-values with the pairwise maximums of the null p-values having known convex distribution functions, each providing uniformly larger critical values than Seneta-Chen's. The desired convexity property of p-values is shown for many multivariate distributions. Extensive numerical and simulation studies reveal that our proposed procedures are better choices than Seneta-Chen's for improving Holm's method, especially when there are high pairwise dependencies among the test statistics.

3.3 Seneta-Chen's Modified Holm's Procedure

Let $P_{(1)} \leq \cdots \leq P_{(n)}$ be the ordered versions of p-values available for testing n null hypotheses, with $H_{(i)}$ being the null hypothesis corresponding to $P_{(i)}, i = 1, \ldots, n$. Let each null p-value be distributed as $U(0, 1)$. Then, Holm's method controlling the familywise error rate at a pre-specified level α is a step-down procedure with the critical values $\alpha_i = \alpha/(n-i+1)$, $i = 1, \ldots, n$, that is, it rejects $H_{(i)}$ for all $i \leq R = \max\{i : P_{(j)} \leq \alpha_j = \alpha/(n-j+1) \text{ for all } j \leq i\}$, provided the maximum exists; otherwise, it rejects none.

With n_0 (≥ 1) true null hypotheses, the familywise error rate (FWER) of a step-down procedure with any critical values $\alpha_1 \leq \cdots \leq \alpha_n$ satisfies the following inequality:

$$\text{FWER} \leq \max_{I_{n_0} \in C_{n_0}} pr \left(\min_{j \in I_{n_0}} P_j \leq \alpha_{n-n_0+1} \right),$$

where I_{n_0} is the set of indices of the n_0 true null hypotheses and C_{n_0} is the collection of all such sets. Hence, the α_i's providing the familywise error rate control by this method at level α can be determined by finding α_{n-n_0+1}, for each $1 \leq n_0 \leq n$, such that $\max_{I_{n_0} \in C_{n_0}} pr \left(\min_{j \in I_{n_0}} P_j \leq \alpha_{n-n_0+1} \right) \leq \alpha$. The Bonferroni inequality gives $pr \left(\min_{j \in I_{n_0}} P_j \leq \alpha_{n-n_0+1} \right) \leq n_0 \alpha_{n-n_0+1}$, the right-hand side of which, when bounded from above by α, yields $\alpha_{n-n_0+1} = \alpha/n_0$ for $n_0 = 1, \ldots, n$, that is, $\alpha_i = \alpha/(n-i+1)$ for $i = 1, \ldots, n$. These are Holm's original critical values.

Seneta and Chen (2005) sharpened these critical values by using the following inequality due to Kounias (1968) in terms of the pairwise distributions of the p-values:

$$\max_{I_{n_0} \in C_{n_0}} pr \left(\min_{j \in I_{n_0}} P_j \leq \alpha_{n-n_0+1} \right) \leq n_0 \alpha_{n-n_0+1} - (n_0 - 1) \beta_{n_0}(\alpha_{n-n_0+1}),$$

where

$$\beta_{n_0}(\alpha_{n-n_0+1}) = \frac{1}{n_0 - 1} \min_{I_{n_0} \in C_{n_0}} \max_{j \in I_{n_0}} \left\{ \sum_{k(\neq j) \in I_{n_0}} pr \left(P_j \leq \alpha_{n-n_0+1}, \, P_k \leq \alpha_{n-n_0+1} \right) \right\}$$

for $n_0 = 2, \ldots, n$. Let

$$G_{n_0}(\alpha_{n-n_0+1}) = n_0 \alpha_{n-n_0+1} - (n_0 - 1) \beta_{n_0}(\alpha_{n-n_0+1}).$$

They suggested using $\alpha_{n-n_0+1} = \left\{ \alpha + (n_0 - 1) \beta_{n_0}(\alpha/n_0) \right\} / n_0$ as a solution to the inequality $G_{n_0}(\alpha_{n-n_0+1}) \leq \alpha$, tighter than Bonferroni's, for each $n_0 = 1, \ldots, n$, before proposing their modified version of Holm's critical values maintaining the non-decreasing property as follows:

$$\alpha_i = \min \left\{ \frac{\alpha + (n-i) \beta_{n-i+1}(\frac{\alpha}{n-i+1})}{n-i+1}, \frac{\alpha}{n-i} \right\}, \, i = 1, \ldots, n. \qquad (3.1)$$

3.4 Proposed Modifications of Holm's Procedure

We find solutions of the form $u = c\alpha/n_0$ to the inequality $G_{n_0}(u) \leq \alpha$ that are better, i.e., larger, than Seneta-Chen's solution, in which $c = \{\alpha + (n_0 - 1)\beta_{n_0}(\alpha/n_0)\}/\alpha$, under the following assumption:

Assumption 1. *The probability pr* $\{\max(P_j, P_k) \leq u\}$ *is convex in* $u \in (0,1)$, *for each* $j, k \in I_{n_0}$ *such that* $j \neq k$.

The assumption ensures the convexity of $\beta_{n_0}(u)$, for each fixed $n_0 \geq 2$, since the sum as well as the maximum of multiple convex functions is also convex, and so the concavity of $G_{n_0}(u)$ greatly facilitates finding the desired c.

We propose two types of Holm's modification under Assumption 1, one without any additional condition on the null p-values and the other assuming that the null p-values are also exchangeable.

3.4.1 *Non-exchangeable null p-values*

The concavity of $G_{n_0}(u)$ in $u \in (0,1)$ along with $G_{n_0}(0) = 0$ yield $G_{n_0}(c\alpha/n_0) \leq cG_{n_0}(\alpha/n_0)$, for all $c \geq 1$, and hence $c = \alpha/G_{n_0}(\alpha/n_0)$ gives us a solution to the inequality $G_{n_0}(c\alpha/n_0) \leq \alpha$. This c equals $\alpha/\{\alpha - (n_0-1)\beta_{n_0}(\alpha/n_0)\}$, which is clearly larger than $\{\alpha + (n_0-1)\beta_{n_0}(\alpha/n_0)\}/\alpha$, as desired. Thus, our proposed solution to the inequality $G_{n_0}(\alpha_{n-n_0+1}) \leq \alpha$ is $\alpha_{n-n_0+1} = \alpha^2/\{n_0 G_{n_0}(\alpha/n_0)\}$, which leads to the following theorem as one of our main results in this paper.

Theorem 1. *Let*

$$\widetilde{\alpha}_i = \frac{\alpha/(n-i+1)}{G_{n-i+1}\{\alpha/(n-i+1)\}}\alpha, \tag{3.2}$$

$i = 1, \ldots, n$. *Then, the step-down procedure based on the critical values* α'_i, $i = 1, \ldots, n$, *where* $\alpha'_i = \min\{\widetilde{\alpha}_i, \alpha'_{i+1}\}$ *for* $i = 1, \ldots, n-1$, *and* $= \widetilde{\alpha}_n$ *for* $i = n$, *provides tighter control of the familywise error rate than that based on Seneta-Chen's proposed modification given in (3.1) under Assumption 1.*

The procedure in Theorem 1 is one of our proposed modifications of Holm's procedure. Although $\widetilde{\alpha}_i > \alpha/(n-i+1)$, it may not be non-decreasing in i, and hence we could not consider $\widetilde{\alpha}_i$, $i = 1, \ldots, n$, themselves as the critical values in our proposed step-down procedure. We had to modify them as in this theorem to bring the non-decreasing property into them. However, this is not going to be an issue if the null p-values are exchangeable, since in this case $\widetilde{\alpha}_i$ is non-decreasing in i, as will be shown in Section 3.8. Thus, we have the following:

Remark 1. With exchangeable null p-values having a common known distribution function H of the pairwise maximums, the step-down procedure using the critical values in (3.2) with $G_{n_0}(u) = n_0 u - (n_0 - 1)H(u)$ is proposed as our improved Holm's procedure, instead of the one in Theorem 1, under Assumption 1.

In fact, under the exchangeability of the null p-values, we can obtain a better, i.e., larger, solution of the form $c\alpha/n_0$, $c \geq 1$, to the inequality $G_{n_0}(\alpha_{n-n_0+1}) \leq \alpha$ than what we consider in constructing the step-down method in Remark 1. This is the topic of the next sub-section.

3.4.2 *Exchangeable null p-values*

Using the Taylor expansion of $G_{n_0}(c\alpha/n_0)$ around $c = 1$, we get

$$\begin{aligned} G_{n_0}(c\alpha/n_0) &\leq G_{n_0}(\alpha/n_0) + (c-1)G'_{n_0}(\alpha/n_0)\alpha/n_0 \\ &= \alpha - (n_0 - 1)H(\alpha/n_0) + (c-1)\{n_0 - (n_0-1)h(\alpha/n_0)\}\alpha/n_0, \end{aligned} \tag{3.3}$$

for any $c \geq 1$, where G'_{n_0} and h are the derivative of G_{n_0} and the density of H, respectively. Equating the right-hand side of (3.3) to α and solving the resulting equation in c, we then obtain the following solution to $G_{n_0}(\alpha_{n-n_0+1}) \leq \alpha$:

$$\alpha_{n-n_0+1} = \frac{\alpha}{n_0} + \frac{\frac{n_0-1}{n_0}H(\alpha/n_0)}{1 - \frac{n_0-1}{n_0}h(\alpha/n_0)}. \tag{3.4}$$

Since $H(u)$ is convex in $u \in (0,1)$ and $H(0) = 0$, we have $H(u) \leq uh(u)$ for all $u \in (0,1)$, and hence the solution in (3.4) is greater than or equal to

$$\frac{\alpha}{n_0} + \frac{\frac{n_0-1}{n_0}H(\alpha/n_0)}{1 - \frac{n_0-1}{\alpha}H(\alpha/n_0)} = \frac{\alpha}{\alpha - (n_0-1)H(\alpha/n_0)}\frac{\alpha}{n_0},$$

which is the solution in (3.2), as we have said following Remark 1. Moreover, as we show in Section 3.8, the critical values

$$\alpha_i^* = \frac{\alpha}{n-i+1} + \frac{\frac{n-i}{n-i+1}H(\alpha/n-i+1)}{1 - \frac{n-i}{n-i+1}h(\alpha/n-i+1)}, \; i = 1, \ldots, n, \tag{3.5}$$

as suggested by the solution in (3.4) are increasing in i if the chosen α satisfies the following:

Condition 1. *The α is such that $h(\alpha) \leq 1$.*

Thus, we have our next main result in the following:

Theorem 2. *Let the null p-values be exchangeable with their pairwise maximums having the common distribution function H with the density h. Then, the step-down procedure based on the modified Holm's critical values in (3.5) provides tighter control of the familywise error rate than that mentioned in Remark 1 under Assumption 1 if Condition 1 holds.*

Remark 2. We show in the next section that Condition 1 holds for commonly chosen values of α in many multiple testing problems. However, if this condition is not satisfied, the critical values of the procedure in Theorem 2 can be modified as in Theorem 1 to ensure the monotonicity.

3.5 Examples

Suppose that the p-values are generated from some continuous test statistics $X_1, \ldots, X_n$ through their common marginal null distribution function F. Let $P_i = 1 - F(X_i)$, that is, we have right-tailed test based on each X_i. Then, with $x(t) = F^{-1}(1-t)$, we have $H_{jk}(t) = pr\left\{\max(P_j, P_k) \leq t\right\} =$

$pr\{X_j \geq x(t), X_k \geq x(t)\}$, and so the density $h_{jk}(t)$, which is the derivative of $H_{jk}(t)$, of $\max(P_j, P_k)$ is given by, with f being the density of F,

$$\begin{aligned} & -pr\{X_j \geq x(t) \mid X_k = x(t)\} f\{x(t)\}x'(t) - pr\{X_k \geq x(t) \mid X_j = x(t)\} f\{x(t)\}x'(t) \\ = \; & pr\{X_j \geq x(t) \mid X_k = x(t)\} + pr\{X_k \geq x(t) \mid X_j = x(t)\}, \end{aligned}$$

since $f\{x(t)\}x'(t) = -1$. Therefore, the underlying convexity condition holds for such p-values when the X_i's have the multivariate distribution that satisfies the following property:

Property 1. *The conditional probability* $pr\{X_j \geq x \mid X_k = x\}$ *is decreasing in* x *under the joint null distribution of* (X_j, X_k)*, for any* $1 \leq j < k \leq n$.

Remark 3. If the p-values correspond to left-tailed tests based on the X_i's, that is, if $P_i = F(X_i)$, it is easy to see that Assumption 1 still holds under Property 1.

The following property provides an assurance for Condition 1 to hold in case of exchangeable null test statistics with the common density of the pairwise maximums of the null p-values being given by

$$h(t) = 2\,pr\left\{X_1 \geq F^{-1}(1-t) \mid X_2 = F^{-1}(1-t)\right\},$$

where (X_1, X_2) is any pair of these statistics.

Property 2. *There exists an* $\alpha_0 \in (0,1)$ *such that* $h(\alpha) \leq 1$ *for all* $0 < \alpha \leq \alpha_0$.

We now give examples of multivariate distributions arising in some standard multiple testing problems exhibiting Property 1 for Assumption 1 to hold and Property 2, with some α_0 for the typically chosen values of α, to satisfy Condition 1 in the exchangeable case.

3.5.1 *Example 1. Multivariate and absolute-valued multivariate normals*

Let $X_1, \ldots, X_n$ be test statistics jointly distributed as multivariate normal with $E(X_i) = \mu_i$, $\text{var}(X_i) = 1$, and $\text{corr}(X_i, X_j) = \rho_{ij}$, and are available for testing $\mu_i = 0$, $i = 1, \ldots, n$.

Since $X_j \mid X_k = x \sim N(\rho_{jk}x, 1-\rho_{jk}^2)$, we have

$$pr\left\{X_j \geq x \mid X_k = x\right\} = 1 - \Phi\left\{x\left(\frac{1-\rho_{jk}}{1+\rho_{jk}}\right)^{\frac{1}{2}}\right\},$$

where Φ is the cdf of $N(0,1)$. This is decreasing in x, and hence Property 1 holds when testing $\mu_i = 0$ against $\mu_i > 0$ simultaneously for $i = 1, \ldots, n$. Property 2, which is in the exchangeable case with $\rho_{jk} = \rho$, also holds with $\alpha_0 = 1/2$, since $x = \Phi^{-1}(1-\alpha) \geq 0$, for $0 < \alpha \leq 1/2$, making the above conditional probability less than or equal to $1/2$ as desired.

In terms of the $|X_i|$'s, we see that

$$pr\left(|X_j| \geq x \;\middle|\; |X_k| = x\right) = 2 - \Phi\left\{x\left(\frac{1-\rho_{jk}}{1+\rho_{jk}}\right)^{\frac{1}{2}}\right\} - \Phi\left\{x\left(\frac{1+\rho_{jk}}{1-\rho_{jk}}\right)^{\frac{1}{2}}\right\}, \tag{3.6}$$

which is decreasing in $x \geq 0$. Hence, Property 1 holds when testing $\mu_i = 0$ against $\mu_i \neq 0$ simultaneously for $i = 1, \ldots, n$. If $x \geq 1$, the right-hand side of (3.6) is less than or equal to

$$2 - \Phi\left\{\left(\frac{1-\rho_{jk}}{1+\rho_{jk}}\right)^{\frac{1}{2}}\right\} - \Phi\left\{\left(\frac{1+\rho_{jk}}{1-\rho_{jk}}\right)^{\frac{1}{2}}\right\},$$

which is increasing in $|\rho_{jk}|$, as shown in Section 3.8, and so it is less than or equal to $2 - \Phi(\infty) - \Phi(0) = 1/2$. Since here $x = \Phi^{-1}(1-\alpha/2)$, Property 2, which is in the exchangeable case with $\rho_{jk} = \rho$, is also seen to hold with $\alpha_0 = 2\{1 - \Phi(1)\} \approx 0.3173$.

3.5.2 *Example 2. Multivariate and absolute-valued multivariate t's*

Consider the same testing problems as in Example 1 based on t test statistics $T_i = \nu^{\frac{1}{2}}X_i/Y^{\frac{1}{2}}$, where the X_i's are distributed as in that example, but now with $\text{var}(X_i) = \sigma^2$, $i = 1, \ldots, n$, for some unknown σ^2, and Y is distributed independently of the X_i's as $\sigma^2\chi^2_\nu$.

Since the joint null distribution of the pair (T_j, T_k) is central bivariate t with ν degrees of freedom and the associated correlation ρ_{jk}, we have from the result that $(\nu+1)^{\frac{1}{2}}(T_j - \rho_{jk}x)/\left\{(\nu+x^2)(1-\rho_{jk}^2)\right\}^{\frac{1}{2}} \sim t_{\nu+1}$, conditional on $T_k = x$, (see, for instance, Kotz and Nadarajah, 2004),

$$pr\,(T_j \geq x \mid T_k = x) = 1 - \Psi_{\nu+1}\left[x\left\{\frac{(\nu+1)(1-\rho_{jk})}{(\nu+x^2)(1+\rho_{jk})}\right\}^{\frac{1}{2}}\right],$$

where $\Psi_{\nu+1}$ is the cumulative distribution function of $t_{\nu+1}$, the central t with $\nu+1$ degrees of freedom. This is decreasing in x, and hence Property 1 holds when testing $\mu_i = 0$ against $\mu_i > 0$ simultaneously for $i = 1, \ldots, n$. Property 2, which is in the exchangeable case with $\rho_{jk} = \rho$ also holds in this context with $\alpha_0 = 1/2$, since in this case $x = \Psi_\nu^{-1}(1-\alpha) \geq 0$, for $0 < \alpha \leq 1/2$, making the above conditional probability less than or equal to 1/2.

In terms of the $|T_i|$'s, we see that

$$\begin{aligned} & pr\left(|T_j| \geq x \;\middle|\; |T_k| = x\right) \\ = \;\; & 2 - \Psi_{\nu+1}\left[x\left\{\frac{(\nu+1)(1-\rho_{jk})}{(\nu+x^2)(1+\rho_{jk})}\right\}^{\frac{1}{2}}\right] - \Psi_{\nu+1}\left[x\left\{\frac{(\nu+1)(1+\rho_{jk})}{(\nu+x^2)(1-\rho_{jk})}\right\}^{\frac{1}{2}}\right], \end{aligned} \tag{3.7}$$

which is decreasing in $x \geq 0$. Hence, Property 1 holds when testing against $\mu_i = 0$ against $\mu_i \neq 0$ simultaneously for $i = 1, \ldots, n$. If $x \geq 1$, the right-hand side of (3.7) is less than or equal to

$$2 - \Psi_{\nu+1}\left\{\left(\frac{1-\rho_{jk}}{1+\rho_{jk}}\right)^{\frac{1}{2}}\right\} - \Psi_{\nu+1}\left\{\left(\frac{1+\rho_{jk}}{1-\rho_{jk}}\right)^{\frac{1}{2}}\right\},$$

which is increasing in $|\rho_{jk}|$, as will be shown in Section 3.8, and hence less than or equal to $2 - \Psi_{\nu+1}(\infty) - \Psi_{\nu+1}(0) = 1/2$, as desired for Property 2 to hold in the exchangeable case. Here, $x = \Psi_\nu^{-1}(1-\alpha/2)$, and so Property 2 is seen to hold if $\alpha \leq 2\{1 - \Psi_\nu(1)\}$. Since $1 - \Psi_\nu(1) \geq 1 - \Phi(1)$, again the desired $\alpha_0 = 0.3173$.

Remark 4. *The ranges of α-values guaranteeing Property 2 in Example 2 are shown to contain typically used α in practice, but it should be noted that they can be widened.*

3.6 Numerical and Simulation Investigations

Numerical and simulation studies were conducted in the exchangeable case to investigate the extent of improvements that our proposed methods can offer over Seneta-Chen's and Hochberg's (Hochberg, 1988). The studies considered $n = 8$ two-sided tests at $\alpha = 0.05$ in the setting of Example 1, applied the method described in Seneta and Chen (2005) to generate the data, and used 2 million independent replications in all simulations. We make the comparisons, as in Seneta and Chen (2005), in terms of Type I error rate, which is the familywise error rate in the weak sense, and minimal power, which is the probability of rejecting at least one false nulls.

Table 3.1: Selected critical values in the order of Seneta-Chen's, Procedure 1 and Procedure 2

i	2	5	8
α/i	0.025	0.01	0.00625
$\rho = 0.0$	(0.02531, 0.02532, 0.02532)	(0.01008, 0.01008, 0.01008)	(0.00628, 0.00628, 0.00628)
0.3	(0.02580, 0.02583, 0.02584)	(0.01030, 0.01031, 0.01031)	(0.00641, 0.00641, 0.00641)
0.5	(0.02676, 0.02690, 0.02695)	(0.01079, 0.01086, 0.01089)	(0.00670, 0.00674, 0.00675)
0.7	(0.02851, 0.02909, 0.02925)	(0.01182, 0.01223, 0.01235)	(0.00737, 0.00762, 0.00769)
0.9	(0.03192, 0.03457, 0.03494)	(0.01406, 0.01685, 0.01731)	(0.00891, 0.01089, 0.01122)
1.0	(0.03750, 0.05000, 0.05000)	(0.01800, 0.05000, 0.05000)	(0.01172, 0.05000, 0.05000)

Table 3.2: Simulated Type I error rate and minimal power for $n = 8$ and $\alpha = 0.05$

	$\delta = 0$			$\delta = 1$			$\delta = 2$		
Procedure \ ρ	0.5	0.7	0.9	0.5	0.7	0.9	0.5	0.7	0.9
Hochberg	0.0401	0.0312	0.0209	0.1902	0.1460	0.1014	0.6278	0.5220	0.4056
Seneta-Chen	0.0427	0.0360	0.0257	0.1972	0.1592	0.1149	0.6368	0.5433	0.4317
Procedure 1	0.0429	0.0371	0.0307	0.1978	0.1622	0.1293	0.6376	0.5481	0.4599
Procedure 2	0.0430	0.0374	0.0315	0.1980	0.1630	0.1316	0.6380	0.5494	0.4642

Table 3.1 shows that the critical values of the proposed procedures are always larger than Seneta-Chen's and Holm's as expected, but their differences become more and more significant when the correlation becomes large. Also, as seen from Table 3.2, the Type I error rates of the proposed procedures, $\delta = 0$, are closer to the pre-specified level 0.05 than those of Seneta-Chen and Holm, especially when the correlation ρ is large. In terms of the power, $\delta = 1$ and 2, the proposed procedures are always more powerful than Seneta-Chen and Hochberg's, with the power difference becoming much more evident when ρ is large.

3.7 Concluding Remarks

Hochberg's procedure, the step-up analog of Holm's, is a valid familywise error rate controlling procedure under the distributional settings in Section 3.4 (see, for instance, Sarkar and Chang, 1997). However, the idea of improving it incorporating pairwise dependencies by using the step-up analogs of the proposed step-down procedures will not work. This can be seen from the $n = 2$ case, where $\alpha_1 = \alpha^2/[2\{\alpha - H(\alpha/2)\}]$ and $\alpha_2 = \alpha$ are the critical values that don't control the familywise error rate at α when the p-values are independent, since $\alpha_1 > \alpha/2$. The proposed methods incorporating pairwise dependencies, even though they are step-down, provide reasonably good alternatives to Hochberg's, as seem to be suggested by Table 2.

Developing familywise error rate controlling procedures under some commonly used distributional models effectively utilizing dependencies among the test statistics through their pairwise distributions, rather than the joint distribution that can be extremely difficult to utilize with large number of tests, as seen in the procedure of Dunnett and Tamhane (1992), and allowing explicit formulation of the critical values has been the primary goal of this paper. We have achieved this goal in terms of step-down procedures, having noted above its failure in terms of the corresponding step-up ones.

The examples of bivariate distributions having the desired convexity property given in Section 3.4 are those that are commonly seen in applications of Holm's procedure. The same property is seen in many other bivariate distributions, including certain families of location- and scale- mixture distributions, e.g., bivariate Gamma and F, that appear in multiple testing contexts (Sarkar and Chang, 1997), and family of distributions corresponding to Archimedean copulas (Nelson, 2007) used in other contexts, such as in actuarial science and finance. The convexity property for these other bivariate distributions are shown in the Supplementary Material.

Estimating G or H in general is beyond the scope of the paper, but we demonstrate in the Supplementary Material how in practice one could do so and check the desired convexity and other properties before calculating the α_i's.

3.8 Proofs of Some Results

Result. The critical values $\widetilde{\alpha}_i$ in (3.2) are increasing in i under Assumption 1 if the null p-values are exchangeable.

Proof. The result follows from the fact that $\alpha/\{mG_m(\alpha/m)\}$, where $G_m(u) = mu - (m-1)H(u)$, is decreasing in $m = 1, \ldots, n$, since (i) $G_m(u)$ is increasing in m (for fixed u) because $G_{m+1}(u) - G_m(u) = u - H(u) \geq 0$, and (ii) $G_m(u)/u$ is decreasing in $u \in (0,1)$ (for fixed m) since $G_m(u)$ is concave in $u \in (0,1)$ and $G_m(0) = 0$. □

Result. The critical values α_i^* in (3.5) are increasing in $i = 1, \ldots, n$, if Condition 1 holds.

Proof. The function $\alpha u + (1-u)H(\alpha u)/\{1-(1-u)h(\alpha u)\}$ is increasing in $u \in (0,1)$, since its derivative, $\{1-(1-u)h(\alpha u)\}^{-2}\,[\alpha\{1-h(\alpha u)\} + \alpha u h(\alpha u) - H(\alpha u) + \alpha(1-u)^2 h'(\alpha u)H(\alpha u)]$, is non-negative due to the following: (i) $h'(\alpha u) \geq 0$ since $H(u)$ is convex in $u \in (0,1)$; (ii) $H(\alpha u) \leq \alpha u h(\alpha u)$ since

$H(u) \geq 0$ is convex in $u \in (0,1)$ and $H(0) = 0$; and (iii) $h(\alpha u) \leq h(\alpha) \leq 1$ since $h(u)$ is increasing in $u \in (0,1)$. Thus the result follows. □

Result. The following function

$$\Phi\left\{\left(\frac{1+\rho}{1-\rho}\right)^{\frac{1}{2}}\right\} + \Phi\left\{\left(\frac{1-\rho}{1+\rho}\right)^{\frac{1}{2}}\right\} \tag{3.8}$$

is decreasing in $|\rho| \in [0,1)$.

Proof. Let's assume without any loss of generality that $\rho \geq 0$. The first derivative of the function in (3.8) with respect to ρ is $(2\pi)^{-\frac{1}{2}}(1-\rho^2)^{-1}$ times

$$\left(\frac{1+\rho}{1-\rho}\right)^{\frac{1}{2}} \exp\left\{-\frac{1+\rho}{2(1-\rho)}\right\} - \left(\frac{1-\rho}{1+\rho}\right)^{\frac{1}{2}} \exp\left\{-\frac{1-\rho}{2(1+\rho)}\right\}.$$

This is non-positive since $\log\{(1-\rho)/(1+\rho)\} + 2\rho/(1-\rho^2)$ is increasing in $\rho \in [0,1)$ and hence ≥ 0. Thus, the result follows. □

Result. The function

$$\Psi_{\nu+1}\left\{\left(\frac{1+\rho}{1-\rho}\right)^{\frac{1}{2}}\right\} + \Psi_{\nu+1}\left\{\left(\frac{1-\rho}{1+\rho}\right)^{\frac{1}{2}}\right\} \tag{3.9}$$

is decreasing in $|\rho| \in [0,1)$.

Proof. Assuming without any loss of generality that $\rho \geq 0$, we see that the first derivative of the function in (3.9) with respect to ρ is $\Gamma\{(\nu+2)/2\}/\,\Gamma\{(\nu+1)/2\}\,\{(\nu+1)\pi\}^{-\frac{1}{2}}\,(1-\rho^2)^{-1}$ times

$$\left(\frac{1+\rho}{1-\rho}\right)^{\frac{1}{2}} \left\{1 + \frac{1+\rho}{(\nu+1)(1-\rho)}\right\}^{-\frac{\nu+2}{2}} - \left(\frac{1-\rho}{1+\rho}\right)^{\frac{1}{2}} \left\{1 + \frac{1-\rho}{(\nu+1)(1+\rho)}\right\}^{-\frac{\nu+2}{2}}.$$

This is non-positive, which can be checked from the result that $(\nu+2)\log\{(\nu+2+\nu\rho)/(\nu+2-\nu\rho)\} - \nu\log\{(1+\rho)/(1-\rho)\}$ is decreasing in $\rho \in [0,1)$ and hence ≤ 0. Thus, the result is proved. □

3.9 Supplementary Material

3.9.1 *Other distributions satisfying assumption 1*

The following are examples of distributions other than those given in Section 3.4 for the underlying null test statistics or p-values for which Assumption 1 holds.

1. Multivariate Gamma. Let $X_i = Y_0 + Y_i$, $i = 1, \ldots, n$, where Y_i, $i = 0, 1, \ldots, n$, are independent with $Y_0 \sim Gamma(\alpha_0, \beta)$, where $\alpha_0 \geq 1$, and $Y_i \sim Gamma(\alpha, \beta)$ for $i = 1, \ldots, n$.

2. Multivariate F. Let $X_i = Y_i/Y_0$, $i = 1, \ldots, n$, where Y_i, $i = 0, 1, \ldots, n$, are independent with $Y_0 \sim \nu_0 \chi^2_{\nu_0}$ and $Y_i \sim \nu \chi^2_\nu$ for $i = 1, \ldots, n$.

3. Archimedean Copula. Let the distribution of the p-values generated from the test statistics be assumed to be such that the pairwise joint distribution of the null p-values can be modelled by an Archimedean copula. A bivariate copula, which is the joint cumulative distribution function of a pair of random variables on a unit square with unform marginals, is said to be Archimedean if it can be expressed by $C(u, v) = \phi^{-1}\{\phi(u) + \phi(v)\}$, $0 < u, v < 1$, for some convex decreasing function ϕ, called generator, satisfying $\phi(1) = 0$ $\{\phi^{-1}(u) = 0$ if $u > \phi(0)\}$. Following are some well-known systems of bivariate distributions belong to this class:

(a) Clayton copula:
$C_\theta(u, v) = \left\{\max(u^{-\theta} + v^{-\theta} - 1,\ 0)\right\}^{-1/\theta}, \quad \theta \in [-1, \infty) \setminus \{0\}.$

(b) Gumbel copula:
$C_\theta(u, v) = \exp\left[-\left\{(-\log u)^\theta + (-\log v)^\theta\right\}^{1/\theta}\right], \quad \theta \in [1, \infty).$

(c) Frank copula:
$C_\theta(u, v) = -\log\left[1 + \{\exp(-\theta u) - 1\}\{\exp(-\theta v) - 1\}/\{\exp(-\theta) - 1\}\right]/\theta, \ \ \theta \in (-\infty, \infty) \setminus \{0\}.$

(d) Joe copula:
$C_\theta(u,v) = 1 - \left\{(1-u)^\theta + (1-v)^\theta - (1-u)^\theta(1-v)^\theta\right\}^{1/\theta}, \quad \theta \in [1,\infty)$.

(e) Ali-Mikhail-Haq copula:
$C_\theta(u,v) = uv/\{1-\theta(1-u)(1-v)\}, \quad \theta \in [-1,1)$.

The following three lemmas prove the desired convexity property for all of the above distributions. While we prove the property for multivariate Gamma and F by establishing it for certain general families of location- and scale-mixture distributions in Lemmas 1 and 2, respectively, we do it individually for each of the Archimedean copulas in Lemma 3.

Proofs of Lemmas 1 and 2 rely on the following important result: Let X be a random variable with the density, $f(x,\theta)$, at x depending on parameter θ. Then, the expectation of an increasing function of X is increasing in θ if $f(x,\theta)$ is totally positive of order two in (x,θ); that is, $f(x,\theta)f(x',\theta') \geq f(x',\theta)f(x,\theta')$ for all $x \leq x', \theta \leq \theta'$ (Karlin, 1968).

Lemma 1. *Let the random variables $X_1,\ldots,X_n$ be such that, given $Y = y$, they are independent and identically distributed with a common density $\phi_1(x-y)$ and $Y \sim \phi_2(y)$, for some densities ϕ_1 and ϕ_2. Then, Property 1 holds for $X_1,\ldots,X_n$ if $\phi_2(x-y)$ is totally positive of order two in (x,y).*

Proof. We prove the lemma only for the pair (X_1,X_2) without any loss of generality. Let $\Phi_1(x-y) = pr(X_i \leq x \,|\, Y = y)$, and $f(x) = \int \phi_1(x-y)\phi_2(y)dy = \int \phi_1(y)\phi_2(x-y)dy$ be the common density of X_i, for $i = 1,2$. Then, we have

$$\begin{aligned} pr(X_1 \leq x \,|\, X_2 = x) &= \{f(x)\}^{-1} \int \Phi_1(x-y)\phi_1(x-y)\phi_2(y)dy \\ &= \{f(x)\}^{-1} \int \Phi_1(y)\phi_1(y)\phi_2(x-y)dy \\ &= \int \Phi_1(y)\phi^*(y,x)dy, \end{aligned}$$

where $\phi^*(y,x) = \phi_1(y)\phi_2(x-y)/f(x)$. Since the density $\phi^*(y,x)$, with x treated as parameter, is totally positive of order two in (y,x) and $\Phi_1(y)$ is increasing in y, the above integral is increasing in x. This proves the lemma.

□

Remark 5. Multivariate Gamma belongs to the family of distributions considered in Lemma 1. It is easy to see that here the density of Y_0 at y, say $\phi_2(y)$, is such that $\phi_2(x-y)$ is totally positive of order two in (x,y). Hence, these X_i's that jointly have multivariate gamma distribution satisfy Property 1.

Lemma 2. *Let the positive valued random variables $X_1,\ldots,X_n$ be such that, given $Y=y$, where Y is also positive valued, they are independent and identically distributed with a common density $y\psi_1(yx)$ and $Y\sim\psi_2(y)$, for some densities ψ_1 and ψ_2. Then, Property 1 holds for $X_1,\ldots,X_n$ if $\psi_2(x/y)$ is totally positive of order two in (x,y).*

Proof. As in Lemma 1, we prove the lemma only for the pair (X_1,X_2). Let $pr(X_i\le x\,|\,Y=y)=\Psi_1(yx)$, and $f(x)=\int y\psi_1(yx)\psi_2(y)dy = x^{-2}\int y\psi_1(y)\psi_2(y/x)dy$ be the common density of X_i, for $i=1,2$.

$$\begin{aligned}pr(X_1\le x\,|\,X_2=x) &= \{f(x)\}^{-1}\int \Psi_1(yx)y\psi_1(yx)\psi_2(y)dy\\ &= \{f(x)\}^{-1}x^{-2}\int \Psi_1(y)y\psi_1(y)\psi_2(y/x)dy\\ &= \int \Psi_1(y)\psi^*(y,x)dy,\end{aligned}$$

where $\psi^*(y,x)=x^{-2}y\psi_1(y)\psi_2(y/x)/f(x)$. Since the density $\psi^*(y,x)$, with x treated as parameter, is totally positive of order two in (y,x) and $\Psi_1(y)$ is increasing in y, the above integral is increasing in x. This proves the lemma.

□

Remark 6. Multivariate F belongs to the so-called scale mixture family of distributions considered in Lemma 2. It is easy to see that the density of $\chi^2_{\nu_0}$ at y, say $\psi_2(y)$, is such that $\psi_2(y/x)$ is totally positive of order two is (y,x). Hence, these X_i's that jointly have multivariate F distribution satisfy Property 1.

Lemma 3. *The Archimedean copulas listed above satisfy Property 1.*

Proof. The lemma will be proved by showing that $H'_\theta(u)$, the derivative of $H_\theta(u) = C_\theta(u,u)$ with respect to u, is non-decreasing in $u \in (0,1)$, and thus proving the desired convexity result, for each of these copulas.

(a). Clayton copula: $H_\theta(u) = (2u^{-\theta} - 1)^{-1/\theta}$ if $2u^{-\theta} \geq 1$; otherwise $= 0$, $\theta \in [-1, \infty) \setminus \{0\}$.

For this copula, $H'_\theta(u) = 2/\{(2u^{-\theta} - 1)^{1/\theta+1} u^{\theta+1}\}$, which is non-decreasing in $u \in (0,1)$, since the denominator term has the following derivative, $-(1+\theta)(2u^{-\theta} - 1)^{1/\theta} u^{\theta}$, which is ≤ 0, for $u \in (0,1)$.

(b). Gumbel copula: $H_\theta(u) = \exp[-\{2(-\ln u)^\theta\}^{1/\theta}] = u^{2^{1/\theta}}$, $\theta \geq 1$.

Here, $H'_\theta(u) = 2^{1/\theta} u^{2^{1/\theta}-1}$, which is clearly non-decreasing in $u \in (0,1)$.

(c). Frank copula: $H_\theta(u) = -\theta^{-1} \log\left[\{\exp(-u\theta) - 1\}^2 / \{\exp(-\theta) - 1\} + 1\right]$, $\theta \in (-\infty, \infty) \setminus \{0\}$.

For this copula,

$$\begin{aligned} H'_\theta(u) &= \frac{2\exp(-2\theta u) - 2\exp(-\theta u)}{\exp(-2\theta u) - 2\exp(-\theta u) + \exp(-\theta)} \\ &= \frac{1}{1 - \left(\frac{1}{2}\right)\left\{\frac{\exp(-2\theta u) - \exp(-\theta)}{\exp(-2\theta u) - \exp(-\theta u)}\right\}}. \end{aligned}$$

It is easy to check that the term $\{\exp(-2\theta u) - \exp(-\theta)\}/\{\exp(-2\theta u) - \exp(-\theta u)\}$ is non-decreasing in $u \in (0,1)$, and so is $H'_\theta(u)$.

(d). Joe copula: $H_\theta(u) = 1 - \{2(1-u)^\theta - (1-u)^{2\theta}\}^{1/\theta}$, $\theta \geq 1$.

For this copula, $H'_\theta(u) = 2\{2 - (1-u)^\theta\}^{1/\theta}\{(1-u)^\theta - 1\}/\{(1-u)^\theta - 2\}$, which is non-decreasing in $u \in (0,1)$, since both of the terms $\{2 - (1-u)^\theta\}^{1/\theta}$ and $\{(1-u)^\theta - 1\}/\{(1-u)^\theta - 2\}$ are so and are non-negative.

(e). Ali-Mikhail-Haq copula: $H_\theta(u) = u^2/\{1 - \theta(1-u)^2\}$, $\theta \in [-1, 1)$.

For this copula, $H'_\theta(u) = 2u\{1 - \theta(1-u)\}/\{1 - \theta(1-u)^2\}^2$. Let $G_\theta(u) = H'_\theta(1-u)$. Then, we note that $G'_\theta(u) = 2y_\theta(u)/\{(1 - \theta u^2)^3\}$, where $y_\theta(u) = -1 - \theta + 6\theta u - 3\theta u^2 - 3\theta^2 u^2 + 2\theta^2 u^3$. Since $y'_\theta(u) = 6\theta(1-u)(1-\theta u)$, we see the following:

Case 1: $\theta \geq 0$. The function $y_\theta(u)$ is non-decreasing in $u \in (0, 1)$, and so takes the maximum value at $u = 1$, which is $-(1-\theta)^2 < 0$, implying that $y_\theta(u) < 0$ on $(0, 1)$.

Case 2: $\theta < 0$. The function $y_\theta(u)$ is monotonically decreasing in $u \in (0, 1)$, and takes the maximum value at $u = 0$, which is $-(1+\theta) \leq 0$, implying that $y_\theta(u) \leq 0$ on $(0, 1)$.

Thus, $G_\theta(u)$ is non-increasing, and hence $H'_\theta(u)$ is non-decreasing, in $u \in (0, 1)$. □

3.9.2 *Estimating H and checking its convexity from Data*

To illustrate how to implement the proposed procedures in practice without making any distributional assumptions that would allow a known form for G or H with the concavity or convexity condition, we consider analyzing a commonly used gene expression data from the leukemia microarray study of Golub et al. (1999). The data consist of 3051 gene expression levels across 38 tumor mRNA samples, of which 27 are of acute lymphoblastic leukemia and 11 are of acute myeloid leukemia. The data were log-transformed and normalized.

The goal of the study is to determine which genes are differentially expressed by testing $H_{0i} : \mu_{1i} = \mu_{2i}$ against $H'_{0i} : \mu_{1i} \neq \mu_{2i}$ simultaneously for $i = 1, \ldots, 3051$, where μ_{1i} and μ_{2i} are the gene specific mean expressions respectively for the acute lymphoblastic leukemia type and the acute myeloid leukemia type. For ease of illustration, we consider using the subset of the data that contain the first $n = 20$ gene expression levels.

We shall use two-sample t-test statistics for testing the 20 hypotheses and

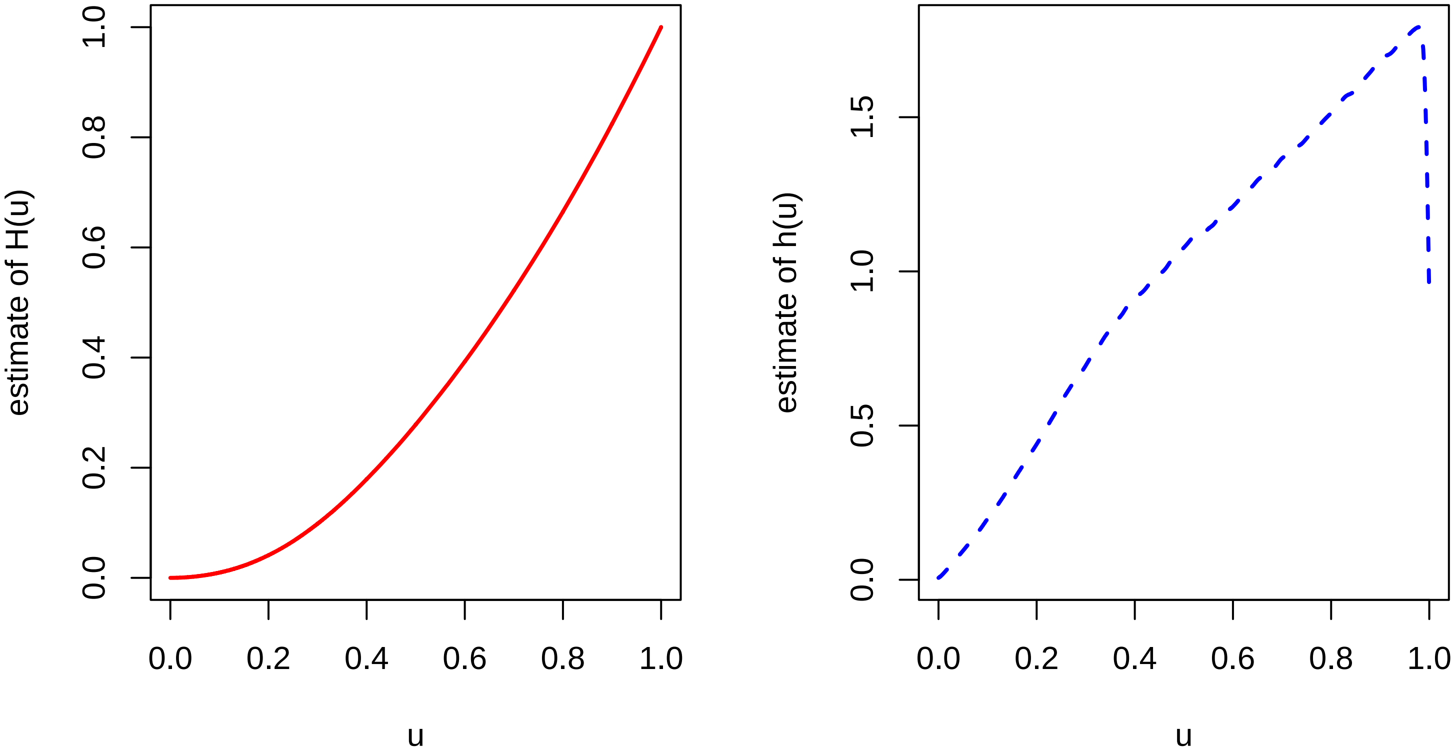
Table 3.3: Estimated values of both H and h and the calculated critical values in the order of Holm's procedure, Procedures 1 and 2 for $n = 20$ and $\alpha = 0.05$.

i	$\widehat{H}\{\alpha/(n-i+1)\}$	$\widehat{h}\{\alpha/(n-i+1)\}$	$\alpha/(n-i+1)$	$\tilde{\alpha}_i$	α_i^*
1	0.0000069	0.0090777	0.0025000	0.0025066	0.0025066
2	0.0000078	0.0092484	0.0026316	0.0026390	0.0026390
3	0.0000085	0.0094381	0.0027778	0.0027859	0.0027859
4	0.0000096	0.0096500	0.0029412	0.0029503	0.0029503
5	0.0000112	0.0098885	0.0031250	0.0031355	0.0031356
6	0.0000125	0.0101588	0.0033333	0.0033450	0.0033451
7	0.0000144	0.0104677	0.0035714	0.0035849	0.0035850
8	0.0000169	0.0108241	0.0038462	0.0038618	0.0038619
9	0.0000192	0.0112399	0.0041667	0.0041843	0.0041844
10	0.0000225	0.0117863	0.0045455	0.0045660	0.0045661
11	0.0000275	0.0124501	0.0050000	0.0050249	0.0050251
12	0.0000342	0.0132615	0.0055556	0.0055861	0.0055863
13	0.0000429	0.0142756	0.0062500	0.0062878	0.0062880
14	0.0000555	0.0156962	0.0071429	0.0071908	0.0071911
15	0.0000741	0.0175943	0.0083333	0.0083955	0.0083960
16	0.0001061	0.0204084	0.0100000	0.0100856	0.0100863
17	0.0001611	0.0247778	0.0125000	0.0126220	0.0126231
18	0.0002785	0.0322859	0.0166667	0.0168544	0.0168564
19	0.0006144	0.0474240	0.0250000	0.0253110	0.0253146
20	0.0023727	0.0942955	0.0500000	0.0500000	0.0500000

generating the corresponding p-values, $P_i, i = 1, \ldots, 20$. While applying our proposed procedures to the above p-values, we assume that the true null p-values are exchangeable; that is, we use Procedures 1 and 2 whose critical values are given in (3.2) and (3.5) respectively. To determine the critical values of Procedure 1, we need only to estimate the values of $H\{\alpha/(n-i+1)\}$; whereas for Procedure 2, we need to estimate as well the values of $h\{\alpha/(n-i+1)\}$, for $i = 1, \ldots, n = 10$.

In the two-group experimental setting, we used the permutation approach described in Dudoit and van der Laan (2008, §2) to generate the distribution of the true null p-values. We considered generating $B = 100,000$ permutations between the two groups that correspond to the two leukemia types. For each permutated data, we used the two-sample t-test to calculate the corresponding p-value $P_i^{(b)}$ for each gene, where $i = 1, \ldots, n = 20$ and $b = 1, \ldots, B$. Then, for each pair of p-values, $(P_i^{(b)}, P_j^{(b)})$, where $1 \leq i < j \leq n = 20$, we calculated their maximum value, $\widetilde{P}_k^{(b)} = \max(P_i^{(b)}, P_j^{(b)})$, where $k = 1, \ldots, N$ and $N = n(n-1)/2$. Based on these calculated values, $\widetilde{P}_k^{(b)}, k = 1, \ldots, N$ and $b = 1, \ldots, B$, we computed its empirical distribution $\widehat{H}$, an estimate of H. Also, we derived an estimate $\widehat{h}$ of h using the R function density. Thus, we obtained the estimated values of both $H\{\alpha/(n-i+1)\}$ and $h\{\alpha/(n-i+1)\}$ for $i = 1, \ldots, n$, and computed the critical values of Procedures 1 and 2, which are presented in Table 3.3.

We then generated the plots of $\widehat{H}$ and $\widehat{h}$, and graphically checked whether H is convex and h is increasing in u on an interval $(0, \alpha_0)$, with $\alpha_0 > \alpha$. As seen from Figure 1, $\widehat{H}$ is indeed convex on $(0, 1)$ and $\widehat{h}$ is increasing on $(0, 0.9)$. Thus, the desired conditions for H and h are satisfied in this real data example.


estimate of H(u)
0.0
0.2
0.4
0.6
0.8
1.0
u
0.0
0.2
0.4
0.6
0.8
1.0
estimate of h(u)
0.0
0.5
1.0
1.5
u
0.0
0.2
0.4
0.6
0.8
1.0

The above example illustrates how the distribution function H and the density function h of the pairwise maximums of null p-values can be estimated from data by an appropriate resampling method, and doing so, the assumptions pertaining to H and h can be checked graphically.

CHAPTER 4

EXTENDING STEPDOWN MULTIPLE TESTING PROCEDURES TO GROUP-SEQUENTIAL SETTING

4.1 Research Objectives and Findings

This chapter is focused on the development of a group-sequential extension of Dunnett & Tamhane's stepdown procedure (1991), which is a classical parametric stepdown procedure fully utilizing hypothesis-wise correlation of test statistics. It is assumed that hypothesis-wise dependency structure is fully known *a priori* or estimable. We propose a group-sequential version of the weighted parametric stepdown procedure. We will show that the proposed group-sequential weighted parametric procedure is stepdown, sequentially rejective at each of the analysis time points, and therefore relatively simple to use. And we will show that the proposed group-sequential procedure inherits fully the *actual* significance level from its fixed-sample counterpart.

Note that the fixed sample Dunnett & Tamhane's stepdown procedure (1991) belongs to the Union Intersection (UI) test procedure. Chapters 4 and 5 are exclusively focused on the UI test procedure in group-sequential setting, where the clinical objective is to claim rejection on at least one of the hypotheses (e.g, endpoints). Note that the Intersection Union (IU) test procedure in group-sequential setting is out of scope of our current research. We have seen that there is a growing body of literature, in recent years, on the IU type of group-sequential trial that is designed to reject all of the hypotheses (e.g, endpoints). The nomenclatures in the clinical trial literature are differentiated as being "multiple primary endpoints" vs "multiple co-primary endpoints" coined by Often et al. (2007). Our proposed methods in Chapters 4 and 5 are applied to group-sequential trials with "multiple primary endpoints" , as opposed to "multiple co-primary endpoints".

4.2 The State-of-the-Art Methodologies

Building on some earlier developments (Bretz et al. 2009; Dmitrienko et al. 2003, 2006, 2008), Bretz et al (2011) refined a generalised class of sequentially rejective weighted Bonferroni MTPs via graphical approaches. The generalised class of MTPs dissociates the underlying weighting strategy from the stepdown procedure, thus providing a large degree of flexibility for applications. With their systematic treatment, many seemingly different MTPs, such as fixed sequence procedure, fallback procedure, gatekeeping procedure can be placed into one framework due to having some common features.

Recently, Maurer & Bretz (2013) developed a class of weighted group-sequential Bonferroni procedures for testing multiple endpoints in clinical trials, where it is desirable to yield as many rejections as possible. The class, which is of union intersection (UI) test, maintains a strong control of the FWER at a pre-specified level α across all analysis time points and all hypotheses. The class, while conveniently based on the use of individual marginal p-values, does not take into account hypothesis-wise dependency structure of test statistics. Analogous to its fixed-sample counterpart, the class of the group-sequential procedures is conservative under positively dependent test statistics, severely being so under highly positive correlated test statistics, and with large number of hypotheses. In this chapter, we provide some further development by incorporating hypothesis-wise dependencies.

4.3 A Weighted Parametric Stepdown Procedure for fixed-sample Setting

Consider, in the fixed-sample setting, the problem of simultaneous testing of m hypotheses. Let $I = \{1, 2, \ldots, m\}$ denote the index set of the null hypotheses $\{H_1, H_2, \ldots, H_m\}$, with the cardinality $|I| = m$. A set of standard assumptions are as follows. The hypotheses satisfy the free combination condition, i.e, for any subset $J \subset I$ the simultaneous truth of $H_i, i \in J$, and falsity of the remaining hypotheses is possible. Univariate test statistics X_i for $H_i, i \in I$ is readily available, and test statistics $(X_1, X_2, \ldots, X_m)$ is jointly distributed as m-variate normal or t, with a zero mean vector and a known positive definite correlation matrix Σ under the null hypotheses. Note that it is not necessary to assume equal-correlation. Let P_i be a random variable taking on null p-values based on univariate (marginal) test statistics X_i. We assume $P_i \sim U(0, 1), i \in I$.

Dunnett & Tamhane stepdown procedure (1991), which is unweighted, was originally developed for comparing several treatments with a control in one-way layouts. Through the utilization of a fully parametric joint distribution of test statistics, Dunnett's stepdown procedure controls the FWER strongly and exactly at a pre-specified level α. The procedure is also applicable in the testing of multiple clinical endpoints. In the context of multiple endpoints, hypotheses are sometimes unequally weighted to reflect their varying degree of ‘importance’, thus calling for the development of a weighted version of Dunnett’s stepdown procedure.

When a weighting scheme is applied to a parametric closed testing procedure, loss of consonancy generally occurs. However, there exists a simple and popular weighting scheme that ensures consonancy of the parametric closed test procedure. In this paper, we show that Holm’s weighting scheme (1979), when applied to a parametric closed testing procedure, will ensure consonancy, resulting in a procedure which we term “weighted Dunnett’s stepdown procedure”. The original Dunnett’s stepdown procedure (1991) can be viewed as a special case of the “weighted”, with the initial weight allocated to each hypothesis being equally $1/m$.

4.3.1 *Holm’s weighting scheme (1979)*

Holm’s weighting scheme is graphically illustrated in the “generalized sequentially rejective procedure” he originally proposed (1979). The procedure is usually known as the weighted Holm’s procedure in the literature. It operates as follows in the context of testing a family of m hypotheses. Let J denote the index set of the m null hypotheses. $J = \{1, \ldots, m\}$. Assume a collection of initial weights $w_i(J)$, $i \in J$, with $0 < w_i(J) \leq 1$ and $\sum_{i \in J} w_i(J) = 1$, for the family. At step 1, the procedure compares each p_i with $w_i(J)\alpha, i \in J$. If $p_i \leq w_i(J)\alpha$, then the corresponding H_i is rejected. Let $R = \{H_i, i \in J : p_i < w_i(J)\alpha\}$. So, R stands for the set of rejected hypotheses by Step 1’s tests. The procedure then proceeds to Step 2 for testing of the non-rejected hypotheses in set $J \setminus R$, using a collection of updated weights $w_i(J \setminus R) = w_i(J)/(1 - \sum_{r \in R} w_r(J))$, $i \in J \setminus R$. At Step 2, the procedure compares each p_i with the updated critical value $w_i(J \setminus R)\alpha$, $i \in J \setminus R$. It continues like Step 1 and 2 until the first non-rejection.

A Bonferroni-based closed testing procedure assumes little restriction on the allocation and re-allocation of weights for the procedure to be consonant, hence sequentially rejective. So there is a large degree of freedom in selecting a particular weighting scheme for the Bonferroni-based procedure to suit a specific study objective. In contrast, to preserve consonancy for a parametric-based closed testing procedure, the restriction has to be made stronger, and one has to sacrifice the nearly unconstrained choice of weighting schemes. Specific examples are given in Bretz et al. 2011 (p. 901) as to how some weighting strategies fail to preserve consonancy in parametric-based closed testing procedures. However, we find that Holm's weighting scheme preserves consonancy in parametric-based closed testing procedures. Intuitively, Holm's weighting scheme tends to not deviate much from the initial ratios of the non-zero weights (representing the initial relative importance), so it is generally suitable, for instance, for the testing of several primary endpoints, as opposed to the testing of hierarchical ordered endpoints (primary and secondary endpoints).

4.3.2 *A weighted parametric test for intersection hypothesis*

For testing of an intersection (null) hypothesis $H_J = \bigcap_{i \in J} H_i$, where $J \subseteq I = \{1, 2, \ldots, m\}$, at a pre-specified significance level α, a parametric test utilizes the joint distribution of test statistics, resulting in a larger critical region than that of a Bonferroni test, and thus is more powerful. Consider a weighted parametric test for H_J as follows. H_J is rejected if $p_i < \xi_J w_i(J)\alpha$ for some $i \in J$, where $w_i(J)$ is the weight updated for H_i, and ξ_J, with $\xi_J \geq 1$, is the exact solution to the following equation under the intersection hypothesis H_J:

$$Pr\left(\bigcap_{i \in J}\{P_i \geq \xi_J w_i(J)\alpha\}\right) = 1 - \alpha \tag{4.1a}$$

If $\sum_{i=1}^{|J|} \mathcal{I}(p_i < \xi_J w_i(J)\alpha) \geq 0$, then H_J is rejected. Note that $\mathcal{I}(\cdot)$ denotes the indicator function.

Lemma 4. *In the above setting, Holm's weighting scheme when employed in*

α-exhaustive parametric tests for all subset intersection hypotheses will ensure monotonicity condition as defined in the following: $\xi_J w_i(J) \leq \xi_{J'} w_i(J')$ for all $J' \subseteq J \subseteq I$ and $i \in J'$.

Note that the definition of monotonicity condition is per Bretz et al. 2011 (p. 901).

Proof. Without loss of generality, assume that J' is a proper subset of J; that is, $J' = J \setminus R$, where $R \neq \emptyset$. Let ξ_J be the solution to equation (4.1a), and $\xi_{J'}$ be the solution to the following equation:

$$Pr\left(\bigcap_{i \in J \setminus R} \{P_i \geq \xi_{J'} w_i(J \setminus R)\alpha\}\right) = 1 - \alpha \tag{4.1b}$$

where $w_i(J \setminus R) = w_i(J)/(1 - \sum_{r \in R} w_r)$, for $i \in J \setminus R$.

Using (4.1a) and (4.1b), we have the following equality (4.1c):

$$Pr\left(\bigcap_{i \in J} \{P_i \geq \xi_J w_i(J)\alpha\}\right) = 1 - \alpha = Pr\left(\bigcap_{i \in J \setminus R} \{P_i \geq \xi_{J'} w_i(J \setminus R)\alpha\}\right) \tag{4.1c}$$

And since

$$Pr\left(\bigcap_{i \in J \setminus R} \{P_i \geq \xi_J w_i(J)\alpha\}\right) \geq Pr\left(\bigcap_{i \in J} \{P_i \geq \xi_J w_i(J)\alpha\}\right)$$

We have the following inequality expression (4.1d):

$$Pr\left(\bigcap_{i \in J \setminus R} \{P_i \geq \xi_J w_i(J)\alpha\}\right) \geq Pr\left(\bigcap_{i \in J \setminus R} \{P_i \geq \xi_{J'} w_i(J \setminus R)\alpha\}\right) \tag{4.1d}$$

Let η denote the ratio:

$$\eta = \frac{\xi_{J'} w_i(J \setminus R)\alpha}{\xi_J w_i(J)\alpha} = \frac{\xi_{J'}}{\xi_J(1 - \sum_{r \in R} w_r)}, \text{ for all } i \in J \setminus R. \tag{4.1e}$$

Then plugging (4.1e) into (4.1d) and simplify, we get

$$Pr\left(\bigcap_{i\in J\setminus R}\{P_i \geq \xi_J w_i(J)\alpha\}\right) \geq Pr\left(\bigcap_{i\in J\setminus R}\{P_i \geq \eta\cdot\xi_J w_i(J)\alpha\}\right) \tag{4.1f}$$

If $\eta < 1$, we would have $\{P_i \geq \xi_J w_i(J)\alpha\} \subset \{P_i \geq \eta\cdot\xi_J w_i(J)\alpha\}$, for all $i \in J \setminus R$. Thus leading to $Pr\left(\bigcap_{i\in J\setminus R}\{P_i \geq \xi_J w_i(J)\alpha\}\right) < Pr\left(\bigcap_{i\in J\setminus R}\{P_i \geq \eta\cdot\xi_J w_i(J)\alpha\}\right)$, which is a contradiction to inequality (4.1f). Then it follows that $\eta \geq 1$. □

Remark 7. *Lemma 4 can be stated for the commonly used model settings, such as the multivariate normal or t test statistics. But this particular distributional assumption is not required. And equal-correlation (exchangeability of the null p-values) is not needed either.*

4.3.3 *A weighted parametric stepdown multiple comparison procedure*

Lemma 4 ensures consonancy of the parametric-based closed testing procedure. We now briefly outline the steps of the proposed weighted parametric procedure as the following. Start with testing of the global intersection hypothesis H_I using a level-α critical boundary $\xi_I w_i(I)\alpha$, $i \in I$. If H_I is rejected, then all the elementary hypotheses in the set $R = \{H_i, i \in I : p_i < \xi_I w_i(I)\alpha\}$ are rejected. Continue testing the reduced (or nested) intersection hypothesis $H_{I\setminus R}$ that is formed by the not-yet rejected hypotheses, using an updated level-α critical boundary $\xi_{I\setminus R} w_i(I \setminus R)\alpha$, $i \in I \setminus R$, and so on until the first non-rejection.

Because it is constructed based on the closure method, the proposed weighted parametric procedure controls the FWER strongly. And, in common with the class of weighted Bonferroni procedures, the proposed weighted parametric procedure remains stepdown, sequentially rejective, requiring at most m steps (tests). In fact, due to all intersection tests being α-exhaustive, the proposed weighted parametric procedure controls the FWER exactly at the pre-specified level α, whereas the weighted Bonferroni procedure is conservative.

4.4 A Weighted Parametric Stepdown Procedure for group-sequential setting

We'll show that the fixed-sample weighted parametric stepdown procedure can be naturally extended into group-sequential setting for application, with desirable properties (such as the sequential rejectiveness) retained.

4.4.1 *The criterion of a well-behaved error spending function*

A group-sequential design is distinctly characterized by the allocation of the Type I error and the Type II error to each of the analysis time points. The error spending approach proposed by Lan & DeMets (1983) is convenient for implementing the group-sequential design and analysis. Parameterized by a pre-specified Type I error rate α for control, an error spending function $\mathbb{A}(\alpha, t)$ is a cumulative Type I error spent function of information fraction t, $0 \leq t \leq 1$. Given an α, $\mathbb{A}(\alpha, t)$ is non-decreasing in t, with $\mathbb{A}(\alpha, 0) = 0$ and $\mathbb{A}(\alpha, 1) = \alpha$. Such function, originally developed for the use in testing of a single-hypothesis, is not adequate for multiple hypotheses. To facilitate our methodological development, we need to use the "well-behaved" error spending functions, which is obtainable by imposing two extra restrictions as the followings: (1) $\mathbb{A}(\alpha, t)$ is strictly increasing in t as opposed to 'non-decreasing'. (2) Given a t, $\mathbb{A}(\alpha, t)$ is strictly increasing in α. Note that these two extra restrictions have no practical significance virtually, since most of the commonly used error spending functions, such as the Wang-Tsiatis class and the power class, are in fact "well-behaved". But the property of strict monotonicity in both α and t has theoretical significance and is critical to our methodological development in what follows.

For an error spending function $\mathbb{A}(\alpha, t)$ that is differentiable with respect to α and t, an equivalent criterion of the "well-behaved" is: $\partial^2\mathbb{A}(\alpha, t)/\partial\alpha\partial t > 0$, for $0 < t \leq 1$, and $0 < \alpha < \alpha^0$, where α^0 is a fixed number $\in (0, 1)$. This criterion comes handy for verifying whether an error spending function is well-behaved or not.

Consider, in a J-stage group-sequential setting, a repeated significance testing of a single hypothesis H_i controlling the family-wise (experimental-wise) error rate at level α, with $0 < \alpha \leq \alpha^0$. Let assume the set of information fractions $\langle t_1, t_2, \ldots, t_J\rangle$ be given, with $0 < t_1 < t_2 < \cdots < t_J = 1$. Let $\alpha_1, \ldots, \alpha_J$ denote the spending levels respectively for analysis $j = 1, \ldots, J$, with $\sum_{j=1}^{J} \alpha_j = \alpha$. The job of an error spending function is to pre-specify (or allocate) with some rules a fraction of α to each j, $j = 1, \ldots, J$. Thus, the allocation of spent levels $\langle\alpha_1, \alpha_2, \ldots, \alpha_J\rangle$ can be treated as a vector function of α. Why would we need an error spending function to be well-behaved as elaborated above? Because such an error spending function will ensure that the vector function $\langle\alpha_1(\alpha), \alpha_2(\alpha), \ldots, \alpha_J(\alpha)\rangle$ be monotonically increasing in α on the domain $(0, \alpha^0)$. And this monotonicity in α will have theoretical significance as we will show later.

For example, the power-type of error spending function $\mathbb{A}(\alpha, t) = \alpha t^\rho$, $\rho > 0$ belongs to the class of the "well-behaved". To see this, $\partial^2\mathbb{A}(\alpha, t)/\partial\alpha\partial t = \rho t^{\rho-1} > 0$ on $(0,\ 1)$, the entire domain of α, for $0 < t \leq 1$. It can also be seen that the pre-specified spending level α_j is strictly increasing in α for all $j = 1, \ldots, J$. Here is the details: Let's calculate the spending levels for analysis $j = 1, 2, \ldots, J$ respectively, assuming a given set of information fractions $\langle t_1, t_2, \ldots, t_J \rangle$, with $0 < t_1 < t_2 < \cdots < t_J = 1$.

$$\begin{aligned}
\alpha_1(\alpha) &= \mathbb{A}(\alpha, t_1) \\
\alpha_2(\alpha) &= \mathbb{A}(\alpha, t_2) - \mathbb{A}(\alpha, t_1) \\
&\vdots \\
\alpha_j(\alpha) &= \mathbb{A}(\alpha, t_j) - \mathbb{A}(\alpha, t_{j-1}) \\
&\vdots \\
\alpha_J(\alpha) &= \alpha - \mathbb{A}(\alpha, t_{J-1})
\end{aligned}$$

It is easy to see that $\alpha_j(\alpha) = \alpha(t_j^\rho - t_{j-1}^\rho)$, which is strictly increasing in α, for all $j = 1, 2, \ldots, J$.

One of the commonly used error spending functions are of the Wang & Tsiatis type, of which the OBF-type (O'Brien-Flemming-type) and the PO-type (Pocock-type) are special cases. The Wang & Tsiatis type of error spending functions is well-behaved. The specific function form approximating the original OBF-type is given as $\mathbb{A}(\alpha, t) = 2(1 - \Phi(\Phi^{-1}(1 - \alpha/2)/\sqrt{t}))$, and that approximating the original PO-type is given as $\mathbb{A}(\alpha, t) = \alpha \ln(1+(e-1)t)$, where e the natural number. It can be proved that the PO-type is well-behaved on the entire domain of α, which is $(0, 1)$, and the OBF-type is well-behaved on the α domain of $(0, 0.3173)$ (it is 0.318 per Maurer & Bretz, 2013, p.313).

Note that Maurer & Bretz (2013, p. 312 - 313) proposed some "well-ordered families of spending functions" in order to facilitate their development of the weighted group-sequential Bonferroni multiple testing procedure. In our view, the restrictions they imposed on families of error spending functions appeared to be complex, and also not necessary. In our view, their definition would have implied that, for example, the OBF-type of error spending function could not in general be used together with the PO-type in the testing of 2 primary endpoints, where one of the endpoints employs the OBF-type, while the other employing the PO-type. Our notion is that the OBF-type and the PO-type can be generally used together in the testing of 2 primary endpoints.

4.4.2 *The standard group-sequential methodology for single hypothesis*

In this section, we'll review with a perspective the standard group-sequential method for a single-hypothesis. And we'll introduce some concepts and notations.

Consider the testing of a single-hypothesis $H_i, i \in I = \{1\}$ using some test statistics (e.g., Wald test statistics) in a J-stage group-sequential setting. Let $P_{i,j}$ be the p-values corresponding to the test statistics $X_{i,j}$ based on cumulative data up to analysis j, $j = 1, \ldots, J$. Under the null hypothesis, the test statistics $\{X_{i,j}, j = 1, \ldots, J\}$ are assumed to be identically distributed as approximately standard normal.

By employing a well-behaved error spending function $\mathbb{F}^i$ for H_i, the allocation of spent levels $\langle \alpha_1, \ldots, \alpha_j, \ldots, \alpha_J \rangle$ across J analyses shall be immediately obtainable, with $\alpha_j \neq 0$ for analysis $j = 1, \ldots, J$, and $\sum_{j=1}^{J} \alpha_j = \alpha$. We shall solve the following set of equations for the univariate (single-hypothesis) critical boundary $\{b_j, j = 1, \ldots, J\}$ that is expressed in terms of nominal significance level. Note that the probability in the equations is evaluated under the null hypothesis.

$$
\begin{aligned}
& pr(P_{i,1} < b_{i,1}) = \alpha_1 \\
& pr(\{P_{i,1} \geq b_{i,1}\} \cap \{P_{i,2} < b_{i,2}\}) = \alpha_2 \\
& \vdots \\
& pr(\cap_{k=1}^{j-1} \{P_{i,k} \geq b_{i,k}\} \, \cap \, \{P_{i,j} < b_{i,j}\}) = \alpha_j \\
& \vdots \\
& pr(\cap_{k=1}^{J-1} \{P_{i,k} \geq b_{i,k}\} \, \cap \, \{P_{i,J} < b_{i,J}\}) = \alpha_J
\end{aligned}
\tag{4.2}
$$

Given a well-behaved error spending function $\mathbb{F}^i$ employed for H_i, the boundary solution $\{b_j, j = 1, \ldots, J\}$ to the above set of equations can be in principle treated as an implicit vector function of α. Furthermore, it can be shown that the solution $\{b_j, j = 1, \ldots, J\}$ is continuously monotonically increasing in α. Thus we define a continuous vector function $\mathbb{F}_i : \mathcal{A} \mapsto \mathcal{B}, \quad b_1 = \mathbb{F}_{i,1}(\alpha), \ldots, b_J = \mathbb{F}_{i,J}(\alpha)$, where

Notation 1.

$$\begin{aligned}
&\mathcal{A} = \{\alpha \in (0, \alpha^0) : 0 < \alpha < \alpha^0\} = (0, \alpha^0) \subset \mathbb{R}^1 \\
&\mathcal{B} = \{(b_1, b_2, \ldots, b_J) \in (0, \alpha^0)^J : 0 < b_1 < \mathbb{F}_{i,1}(\alpha^0), 0 < b_2 < \mathbb{F}_{i,2}(\alpha^0), \ldots, 0 < b_J < \mathbb{F}_{i,J}(\alpha^0)\} \\
&\qquad \subset (0, \alpha^0)^J \subset \mathbb{R}^J
\end{aligned}$$

Note the notational difference here. $\mathbb{F}^i$ is the error spending function employed for hypothesis H_i, whereas the vector function $\{\mathbb{F}_{i,1}(\alpha), \ldots, \mathbb{F}_{i,J}(\alpha)\}$, when evaluated at α, is the $\mathbb{F}^i$-induced, size-α, univariate critical boundary for testing H_i. The argument of this vector function is $\alpha, \ \alpha \in (0, \alpha^0)$. Now consider a reference test $\mathcal{T}_i$, which is the fixed-sample, classical non-sequential test of H_i with a pre-specified nominal Type I error rate α. The reference test $\mathcal{T}_i$ also uses Wald test statistics. Let P_i be the p-value corresponding to Wald test statistics X_i for the reference test $\mathcal{T}_i$.

Due to strict monotonicity (possessed by the well behaved error spending function), there is a 1:1 unique mapping between a point $\alpha \in \mathcal{A}$ and a point $(b_1, b_2, \ldots, b_J) \in \mathcal{B}$. Hence, there exists 1:1 mapping between the critical region of the fixed-sample reference test $\mathcal{T}_i$ and that of the $\mathbb{F}^i$-employed, group-sequential test. To facilitate the development of what follows, we shall invent notations for critical regions.

Notation 2.
The size-α critical region of the fixed-sample reference test for H_i is denoted as $R_i(\alpha) = \{p_i \in \mathcal{A} : 0 \leq P_i < \alpha\}$. The mapped size-$\alpha$ critical region

of the group-sequential test for H_i is denoted as $\bigsqcup_{j=1}^{J} R_{i,j}$, where $R_{i,1}(\alpha) = \{(p_{i,1}, \ldots, p_{i,J}) \in \mathcal{B} : P_{i,1} < \mathbb{F}_{i,1}(\alpha)\}$ for $j = 1$, and for $j \geq 2$, $R_{i,j}(\alpha) = \{(p_{i,1}, \ldots, p_{i,J}) \in \mathcal{B} : \cap_{k=1}^{j-1}\{P_{i,k} \geq \mathbb{F}_{i,k}(\alpha)\} \cap \{P_{i,j} < \mathbb{F}_{i,j}(\alpha)\}\}$

As shown above, the probabilities assigned to events $R_i(\alpha)$ in $\mathcal{A}$ are transferred through the (implicit) functional relationship to events $\sqcup_{j=1}^{J} R_{i,j}(\alpha)$ in $\mathcal{B}$. The corresponding events reside in different probability spaces. Given an α, there is a 1:1 probability mapping by the very construction. Therefore $R_i(\alpha)$ and $\sqcup_{j=1}^{J} R_{i,j}(\alpha)$ are considered as equivalent events, conventionally denoted as $R_i(\alpha) \equiv \sqcup_{j=1}^{J} R_{i,j}(\alpha)$. It follows that equivalent events have the same probability. That is, $pr\{R_i(\alpha)\} = pr\{\bigsqcup_{j=1}^{J} R_{i,j}(\alpha)\} = \alpha$. Note that $R_{i,j}(\alpha) \sqcap R_{i,j'}(\alpha) = \emptyset$, for all $j \neq j'$.

To facilitate proofs in the next section, we shall introduce a notation $S_{i,j}(\alpha)$. Let $S_{i,j}(\alpha) = \sqcup_{k=1}^{j} R_{i,k}(\alpha)$. It can be shown that $\sqcup_{j=1}^{J} R_{i,j}(\alpha) = \cup_{j=1}^{J} S_{i,j}(\alpha)$. e.g. for a two-stage ($J = 2$) group-sequential test of H_i with Type I error rate α, $S_{i,1}(\alpha) = R_{i,1}(\alpha) = \{(p_{i,1}, p_{i,2}) \in \mathcal{B} : P_{i,1} < \mathbb{F}_{i,1}(\alpha)\}$, and $S_{i,2}(\alpha) = R_{i,1}(\alpha) \sqcup R_{i,2}(\alpha) = \{(p_{i,1}, p_{i,2}) \in \mathcal{B} : P_{i,2} < \mathbb{F}_{i,2}(\alpha)\}$.

In summary, we have the following result concerning the standard J-stage group-sequential methodology for testing a single-hypothesis H_i:
$R_i(\alpha) \equiv \cup_{j=1}^{J} S_{i,j}(\alpha)$.

4.4.3 *A group-sequential Bonferroni test for intersection hypotheses*

First recall the fixed-sample Bonferroni test for the global null hypothesis $H_0 = \cap_{i \in I} H_i$ at level-α, where $H_1, \ldots, H_m$ is a collection of m null hypotheses, and $I = \{1, \ldots, m\}$. Suppose that Wald test statistics $X_1, \ldots, X_m$ are used for testing the m hypotheses. Let $p_{(1)} < p_{(2)} < \cdots < p_{(m)}$ be the ordered marginal p-values with $H_{(i)}$ being the null hypothesis corresponding to $p_{(i)}$. The fixed-sample Bonferroni test rejects H_0 if $p_{(1)} \leq \alpha/m$. Otherwise the test accepts H_0.

Consider a way of extending Bonferroni test of $H_0 = \cap_{i \in I} H_i$, $i \in I$ into a two-stage group-sequential setting, where information fractions are assumed given, which can be equal-distanced or not. For notational simplicity, we assume for now that a single spending function $\mathbb{A}$ is employed for all of the individual hypotheses $H_i, i \in I$. Let the vector $\{\mathbb{A}_1(\alpha), \mathbb{A}_2(\alpha)\}$ be the $\mathbb{A}$-induced, size-α, univariate (single-hypothesis) critical boundary, which is the same for all H_i's (because of the same error spending function being used). Let $p_{(1),1} < p_{(2),1} < \cdots < p_{(m),1}$ be the m observed p-values ordered for the first analysis, with $H_{(i)}$ corresponding to $p_{(i),1}$. Let $p_{(1),2} < p_{(2),2} < \cdots < p_{(m),2}$ be the m observed p-values (based on cumulative data) re-ordered for the second (final) analysis, with $H_{(i)}$ corresponding to $p_{(i),2}$. Note that a $H_{(i)}$ for analysis 1 is generally not the same hypothesis as the $H_{(i)}$ for analysis 2. We herein describe a two-stage group-sequential Bonferroni test as follows. The test rejects H_0 if $p_{(1),1} \leq \mathbb{A}_1(\alpha/m)$ or $p_{(1),2} \leq \mathbb{A}_2(\alpha/m)$. Otherwise, the test accepts H_0.

Lemma 5. *The above described two-stage group-sequential Bonferroni test for the global null hypothesis has the same size as that of its fixed-sample counterpart. This holds true irregardless of the type of the error spending function $\mathbb{A}$.*

We need first state an important fundamental result to be used in the proofs for Lemma 5. This result is proved in §4.8.

Result 1.
Suppose $R_i(\alpha) \equiv \cup_{j=1}^{J} S_{i,j}(\alpha)$, $\forall\, i = 1, \ldots, m$, then we have

$$\bigcup_{i=1}^{m} R_i(\alpha) \equiv \bigcup_{i=1}^{m} \left\{ \cup_{j=1}^{J} S_{i,j}(\alpha) \right\}$$

$$\bigcap_{i=1}^{m} R_i(\alpha) \equiv \bigcap_{i=1}^{m} \left\{ \cup_{j=1}^{J} S_{i,j}(\alpha) \right\}$$

We now prove Lemma 5 as follows.

Proof. Consistent with the notations introduced earlier, let $P_{i,1}$, which corresponds to $X_{i,1}$, be the random variable generating (realizing) $p_{i,1}$ and $P_{i,2}$, which corresponds to $X_{i,2}$, be the random variable generating (realizing) $p_{i,2}$. Note that a $H_{(i)}$ for analysis 1 is generally not the same hypothesis as the $H_{(i)}$ for analysis 2.

$$
\begin{aligned}
&pr\left(\{P_{(1),1} \leq \mathbb{A}_1(\alpha/m)\} \quad \cup \quad \{P_{(1),2} \leq \mathbb{A}_2(\alpha/m)\}\right) \\
&= pr\left(\bigcup_{i=1}^{m}\{P_{i,1} \leq \mathbb{A}_1(\alpha/m)\} \quad \cup \quad \bigcup_{i=1}^{m}\{P_{i,2} \leq \mathbb{A}_2(\alpha/m)\}\right) \\
&= pr\left(\bigcup_{i'=1}^{m}\left\{\{P_{i',1} \leq \mathbb{A}_1(\alpha/m)\} \cup \{P_{i',2} \leq \mathbb{A}_2(\alpha/m)\}\right\}\right) \\
&= pr\left(\bigcup_{i'=1}^{m}\left\{S_{i',1}(\alpha/m) \cup S_{i',2}(\alpha/m)\right\}\right) \\
&= pr\left(\bigcup_{i'=1}^{m} R_{i'}(\alpha/m)\right) \\
&= pr\left(R_{(1)}(\alpha/m)\right)
\end{aligned}
$$

□

Remark 8. *Lemma 5 still holds when different hypotheses are prescribed with different error spending functions.*

This can be seen from Result 1, which still holds when different error spending functions are used for different hypotheses.

4.4.4 *A group-sequential parametric test for intersection hypotheses*

In what follows, we describe a J-stage group-sequential parametric test for an intersection hypothesis $H_I = \bigcap_{i \in I} H_i$, where $I = \{1, 2, \ldots, m\}$. let's assume a collection of weights $\{w_i(I), i \in I\}$ and a collection of error spending functions $\{\mathbb{F}^i, i \in I\}$, with w_i and $\mathbb{F}^i$ prescribed for H_i. Reject H_I if there exists an $i \in I$ and a $j, j = 1, \ldots, J$ such that $p_{i,j} \leq \mathbb{F}_{i,j}(\xi_I w_i(I)\alpha)$, where ξ_I is the exact solution to equation (4.1a).

Lemma 6. *The above described J-stage group-sequential parametric test for an intersection hypothesis $H_I = \bigcap_{i \in I} H_i$ is a size-α test. The group-sequential test inherits fully the actual level of significance from its fixed-sample counterpart.*

Proof. The lemma can be verified by calculating the probability of rejecting H_I under the intersection null hypothesis.

$$
\begin{aligned}
& pr\left(\cup_{j=1}^{J} \cup_{i=1}^{m} \{P_{i,j} \leq \mathbb{F}_{i,j}(\xi_I w_i(I)\alpha)\}\right), \text{ where } \xi_I \text{ is the exact solution to (4.1a).} \\
& = pr\left(\cup_{i=1}^{m} \cup_{j=1}^{J} \{P_{i,j} \leq \mathbb{F}_{i,j}(\xi_I w_i(I)\alpha)\}\right) \\
& = pr\left(\cup_{i=1}^{m} \cup_{j=1}^{J} S_{i,j}(\xi_I w_i(I)\alpha)\right) \\
& = pr\left(\cup_{i=1}^{m} R_i(\xi_I w_i(I)\alpha)\right) \quad \text{(by applying Result 1, which concerns equivalency)} \\
& = \alpha \quad \text{(from the result given in (4.1a))}
\end{aligned}
$$

□

4.4.5 *A group-sequential parametric stepdown multiple testing procedure*

The property of monotonicity of critical values in a fixed-sample parametric-based closed testing procedure is well retained in a group-sequential closed testing procedure.

Lemma 7. *The described J-stage group-sequential parametric tests for the family of all subset intersection hypotheses $\{H_K = \bigcap_{i \in K} H_i,\ K \subseteq I\}$ have the desirable property of monotonicity of critical values at each of the analysis time points, $j = 1, \ldots, J$.*

Proof. Without loss of generality, let $i \in K' \subset K \subseteq I$. That is, let H_i be a component hypothesis in both $H_{K'}$ and H_K.

$$w_i(K) < w_i(K'), \text{ which is per Holm's weighting scheme} \quad \Longrightarrow$$
$$\xi_K w_i(K) < \xi_{K'} w_i(K'), \text{ which is per Lemma 4} \quad \Longrightarrow$$
$$\mathbb{F}_{i,j}(\xi_K w_i(K)\alpha) < \mathbb{F}_{i,j}(\xi_{K'} w_i(K')\alpha), \text{ for each analysis time point } j = 1, \ldots, J$$
$$\text{which is per the employment of a well-behaved error spending function } \mathbb{F}^i \text{ for } H_i$$

□

The lemma above implies that there exists a short-cut of the group-sequential closed testing procedure. We therefore construct a group-sequential weighted parametric procedure, sharing the same feature as that of the weighted Bonferroni procedure (Maurer & Bretz 2013) in that it is stepdown, sequentially rejective at each analysis time point, requiring at most $\max(m, J)$ steps (tests) to finish the job (testing of m hypotheses). We outline the following algorithm for performing the proposed J-stage group-sequential parametric stepdown procedure.

Algorithm

0. Set $j = 1$, $I = \{1, 2, \ldots, m\}$, $\{w_i(I), i \in I\}$, $\{\mathbb{F}^i, i \in I\}$.

1. Given $\{w_i(I), i \in I\}$, compute ξ_I from the equation (4.1a), and then compute the critical boundary (nominal significance level) $\{\mathbb{F}_{i,j}(\xi_I w_i(I)\alpha), i \in I\}$.

2. Construct an index set $W = \{i \in I : p_{i,j} < \mathbb{F}_{i,j}(\xi_I w_i(I)\alpha)\}$, where $p_{i,j}$ is the unadjusted observed p-value for H_i at analysis j.

3. If $W \neq \emptyset$, then reject $H_i, i \in W$. And update with $I \to I \setminus W$, and correspondingly $w_i(I) \to w_i(I \setminus W)$. Go to Step 1.

4. If $W = \emptyset$, and $j < J$ then the trial can be continued with $j \to j + 1$. Go to Step 2; otherwise stop.

5. If $|I| \geq 1$, go to Step 1; otherwise stop.

Theorem 3. *The proposed group-sequential weighted parametric multiple testing procedure controls the FWER strongly and also exactly at the pre-specified significance level α.*

Proof. This can be proved by a straightforward application of closure testing principle in group-sequential setting, since (i) (i) the group-sequential weighted parametric tests for all intersection hypotheses $H_K = \bigcap_{i \in K} H_i$, $K \subseteq I$, are of size-α, due to Lemma 6, and (ii) the closed testing procedure retains consonancy at each analysis time point, due to Lemma 7. □

Corollary 1. *The proposed group-sequential weighted parametric multiple testing procedure inherits fully the actual level of significance from its fixed sample counterpart.*

Proof. The proposed group-sequential weighted parametric multiple testing procedure is of size-α, which is the same as that of its fixed-sample weighted parametric multiple testing procedure (which is the weighted Dunnett's step-down procedure). □

Corollary 2. *The proposed group-sequential weighted parametric multiple testing procedure is more powerful than the group-sequential weighted Bonferroni-Holm's multiple testing procedure (Maurer & Bretz 2013), under Holm's weighting scheme, or in general under any other weighting schemes that maintain the monotonicity condition (Bretz et al. 2011, p. 901) in the fixed-sample settings.*

Proof. Under Holm's weighting scheme, the group-sequential weighted Bonferroni-Holm's procedure is a special case of our proposed group-sequential weighted parametric procedure. The former always has the scalar $\xi = 1$ (see 4.1a) as default, while the latter always has the scalar $\xi > 1$ (which is the solution to the equation 4.1a) by fully utilizing the hypothesis-wise dependencies. □

4.5 Simulation Studies

We have carried out some simulation of the Type I error rate and power properties of the two competing procedures, namely, the group-sequential

Holm's procedure (Ye et al. 2012) and the group-sequential parametric procedure we proposed, by testing multiple hypotheses in a two-stage group-sequential setting. The simulation is based on one-sided tests. The nominal FWER α is set at 0.05. For simplicity, we have the following setups: (i) Equi-correlated Wald test statistics is used, with a common correlation coefficient ρ. (ii) Equal-distanced information fractions is assumed (i.e, equal sample sizes for stage 1 and stage 2). (iii) An error spending function of the same type is applied to all the hypotheses. (Note that both of the two competing procedures are well applicable if different types of error spending function are employed for different hypotheses, as long as the type of error spending function is pre-specified for each hypothesis and remains unchanged during the course of analyses).

In this section, power refers to the minimal power, which is defined as the probability of rejecting at least one false null hypotheses. All the simulations conducted use 2 million independent replications per simulation run, and our estimates are likely correct to the third decimal places. Our computation and simulation is carried out using R version 3.0.3. Without exception(s), all boundary values (in terms of nominal significance levels) are conveniently computed by using R function *gsDesign* contained in the R package *gsDesign* (Anderson 2011).

4.5.1 *Group-sequential testing of 3 unequally weighted hypotheses*

This section presents simulation results obtained from a two-stage group-sequential testing of 3 unequally weighted hypotheses. An initial weight vector $\{0.5, 0.3, 0.2\}$, denoting the fractions of significance level for 3 elementary hypotheses, is assigned to $\{H_1, H_2, H_3\}$. Tabulated in Table 4.1 is the simulated Type-I error rate of the group-sequential Holm's procedure. The variant 'GSHv' is used. See Ye et al (2012) who proposed 2 variants, which are 'GSHv' and 'GSHf'. The critical boundary based on 'GSHv' is conveniently obtainable by invoking R function *gsDesign*. Tabulated in Table 4.2 is the simulated Type I error rates of the group-sequential weighted parametric procedure, which is proposed in this paper. Note that the middle column of tables show the Type I error rates under the OBF-type (which is applied to all the 3 hypotheses), and the right column shows the Type-I error rates under the PO-type (which is applied to all the 3 hypotheses).

Table 4.1: Simulated Type I error rates of the group-sequential Holm's procedure

Testing a family of 3 hypotheses, with initial weight vector $\{0.5, 0.3, 0.2\}$, at the nominal significance level $\alpha = 0.05$

$\rho \ \backslash \ Spending\, Function$	OBF-type	PO-type
$\rho = 0.0$	0.04943	0.04908
0.3	0.04685	0.04673
0.5	0.04347	0.04373
0.7	0.03851	0.03871
0.9	0.03032	0.03107

Under the complete null configuration.

Table 4.2: Simulated Type I error rates of the group-sequential weighted parametric procedure

Testing a family of 3 hypotheses, with initial weight vector $\{0.5, 0.3, 0.2\}$, at the nominal significance level $\alpha = 0.05$.

$\rho \quad \backslash \quad SpendingFunction$	OBF-type	PO-type
$\rho = 0.0$	0.05008	0.04991
0.3	0.04987	0.05010
0.5	0.04972	0.05044
0.7	0.05011	0.05065
0.9	0.05022	0.05044

Under the complete null configuration.

Note that on Table 4.1 and Table 4.2, theoretically, the Type I error rates in the "OBF-type" column are equal to that in the "PO-type" column. This conclusion is reached by applying Lemma 5 and the closure testing principle. We can observe that the simulated error rates are very close between the two columns. This observation appears to land support to the validity of Lemma 5.

Tabulated in Table 4.3 and Table 4.4 are the simulated power, respectively of the group-sequential Holm's procedure, and of the group-sequential weighted parametric procedure. In the tables, μ specifies the the mean vector of the alternative distribution, which is tri-variate normal. Note that, for simplicity, the mean is set the same for all the 3 hypotheses.

Table 4.3: Simulated Power of the group-sequential Holm's procedure

Testing a family of 3 hypotheses, with initial weight vector $\{0.5, 0.3, 0.2\}$, at the nominal significance level $\alpha = 0.05$.

ρ \ *Spending Function*	OBF-type			PO-type		
\ *Alternative Hypothesis*	$\mu = 1$	$\mu = 1.5$	$\mu = 2$	$\mu = 1$	$\mu = 1.5$	$\mu = 2$
$\rho = 0.0$	0.54788	0.86602	0.98431	0.49498	0.82392	0.97408
0.3	0.48571	0.79444	0.95520	0.43965	0.75014	0.93729
0.5	0.44130	0.74290	0.92811	0.40015	0.69810	0.90452
0.7	0.39186	0.68439	0.89282	0.35462	0.63979	0.86347
0.9	0.33058	0.61033	0.84108	0.29640	0.56335	0.80570

Table 4.4: Simulated Power of the group-sequential weighted parametric procedure

Testing a family of 3 hypotheses, with initial weight vector $\{0.5, 0.3, 0.2\}$, at the nominal significance level $\alpha = 0.05$.

ρ \ *Spending Function*	OBF-type			PO-type		
\ *Alternative Hypothesis*	$\mu = 1$	$\mu = 1.5$	$\mu = 2$	$\mu = 1$	$\mu = 1.5$	$\mu = 2$
$\rho = 0.0$	0.55118	0.86808	0.98481	0.49841	0.82728	0.97487
0.3	0.49862	0.80384	0.95864	0.45307	0.76083	0.94115
0.5	0.46839	0.76417	0.93721	0.42747	0.72169	0.91590
0.7	0.44004	0.72711	0.91383	0.40247	0.68505	0.88973
0.9	0.41501	0.69228	0.88893	0.37902	0.65050	0.86148

4.5.2 *Group-sequential testing of 8 equally weighted hypotheses*

This section presents simulation results obtained from a two-stage group-sequential testing of 8 equally weighted hypotheses. Tabulated in Table 4.5 and Table 4.6 are the simulated Type-I error rates, respectively of the group-sequential Holm's procedure, and of the group-sequential weighted parametric procedure.

Table 4.5: Simulated Type I error rates of the group-sequential Holm's procedure

Testing a family of 8 equally weighted hypotheses at the nominal significance level $\alpha = 0.05$

$\rho \quad \backslash \quad Spending\,Function$	OBF-type	PO-type
$\rho = 0.0$	0.04920	0.04875
0.3	0.04412	0.04409
0.5	0.03773	0.03804
0.7	0.02914	0.02967
0.9	0.01736	0.01780

Under the complete null configuration.

Table 4.6: Simulated Type I error rates of the group-sequential weighted parametric procedure

Testing a family of 8 equally weighted hypotheses at the nominal significance level $\alpha = 0.05$.

$\rho \quad \backslash \quad SpendingFunction$	OBF-type	PO-type
$\rho = 0.0$	0.04977	0.05039
0.3	0.05007	0.05076
0.5	0.05045	0.05109
0.7	0.05013	0.05133
0.9	0.05016	0.05080

Under the complete null configuration.

Note that on Table 4.5 and Table 4.6, theoretically, the Type I error rates in the "OBF-type" column are equal to that in the "PO-type" column. This conclusion is reached by applying Lemma 5 and the closure testing principle. We can observe that the simulated error rates are very close between the two columns. This observation appears to land support to the validity of Lemma 5.

Tabulated in Table 4.7 and Table 4.8 are the simulated power, respectively of the group sequential Holm's procedure and of the group-sequential weighted parametric procedure. In the table, μ specifies the the mean vector of the alternative distribution, which is 8-variate normal. Note that, for simplicity, the mean is set the same for all the 8 hypotheses. By comparing Table 4.7 and Table 4.8, we can see that when larger number of hypotheses, say 8 in this example, coupled with moderately or highly positive correlations, the power advantage of the parametric procedure over Holm's procedure is very prominent.

Table 4.7: Simulated Power of the group-sequential Holm's procedure

Testing a family of 8 equally weighted hypotheses at the nominal significance level $\alpha = 0.05$.

$\rho \quad \backslash \quad Spending\,Function$	OBF-type			PO-type		
$\backslash \quad Alternative\,Hypothesis$	$\mu = 1$	$\mu = 1.5$	$\mu = 2$	$\mu = 1$	$\mu = 1.5$	$\mu = 2$
$\rho = 0.0$	0.69919	0.96918	0.99963	0.63206	0.94682	0.99893
0.3	0.55015	0.86337	0.98190	0.49832	0.82431	0.97156
0.5	0.45981	0.77587	0.94813	0.41562	0.73146	0.92861
0.7	0.36701	0.67163	0.89155	0.32996	0.62450	0.86203
0.9	0.25642	0.52979	0.78800	0.22592	0.48175	0.74779

Table 4.8: Simulated Power of the group-sequential weighted parametric procedure

Testing a family of 8 equally weighted hypotheses at the nominal significance level $\alpha = 0.05$.

$\rho \quad \backslash \quad Spending\,Function$	OBF-type			PO-type		
$\backslash \quad Alternative\,Hypothesis$	$\mu = 1$	$\mu = 1.5$	$\mu = 2$	$\mu = 1$	$\mu = 1.5$	$\mu = 2$
$\rho = 0.0$	0.70178	0.96981	0.99965	0.63888	0.94880	0.99901
0.3	0.57586	0.87813	0.98478	0.52884	0.84269	0.97628
0.5	0.51535	0.81590	0.96184	0.47246	0.77655	0.94693
0.7	0.46641	0.75808	0.93206	0.42786	0.71777	0.91145
0.9	0.42394	0.70436	0.89739	0.38888	0.66302	0.87157

4.6 Application

It is generally difficult to estimate the hypothesis-wise correlation of test statistics in the case of multiple primary endpoints in clinical trials. However, there are some situations where the hypothesis-wise correlation is structural. e.g. there is a common component among multiple primary endpoints which are themselves composite endpoints. In this case, the hypothesis-wise correlation can be estimated.

The proposed group-sequential parametric stepdown procedure is applicable to the problem of testing multiple hypotheses of the same endpoint, but associated with different populations. Those different populations are overlapped to some degree.

The original (unweighted) Dunnett & Tamhane step-down procedure (1991) deals with the comparison of several treatments with a common control group (placebo or Best Supportive Care), and thus correlation structure can be reliably estimated. Note that, in this multiple testing setting, free-combination condition (Holm, 1979) is satisfied, because the comparisons are not all-pairwise comparisons, but "with a common control". The proposed group-sequential parametric step-down procedure is applicable to the original type of "unweighted" problem, as well as to the type of "weighted" problem. (i.e. a larger weight is allocated to the most promising treatment group). For example, immunotherapy will replace chemo-therapy as the first-line therapy on many types of solid tumor, such as lung cancer. This revolution in cancer treatment started so quickly that it caught people (including competitors) off-guard. Few has thought of it possible 2 years ago. As a result, some nonregistration chemotherapy trials on solid tumor, which have already started out but have not yet enrolled too many subjects, begin adding an immuno-therapy as a new treatment arm. As a practice in the clinical trial industry, the study protocol will be amended for any addition of treatment arm(s), and such amendment is legitimate before data lock. Operationally speaking, for a nonregistration late-stage trial that has the discipline of implementing multiplicity adjustment, it is still feasible to add a new treatment arm to act upon the latest information, supposed that the trial did not start long ago.

4.7 Concluding Remarks

Where hypothesis-wise dependency structure of test statistics is fully estimable or known *a priori*, the proposed group-sequential parametric stepdown procedure can be applied. For a given level of significance α, the scalar ξ_I (the solution to equation 4.1a) reflects (or quantifies) the magnitude of overall hypothesis-wise correlation, and such a summary quantity ξ_I translates into a 'sized-up' group-sequential critical region (nominal significance level per Maurer & Bretz 2013) for the group-sequential setting. Where the dependency structure is not perfectly known, but partly known for some pairs or blocks, Seneta & Chen's procedure (2005), which is a variant of the Holm's procedure with the incorporation of pairwise dependencies, can be applied for testing m equally weighted multiple endpoints, and Seneta & Chen's procedure can be naturally group-sequentialized in the same way as shown in this chapter.

Virtually all marginal p-value-based, FWER-controlling stepdown multiple testing procedures for the fixed-sample testing can be, by the use of some commonly-used error spending functions, naturally group-sequentialized for application in the group-sequential setting. The group-sequential version retains a short-cut (consonancy) of the closed testing procedure, and results in having exactly the same *actual* level of significance as that of the fixed-sample reference procedure.

4.8 Proof of a Fundamental Result

Result 1 of §4.4.3 The equivalency of two events A and B is defined as: Event A occurs if and only if event B occurs. That is, $A \equiv B$ iff $pr(A \cap \bar{B}) = 0$ and $pr(B \cap \bar{A}) = 0$. Without loss of generality, we prove that: if $A_i \equiv B_i$ and

$A_j \equiv B_j$, then
(i) $(A_i \cup A_j) \equiv (B_i \cup B_j)$.
(ii) $(A_i \cap A_j) \equiv (B_i \cap B_j)$.
For (i), we show that:

$$\begin{aligned}
&pr\{(A_i \cup A_j) \cap \overline{B_i \cup B_j}\} \\
&= pr\{(A_i \cap \overline{B_i \cup B_j}) \cup (A_j \cap \overline{B_i \cup B_j})\} \\
&\leq pr\{A_i \cap \overline{B_i \cup B_j}\} + pr\{A_j \cap \overline{B_i \cup B_j}\} \\
&\quad = pr(A_i \cap \bar{B}_i \cap \bar{B}_j) + pr(A_j \cap \bar{B}_i \cap \bar{B}_j) \\
&\quad = pr(\emptyset \cap \bar{B}_j) + pr(\emptyset \cap \bar{B}_i) \\
&\quad = 0 + 0 \\
&\quad = 0
\end{aligned}$$

Similarly, it can be shown that $pr\{(B_i \cup B_j) \cap \overline{A_i \cup A_j}\} = 0$.

For (ii), we use the result (i), and an easily verifiable result that the complements of two equivalent events are also equivalent.
From $A_i \equiv B_i$, we have $\overline{A_i} \equiv \overline{B_i}$, and from $A_j \equiv B_j$, we have $\overline{A_j} \equiv \overline{B_j}$. Then, we have

$$\begin{aligned}
&(\overline{A_i} \cup \overline{A_j}) \equiv (\overline{B_i} \cup \overline{B_j}) \\
\Rightarrow \quad &\overline{\overline{A_i} \cup \overline{A_j}} \equiv \overline{\overline{B_i} \cup \overline{B_j}} \\
\Rightarrow \quad &A_i \cap A_j \equiv B_i \cap B_j
\end{aligned}$$

CHAPTER 5

EXTENDING STEPUP MULTIPLE TESTING PROCEDURES TO GROUP-SEQUENTIAL SETTING

5.1 Research Objectives and Findings

In this chapter, we propose a group-sequential version of Hochberg's stepup procedure (1988) for testing n equally weighted hypotheses. The proposed procedure is in essence the stepup analogue of the recently proposed group-sequential Holm's stepdown procedure (Ye et al. 2012), and is thus more powerful. This stepup analogue nicely retains a shortcut (which is consonancy) of closed testing procedures in group-sequential setting, and also maintains a strong control of the FWER under some commonly encountered non-negative hypothesis-wise dependency structures of test statistics.

5.2 The Stepup Analogue of the Group-sequential Holm's Stepdown Procedure

Consider the problem of testing n equally weighted hypotheses (endpoints) in a two-stage group-sequential setting, targeting the control of the family-wise error rate at α in the strong sense. As usual, information fractions are assumed given. Let $\mathbb{A}$ be the error spending function employed for each of the n hypotheses. And let the vector $\{\mathbb{A}_1(\alpha), \mathbb{A}_2(\alpha)\}$ be the $\mathbb{A}$-induced, level-α, univariate critical boundary, which is the same for all the H_i's. (because of the same error spending function being used). We'll show that a new procedure, which we call group-sequential Hochberg's procedure, can be simply constructed using the same critical boundary as that of the group-sequential Holm's step-down procedure (Ye et al. 2012).

A two-stage group-sequential Holm's procedure would be conducted in the followings.

At analysis 1, start testing the n hypotheses. Let $p_{(1),1} < p_{(2),1} < \cdots < p_{(n),1}$ be the ordered p-values and $H_{(1)}, \dots, H_{(n)}$ be their corresponding hypotheses. Reject $H_{(i)}$ when, for all $k = 1, \dots, i$,

$$p_{(k),\,1} \leq \mathbb{A}_1 \left(\frac{\alpha}{n-k+1} \right) \tag{5.1a}$$

where $\mathbb{A}_1 \left(\frac{\alpha}{n-k+1} \right)$ is the first component of the $\mathbb{A}$-induced critical boundary with a level of $\frac{\alpha}{n-k+1}$. Let r_1 denote the number of hypotheses rejected for analysis 1. Denote by $n_2 = n - r_1$ the number of hypotheses not-yet rejected.

For analysis 2, test again those un-rejected n_2 hypotheses. Let $p_{(1),2} < p_{(2),2} \cdots < p_{(n_2),2}$ be the n_2 p-values (based on cumulative data) re-ordered for analysis 2, and $H_{(1)}, \dots, H_{(n_2)}$ be their corresponding hypotheses. Reject $H_{(i)}$ when, for all $k = 1, \dots, i$,

$$p_{(k),\,2} \leq \mathbb{A}_2 \left(\frac{\alpha}{n_2-k+1} \right) \tag{5.1b}$$

where $\mathbb{A}_2\left(\frac{\alpha}{n_2-k+1}\right)$ is the second component of the $\mathbb{A}$-induced critical boundary with a level of $\frac{\alpha}{n_2-k+1}$. Let r_2 denote the number of hypotheses rejected for stage 2. The total number of hypotheses thus rejected is $r_1 + r_2$.

Using the same critical boundary, we propose an alternative procedure, which is conducted in the step-up fashion as described in the followings.

For analysis 1, for any $i = n, n-1, \ldots, 1$, if

$$p_{(i),1} \leq \mathbb{A}_1\left(\frac{\alpha}{n-i+1}\right) \tag{5.1c}$$

then reject all $H_{(k)}$ with $k \leq i$. Let r_1 denote the number of hypotheses rejected for analysis 1. Denote by $n_2 = n - r_1$ the number of hypotheses not-yet rejected.

For analysis 2, for any $i = n_2, n_2 - 1, \ldots, 1$, if

$$p_{(i),2} \leq \mathbb{A}_2\left(\frac{\alpha}{n_2-i+1}\right) \tag{5.1d}$$

then reject all $H_{(k)}$ with $k \leq i$. The total number of rejections r_1+r_2, made by the group-sequential step-up analogue, will be equal or greater than that by the step-down procedure, since the rejection region of the step-down procedure is a subset of that of its step-up analogue.

5.2.1 *A group-sequential Simes' test for intersection hypotheses*

Recall the fixed-sample Simes' test for the global null hypothesis $H_0 = \cap_{i\in I} H_i$ at level-α, where $H_1, \ldots, H_m$ is a collection of m null hypotheses, and $I = \{1, \ldots, m\}$. Let $p_{(1)} < p_{(2)} \cdots < p_{(m)}$ be the ordered marginal p-values with $H_{(i)}$ corresponding to $p_{(i)}$, $i \in I$. The fixed-sample Simes' test rejects H_0 if, for any $i \in I$, $p_{(i)} \leq i\alpha/m$. Simes (1986) proved that it is a size-α test for independent test statistics. And Sarkar & Chang (1997) and Sarkar (1998) theoretically proved that the size of the test is less than α for some positively dependent test statistics that arise in many multiple-hypothesis testing situations.

Let's consider an appropriate extension of Simes' test for the global null hypothesis H_0 into a two-stage group-sequential setting, where information fractions for the 2 stages (analyses) are assumed given, which can be equal-distanced or not. For notational simplicity, we assume for now that a single spending function $\mathbb{A}$ is employed for all the individual hypotheses $H_i, i \in I$. Let the vector $\{\mathbb{A}_1(\alpha), \mathbb{A}_2(\alpha)\}$ be the $\mathbb{A}$-induced, size-α, univariate (single-hypothesis) critical boundary, which is the same for all H_i's , obviously because of the same error spending function being used. Let $p_{(1),1} < p_{(2),1} < \cdots < p_{(m),1}$ be the m observed p-values ordered for the first analysis, with $H_{(i)}$ corresponding to $p_{(i),1}$. Let $p_{(1),2} < p_{(2),2} < \cdots < p_{(m),2}$ be the m observed p-values (based on *cumulative* data) re-ordered for the second (final) analysis, with $H_{(i)}$ corresponding to $p_{(i),2}$. Note that a $H_{(i)}$ for analysis 1 is generally not the same hypothesis as the $H_{(i)}$ for analysis 2. We herein describe a two-stage group-sequential Simes' test of H_0 as follows. The test rejects H_0 if, for any $i \in I$, $p_{(i),1} \leq \mathbb{A}_1(i\alpha/m)$ or $p_{(i),2} \leq \mathbb{A}_2(i\alpha/m)$. Otherwise, the test accepts H_0.

Lemma 8. *The above described two-stage group-sequential Simes' test for the global null hypothesis has a size less than that of its fixed-sample counterpart.*

For convenience, the following notations will be used for the proof.

Notation 3.

For the level-α fixed-sample Simes' test of H_0, the critical region is $\bigcup_{i=1}^m R_{(i)}(i\alpha/m)$, where $R_{(i)}(i\alpha/m) = \{(p_1, \ldots, p_m) \in (0,1)^m : P_{(i)} \leq i\alpha/m\}$.

For the corresponding two-stage group-sequential Simes' test, the critical region is $\{\bigcup_{i=1}^m S_{(i),1}(i\alpha/m)\} \cup \{\bigcup_{i=1}^m S_{(i),2}(i\alpha/m)\}$, where
$S_{(i),1}(i\alpha/m) = \{(p_{1,1}, \ldots, p_{m,1}, p_{1,2}, \ldots, p_{m,2}) \in (0,1)^{2m} : P_{(i),1} \leq \mathbb{A}_1(i\alpha/m)\}$ and $S_{(i),2}(i\alpha/m) = \{(p_{1,1}, \ldots, p_{m,1}, p_{1,2}, \ldots, p_{m,2}) \in (0,1)^{2m} : P_{(i),2} \leq \mathbb{A}_2(i\alpha/m)\}$.

To facilitate the proof, we shall introduce a concept of "proper subset in extended sense" as follows.

Definition 1. *Let $A_1 \subset A_2 \cdots \subset A_i \cdots \subset A_n \subset \mathbb{R}^p$ and $B_1 \subset B_2 \cdots \subset B_i \cdots \subset B_n \subset \mathbb{R}^q$, where $\mathbb{R}^p$ is a p-dimensional real space, and $\mathbb{R}^q$ is a q-*

dimensional real space, and $p \neq q$. *Suppose that* $A_i \equiv B_i$, $\forall\, i = 1, \ldots, n$. *Then, with slight abuse of notation, if* $i' < i''$, *we denote* $A_{i'} \subset^* B_{i''}$ *or* $B_{i'} \subset^* A_{i''}$.

It's easy to see that if $A_{i'} \subset^* B_{i''}$, then $pr(A_{i'}) < pr(B_{i''})$. The following type of set operation will be also used in the proof:

$$\left\{ \bigcap_{i=1}^{m} A_i \;\cup\; \bigcap_{i=1}^{m} B_i \right\} \subset \left\{ \bigcap_{i=1}^{m} \{A_i \cup B_i\} \right\}$$

Note that a $H_{(i)}$ for analysis 1 is generally not the same hypothesis as the $H_{(i)}$ for analysis 2.

Proof. With the use of the notations and the definition introduced above, now comes the proof for Lemma 8. We need to prove that:

$$\left\{ \bigcup_{i=1}^{m} S_{(i),1}(i\alpha/m) \quad \cup \quad \bigcup_{i=1}^{m} S_{(i),2}(i\alpha/m) \right\} \quad \subset^* \quad \bigcup_{i=1}^{m} R_{(i)}(i\alpha/m)$$

Specifically, we need to prove the following 3 cases.

(i) For $i = 1$, $\{S_{(1),1}(\alpha/m) \cup S_{(1),2}(\alpha/m)\} \equiv R_{(1)}(\alpha/m)$. For notational convenience, the argument (α/m) will be omitted in subsequent expressions.

(ii) For $i = m$, $\{S_{(m),1}(\alpha) \cup S_{(m),2}(\alpha)\} \subset^* R_{(m)}(\alpha)$. The argument (α) will be omitted in subsequent expressions.

(iii) For any $i = 2, \ldots, m-1$, $\{S_{(i),1}(i\alpha/m) \cup S_{(i),2}(i\alpha/m)\} \subset^* R_{(i)}(i\alpha/m)$. The argument $(i\alpha/m)$ is omitted in subsequent expressions.

Proof for case (i):

$$
\begin{aligned}
& S_{(1),1} \;\cup\; S_{(1),2} \\
&= \bigcup_{k=1}^{m} S_{k,1} \;\cup\; \bigcup_{k=1}^{m} S_{k,2} \\
&= \bigcup_{k'=1}^{m} \{S_{k',1} \cup S_{k',2}\} \\
&\equiv \bigcup_{k'=1}^{m} R_{k'} \\
&= R_{(1)}
\end{aligned}
$$

Proof for case (ii):

$$
\begin{aligned}
& S_{(m),1} \;\cup\; S_{(m),2} \\
&= \bigcap_{k=1}^{m} S_{k,1} \;\cup\; \bigcap_{k=1}^{m} S_{k,2} \\
&\subset \bigcap_{k'=1}^{m} \{S_{k',1} \cup S_{k',2}\} \\
&\equiv \bigcap_{k'=1}^{m} R_{k'} \\
&= R_{(m)}
\end{aligned}
$$

Proof for case (iii): We resort to a 'representative' configuration. For a given i (assuming i is fixed), the total number of unique configurations is $\binom{m}{i}$.

$$\begin{aligned} & S_{(i),1} \;\cup\; S_{(i),2} \\ = & \bigcap_{k=1}^{i} S_{k,1} \;\cup\; \bigcap_{k=1}^{i} S_{k,2} \\ \subset & \bigcap_{k'=1}^{i} \{S_{k',1} \cup S_{k',2}\} \\ \equiv & \bigcap_{k'=1}^{i} R_{k'} \\ = & R_{(i)} \end{aligned}$$

Thus, it is established that, for $m \geq 2$, the critical region of the two-stage group-sequential Simes' test of the global null hypothesis H_0 as described above is a proper subset, in the extended sense, of that of the level-α fixed-sample Simes' test of H_0. It follows that the size of the group-sequential Simes' test is smaller than its fixed-sample counterpart. □

Corollary 3. *If the hypothesis-wise test statistics are independent, or of some positive dependency structures as studied in Sarkar & Chang (1997), the above constructed two-stage group-sequential Simes' test for the global null hypothesis (weakly) controls the Type I error rate at level α.*

5.2.2 *A group-sequential Hochberg's stepup multiple testing procedure*

Theorem 4. *If test statistics are hypothesis-wise independent, or of some positive dependency structures as studied in Sarkar & Chang (1997), the proposed $\mathbb{A}$-employed two-stage group-sequential Hochberg's step-up procedure (5.1c and 5.1d) for testing $n\,(\geq 2)$ hypotheses controls the familywise error rate at level α in the strong sense, where $\mathbb{A}$ is the pre-specified error spending function employed for each $H_i, i = 1, \ldots, n$.*

Proof. We first show that the fixed-sample Hochberg's step-up procedure for testing n hypotheses strongly controls the family-wise error rate (FWER) at level α. Then, in a similar fashion, we show that the two-stage group-sequential Hochberg's step-up procedure strongly controls the FWER at level α.

The fixed-sample Hochberg step-up procedure can be seen as applying the closure principle to Simes' method. Here is to show that fixed-sample Hochberg's step-up procedure controls the FWER strongly at level α as the follows. As usual, order the observed p-values, $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$, with the corresponding hypotheses being $H_{(1)}, H_{(2)} \ldots H_{(n)}$. Let $i = \max\{1 \leq l \leq n : p_{(l)} \leq \frac{\alpha}{n-l+1}\}$.
To prove this, it suffices to establish that $H_{(i)}$ is rejected. (Here, one can convince oneself: if Simes' test rejects all H_J with $i \in J$, $J \subseteq I$, then, if $p_{(i')} \leq p_{(i)}$, Simes' test would as well reject all $H_{J'}$ with $i' \in J'$, $J' \subseteq I$)
Consider such a J and let

$$m_0 = |\{k \in J : p_{(k)} \leq p_{(i)}\}| \geq 1.$$

Simes' test will certainly reject H_J if

$$p_{(i)} \leq \alpha m_0/|J|,$$

and the conclusion follows from

$$\frac{\alpha}{n-i+1} \leq \alpha \frac{m_0}{|J|}$$

since the above inequality is equivalent to $(|J| - m_0) \leq m_0(n-i)$, which holds since $|J| - m_0 \leq n - i$.

Here is to show that the two-stage group-sequential Hochberg's step-up procedure strongly controls the FWER at level α.
Let $i = \max\{1 \leq l \leq n : p_{(l),1} \leq \mathbb{A}_1\left(\frac{\alpha}{n-l+1}\right)\}$. If there does not exist such an i at Analysis 1, then go testing at Analysis 2. If there exists such an i at Analysis 1, it suffices to establish that $H_{(i)}$ is rejected (by the same reasoning given above for the fixed-sample proof. That is, all $H_{(i')}$ whose $p_{(i'),1} \leq p_{(i),1}$

will also be rejected). Hence we need show that group-sequential Simes' test rejects all H_J, with $i \in J$, $J \subseteq I$.

Consider such a J, and let

$$m_{0,1} = |\{k \in J : p_{(k),1} \leq p_{(i),1}\}| \geq 1.$$

Group-sequential Simes' test will certainly reject H_J if

$$p_{(i),1} \leq \mathbb{A}_1 \left(\alpha m_{0,1}/|J|\right),$$

and the conclusion follows from

$$\mathbb{A}_1 \left(\frac{\alpha}{n-i+1}\right) \leq \mathbb{A}_1 \left(\alpha \frac{m_{0,1}}{|J|}\right)$$

since the above inequality is equivalent to $(|J| - m_{0,1}) \leq m_{0,1}(n-i)$, which holds since $|J| - m_{0,1} \leq n - i$.

Now let $n_2 = n - i$, that is, n_2 is the number of un-rejected hypotheses, which are subject to retesting at Analysis 2. Let I_2 be the index set of those n_2 hypotheses.

Let $i_2 = \max\{1 \leq l_2 \leq n_2 : p_{(l_2),2} \leq \mathbb{A}_2 \left(\frac{\alpha}{n_2 - l_2 + 1}\right)\}$. If there exists such an i_2, it suffices to establish that H_{i_2} is rejected. Hence, we need to show that group-sequential Simes' test rejects all H_{J_2}, with $i_2 \in J_2$, $J_2 \subseteq I_2$.

Consider such a J_2, with $J_2 \subseteq I_2$, and let

$$m_{0,2} = |\{k_2 \in J_2 : p_{(k_2),2} \leq p_{(i_2),2}\}| \geq 1.$$

Group-sequential Simes' test will certainly reject H_{J_2} if

$$p_{(i_2),2} \leq \mathbb{A}_2 \left(\alpha m_{0,2}/|J_2|\right),$$

and the conclusion follows from

$$\mathbb{A}_2 \left(\frac{\alpha}{n_2 - i_2 + 1}\right) \leq \mathbb{A}_2 \left(\alpha \frac{m_{0,2}}{|J_2|}\right)$$

since the above inequality is equivalent to $(|J_2| - m_{0,2}) \leq m_{0,2}(n_2 - i_2)$, which holds since $|J_2| - m_{0,2} \leq n_2 - i_2$.

Lastly, one can convince oneself that if group-sequential Simes' test rejects all H_{J_2}, with $i_2 \in J_2$, $J_2 \subseteq I_2$, then group-sequential Simes' test will of course reject all H_J, with $i_2 \in J$, $J \subseteq I$. □

Remark 9. *From the above proof, one can see that group-sequential Hochberg step-up procedure retains a shortcut (which is consonancy) of the closed testing procedure in group-sequential setting.*

5.3 The general case of employing different types of error spending functions for different hypotheses

The development in §5.2 is based on the assumption of a single type of error spending function being employed for all of the n hypotheses. This assumption can be dispensed with. In the general case where different types of spending functions are employed for different hypotheses, the results in §5.2 hold valid as well.

Consider a two-stage group-sequential Simes' test for a global null hypothesis $H_0 = \cap_{i \in I} H_i$, where I is the set of indices of n null hypotheses. Further, suppose different types of spending functions are employed for different hypotheses, with $\mathbb{F}^i$ prescribed for $H_i, i = 1, 2, \ldots, n$. Let $\{\mathbb{F}_1^i(\alpha), \mathbb{F}_2^i(\alpha)\}$ be the $\mathbb{F}^i$-induced level-α group-sequential boundary for H_i, where $\mathbb{F}_1^i(\alpha)$ is the component boundary for stage 1, and $\mathbb{F}_2^i(\alpha)$ for stage 2, for $i = 1, 2, \ldots, n$.

We need to introduce the notation $p_{(j),1} \leq \mathbb{F}_1^{(j)}(j\alpha/n)$ as the followings.

Notation 4. *Let $\{p_{(i),1} \leq \mathbb{F}_1^{(i)}(i\alpha/n)\}$ denote $\{p_{k,1} : \sum_{k=1}^n \mathcal{I}(p_{k,1} \leq \mathbb{F}_1^k(i\alpha/n)) \geq i\}$, where $\mathcal{I}(\cdot)$ is an indicator function. Similarly, let $\{p_{(i),2} \leq \mathbb{F}_2^{(i)}(i\alpha/n)\}$ denote $\{p_{k,2} : \sum_{k=1}^n \mathcal{I}(p_{k,2} \leq \mathbb{F}_2^k(i\alpha/n)) \geq i\}$.*

(For example, here is a alternative way of understanding the notation using the concept of 'event': when $i = 1$, the event $p_{(1),1} \leq \mathbb{F}_1^{(1)}(\alpha/n)$ is evaluated

true if there exist at least one i's such that the event $p_{i,1} \leq F_1^i(\alpha/n)$ *is true. When* $i = 2$*, the event* $p_{(2),1} \leq \mathbb{F}_1^{(2)}(2\alpha/n)$ *is evaluated true if there exist at least 2 i's such that the event* $p_{i,1} \leq F_1^i(2\alpha/n)$ *is evaluated true. etc.)*

Consider constructing a two-stage group sequential Simes' test for the global null hypothesis $H_0 = \cap_{i \in I} H_i$, with error spending function $\mathbb{F}^i$ for $H_i, i \in I = \{1, 2, \ldots, n\}$. The group-sequential Simes' test rejects H_0 if, for any $i \in I$, $p_{(i),1} \leq \mathbb{F}_1^{(i)}(i\alpha/n)$ or $p_{(i),2} \leq \mathbb{F}_2^{(i)}(i\alpha/n)$. Otherwise, the test accepts H_0.

Lemma 9. *The two-stage group-sequential Simes' test, as constructed above with the employment of different types of spending functions for different hypotheses, for testing the intersection null hypothesis* $H_0 = \cap_{i \in I} H_i$*, controls the Type I error rate at level* α*. The size of the group-sequential Simes' test is less than that of its fixed-sample counterpart.*

Proof. The equivalency of events as established in *Result 1* in §4.4.3 holds valid for each H_i, irrespective of which type of spending function being used. Thus, proof is similar as that for Lemma 8. □

Consider the problem of testing of n multiple hypotheses in a two-stage group-sequential setting, where different types of error spending functions are employed for different hypotheses to suit a specific study objective. We shall propose a more generalized group-sequential step-up procedure. We first clarify a notation as followings.

Notation 5. *For a given i, Let* $\{p_{(i),1} \leq \mathbb{F}_1^{(i)}(\alpha/(n-i+1))\}$ *denote* $\{p_{k,1} : \sum_{k=1}^{n} \mathcal{I}(p_{k,1} \leq \mathbb{F}_1^k(\alpha/(n-i+1))) \geq i\}$*, where* $\mathcal{I}(\cdot)$ *is an indicator function. Similarly, let* $\{p_{(i),2} \leq \mathbb{F}_2^{(i)}(\alpha/(n_2-i+1))\}$ *denote* $\{\sum_{k=1}^{n_2} \mathcal{I}(p_{k,2} \leq \mathbb{F}_2^k(\alpha/(n_2 - i+1))) \geq i\}$*. Note that* n_2 *is the number of non-yet rejected hypotheses by stage 1's testing , and thus subject to re-testing at stage 2.*

We propose a more generalized group-sequential Hochberg's stepup procedure as the followings.

For stage 1, for any $i = n, n-1, \dots, 1$, if

$$p_{(i),1} \leq \mathbb{F}_1^{(i)}\left(\frac{\alpha}{n-i+1}\right) \tag{5.2a}$$

then reject all H_k whose $p_{k,1}$ is in the set $\{p_{k,1} : \sum_{k=1}^{n} \mathcal{I}(p_{k,1} \leq \mathbb{F}_1^k(\alpha/(n-i+1))) \geq i\}$.

Now let n_2 be the number of not-yet rejected hypotheses. For stage 2, for any $i = n_2, n_2 - 1, \dots, 1$, if

$$P_{(i),2} \leq \mathbb{F}_2^{(i)}\left(\frac{\alpha}{n_2-i+1}\right) \tag{5.2b}$$

then reject all H_k whose $p_{k,2}$ is in the set $\{p_{k,2} : \sum_{k=1}^{n2} \mathcal{I}(p_{k,2} \leq \mathbb{F}_2^k(\alpha/(n2-i+1))) \geq i\}$.

Theorem 5. *The proposed two-stage group sequential Hochberg's stepup procedure (5.2a and 5.2b) for testing n multiple hypotheses controls the family-wise error rate strongly at level α.*

Proof. Similar to the proof for Theorem 4. □

5.4 A Multi-stage Group-sequential Stepup Multiple Testing Procedure

It can be seen that, the constructed two-stage group-sequential Simes' test (in §5.3) for an intersection hypothesis can be easily generalized, and be applicable in a multi-stage group-sequential setting. Thus, the proposed two-stage group-sequential Hochberg's stepup procedure (in §5.3) can be easily generalized, and be applicable in a multi-stage group-sequential setting.

5.5 Simulation Studies

We have carried out some simulations of the Type I error rate and power properties of the competing procedures by testing a family of 3 correlated hypotheses in a two-stage group-sequential setting. The simulation is based on the one-sided tests. The significance level α was set at 0.05. For simplicity, we have the following setups: (i) Equi-correlated Wald test statistics is used, with a common correlation coefficient ρ. (ii) Equal-distanced information fractions is assumed (i.e, equal sample sizes for stage 1 and stage 2). (iii) A single spending function of the same type is employed to all the hypotheses in the family.

In this section, power refers to the minimal power, which is defined as the probability of rejecting at least one false null hypotheses. All the simulations conducted use 2 million independent replications per simulation run, and our estimates are likely correct to the third decimal places. Our computation and simulation is carried out using R version 3.0.3. And without any exception, all boundary values (nominal significance levels) are conveniently computed by using R function *gsDesign* contained in the R package *gsDesign* (Anderson 2011).

Tabulated in Table 5.1 is the simulated Type I error rates of the two-stage group-sequential Simes' test for a 3-hypotheses intersection hypothesis. The OBF-type and the PO-type of spending functions are respectively employed in the group-sequentializing of the Simes' test.

Table 5.1: Simulated Type I error rates of the group-sequential Simes' test

Testing an intersection null hypothesis $H_0 = \bigcap_{i \in I} H_i$ at level $\alpha = 0.05$, where $|I| = 3$.

$\rho \quad \backslash \quad SpendingFunction$	OBF-type	PO-type
$\rho = 0.0$	0.04986	0.04945
0.3	0.04793	0.04772
0.5	0.04563	0.04553
0.7	0.04221	0.04192
0.9	0.03731	0.03690

Tabulated in Table 5.2 and Table 5.3 are the simulated Type I error rate, respectively of the stepdown, and of the stepup analogue. Note that Table 5.2 and Table 5.3 are based on the same critical boundary. Note that, for the stepdown, the variant 'GSHv' is used. See Ye et al (2012) who proposed 2 variants, which are 'GSHv' and 'GSHf'. The critical boundary based on 'GSHv' is conveniently obtainable by invoking R function *gsDesign*.

Table 5.2: Simulated Type I error rates of the group-sequential Holm's procedure

Testing a family of 3 null hypotheses at level $\alpha = 0.05$.

$\rho \quad \backslash \quad Spending\,Function$	OBF-type	PO-type
$\rho = 0.0$	0.04900	0.04881
0.3	0.04623	0.04646
0.5	0.04303	0.04338
0.7	0.03773	0.03785
0.9	0.02854	0.02907

Footnote: Under the complete null configuration.

Table 5.3: Simulated Type I error rates of the group-sequential Hochberg's stepup procedure

Testing a family of 3 null hypotheses at level $\alpha = 0.05$.

$\rho \quad \backslash \quad SpendingFunction$	OBF-type	PO-type
$\rho = 0.0$	0.04957	0.04919
0.3	0.04746	0.04700
0.5	0.04420	0.04408
0.7	0.03976	0.03969
0.9	0.03485	0.03445

Footnote: Under the complete null configuration.

Tabulated in Table 5.4 and Table 5.5 are the simulated power, respectively of the stepdown, and of the stepup analogue. In the table, μ specifies the the mean vector of the alternative distribution, which is tri-variate normal. Note that the mean is set the same for all the 3 hypotheses.

Table 5.4: Simulated Power of the group-sequential Holm's stepdown procedure

Testing a family of 3 alternative hypotheses at level $\alpha = 0.05$.

$\rho \quad \backslash \quad$ *Spending Function*	OBF-type			PO-type		
$\backslash \quad$ *Alternative Hypothesis*	$\mu = 1$	$\mu = 1.5$	$\mu = 2$	$\mu = 1$	$\mu = 1.5$	$\mu = 2$
$\rho = 0.0$	0.55395	0.87114	0.98545	0.50015	0.82908	0.97563
0.3	0.48879	0.79796	0.95726	0.44188	0.75358	0.93888
0.5	0.44261	0.74563	0.92922	0.40106	0.70026	0.90606
0.7	0.39112	0.68521	0.89360	0.35326	0.63868	0.86410
0.9	0.32223	0.60235	0.83744	0.28845	0.55507	0.80094

Table 5.5: Simulated Power of the group-sequential Hochberg's stepup procedure

Testing a family of 3 alternative hypotheses at level $\alpha = 0.05$.

$\rho \quad \backslash \quad$ *Spending Function*	OBF-type			PO-type		
$\backslash \quad$ *Alternative Hypothesis*	$\mu = 1$	$\mu = 1.5$	$\mu = 2$	$\mu = 1$	$\mu = 1.5$	$\mu = 2$
$\rho = 0.0$	0.56332	0.88030	0.98805	0.50533	0.83748	0.97875
0.3	0.49869	0.80881	0.96164	0.45084	0.76319	0.94409
0.5	0.45448	0.75764	0.93605	0.41073	0.71228	0.91355
0.7	0.40703	0.70247	0.90326	0.36747	0.65589	0.87530
0.9	0.35585	0.63902	0.86089	0.31966	0.59244	0.82748

5.6 Application

The proposed group-sequential Hochberg's stepup procedure can be applied to the problem of testing a family of multiple primary endpoints, if test statistics associated with those endpoints are known to be independent or positively dependent. It can also be applied to the problem of testing multiple hypotheses of the same endpoint associated with different populations. See Ye et al 2012.

Empirically, based on some simulations (Gou & Tamhane, 2013), the application of Simes' test can be greatly expanded to many types of dependency structure of test statistics, including the negative dependency, if the number of hypotheses is 4 or greater. We have conducted extensive independent simulations. Our simulation study lands support for such empirical findings.

The proposed group-sequential procedure can be applied to a type of a commonly seen confirmatory clinical trial, which is typically powered by one primary endpoint, but there are a number of secondary endpoints, which are often non-negatively correlated. The set of secondary hypotheses will be tested at stage j (based on accumulated data up to stage j) only after the primary endpoint is rejected by stage j's test. And it is desirable to reject as many secondary hypotheses as possible (for drug label claims). This application will be elaborated in future communications.

5.7 Concluding Remarks

If test statistics are hypothesis-wise independent or positively dependent, the group-sequential Hochberg's stepup procedure shall be preferred to the group-sequential Holm's stepdown procedure for the testing of n unweighted hypotheses. Power improvement, as shown by simulation studies, is meaningful for moderate or high degree of positive dependency, and marginal for independency or weak dependency. This resembles, to some extent, the relative power advantage of Hochberg's stepup procedure over Holm's stepdown for the fixed-sample testing situations. However, the power advantage (Hochberg's over Holm's) in the applicable fixed-sample setting can be larger (and cannot be smaller) as opposed to that of the group-sequential setting, because the rejection region of the group-sequential Simes' test of intersection hypothesis is a proper subset, in the extended sense, of that of the fixed-sample Simes' test. From this perspective, the group-sequential Hochberg's stepup procedure inherits *partially* the actual level of significance from its fixed-sample counterpart, though still being more powerful than the group-sequential Holm's procedure.

CHAPTER 6

FUTURE RESEARCH

6.1 Extending the Modified Seneta-Chen Procedure to Group-sequential Setting

6.1.1 *Research objectives and findings*

The modified Seneta-Chen procedures as proposed in Chapter 3 can be group-sequentialized for application. The group sequential version of the modified Seneta-Chen procedure is a viable alternative to the fully-parametric stepdown procedure if the hypothesis-wise dependency structure is not fully known. The group sequential Seneta-Chen procedure belongs to the class of group sequential stepdown procedure, thus, it retains the shortcut of closed test procedures and it inherits fully the *actual* significance level from its fixed-sample counterpart.

6.2 Properties pertaining to the Distribution of Null P-values

The convexity property for the distribution of the maxima of a set of null p-values in higher dimensional space will be explored. Conditional distribution will be explored, for possible applications in adaptive trials.

REFERENCES

[1] Armitage, P. and McPherson, C. and Rowe, B. (1969). Repeated significance tests on accumulating data. *J.R.Statist.Soc.A*, **132**, 235-244.

[2] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J.R.Statist.Soc.B*, **57**, 289-300.

[3] Benjamini, Y. and 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**, 60-83.

[4] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann.Statist.*, **29**, 1165-1188.

[5] Bretz, F. and Posch, M. and Glimm, E. and Klinglmueller, F. and Maurer, W. and Rohmeyer, K. (2011) Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes or parametric tests. *Biometrical Journal*, **53**, 894-913.

[6] Bretz, F. and Hothorn, T. and Westfall, P. (2011). Multiple comparisons using R.

[7] Dunnett, C. and Tamhane, A. (1991). Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Statistics in Medicine*, **10**, 939-947.

[8] Dunnett, C. and Tamhane, A. (1992). A step-up multiple test procedure. *J.Amer.Statist.Assoc.*, **87**, 162-170.

[9] Dunnett, C. and Tamhane, A. (1995). Step-up multiple testing of parameters with unequally correlated estimates. *Biometrics*, **51**, 217-227.

[10] Duoit, S. & Van Der Lann, M. (2008). Multiple Testing Procedures with Applications to Genomics. Springer, New York.

[11] Finner, H. and Strassburger,K. (2002). The partitioning principle: a powerful tool in multiple decision theory. *Ann.Statist.*, **30**, 1194-1213.

[12] Finner, H. and Gontscharuk, V. (2009). Controlling the familywise error rate with plug-in estimator for the proportion of true null hypotheses. *J.R.Statist.Soc.B*,**71**, 1031-1048.

[13] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

[14] Gou, J., Tamhane, A., Xi, D. and Rom, D. (2014). A class of improved hybrid Hochberg-Hommel type step-up multiple test procedures. *Biometrika*, **101**, 899-911.

[15] Guo, W.(2009). A note on adaptive Bonferroni and Holm procedures under dependence. *Biometrika*, **96**, 1012-1018.

[16] Hochberg, Y. and Tamhane, A. (1987). Multiple comparison Procedures. John Wiley & Sons, Inc.

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

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

[19] Jennison, C. and Turnbull, B.W. (2000). Group sequential methods applications to clinical trials. Chapman & Hall/CRC.

[20] Jennison, C. and Turnbull, B.W. (1993c). Group sequential tests for bivariate response: Interim analyses of clinical trialss with both efficacy and safety endpoints. *Biometrics*, **49**,741-752.

[21] Jennison, C. and Turnbull, B.W. (1997a). Group sequential analysis incorporating covariate information. *J.Amer.Statist.Assoc.*, **92**, 1330-1341.

[22] Jennison, C. and Turnbull, B.W. (2006). Adaptive and non-adaptive group sequential tests. *Biometrica*, **93**(**1**), 1-21.

[23] Karlin, S. (1968). Total Positivity. Stanford Univeristy Press, Stanford.

[24] Kotz, S. and Nadarajah, S. (2004). *Multivariate t distributions and their applications*. Cambridge University Press, Cambridge, United Kingdom.

[25] Kounias, E.G. (1968). Bounds for the probability of a union, with applications. *The Annals of Mathematical Statistics*, **39**, 2154-2158.

[26] Liu, F. and Sarkar, S.K. (2009). A note on estimating the false discovery rate under mixture model. *J.Stat.Plan.Inference*, **140**, 1601-1609.

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

[28] Maurer,W. and Bretz, F (2011). Multiple and Repeated testing of primary, co-primary and secondary hypotheses. *Statistics in Biopharmaceutical Research*, **3**, 336-352.

[29] Maurer,W. and Bretz, F (2013). Multiple testing in group sequential trials using graphical approaches. *Statistics in Biopharmaceutical Research*, **5** : **4**, 311-320.

[30] Millen, B.A. and Dmitrienko, A. (2011). Chain procedures: a class of flexible closed testing procedures with clinical trial applications. *Statistics in Biopharmaceutical Research*, **3**, 14-30.

[31] Nelsen, R. (2007). An Introduction to Copulas. Springer, New York.

[32] Odeh, R. (1982), Tables of percentage points of the distribution of the maximum absolute value of equally correlated normal random variable. *Commun. Statist. Simula. Compu*, **11**, 65-87.

[33] Rom, D.M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, **77**, 663-665.

[34] Sarkar, S.K. (1998). Some probability inequalities for ordered MTP$_2$ random variables: A proof of the Simes conjecture. *Ann. Statist.*, **26**, 494-504.

[35] Sarkar, S.K. (2008). Generalizing Simes' test and Hochberg's stepup procedure. *Ann. Statist.*, **36**, 337-363.

[36] Sarkar, S.K. and Chang C. (1997). The simes method for multiple hypothesis testing with positively dependent test statistics. *J.Amer.Statist.Assoc.*, **92**, 1601-1608.

[37] Sarkar, S.K., Guo, W. and Finner, H. (2012). On adaptive procedures controlling the familywise error rate. *J.Stat.Plan.Inference*, **142**, 65-78.

[38] Sarkar, S.K. (2008). On methods controlling the false discovery rate. *The Indian Journal of Statistics*, **70**, 135-168.

[39] Sarkar, S.K. (2002). Some results on false discovery rate in stepwise multiple testing procedures.*Ann.Statist.*, **30**, 239-257.

[40] Sarkar, S.K., Fu, Y., and Guo, W. (2015). On improving Holm's procedure using pairwise dependencies. Submitted.

[41] Seneta, E. , Chen, J.T. (2005). Simple stepwise tests of hypotheses and multiple comparisons. *Int. Stat. Rev.*, **73**, 21-34.

[42] Storey, J. (2002). A direct approach to false discovery rates. *J.R.Statist.Soc.B*, **64**, 479-498.

[43] Tang, D-I. and Geller, N.L, and Pocock, S.J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics*, **49**, 23-30.

[44] Tang, D-I. and Geller, N.L. (1999). Closed testing procedures for group sequential clinical trials with multiple endpoints. *Biometrics*, **55**, 1188-1192.

[45] Tong, Y.L. (1990). The Multivariate normal distribution. Springer Series in Statistics.

[46] Wald, A. (1947). Sequential analysis. John Wiley & Sons, Inc.

[47] Westfall, P. and Tobias, R. and Rom, D. and Wolfinger, R and Hochberg, Y. (1999). Multiple comparisons and multiple tests using the SAS system.

[48] Ye, Y. and Li, A. and Liu, L. and Yao, B. (2013). A group sequential holm procedure with multiple primary endpoints. *Statistics in Medicine*, **32**, 1112-1124.