FWER controlling procedures for testing multiple hypotheses with hierarchical structure and applications in clinical trials

Zhiying Qiu
New Jersey Institute of Technology

Follow this and additional works at: https://digitalcommons.njit.edu/dissertations

Part of the Mathematics Commons

Recommended Citation
Qiu, Zhiying, "FWER controlling procedures for testing multiple hypotheses with hierarchical structure and applications in clinical trials" (2014). Dissertations. 162.
https://digitalcommons.njit.edu/dissertations/162

This Dissertation is brought to you for free and open access by the Electronic Theses and Dissertations at Digital Commons @ NJIT. It has been accepted for inclusion in Dissertations by an authorized administrator of Digital Commons @ NJIT. For more information, please contact digitalcommons@njit.edu.
Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be “used for any purpose other than private study, scholarship, or research.” If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of “fair use” that user may be liable for copyright infringement.

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation.

Printing note: If you do not wish to print this page, then select “Pages from: first page # to: last page #” on the print dialog screen.
The Van Houten library has removed some of the personal information and all signatures from the approval page and biographical sketches of theses and dissertations in order to protect the identity of NJIT graduates and faculty.
ABSTRACT

FWER CONTROLLING PROCEDURES FOR TESTING MULTIPLE HYPOTHESES WITH HIERARCHICAL STRUCTURE AND APPLICATIONS IN CLINICAL TRIALS

by
Zhiying Qiu

In applications of clinical trials, the hypotheses to be tested often exhibit a hierarchical structure, and are usually hierarchically ordered based on their importance, clinical relevance, or dose concentration, etc. Thus, they are tested in a pre-defined fixed sequence. In some more complex cases, the hypotheses to be tested are hierarchically grouped into several families, and thus the families are tested in a sequential order. Although such problems of structured multiple testing have received much attention and several popular FWER controlling procedures, such as conventional fixed sequence procedure, fallback procedure and gatekeeping procedure have been introduced, not much progress has been made yet advancing their theory and methods.

This research contributes to the development of theory and methods of multiple testing problems under structured hypotheses in the following aspects. First, a class of generalized fixed sequence procedures is introduced for testing a single family of hypotheses, which allow each hypothesis to be tested even though some early hypotheses in the sequence are not rejected. A condition of a given generalized fixed sequence procedure can strongly control the FWER under arbitrary dependence is proposed. Based on the condition, three special generalized fixed sequence procedures controlling the FWER are developed. Through extensive simulation studies, the advantages of proposed procedures are shown over the existing FWER controlling procedures in terms of the FWER control and power. When the pairwise joint distributions of the true null p-values are known, these procedures can be improved further by incorporating such pairwise correlation information while maintaining the control of the FWER.
Secondly, a family-based graphical approach is proposed to construct general stepwise multilevel family-based procedures for testing multiple hierarchically ordered families of hypotheses. The resulting procedures can be elegantly represented by directed acyclic graphs. Though some examples, it is shown that the proposed family-based graphical approach can present the testing strategy simpler and more efficiently than the existing hypothesis-based graphical approach.

Thirdly, a Bonferroni-based gatekeeping procedure with retesting option for testing hierarchically ordered families of hypotheses is proposed. By this procedure, each family of hypotheses is repeatedly tested using Bonferroni procedure with updated local critical values. It is proved that the proposed procedure can strongly control the global FWER under arbitrary dependence.

Lastly, a multilevel partial hierarchical procedure in dealing with the problem of testing multiple families of hypotheses with partially ordered hierarchical structure is introduced. One hypothesis in current family is of interest only if some hypotheses in the previous families satisfy certain conditions. It is shown that the proposed procedure can control the FWER strongly at level $\alpha$ under the assumption that $p$-values of hypotheses from different families are independent.
FWER CONTROLLING PROCEDURES FOR TESTING MULTIPLE HYPOTHESES WITH HIERARCHICAL STRUCTURE AND APPLICATIONS IN CLINICAL TRIALS

by
Zhiying Qiu

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Mathematical Sciences

Department of Mathematical Sciences

May 2014
BIOGRAPHICAL SKETCH

Author: Zhiying Qiu
Degree: Doctor of Philosophy
Date: May 2014

Undergraduate and Graduate Education:

• Doctor of Philosophy in Mathematical Sciences,
  New Jersey Institute of Technology, Newark, NJ, 2014

• Master of Science in Applied Statistics,
  New Jersey Institute of Technology, Newark, NJ, 2010

• Bachelor of Engineering in Electronic Information Engineering,
  Zhejiang University of Technology, Hangzhou, Zhejiang, China, 2008

Major: Mathematical Sciences

Presentations and Publications:


This dissertation is dedicated to my beloved parents, Zulong Qiu and Min Zhao, for supporting me all the way.
ACKNOWLEDGMENT

First of all, I would like to sincerely thank my dissertation advisor, Dr. Wenge Guo, for his tremendous support and guidance. I am deeply grateful for his patience and encouragement throughout my four years graduate study. Without his guidance, I could not have possibly finished this dissertation.

I would also like to thank my other committee members, Dr. Sunil Dhar, Dr. Sundarraman Subramanian, Dr. Ji Meng Loh, and Dr. Zhi Wei.

Additionally, I would like to thank all the other faculty and Phd students in the statistics group for all the help they have been giving me during all these years.

Last but not the least, I am enormously grateful to my husband, Yang Li, for his unconditional love and support.
### TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction to Multiple Hypothesis Testing</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1 Error Rates and Power</td>
<td>2</td>
</tr>
<tr>
<td>1.1.2 Closure Principle</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Single Family FWER Controlling Procedures</td>
<td>4</td>
</tr>
<tr>
<td>1.2.1 Procedures based on Ordered p-values</td>
<td>4</td>
</tr>
<tr>
<td>1.2.2 Procedures based on a Pre-Ordered Sequence of Hypotheses</td>
<td>5</td>
</tr>
<tr>
<td>1.2.3 Graphical Visualization</td>
<td>6</td>
</tr>
<tr>
<td>1.2.4 Distributional Assumption</td>
<td>6</td>
</tr>
<tr>
<td>1.3 Multiple Families FWER Controlling Procedures</td>
<td>7</td>
</tr>
<tr>
<td>1.3.1 Procedures based on Multiple Hierarchically Ordered Families</td>
<td>8</td>
</tr>
<tr>
<td>1.3.2 Procedures based on Multiple Families with Special Hierarchical Structure</td>
<td>10</td>
</tr>
<tr>
<td>1.4 Motivation and Outline</td>
<td>11</td>
</tr>
<tr>
<td>2 A CLASS OF GENERALIZED FIXED SEQUENCE PROCEDURES FOR CONTROLLING THE FWER</td>
<td>13</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>13</td>
</tr>
<tr>
<td>2.2 Preliminary</td>
<td>15</td>
</tr>
<tr>
<td>2.3 Main Theoretical Results</td>
<td>17</td>
</tr>
<tr>
<td>2.4 Procedures under Arbitrary Dependence</td>
<td>19</td>
</tr>
<tr>
<td>2.5 Generalized Fixed Sequence Procedure as a Closed Test</td>
<td>21</td>
</tr>
<tr>
<td>2.6 Numerical Findings</td>
<td>23</td>
</tr>
<tr>
<td>2.7 A Clinical Trial Example</td>
<td>32</td>
</tr>
<tr>
<td>2.8 Further Improvement</td>
<td>34</td>
</tr>
<tr>
<td>2.9 Concluding Remarks</td>
<td>40</td>
</tr>
<tr>
<td>Chapter</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------------------------------</td>
</tr>
<tr>
<td>3</td>
<td>FAMILY-BASED GRAPHICAL APPROACH FOR SEQUENTIALLY TESTING HIERARCHICALLY ORDERED FAMILIES OF HYPOTHESES</td>
</tr>
<tr>
<td>3.1</td>
<td>Motivation</td>
</tr>
<tr>
<td>3.2</td>
<td>Preliminary</td>
</tr>
<tr>
<td>3.3</td>
<td>Main Theoretical Results</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Two-Level Family-based Graphical Approach with Four Families</td>
</tr>
<tr>
<td>3.3.2</td>
<td>General Multilevel Family-based Graphical Approach</td>
</tr>
<tr>
<td>3.4</td>
<td>Discussions</td>
</tr>
<tr>
<td>3.5</td>
<td>A Clinical Trial Example</td>
</tr>
<tr>
<td>3.6</td>
<td>Concluding Remarks</td>
</tr>
<tr>
<td>4</td>
<td>BONFERRONI-BASED GATEKEEPING PROCEDURE WITH RETESTING OPTION</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>4.2</td>
<td>Preliminary</td>
</tr>
<tr>
<td>4.3</td>
<td>Main Results</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Two-Family Problem</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Multi-Family Problem</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Global Familywise Error Rate Control</td>
</tr>
<tr>
<td>4.4</td>
<td>Discussions</td>
</tr>
<tr>
<td>4.5</td>
<td>Clinical Trial Examples</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Two-Family Problem</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Three-Family Problem</td>
</tr>
<tr>
<td>4.6</td>
<td>Numerical Findings</td>
</tr>
<tr>
<td>4.7</td>
<td>An Extension</td>
</tr>
<tr>
<td>4.8</td>
<td>Concluding Remarks</td>
</tr>
<tr>
<td>5</td>
<td>PROCEDURES FOR TESTING MULTIPLE FAMILIES WITH PARTIALLY ORDERED HIERARCHICAL STRUCTURE</td>
</tr>
</tbody>
</table>
# TABLE OF CONTENTS

(Continued)

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Motivation</td>
<td>86</td>
</tr>
<tr>
<td>5.2 Preliminary</td>
<td>91</td>
</tr>
<tr>
<td>5.3 Main Theoretical Results</td>
<td>92</td>
</tr>
<tr>
<td>5.3.1 General Multilevel Partial Hierarchical Procedure</td>
<td>92</td>
</tr>
<tr>
<td>5.4 Illustration Analysis</td>
<td>95</td>
</tr>
<tr>
<td>5.5 Concluding Remarks</td>
<td>99</td>
</tr>
<tr>
<td>6 CONCLUSION AND FUTURE WORK</td>
<td>100</td>
</tr>
<tr>
<td>APPENDIX A PROOFS FOR FIXED SEQUENCE PROCEDURE</td>
<td>103</td>
</tr>
<tr>
<td>A.1 Proof of Proposition 2.1</td>
<td>103</td>
</tr>
<tr>
<td>A.2 Proof of Theorem 2.1.(ii)</td>
<td>104</td>
</tr>
<tr>
<td>A.3 Proof of Theorem 2.3</td>
<td>106</td>
</tr>
<tr>
<td>APPENDIX B PROOFS FOR FAMILY-BASED GRAPHICAL APPROACH</td>
<td>108</td>
</tr>
<tr>
<td>B.1 Proof of Theorem 3.1</td>
<td>108</td>
</tr>
<tr>
<td>B.2 Proof of Theorem 3.2</td>
<td>110</td>
</tr>
<tr>
<td>APPENDIX C PROOFS FOR RETESTING PROCEDURE</td>
<td>113</td>
</tr>
<tr>
<td>C.1 Proof of Theorem 4.1</td>
<td>113</td>
</tr>
<tr>
<td>C.2 Proof of Theorem 4.2</td>
<td>117</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>120</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Local Critical Values for Any Intersection Hypotheses of Three Hypotheses using Procedure A1-A3 (PA1-PA3), Conventional fixed sequence procedure (FS), Fallback procedure (FB) and Holm’s procedure (HM)</td>
<td>23</td>
</tr>
<tr>
<td>2.2 Comparison Results of Six Procedures in the Hypertension Trial Example</td>
<td>33</td>
</tr>
<tr>
<td>2.3 Percentage of Improvement for Critical Values of Procedure A1 and Procedure B1 with ( n = 10 )</td>
<td>37</td>
</tr>
<tr>
<td>2.4 Percentage of Improvement for Critical Values of Procedure A2 and Procedure B2 with ( n = 10, \beta = 0.5 )</td>
<td>38</td>
</tr>
<tr>
<td>2.5 Percentage of Improvement for Critical Values of Procedure A3 and Procedure B3 with ( n = 5 )</td>
<td>39</td>
</tr>
<tr>
<td>3.1 Two-level and Three-level Family based Procedures in the Type II Diabetes Clinical Trial</td>
<td>57</td>
</tr>
<tr>
<td>4.1 Comparison Results of Three Procedures in the EPHESUS Trial Example</td>
<td>72</td>
</tr>
<tr>
<td>4.2 Comparison Results of Three Procedures in the Dose-Response Trial Example</td>
<td>75</td>
</tr>
<tr>
<td>4.3 Simulated Global FWERs and Average Powers for Four Procedures under Equal Correlation ( \rho ) for ( d_1 = 1, d_2 = 0 )</td>
<td>77</td>
</tr>
<tr>
<td>4.4 Simulated Global FWERs and Average Powers for Four Procedures under Equal Correlation ( \rho ) for ( d_1 = 0.8, d_2 = 0.2 )</td>
<td>78</td>
</tr>
<tr>
<td>4.5 Simulated Global FWERs and Average Powers for Four Procedures under Equal Correlation ( \rho ) for ( d_1 = 0.5, d_2 = 0.5 )</td>
<td>78</td>
</tr>
<tr>
<td>4.6 Simulated Global FWERs and Average Powers for Four Procedures under Equal Correlation ( \rho ) for ( d_1 = 0.2, d_2 = 0.8 )</td>
<td>79</td>
</tr>
<tr>
<td>4.7 Simulated Global FWERs and Average Powers for Four Procedures under Equal Correlation ( \rho ) for ( d_1 = 0, d_2 = 1 )</td>
<td>79</td>
</tr>
<tr>
<td>5.1 Comparison Results of Proposed Procedure, Multistage Gatekeeping Procedure and Tree-Structure Gatekeeping Procedure in the Motivation Example 1</td>
<td>96</td>
</tr>
<tr>
<td>5.2 Comparison Results of Proposed Procedure, Multistage Gatekeeping Procedure and Tree-Structure Gatekeeping Procedure in the Motivation Example 2</td>
<td>98</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>2.1</td>
<td>Simulated FWERs of six procedures with dependent p-values generated from equally correlated multivariate normal test statistics with correlation $\rho$ for $n = 20, \alpha = 0.05, d = \sqrt{10}, N = 0, \beta = 0.5, \gamma = 0.5$.</td>
</tr>
<tr>
<td>2.2</td>
<td>Simulated powers of six procedures with dependent p-values generated from equally correlated multivariate normal test statistics with correlation $\rho$ for $n = 20, \alpha = 0.05, d = \sqrt{10}, N = 0, \beta = 0.5, \gamma = 0.5$.</td>
</tr>
<tr>
<td>2.3</td>
<td>Simulated FWERs of six procedures with dependent p-values generated from equally correlated multivariate normal test statistics with correlation $\rho$ for $n = 20, \alpha = 0.05, d = \sqrt{10}, N = 2, \beta = 0.5, \gamma = 0.5$.</td>
</tr>
<tr>
<td>2.4</td>
<td>Simulated powers of six procedures with dependent p-values generated from equally correlated multivariate normal test statistics with correlation $\rho$ for $n = 20, \alpha = 0.05, d = \sqrt{10}, N = 2, \beta = 0.5, \gamma = 0.5$.</td>
</tr>
<tr>
<td>2.5</td>
<td>Simulated FWERs of six procedures with dependent p-values generated from equally correlated multivariate normal test statistics with correlation $\rho$ for $n = 20, \alpha = 0.05, d = \sqrt{10}, N = 4, \beta = 0.5, \gamma = 0.5$.</td>
</tr>
<tr>
<td>2.6</td>
<td>Simulated powers of six procedures with dependent p-values generated from equally correlated multivariate normal test statistics with correlation $\rho$ for $n = 20, \alpha = 0.05, d = \sqrt{10}, N = 4, \beta = 0.5, \gamma = 0.5$.</td>
</tr>
<tr>
<td>3.1</td>
<td>Hypothesis-based graphical visualization of the Bonferroni-Holm procedure with three hypotheses.</td>
</tr>
<tr>
<td>3.2</td>
<td>Hypothesis-based graphical visualization of gatekeeping procedure with Bonferroni-Holm procedure as gatekeeper.</td>
</tr>
<tr>
<td>3.3</td>
<td>Family-based graphical visualization of gatekeeping procedure with Bonferroni-Holm procedure as gatekeeper.</td>
</tr>
<tr>
<td>3.4</td>
<td>General graphical representation of family-based graphical approach.</td>
</tr>
<tr>
<td>3.5</td>
<td>Graph for two level family-based procedure with m=4.</td>
</tr>
<tr>
<td>3.6</td>
<td>Family-based graphical visualization of Case 1.</td>
</tr>
<tr>
<td>3.7</td>
<td>Family-based graphical visualization of Case 2.</td>
</tr>
<tr>
<td>3.8</td>
<td>Family-based graphical visualization of Case 3.</td>
</tr>
<tr>
<td>3.9</td>
<td>Family-based graph (a) and hypothesis-based graph (b) for Procedure 1 in the Type II diabetes clinical trial.</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>4.1</td>
<td>Three-family Bonferroni-based gatekeeping procedure with retesting option.</td>
</tr>
<tr>
<td>4.2</td>
<td>Two-family Bonferroni-based gatekeeping procedure with retesting option.</td>
</tr>
<tr>
<td>4.3</td>
<td>Graphical visualization of Case 1 with ( m ) families of hypotheses.</td>
</tr>
<tr>
<td>4.4</td>
<td>Graphical visualization of Case 2 with ( m ) families of hypotheses.</td>
</tr>
<tr>
<td>4.5</td>
<td>Graphical visualization of the two-family clinical trial problem.</td>
</tr>
<tr>
<td>4.6</td>
<td>Graphical visualization of three-family clinical trial problem.</td>
</tr>
<tr>
<td>4.7</td>
<td>Two levels of four families generalized multistage Bonferroni procedure</td>
</tr>
<tr>
<td></td>
<td>with retesting.</td>
</tr>
<tr>
<td>5.1</td>
<td>Graphical visualization of logical structure for multiple testing problem</td>
</tr>
<tr>
<td></td>
<td>in Example 1.</td>
</tr>
<tr>
<td>5.2</td>
<td>Graphical visualization of logical structure for multiple testing problem</td>
</tr>
<tr>
<td></td>
<td>in Example 2.</td>
</tr>
</tbody>
</table>
In modern scientific research, it becomes increasingly common to address multiple scientific questions concurrently. Accordingly, the problem of simultaneously testing multiple hypotheses is normally involved. When multiple hypotheses are tested without adjustment of significant levels, the overall error rate can be easily out of control. Therefore, several multiple testing procedures have been introduced in the literature to ensure appropriate control of the overall error rate. In clinical trials, multiple testing problems are always involved due to multiple treatments, multiple doses, multiple endpoints, etc. Especially, the multiple hypotheses are usually hierarchically ordered or logically related. For example, the efficiency of a new drug usually evaluated by multiple endpoints while the endpoints are generally classified into primary, secondary and sometimes tertiary endpoint which forms a natural hierarchal structure. Therefore, the objective of our research is to develop new multiple testing procedures taking the intrinsic hierarchical structure within the hypotheses into account.

1.1 Introduction to Multiple Hypothesis Testing
A standard multiple testing problem involves testing many hypotheses simultaneously. Each hypothesis is associated with a test statistic, and large test statistics provide evidence against the null hypotheses. Some important aspects of multiple testing are discussed below.
1.1.1 Error Rates and Power

Before dealing with multiple testing problems, it is essential to choose an appropriate overall measure of type I error rate. The overall measure is not unique. Several commonly used error rates have been defined in the literature as follows.

- **Comparisonwise error rate** (CWER). The CWER is defined as
  \[
  \text{CWER} = \frac{E(\text{Number of false rejections})}{\text{Total number of hypotheses}}. 
  \]

- **Familywise error rate** (FWER). The FWER is defined as
  \[
  \text{FWER} = \Pr\{\text{at least one false rejection}\}. 
  \]

- **False discovery rate** (FDR). The FDR is defined as
  \[
  \text{FDR} = E\left(\frac{\text{Number of false rejections}}{\text{Total number of rejections}}\right). 
  \]

In clinical trials, multiple testing procedures are usually required to control the FWER at a pre-specified level \(\alpha\). In general, there are two kinds of control, *weak* control and *strong* control. The *weak* control means that the FWER is controlled when all hypotheses are simultaneously true, while *strong* control means the FWER is controlled under any combination of true and false null hypotheses. Strong control of the FWER is mandated by regulators in all confirmatory clinical trials (CPMP, 2002). Therefore, in this thesis, we only consider the FWER control in the strong sense.

In addition to the FWER control, one also needs to define *power* to measure the performance of a multiple testing procedure. The definition of power is also not unique in multiple testing. One of the most widely used, which is also used throughout this proposal, is called *average power* (Benjamini and Liu, 1999; Storey, 2002; Cai and
Sarkar, 2005). It is defined as

$$\text{Average Power} = \frac{E(\text{Number of correct rejections})}{\text{Number of false nulls}}.$$ 

1.1.2 Closure Principle

There are several testing principles on which the FWER controlling multiple testing procedures are based, such as closure principle (Marcus, Peritz and Gabriel, 1976), partitioning principle (Stefansson, Kim and Hsu, 1988; Finner, Strassburger, 2002), sequential rejection principle (Goeman and Solari, 2010). The closure principle is recognized as one of the most fundamental principle of FWER control. It provides a foundation of many multiple testing procedures. The closure principle states that a hypothesis is rejected in the context of a multiple testing if and only if all intersection hypotheses containing this hypothesis are rejected by the corresponding local tests in the context of single test. Any procedures based on this principle are called closed testing procedures and it can be shown that those procedures strongly control the FWER at level \( \alpha \). The process of constructing a closed testing procedure for testing \( n \) hypotheses \( H_1, \cdots, H_n \) is as follows:

- Form the closed family of hypotheses by including all intersection hypotheses defined as

$$H_I = \bigcap_{i \in I} H_i,$$

where \( I \) denotes an non-empty index set such that \( I \subseteq \{1, \cdots, n\} \).

- Test each intersection hypothesis in the closed family by a suitable \( \alpha \) level test. A hypothesis \( H_i \) can be rejected if and only if for any index set \( I \) containing index \( i \), \( H_I \) is rejected by the corresponding \( \alpha \)-level test.

It is easy to see that the closure principle requires to construct local \( \alpha \)-level tests for every subset intersection hypothesis. Therefore, the number of tests increases
exponentially with $n$. Apparently, some shortcuts need to be developed in order to reduce the computational complexity.

1.2 Single Family FWER Controlling Procedures

In this section, we consider sequentially testing $n$ hypotheses $H_1, \cdots, H_n$ as one single family. The hypotheses to be tested are either ordered based on the given $p$-values or pre-ordered based on the prior information. The existing FWER controlling procedures in the literature generally can be divided into two groups. One group includes procedures based on ordered $p$-values of hypotheses. The other group includes procedures based on a pre-ordered sequence of hypotheses.

1.2.1 Procedures based on Ordered $p$-values

The $p$-value ordered stepwise procedures are described by using a sequence of non-decreasing critical values $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_n$. There are two types of stepwise procedure, i.e., step-down procedure and step-up procedure. Suppose the marginal $p$-values associated to the $n$ tested hypotheses are $P_1, \cdots, P_n$. Let the ordered $p$-values be denoted as $P_{(1)}, \cdots, P_{(n)}$ with associated hypotheses $H_{(1)}, \cdots, H_{(n)}$.

- Step-down procedure. A step-down procedure starts with $H_{(1)}$ and gradually steps down to $H_{(n)}$. If $P_{(1)} > \alpha_1$, accept all the $n$ hypotheses. Otherwise, reject $H_{(1)}, \cdots, H_{(j)}$ where $1 \leq j \leq n$ is the largest index satisfying $P_{(1)} \leq \alpha_1, \cdots, P_{(j)} \leq \alpha_j$. The rest hypotheses are accepted automatically. The typical example is Holm’s procedure (Holm, 1979) which is as follows.

  *Holm’s procedure.* The Holm’s procedure is a step-down procedure with the critical value $\alpha_i = \frac{\alpha}{n-i+1}$ for $i = 1, \cdots, n$.

- Step-up procedure. A step-up procedure begins with $H_{(n)}$ and gradually steps up to $H_{(1)}$. If $P_{(n)} \leq \alpha_n$, all hypotheses are directly rejected. Otherwise, accept $H_{(j)}, \cdots, H_{(n)}$ and reject $H_{(1)}, \cdots, H_{(j-1)}$ where $1 \leq j \leq n$ is the smallest index
satisfying $P_{(n)} > \alpha_n, \ldots, P_{(j)} > \alpha_j$. The typical example is Hochberg procedure (Hochberg, 1988) which is as follows.

**Hochberg procedure.** The Hochberg procedure is a step-up procedure with the same set of critical values as Holm’s Procedure, that is, $\alpha_i = \frac{\alpha}{n-i+1}$ for $i = 1, \ldots, n$.

Note that there is one special case of either step-down procedure or step-up procedure, named single-step procedure. A single-step procedure rejects $H_i$ if $P_i \leq c$ for $i = 1, \ldots, n$ where $c$ is a pre-fixed constant. The typical example is Bonferroni procedure which is as follows.

**Bonferroni procedure.** The Bonferroni procedure is a single-step procedure with a equal critical value $\alpha/n$ for each hypothesis. A variant of the Bonferroni procedure is the weighted Bonferroni procedure. Given weights $\omega_1, \ldots, \omega_n$ such that $\sum_{i=1}^n \omega_i = 1$, the weighted Bonferroni procedure rejects $H_i$ if $P_i \leq \omega_i \alpha$.

### 1.2.2 Procedures based on a Pre-Ordered Sequence of Hypotheses

Unlike the $p$-value ordered stepwise procedures, the fixed sequence procedures assume that the order of $n$ hypotheses, $H_1, \ldots, H_n$, to be tested is pre-specified. This order is usually determined based on the importance of the hypotheses. Two widely used fixed sequence procedures are conventional fixed sequence procedure (Maurer et al., 1995; Westfall and Krishen, 2001) and fallback procedure (Wiens, 2003; Wiens and Dmitrienko, 2005). The relevant research have been done by Wiens and Dmitrienko (2005), Hommel and Kropf (2005), Dmitrienko, Wiens and Westfall (2006), Hommel, Bretz and Maurer (2007), Hommel and Bretz (2008) and Bretz et al. (2009). The two fixed sequence procedures are described in the following.

- **Conventional fixed sequence procedure.** The procedure sequentially tests each hypothesis and rejects $H_i$ if $P_i \leq \alpha$ and all previous hypotheses $H_1, \ldots, H_{i-1}$ are rejected for each $i = 1, \ldots, n$. 
• **Fallback procedure.** Suppose a series of weights $\omega_1, \cdots, \omega_n$ such that $\sum_{i=1}^{n} \omega = 1$ are pre-fixed. The procedure rejects $H_i$ if $P_i \leq \alpha_i$ where $\alpha_i = \omega_i \alpha + \alpha_{i-1}$ if $H_{i-1}$ is rejected and $\alpha_i = \omega_i \alpha$ otherwise. Unlike conventional fixed sequence procedure, the fallback procedure tests every hypothesis.

1.2.3 **Graphical Visualization**

In order to make the construction of multiple testing procedure be easily explained, Bretz et al. (2009) suggested a graphical approach by which most of the single family FWER controlling procedures can be elegantly represented by some graphs. In this graphical approach, hypotheses $H_i$ are represented by vertices with initial critical value $\alpha_i$ satisfying $\sum_{i=1}^{n} \alpha_i = \alpha$. There is an edge with a weight between any two vertices. The weights are pre-specified according to a $n \times n$ transition matrix. According to the graph, once a hypothesis $H_i$ is rejected, its critical level $\alpha_i$ will be added to the remaining hypotheses according to predefined weights on the edges between the vertices. Then, the transition matrix between the remaining hypotheses is updated by a specific algorithm. The whole process stops when no further hypothesis can be rejected. The graphical approach makes the multiple testing strategies explicitly and easily to explain.

1.2.4 **Distributional Assumption**

In the literature, according to the assumption of joint distribution of the test statistics, multiple testing procedures can also be divided as the following three classes.

- Procedures do not make any assumption about the joint distribution of the test statistics. These procedures only depend on univariate $p$-values, thus can be referred to as $p$-value based procedures.

- Procedures are developed under specific distributional assumptions about the test statistics. For example, the test statistics are assumed to be a multivariate
normal. The procedures under such assumption are referred to as parametric procedures.

- Procedures are developed by approximating the joint distribution of the test statistics by using resampling-based methods.

### 1.3 Multiple Families FWER Controlling Procedures

In clinical trial research, it is becoming increasingly common to consider complex multiplicity problems due to hierarchically ordered multiple objectives. For example, there are usually multiple endpoints of interest in clinical trials and these endpoints are generally classified as primary, secondary and sometimes tertiary endpoints which form a natural hierarchical structure. Thus, the hypotheses formulated to address these objectives are grouped into multiple families and these families are tested in a sequential manner. To test a fixed sequence of families of hypotheses, Maurer, Hothorn and Lehmacher (1995) and Bauer et al. (1998) introduced a convenient and efficient testing strategy known as gatekeeping procedure. The basic idea of the gatekeeping strategy is that the families are tested sequentially an each early family works as a gatekeeper for the subsequent families. Only when one or more hypotheses are rejected in early families, the subsequent families can be tested. Basically, there are two types of gatekeeping procedures. One is serial gatekeeping (Westfall and Krishen, 2001) and the other is parallel gatekeeping (Dmitrienko, Offen and Westfall, 2003). A serial gatekeeping procedure tests a family only if all hypotheses in the previous family are rejected. A parallel gatekeeping procedure tests a family only if at least one hypothesis in the previous family is rejected. Based on these two types of gatekeeping procedures, other types of gatekeeping procedures also have been introduced in the literature, including general multistage procedure (Dmitrienko, Tamhane and Wiens, 2008), union closure procedure (Kim, Entsua and Shults, 2011), superchain procedure (Dmitrienko and Kordzakhia, 2012). Except for the
above simple multiple families testing problems, more complicated problems of testing multiple families with partial hierarchical structure, such as tree structure, are also discussed in the literature including tree-structure procedure (Dmitrienko, Wiens, Tamhance and Wang, 2007; Dmitrienko, Tamhane, Liu and Wiens, 2008) and partial hierarchically ordered procedure (Maurer, Glimm and Bretz, 2011).

Consider a multiple testing problem which \( n \geq 2 \) hypotheses grouped into \( m \) families \( F_1, \cdots, F_m \) based on their hierarchical logical relationships, where \( F_i, i = 1, \cdots, m \) consists of \( n_i \) hypotheses denoted by \( H_{i1}, \cdots, H_{in_i} \) and \( \sum_{i=1}^{m} n_i = n \). Generally, when testing the multiple families of all \( n \) hypotheses, the overall FWER needs to be strongly controlled at a pre-specified level \( \alpha \), that is, the FWER needs to be controlled at level \( \alpha \) for the whole family \( F = \bigcup_{i=1}^{m} F_i \) of all \( n \) hypotheses regardless of which and how many null hypotheses in the \( m \) families are true.

1.3.1 Procedures based on Multiple Hierarchically Ordered Families

- **Serial gatekeeping procedure.** For serial gatekeeping procedure, each family can be tested using any method that controls the FWER at level \( \alpha \) if and only if all of the null hypotheses in the previous families are rejected. It is easy to see that the serial gatekeeping procedure is naturally stepwise without any adjustments towards the tests in each step. However, it is restrictive since the whole testing will stop once one acceptance occurs within one family.

- **Parallel gatekeeping procedure.** The parallel gatekeeping procedure tests hypotheses in \( F_{i+1} \) if and only if at least one hypothesis in \( F_i \) \((i = 1, \cdots, m-1)\) is rejected. It is less restrictive than serial gatekeeping procedure. However, it is derived using the closure principle and thus, the computation is complicated which involves testing up to \( 2^n - 1 \) intersection hypotheses.

- **Multistage gatekeeping procedure.** In order to avoid the complex computation issue of parallel gatekeeping procedures, Dmitrienko, Tamhane, Wang and
Chen (2006), Guibaud (2007) and Dmitrienko, Tamhane and Wiens (2008) introduced simple stepwise approaches of developing gatekeeping strategies, which is different from closure principle. Dmitrienko et al. (2008) captured the essence of the above work and introduced a general multistage gatekeeping procedure by using the notion of an error rate function. Consider a single family of hypotheses, $F = \{H_1, \cdots, H_n\}$. For any $I \subseteq N = \{1, 2, \cdots, n\}$, the error rate function is defined as

$$e(I) = \sup_{H_I} \Pr \left\{ \bigcup_{i \in I} (\text{reject } H_i) | H_I \right\}$$

where $H_I = \cap_{i \in I} H_i$ is the intersection of hypotheses $H_i$ with $i \in I$. It is easy to see that the supreme of the probability is taken over the entire null space defined by $H_I$, including any false hypothesis $H_i$ that $i \notin I$. Generally, $e(I)$ is difficult to calculate and a computable upper bound $e^*(I)$ is often used to replace $e(I)$. For $e^*(I)$, the following requirements are imposed: (1) $e^*(I) = 0$ if $I = \emptyset$; (2) $e^*(I) \leq e^*(J)$ if $I \subseteq J$; (3) $e^*(F) = \alpha$. When testing a fixed sequence of families, error rate function is used to quantify the amount of critical value for current family that can not be transferred to the subsequent families. Let $A_i$ denote the index set corresponding to the accepted hypotheses in $F_i$ and $e_i(I)$ denote the error rate function for the test used in $F_i, i = 1, \cdots, m-1$, the procedure starts to test $F_1$ at level $\alpha_1 = \alpha$ and continues to test any subsequent family $F_i, i = 2, \cdots, m-1$ at level $\alpha_i = \alpha_{i-1} - e^*(A_{i-1})$. In addition, each family (except the last one) should use a multiple testing procedure satisfying the condition that its corresponding error rate function is strictly less than $\alpha$ unless all hypotheses are accepted. This condition is known as the separability condition and those multiple testing procedures satisfying this condition are called separable procedures. Note that among those well known procedures, Bonferroni procedure is separable but other procedures
such as Holms, Hochberg, Fallback procedures, etc. are not. Therefore, the truncated versions of these standard procedures, i.e., a combination of Bonferroni procedure and these standard procedures are proposed which can satisfy the separability condition (see Dmitrienko, Tamhane and Wiens, 2008; Brechemacher, Xu, Dmitrienko and Tamhane, 2010).

- **Union closure procedure.** The union closure procedure generalizes the closed method for testing a fixed sequence of families of hypotheses. It assumes that one null hypothesis in the current family can be rejected by the closed method in the union of all previous families of hypotheses (including the current family) can also be rejected in the subsequent families (see Kim, Entsuah and Shults, 2011).

- **Superchain procedure.** The superchain procedure takes into account the logical relationships among multiple families of null hypotheses. It tests the families simultaneously rather than sequentially. For one family, if at least one null hypothesis is rejected, then an amount of its significant level is transferred to the other families. By using this procedure, one family may have chances to be retested at a increasing critical value after testing all families simultaneously.

### 1.3.2 Procedures based on Multiple Families with Special Hierarchical Structure

- **Tree-structured gatekeeping procedure.** When the hypotheses to be tested are formulated as a tree structure where each leaf corresponds to an individual hypothesis, a tree-structured gatekeeping procedure is introduced which is a hybrid procedure unifying the ideas of serial gatekeeping and parallel gatekeeping. It is derived based on the closure principle and uses weighted Bonferroni procedure for all intersection hypotheses.
• Partial hierarchically ordered procedure. In a simple setting of two families $F_1$ and $F_2$, Maurer, Glimm and Bretz (2011) considered a general hierarchical structure which involves a “parent-descendant” relation between the hypotheses in $F_1$ and $F_2$ in the sense that a descendant hypothesis in $F_2$ is only of interest if one of the respective “parent” hypotheses in $F_1$ is significant. Each hypothesis in $F_2$ has at least one “parent” hypothesis in $F_1$ and each hypothesis in $F_1$ also have at least one “descendant” hypothesis in $F_2$. Suppose the initially assigned critical value for $F_1$ and $F_2$ are $\alpha_1$ and $\alpha_2$. After testing $F_1$ at level $\alpha_1$, in stead of testing the whole family $F_2$, the procedure only test the hypotheses in $F_2$ with rejected parent hypotheses in $F_1$.

1.4 Motivation and Outline

The applications of clinical trials usually involve multiple objectives with hierarchical structure. The formulated hypotheses to address these objectives are usually hierarchically ordered in advance based on their clinical importance, clinical relevance, or dose concentration, etc. In more complex cases, the hypotheses are grouped into several families, and these families have hierarchical relationships. In modern drug development, the problem of constructing multiple testing procedures to test multiple hierarchically related hypotheses has received much attention and several relevant FWER controlling procedures, such as conventional fixed sequence, fallback and gatekeeping procedures have been introduced in the past decade. However, many theoretical and methodological issues related to this kind of multiple testing problems still remain to be fully investigated. In this thesis, our research aims to develop multiple testing methods for controlling FWER by exploiting the intrinsic hierarchical logical structure among the hypotheses.

This thesis is outlined as follows. Chapter 1 provides some basic concepts and results on multiple testing. Chapter 2-5 contains our contributions. In Chapter 2, we
propose a class of generalized fixed sequence procedures which can strongly control the FWER at level \( \alpha \) in dealing with a single family of hypotheses. And numerical analysis for performance comparisons between our proposed procedures and some existing multiple testing procedures is also presented. In Chapter 3, we propose a family-based graphical approach to construct stepwise multilevel family-based procedure for testing multiple hierarchically ordered families of hypotheses. In Chapter 4, we propose a Bonferroni - based gatekeeping procedure with retesting option which allows families of hypotheses to be sequentially tested more than once. In Chapter 5, we introduce a general multilevel partial hierarchical procedure in dealing with hypotheses with special hierarchical structure, such as tree structure. Chapter 6 discusses the possible future work.
CHAPTER 2

A CLASS OF GENERALIZED FIXED SEQUENCE PROCEDURES
FOR CONTROLLING THE FWER

2.1 Introduction

In applications of clinical trials, the hypotheses to be tested are often hierarchically ordered based on their importance, clinical relevance, or dose concentration, etc., and thus are tested in a pre-defined sequential order. Although the problem of fixed sequence multiple testing has received much attention and several popular FWER controlling procedures, such as the conventional fixed sequence procedure and fallback procedure, have been introduced, not much progress has been made yet advancing its theory and methods (Dmitrienko et al. 2009, 2013).

In this chapter, we focus on developing new multiple testing procedures to deal with the situation in which the hypotheses to be tested are pre-ordered based on prior knowledge. Procedures which operate on such multiple testing problems are known as fixed sequence procedures. Maurer, Hothorn and Lehmacher (1995) introduced the first fixed sequence multiple testing procedure, which we will refer to as the conventional fixed sequence procedure. In this procedure, each hypothesis is tested at pre-specified level $\alpha$ as long as all of the previous hypotheses have been rejected. It is proved that the procedure strongly controls the FWER at level $\alpha$ under arbitrary dependence. However, the main issue with this procedure is that it does not allow any acceptances. Once a hypothesis is not rejected, the remaining hypotheses will have no chance to be tested. Therefore, the procedure will perform poorly if one of the early hypotheses are insignificant. To deal with this issue, Wiens (2003) and Wiens and Dmitrienko (2005) introduced another popular fixed sequence procedure – the fallback procedure, in which the remaining hypotheses have a chance to be
tested, even if an acceptance occurs. And later, several authors have proposed various extensions of the fallback procedure in order to improve its power including Li and Mehrotra (2008) and Huque and Alosh (2008). Compared to the conventional fixed sequence procedure, the fallback procedure is more flexible in the sense that every hypothesis has a chance to be tested. However, Hommel and Bretz (2008) showed that in certain situations, the fallback procedure might violate the inherent hierarchical relationships among the hypotheses. For any two hypotheses, the earlier important hypothesis may have less chance to be rejected than the later one, even if their $p$-values are the same. This is not desired for a good multiple testing procedure. Subsequent works on developing more desirable and more powerful procedures for addressing the problem of fixed sequence multiple testing have been done by many authors, including Hommel and Kropf (2005), Rosenbaum (2008), Millen and Dmitrienko (2011), etc. In addition, Hommel and Kropf (2005) introduced a specific fixed sequence procedure, which allows a pre-specified number $k$ of acceptances and has the same critical value $\alpha/k$. For a detailed review of recent developments in this area of research, see Wiens and Dmitrienko (2010) and Dmitrienko et al. (2013), and for applications of fixed sequence multiple testing procedures in different fields, see Alosh and Huque (2009, 2010) and Tu et al. (2012).

In this chapter, a main goal is to develop new theory and methods for addressing the problem of fixed sequence multiple testing. Based on the similar idea of Hommel and Kropf (2005) that the procedure allows a pre-specified number of acceptances, We firstly introduce a more general procedure, termed as generalized fixed sequence procedure, whose critical values are defined by using a function of the numbers of rejections and acceptances, and which allows each hypothesis to be tested even if earlier hypotheses are not rejected. We then discuss a configuration, which we call the Dirac-Ordered configuration, under which the FWER of the procedure attains the maximum among all the configurations having the same joint distribution for the
true null \( p \)-values. Based on this configuration, we present a sufficient condition for FWER control of a generalized fixed sequence procedure under arbitrary dependence. Based on the condition, we develop three new fixed sequence procedures controlling the FWER. To better evaluate the proposed procedure, we illustrate the generalized fixed sequence procedures as closed testing procedures.

The rest of the chapter is organized as follows. We present some basic notations and generalize the conventional fixed sequence procedure in Section 2.2. We construct the least favorable configuration for the aforementioned procedure and present a sufficient condition for the FWER control of such a procedure in Section 2.3. In Section 2.4, we introduce three new fixed sequence procedures based on this condition. And in Section 2.5, we illustrate our proposed procedures as closed testing procedures. Extensive simulation studies and a real data analysis are respectively performed in Section 2.6 and 2.7 to evaluate the performances of the proposed procedures. In Section 2.8, we further improve the aforementioned procedures by incorporating pairwise correlation information of the true null \( p \)-values. Some concluding remarks are made in Section 2.9 and proofs of almost all results are given in the Appendix A.

## 2.2 Preliminary

In this section, we present some basic notations and generalize the concept of the conventional fixed sequence procedure. Suppose \( H_i, i = 1, \cdots, n \), are \( n \) null hypotheses which are pre-ordered based on prior knowledge and are to be tested based on their respective \( p \)-values \( P_i, i = 1, \ldots, n \). Among these \( n \) hypotheses, let \( n_0 \) of them be true null hypotheses and \( n_1 \) be false. For notational convenience, let \( \hat{H}_i \) denote the \( i^{th} \) true null hypothesis and \( \hat{P}_i \) denote the corresponding \( p \)-value. Likewise, let \( \tilde{H}_i \) denote the \( i^{th} \) false null hypothesis and \( \tilde{P}_i \) denote the corresponding \( p \)-value. Define the familywise error rate (FWER) as the probability of incorrectly rejecting at least one true null hypothesis. In this paper, the true null \( p \)-values are always
assumed to be stochastically greater than or equal to uniform distribution on $[0,1]$. That is, for $u \in [0,1],\quad \Pr\{\hat{P}_i \leq u\} \leq u, \ i = 1, \ldots, n_0. \quad (2.1)$

For existing $p$-value based stepwise methods, such as the Holm procedure (Holm, 1979) and Hochberg procedure (Hochberg, 1988), the hypotheses are ordered and tested based on the corresponding $p$-values. Instead, for fixed sequence methods, the hypotheses are ordered based on prior knowledge and tested based on the $p$-values.

Note that for the conventional fixed sequence procedure, a main drawback is that it does not allow any acceptance. In the following, we generalize the concept of the conventional fixed sequence multiple testing procedure so that even though some acceptances occur, the remaining hypotheses still have chance to be tested.

**Definition 2.1.** [Generalized Fixed Sequence Procedure] Given a function $\alpha(s,t)$ defined on $s = 0, \cdots, n-1$ and $t = 0, \cdots, n-1$, consider testing $H_i, \ i = 1, \ldots, n$. $H_i$ is rejected iff $P_i \leq \alpha(s_{i-1}, t_{i-1})$, where $s_{i-1}$ and $t_{i-1}$ are, respectively, the numbers of rejected and accepted hypotheses when testing $H_1, \ldots, H_{i-1}$, with $s_0 = t_0 = 0$.

The function $\alpha(s,t)$ is termed the critical value function throughout the chapter.

**Remark 2.1.** It is easy to see that when $\alpha(s,t) = \alpha$ if $t = 0$ and $\alpha(s,t) = 0$ if $t > 0$, the generalized fixed sequence procedure reduces to the conventional fixed sequence procedure in Maurer et al. (1995). Besides, when the critical value function is given in the form of

$$\alpha(s,t) = \begin{cases} \frac{c}{k}, & \text{if } t = 0, \ldots, k-1, \\ 0, & \text{if } t = k, \ldots, n-1, \end{cases}$$

where $k$ is a pre-specified integer with $0 < k < n$, the corresponding procedure reduces to the fixed sequence procedure introduced by Hommel and Kropf (2005),
which allows a pre-specified number of acceptances. For the fallback procedure in Wiens and Dmitrienko (2005), since the critical value for each hypothesis depends on the specific profile of previously tested hypotheses rather than the number of rejections or acceptances among the previous hypotheses. That is, the fallback procedure is not a kind of generalized fixed sequence procedure.

2.3 Main Theoretical Results

We will introduce in this section a sufficient condition on the critical value function for which the generalized fixed sequence procedure strongly controls the FWER at level $\alpha$ under arbitrary dependence. Before presenting the condition, for any configuration $P$ of the tested hypotheses $(H_1, \ldots, H_n)$ and the corresponding $p$-values $(P_1, \ldots, P_n)$, we introduce a corresponding configuration described as follows: (i) the true null $p$-values $\hat{P}_i, i = 1, \ldots, n_0$, have the same joint distribution as in the configuration $P$, (ii) the false null $p$-values $\tilde{P}_i = 0, i = 1, \ldots, n_1$, with probability 1, (iii) the order of the hypotheses to be tested, $H_1, \ldots, H_n$, is rearranged such that the false null hypotheses are tested before the true null hypotheses so that the order is $\tilde{H}_1, \ldots, \tilde{H}_{n_1}, \hat{H}_1, \ldots, \hat{H}_{n_0}$. This configuration is termed as a Dirac-Ordered (DO$_P$) configuration of $P$ throughout the paper and the FWER under this configuration is denoted by FWER$_{DO_P}$. The following proposition shows that the FWER of the generalized fixed sequence procedure is larger under the Dirac-Ordered configuration than the original configuration. Thus, in order to prove the FWER control of the generalized fixed sequence procedure, it is enough to show its FWER control under the Dirac-Ordered configuration.

**Proposition 2.1.** Consider a generalized fixed sequence procedure with critical value function $\alpha(s, t), s = 0, \ldots, n - 1, t = 0, \ldots, n - 1$. If $\alpha(s, t)$ is increasing in $s$ and decreasing in $t$, then the FWER of this procedure under any configuration $P$, FWER$_P$,
satisfies the following inequality:

\[ \text{FWER}_P \leq \text{FWER}_{DO_P}. \]  

(2.2)

For the proof of Proposition 2.1, see Appendix A.

**Remark 2.2.** The aforementioned Dirac-Ordered configuration is similar to the Dirac-Uniform configuration introduced in Finner and Roters (2001). The Dirac-Uniform configuration assumes independent \( p \)-values where the true null \( p \)-values are \( U(0, 1) \) and the false null \( p \)-values are zero with probability 1. However, in the Dirac-Ordered configuration, no independence assumption is made on the \( p \)-values but instead the order of hypotheses are taken into account.

Based on the Dirac-Ordered configuration, we now present a sufficient condition of a given generalized fixed sequence procedure strongly controlling the FWER under arbitrary dependence. By arbitrary dependence, we mean that the \( p \)-values do not have any type of known dependence structure.

**Theorem 2.1.** Consider a generalized fixed sequence procedure with the critical value function \( \alpha(s, t) \), where \( \alpha(s, t) \) is increasing in \( s \) and decreasing in \( t \).

(i) The generalized fixed sequence procedure strongly controls the FWER at level \( \alpha \) under arbitrary dependence if

\[ \sum_{t=0}^{n-s-1} \alpha(s, t) \leq \alpha \quad \text{for } s = 0, \ldots, n-1. \]  

(2.3)

(ii) The FWER control is sharp in the sense that there exists a joint distribution for \( (P_1, \ldots, P_n) \) for which the FWER of this procedure is exactly \( \alpha \), if there exists an \( s^* \) with \( 0 \leq s^* \leq n - 1 \) such that (3) becomes an equality when \( s = s^* \).

**Proof.** Based on Proposition 2.1, it is enough to show that for any configuration \( P \), \( \text{FWER}_{DO_P} \leq \alpha \). With the probabilities evaluated under the Dirac-Ordered
configuration of \( P \), we have

\[
\text{FWER}_{\text{DO}} = \Pr\{\hat{P}_1 \leq \alpha(n_1, 0)\}
+ \sum_{t=1}^{n_0 - 1} \Pr\{\hat{P}_1 > \alpha(n_1, 0), \ldots, \hat{P}_t > \alpha(n_1, t - 1), \hat{P}_{t+1} \leq \alpha(n_1, t)\}
\leq \sum_{t=0}^{n_0 - 1} \Pr\{\hat{P}_{t+1} \leq \alpha(n_1, t)\}
\leq \sum_{t=0}^{n_0 - 1} \alpha(n_1, t) = \sum_{t=0}^{n-n_1-1} \alpha(n_1, t) \leq \alpha.
\]

The second inequality follows from (2.1) and the last one follows from (2.3).

For the proof of (ii), see Appendix A.

\[\square\]

2.4 Procedures under Arbitrary Dependence

Theorem 2.1 provides a general approach for constructing FWER controlling fixed sequence procedures under arbitrary dependence. We can develop different kinds of fixed sequence procedures by choosing various kinds of critical value functions satisfying (3). In the following, we propose three special fixed sequence procedures based on three different types of critical value functions.

First, we consider the case where the critical value function \( \alpha(s, t) \) increases with \( s \) but stays constant with respect to \( t \). Thus, the procedure rewards the successful rejection of a hypothesis by increasing the critical values for the remaining hypotheses to be tested. But once the hypothesis fails to be rejected, no penalty towards those critical values is made.

PROCEDURE A1. Test the hypotheses according to the generalized fixed sequence procedure with the critical value function

\[
\alpha(s, t) = \frac{\alpha}{n - s} \quad \text{for } 0 \leq s, t \leq n - 1. \tag{2.4}
\]
Remark 2.3. It is easy to see that Procedure A1 is similar to the Holm procedure in the sense that they have similar critical value functions. But the Holm procedure stops on the first accepted hypothesis; whereas, Procedure A1 continues to test all the remaining hypotheses even if a hypothesis fails to be rejected.

Second, we consider the case where the critical value function $\alpha(s, t)$ is constant in $s$ but decreasing in $t$. Specifically, we let $\alpha(s, t)$ decrease in $t$ at a constant rate $\beta$. Thus, in contrast to Procedure A1, this procedure punishes the failure to reject a hypothesis by decreasing the critical values for the remaining hypotheses to be tested, but no reward is made for successful rejections.

**Procedure A2.** Test the hypotheses according to the generalized fixed sequence procedure with critical value function

$$\alpha(s, t) = \frac{1 - \beta}{1 - \beta^n} \beta^t \alpha \text{ for } 0 \leq s, t \leq n - 1,$$

where $\beta$ is a pre-specified constant satisfying $0 \leq \beta < 1$.

Remark 2.4. In Procedure A2, when $\beta = 0$, its critical values are always equal to $\alpha$ for $t = 0$, the critical value of the conventional fixed sequence procedure. On the other hand, as $\beta$ approaches to 1, its critical values approach to $\alpha/n$, the critical value of the Bonferoni procedure.

Finally, we develop a fixed sequence procedure which combines the ideas of Procedures A1 and A2 so that this procedure rewards rejections and punishes acceptances. To construct its critical value function $\alpha(s, t)$, we start by assuming $\alpha(s, t)$ decreases by a constant $c$ for each extra acceptance such that $\alpha(s, t - 1) - \alpha(s, t) = c$ for $1 \leq t \leq n - 1$. Thus, $\alpha(s, t) = \alpha(s, 0) - tc$. In order to satisfy (2.3), it must be the case that

$$\alpha(s, 0) \leq \frac{\alpha}{n - s} + \frac{n - s - 1}{2} c.$$

(2.6)
Furthermore, by taking the derivative of (2.6) with respect to $s$, one can see that $\alpha(s, 0)$, and hence $\alpha(s, t)$, is increasing in $s$ if and only if $c \leq 2\alpha/(n-s)^2$. By taking $c = 2\alpha/n^2$, we obtain the following procedure.

PROCEEDURE A3. Test the hypotheses according to the generalized fixed sequence procedure with critical value function

$$\alpha(s, t) = \left( \frac{1}{n-s} + \frac{n-s-1}{n^2} - \frac{2t}{n^2} \right) \alpha \text{ for } 0 \leq s, t \leq n-1. \quad (2.7)$$

Remark 2.5. In Procedure A3, when a hypothesis is accepted, the critical values for the remaining hypotheses are reduced by constant $2\alpha/n^2$. On the other hand, when a hypothesis is rejected, the critical values increase by $\frac{\alpha}{(n-s+1)(n-s)} - \frac{\alpha}{n^2}$, which depends on the corresponding number of rejections $s$.

It is easy to see that $\alpha(s, t)$ defined in (2.4), (2.5) and (2.7) are all increasing in $s$, decreasing in $t$, and satisfy (2.3) with equality for all values of $s$. Thus, we have the following result.

Theorem 2.2. Procedure A1, A2, and A3 all strongly control the FWER at level $\alpha$ under arbitrary dependence and their FWER controls are sharp in the sense that for each of these procedures, there exists a joint distribution for $(P_1, \ldots, P_n)$ for which its FWER is exactly equal to $\alpha$.

2.5 Generalized Fixed Sequence Procedure as a Closed Test

Suppose the critical value function $\alpha(s, t)$ is given. The generalized fixed sequence procedure with the critical value function $\alpha(s, t)$ can be illustrated as a closed testing procedure defined as follows. For any non-empty index set $I \subseteq \{1, \ldots, n\}$, consider an intersection hypothesis defined as $H_I = \bigcap_{i \in I} H_i$ and a local test based on the $p$-values for testing $H_I$: $H_I$ is rejected if $P_j \leq \alpha(s^*_j, t^*_j)$ for at least one $j \in I$. 

21
where
\[ s_{j-1}^* = \sum_{k<j} I(k \notin I), \]
\[ t_{j-1}^* = \sum_{k<j} I(k \in I), \]
and \( I(\cdot) \) is a indicator function. Here, \( \alpha(s_{j-1}^*, t_{j-1}^*) \) is termed as local critical values of the above local test. Based on such local tests, we can define a closed testing procedure by using the closure principle (Marcus et al., 1976). For these two procedures, we have the following theorem.

**Theorem 2.3.** The generalized fixed sequence procedure and the aforementioned closed testing procedure are equivalent for arbitrary number of hypotheses.

For the proof of Theorem 2.3, see Appendix A.

To better evaluate the performance of the aforementioned three proposed procedures A1-A3, we illustrate them as closed testing procedures and compare their local critical values of testing intersection hypotheses with those of three commonly used multiple testing procedures, Holm’s procedure, the conventional fixed sequence procedure, and the fallback procedure, which can also be illustrated as closed testing procedures. Table 2.1 lists the local critical values of the aforementioned six procedures in the case of three hypotheses. For the fallback procedure, the weights are set to be equal and for Procedure A2, \( \beta \) is set 0.5. It is easy to see from Table 2.1 that there is no procedure which is uniformly more powerful than others. For Procedure A1, its local critical values are smaller than those of Holm’s procedure but are comparable with those of fallback procedure. For the fallback procedure, its local critical values for lower-ranked hypotheses are sometimes larger than higher-ranked hypotheses. This is counter-intuitive. For our proposed procedures, contrary to the fallback procedure, the local critical values for higher-rank hypotheses are always...
Table 2.1 Local Critical Values for Any Intersection Hypotheses of Three Hypotheses using Procedure A1-A3 (PA1-PA3), Conventional fixed sequence procedure (FS), Fallback procedure (FB) and Holm’s procedure (HM)

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>PA1</th>
<th>PA2</th>
<th>PA3</th>
<th>FS</th>
<th>FB</th>
<th>HM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_1 \cap H_2 \cap H_3$</td>
<td>$\frac{\alpha}{3}$, $\frac{\alpha}{3}$, $\frac{\alpha}{3}$</td>
<td>$\frac{4\alpha}{7}$, $\frac{2\alpha}{7}$, $\frac{\alpha}{7}$</td>
<td>$\frac{5\alpha}{9}$, $\frac{\alpha}{3}$, $\frac{\alpha}{9}$</td>
<td>$\alpha$, $0$, $0$</td>
<td>$\frac{\alpha}{3}$, $\frac{\alpha}{3}$, $\frac{\alpha}{3}$</td>
<td>$\frac{\alpha}{3}$, $\frac{\alpha}{3}$, $\frac{\alpha}{3}$</td>
</tr>
<tr>
<td>$H_1 \cap H_2$</td>
<td>$\frac{\alpha}{3}$, $\frac{\alpha}{3}$</td>
<td>$\frac{4\alpha}{7}$, $\frac{2\alpha}{7}$</td>
<td>$\frac{5\alpha}{9}$, $\frac{\alpha}{3}$</td>
<td>$\alpha$, $0$</td>
<td>$\frac{\alpha}{3}$, $\frac{\alpha}{3}$</td>
<td>$\frac{\alpha}{2}$, $\frac{\alpha}{2}$</td>
</tr>
<tr>
<td>$H_1 \cap H_3$</td>
<td>$\frac{\alpha}{3}$, $\frac{\alpha}{2}$</td>
<td>$\frac{4\alpha}{7}$, $\frac{2\alpha}{7}$</td>
<td>$\frac{5\alpha}{9}$, $\frac{\alpha}{7}$</td>
<td>$\alpha$, $0$</td>
<td>$\frac{\alpha}{3}$, $\frac{2\alpha}{3}$</td>
<td>$\frac{\alpha}{2}$, $\frac{\alpha}{2}$</td>
</tr>
<tr>
<td>$H_2 \cap H_3$</td>
<td>$\frac{\alpha}{2}$, $\frac{\alpha}{2}$</td>
<td>$\frac{4\alpha}{7}$, $\frac{2\alpha}{7}$</td>
<td>$\frac{11\alpha}{18}$, $\frac{7\alpha}{18}$</td>
<td>$\alpha$, $0$</td>
<td>$\frac{2\alpha}{3}$, $\frac{\alpha}{3}$</td>
<td>$\frac{\alpha}{2}$, $\frac{\alpha}{2}$</td>
</tr>
<tr>
<td>$H_1$</td>
<td>$\frac{\alpha}{3}$</td>
<td>$\frac{4\alpha}{7}$</td>
<td>$\frac{5\alpha}{9}$</td>
<td>$\alpha$</td>
<td>$\frac{\alpha}{3}$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>$H_2$</td>
<td>$\frac{\alpha}{2}$</td>
<td>$\frac{4\alpha}{7}$</td>
<td>$\frac{11\alpha}{18}$</td>
<td>$\alpha$</td>
<td>$\frac{2\alpha}{3}$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>$H_3$</td>
<td>$\alpha$</td>
<td>$\frac{4\alpha}{7}$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
<td>$\alpha$</td>
</tr>
</tbody>
</table>

Note: For PA2, $\beta = 0.5$. For FB, initial weights are equal.

larger than the lower-rank hypotheses. This implies that more important hypotheses have larger chances to be rejected.

2.6 Numerical Findings

In this section, simulation studies were performed to investigate the performances of the proposed Procedures A1–A3 in terms of the FWER control and power compared to the existing Holm, conventional fixed-sequence and fallback procedures with respect to the correlation $\rho$ among test statistics, the proportion $\pi_0$ of true null hypotheses among all tested hypotheses. For fixed sequence procedures, we consider a pre-specified testing order for which early hypotheses are a mixture of $n_1$ false null hypotheses and a fixed $m$ true null hypotheses. When $m = 0$, it implies an ideal order for fixed sequence procedures in which all the false null hypotheses are ordered in front of true null hypotheses. When $m > 0$, it implies $m$ true null hypotheses are
mistakenly ordered compared to the aforementioned ideal order and we say there are \( m \) ordering mistakes in the testing order.

To simulate the values of FWER and average power (Westfall and Krishen, 2001), the expected proportion of false nulls that are rejected, for each of the aforementioned procedures, we first generated \( n \) dependent normal random variables \( T_i \sim N(\mu_i, 1), i = 1, \ldots, n \), with \( n_0 (= \pi_0 n) \) of the \( \mu_i \)'s being equal to 0 and the rest being equal to \( d = \sqrt{10} \), and an equicorrelation matrix with correlation \( \rho \). We then applied each aforementioned procedure to the generated data to test \( H_i : \mu_i = 0 \) against \( K_i : \mu_i \neq 0 \) simultaneously for \( i = 1, \ldots, n \), at level \( \alpha = 0.05 \). The above steps were repeated for 5,000 times.

In the simulation, the \( p \)-value \( P_i \) corresponding to the hypothesis \( H_i \) was calculated by \( P_i = 2(1 - \Phi(T_i)), i = 1, \ldots, n \), where \( \Phi(\cdot) \) is the cdf of \( N(0, 1) \). For those fixed sequence procedures, the order of the tested hypotheses was specified as follows: let the first \( n_1 \) hypotheses be false nulls, randomly insert \( m \) true null hypotheses among the \( n_1 \) false nulls indicating \( m \) ordering mistakes, and let the last \( n_0 - m \) hypotheses be true nulls. Specifically, for the fallback procedure, the pre-specified weights \( w_i, i = 1, \ldots, n \), for the \( n \) hypotheses are chosen to be an equally decreasing geometric sequence with a decreasing rate \( \gamma \) and a sum equal to one, that is, \( w_i = \frac{\gamma^{i-1}(1-\gamma)}{1-\gamma^n} \). Note that when \( \gamma = 1 \), the hypotheses are equally weighted and when \( \gamma = 0 \), the fallback procedure reduces to the conventional fixed sequence procedure. Finally, for notational convenience, the proposed Procedures A1-A3 are labeled PA1, PA2, and PA3, and the existing Holm, conventional fixed-sequence and fallback procedures are labeled HM, FS and FB, respectively.

In the simulation, we set \( n = 20, \pi_0 = 0.2, 0.4, 0.6 \) or 0.8, and \( m = 0, 2 \) or 4 for all aforementioned procedures. Specifically, we set \( \beta = 0.5 \) for Procedure A2 and \( \gamma = 0.5 \) for the fallback procedure.
Figures 2.1-2.3 present a comparison of the simulated FWERs of the proposed Procedures A1-A3 and the existing Holm, conventional fixed-sequence and fallback procedures. As seen from these figures, the FWERs of all these procedures are controlled at level $\alpha$ under different simulation scenarios. Specifically, when $m = 0$, as seen from Figure 1, the FWERs of Procedures A1-A3 and Holm procedure gradually turn into being decreasing with increasing $\pi_0$ and $\rho$. On the contrary, the FWERs of the conventional fixed sequence and fallback procedures are increasing in $\rho$ except when $\pi_0$ is large. When $m > 0$, as seen from Figure 2.3 and 2.5, the FWERs of all 6 procedures tend to become decreasing with increasing $\rho$.

Figures 2.4-2.6 present a comparison of the simulated average powers of the aforementioned 6 procedures. When $m = 0$, as seen from Figure 2.2, the powers of Procedure A2, conventional fixed sequence and Fallback procedures are increasing in $\rho$ and other 3 procedures perform steadily for different $\rho$. Among these 6 procedures, Procedure A2 is the most powerful for small or moderate $\pi_0$ and $\rho$; however, when $\rho$ is large, it is slightly less powerful than the conventional fixed-sequence and fallback procedures. When $m > 0$, as seen from Figure 2.4 and 2.6, the proposed Procedures A1-A3 perform very well for different values of $\pi_0$ and $\rho$. Among Procedures A1-A3, Procedure A2 or A3 are always slightly more powerful than Procedure A1 under different scenarios and Procedure A2 and A3 are comparable.

Summarizing the above observations, the aforementioned six procedures control the FWER well in each setting, however, there is not an uniformly powerful procedure. Compared to the Holm’s procedure, Procedure A1 is slightly less powerful, but it is easy to implement. If the test statistics are weakly correlated, Procedure A1, A3 and Holm’s procedure will be good choices, however, if the test statistics are highly correlated, Procedure A2 and fallback procedure will more likely perform better than other procedures.
Figure 2.1 Simulated FWERs of 6 procedures (PA1 - ···; PA2 - −−−; PA3 - − ■ −; FB - − × −; FS - −−; HM - −▲−) with dependent p-values generated from multivariate normal test statistics with common correlation $\rho$ for $n = 20, \alpha = 0.05, d = \sqrt{10}, N = 0, \beta = 0.5, \gamma = 0.5$. 

(a) $\pi_0 = 0.2$

(b) $\pi_0 = 0.4$

(c) $\pi_0 = 0.6$

(d) $\pi_0 = 0.8$
Figure 2.2  Simulated powers of six procedures (PA1 - ···; PA2 - −−−; PA3 - −■−; FB - −×−; FS - −; HM −▲−) with dependent p-values generated from equally correlated multivariate normal test statistics with correlation \( \rho \) for \( n = 20, \alpha = 0.05, d = \sqrt{10}, N = 0, \beta = 0.5, \gamma = 0.5 \).
Figure 2.3  Simulated FWERs of six procedures (PA1 - ···; PA2 - −−−; PA3 - -■--; FB - −×--; FS - ---; HM −▲−) with dependent $p$-values generated from equally correlated multivariate normal test statistics with correlation $\rho$ for $n = 20, \alpha = 0.05, d = \sqrt{10}, N = 2, \beta = 0.5, \gamma = 0.5$. 

(a) $\pi_0 = 0.2$

(b) $\pi_0 = 0.4$

(c) $\pi_0 = 0.6$

(d) $\pi_0 = 0.8$
Figure 2.4 Simulated powers of six procedures (PA1 - ⋯; PA2 - −−−; PA3 - ■; FB - ×; FS - ; HM - ▲) with dependent p-values generated from equally correlated multivariate normal test statistics with correlation \( \rho \) for \( n = 20, \alpha = 0.05, d = \sqrt{10}, N = 2, \beta = 0.5, \gamma = 0.5 \).
Figure 2.5  Simulated FWERs of six procedures (PA1 - ···; PA2 - −−−; PA3 - ■−−; FB - ×−−; FS - __; HM - ▲−−) with dependent $p$-values generated from equally correlated multivariate normal test statistics with correlation $\rho$ for $n = 20, \alpha = 0.05, d = \sqrt{10}, N = 4, \beta = 0.5, \gamma = 0.5$. 
Figure 2.6 Simulated powers of six procedures (PA1 - · · ·; PA2 - −−−; PA3 - −■−; FB - −×−; FS - −−; HM - −▲−) with dependent $p$-values generated from equally correlated multivariate normal test statistics with correlation $\rho$ for $n = 20, \alpha = 0.05, d = \sqrt{10}, N = 4, \beta = 0.5, \gamma = 0.5$. 

(a) $\pi_0 = 0.2$

(b) $\pi_0 = 0.4$

(c) $\pi_0 = 0.6$

(d) $\pi_0 = 0.8$
2.7 A Clinical Trial Example

We revisited a hypertension trial example analyzed in Dmitrienko et al. (2006). The purpose of this clinical trial was to test the efficacy and safety of four doses of an investigational drug versus placebo. The four doses, from the lowest to highest doses, were respectively labeled D1, D2, D3, and D4, and the placebo was labeled P. The primary endpoint was the reduction in diastolic blood pressure (measured in mm Hg). There are eight hypotheses including four dose-placebo contrasts and four pairwise contrasts. Since high doses were expected to be more efficacious than low doses, high dose-placebo contrasts (D4 vs. P, D3 vs. P) were tested before testing low dose-placebo contrasts (D2 vs. P, D1 vs. P). After testing these four dose-placebo comparisons, four pairwise comparisons were tested in an order of D4 vs D1, D4 vs. D2, D3 vs. D1, and D3 vs. D2. We pre-specified $\alpha = 0.05$ and applied the three newly proposed Procedures A1-A3 and three existing procedures Holm, conventional fixed sequence and fallback procedures to this example. Same as in Section 2.6, the pre-specified weights for fallback procedure are chosen to be a decreasing sequence with equally decreasing rate $\gamma$ and $\gamma$ is set to be 0.1, 0.5 or 0.9. Table 2.2 lists the raw $p$-values of the eight hypotheses and the test results using the aforementioned six procedures.

As seen from Table 2.2, Procedure A2 performs the best rejecting five null hypotheses at level 0.05. In contrast, the conventional fixed sequence procedure, Holm procedure, Procedure A1, and Procedure A3 only reject three null hypotheses. For the fallback procedure, its testing results depend on the pre-specified weights. When the equally decreasing rate $\gamma$ of weights is set to be 0.1, 0.5 or 0.9, it rejects 3, 4 or 3 hypotheses, respectively.
Table 2.2  Comparison Results of Six Procedures in the Hypertension Trial Example

<table>
<thead>
<tr>
<th></th>
<th>Raw $p$-value</th>
<th>PA1</th>
<th>PA2</th>
<th>PA3</th>
<th>HM</th>
<th>FS</th>
<th>FB1</th>
<th>FB2</th>
<th>FB3</th>
</tr>
</thead>
<tbody>
<tr>
<td>D4-P</td>
<td>0.0008</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>D3-P</td>
<td>0.0135</td>
<td>NR</td>
<td>R</td>
<td>R</td>
<td>NR</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>NR</td>
</tr>
<tr>
<td>D2-P</td>
<td>0.0197</td>
<td>NR</td>
<td>R</td>
<td>NR</td>
<td>NR</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>NR</td>
</tr>
<tr>
<td>D1-P</td>
<td>0.7237</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>D4-D1</td>
<td>0.0003</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>NR</td>
<td>NR</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>D4-D2</td>
<td>0.2779</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>D3-D1</td>
<td>0.0054</td>
<td>R</td>
<td>R</td>
<td>NR</td>
<td>R</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>R</td>
</tr>
<tr>
<td>D3-D2</td>
<td>0.8473</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>Rejection number</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Note: P = Placebo and D1-D4 denote four doses of the investigational drug. PA1-PA3 = Proposed Procedure A1-A3, HM = Holm procedure, FS = Conventional Fixed Sequence Procedure, FB1-FB3 = Fallback Procedure with $\gamma = 0.1, 0.5$ and 0.9. For PA2, $\beta = 0.5$ and for FB1-FB3, $\gamma$ denotes the equally decreasing rate of weights assigned to 8 hypotheses. The overall Type I error rate is $\alpha = 0.05$. 

33
2.8 Further Improvement

In the preceding sections, only the marginal distributions of the null \( p \)-values are used when developing the newly introduced Procedures A1-A3. However, in practice, the null \( p \)-values often have a known common pairwise joint distribution, and it would be worthwhile to consider further improving the aforementioned procedures by explicitly utilizing such additional dependence information, which could potentially produce more powerful FWER controlling procedures than Procedures A1-A3. So, with that in mind, we present some improved results here under the following assumption.

**Assumption 2.1.** The null \( p \)-values \( \hat{P}_1, \ldots, \hat{P}_{n_0} \) have a known common pairwise joint distribution function \( F(u,v) = \Pr(\hat{P}_i \leq u, \hat{P}_j \leq v) \).

Under Assumption 1, Theorem 2.1 can be further strengthened as follows.

**Theorem 2.4.** Under Assumption 2.1, the generalized fixed sequence procedure with critical value function \( \alpha(s,t) \) strongly controls the FWER at level \( \alpha \) if for any \( 0 \leq s, t \leq n - 1 \),

\[
\sum_{t=0}^{n-s-1} \alpha(s,t) - \sum_{t=1}^{n-s-1} F(\alpha(s,t-1), \alpha(s,t)) \leq \alpha, \tag{2.8}
\]

where \( \alpha(s,t) \) is increasing in \( s \) and decreasing in \( t \).

For proof of Theorem 2.4, see Appendix.

**Remark 2.6.** The amount of improvement of the critical values of the aforementioned procedure depends on the pairwise joint cdf \( F(u,v) \). Assume \( \hat{P}_i \sim U(0,1), i = 1, \ldots, n_0 \), then under perfect positive correlation where \( F(\alpha(s,t-1), \alpha(s,t)) = \alpha(s,t) \), (2.8) reduces to \( \alpha(s,t) \leq \alpha \), which is a remarkable improvement on (3). On the other hand, under independence where \( F(\alpha(s,t-1), \alpha(s,t)) = \alpha(s,t-1)\alpha(s,t) \), there is only a limited improvement.

Based on Theorem 2.4, we can respectively develop improved versions of Procedures A1-A3 as follows.
Procedure B1. Test the hypotheses according to the generalized fixed sequence procedure with critical value function $\alpha(s, t) = \alpha(s, 0)$ for $s, t = 0,\ldots, n - 1$ and $\alpha(s, 0)$ satisfies the following equation for $s = 0,\ldots, n - 1$,$$(n - s)\alpha(s, 0) - (n - s - 1)F(\alpha(s, 0), \alpha(s, 0)) = \alpha. \quad (2.9)$$

Procedure B2. Test the hypotheses according to the generalized fixed sequence procedure with critical value function $\alpha(s, t) = \alpha(0, 0)\beta^t$ for $s, t = 0,\ldots, n - 1$ and $\alpha(0, 0)$ satisfies the following equation,$$
\frac{1 - \beta^n}{1 - \beta} \alpha(0, 0) - \sum_{t=0}^{n-1} F(\alpha(0, 0)\beta^{t-1}, \alpha(0, 0)\beta^t) = \alpha, \quad (2.10)$$where $\beta$ is a pre-specified constant satisfying $0 \leq \beta < 1$.

Procedure B3. Test the hypotheses according to the generalized fixed sequence procedure with critical value function $\alpha(s, t) = \alpha(s, 0) - \frac{2t\alpha}{n^2}$ for $s, t = 0,\ldots, n - 1$ and $\alpha(s, 0)$ satisfies the following equation for $s = 0,\ldots, n - 1$,$$
(n - s) \left( \alpha(s, 0) - \frac{(n - s - 1)\alpha}{n^2} \right) - \sum_{t=1}^{n-t-1} F\left( \alpha(s, 0) - \frac{2(t - 1)\alpha}{n^2}, \alpha(s, 0) - \frac{2t\alpha}{n^2} \right)
= \alpha. \quad (2.11)$$

It is easy to see that the critical value functions for Procedures B1-B3 are all decreasing in $t$. And, it can be shown that equations (2.9)-(2.11) all have solutions for any cdf $F(u, v)$, and even have unique solutions if certain conditions are imposed on $F(u, v)$, for example, $F(u, v)$ is the cdf of a bivariate normal distribution. There is no guarantee that a solution for $\alpha(s, 0)$ in (9) and (11) is increasing in $s$. If it is not, a minor adjustment of $\alpha(s, 0)$ can always be made to force $\alpha(s, 0)$ to be increasing, although the resulting $\alpha(s, t)$ becomes a little smaller. For simplicity, in the following discussions, we assume that (2.9)-(2.11) all have unique solutions and the resulting critical value functions $\alpha(s, t)$ of Procedures B1-B3 are increasing in $s$ and decreasing.
in $t$. Finally, we need to point out that it is typically not possible to obtain closed form solutions for (2.9)-(2.11). Instead, these solutions can be approximated numerically by using the bisection method (Arfken, 1985).

Remark 2.7. The critical value functions in Procedures B1-B3 maintain the same monotonicity properties as their corresponding Procedures A1-A3, respectively. For example, the critical value function in Procedure B1 is increasing in $s$ and constant in $t$, and the critical value function in Procedure B2, like Procedure A2, decreases by the constant rate $\beta$ for every unit increase in $t$.

Theorem 2.5. Under Assumption 2.1, Procedures B1-B3 strongly control the FWER at $\alpha$.

Proof. For Procedures B1-B3, all of their critical value functions $\alpha(s, t)$ satisfy (2.8). Based on Theorem 2.4, the proof is complete. \hfill \square

In order to show the improvements of critical values of Procedures B1-B3 over Procedures A1-A3, we performed some numerical calculations to illustrate the gains of critical values of Procedures B1-B3 over Procedures A1-A3, respectively. We consider $n$ two-sided hypothesis tests and assume that any pair of test statistics associated with true null hypotheses follows bivariate normal distribution with common pairwise correlation $\rho$. We set the parameter $\beta$ in procedure A2 and B2 to be $\beta = 0.5$. Table 2.3 summarizes the numerical results of calculating the critical values of procedure A1 and B1 for $n = 10, \rho = 0.3, 0.5, 0.8$ and the improvement percentage of the critical values of Procedure B1 over A1. Table 2.4 and 2.5 show the similar comparison results for Procedure A2 vs B2 with the same values of $n$ and $\rho$ as in Table 2.3, and Procedure A3 vs B3 with $n = 5, \rho = 0.5$ and 0.8. As seen from these three tables, when $\rho$ is small, the percentages of improvement of critical values are pretty small and are generally no more than 2%. However, when $\rho$ is large, the improvements are remarkable and some are even over 30%.
Table 2.3  Percentage of Improvement for Critical Values of Procedure A1 and Procedure B1 with $n = 10$

<table>
<thead>
<tr>
<th>s</th>
<th>Procedure A1</th>
<th>Procedure B1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho = 0.2$</td>
<td>$\rho = 0.5$</td>
</tr>
<tr>
<td>0</td>
<td>0.005</td>
<td>0.005060 (1.2)</td>
</tr>
<tr>
<td>1</td>
<td>0.005556</td>
<td>0.005627 (1.3)</td>
</tr>
<tr>
<td>2</td>
<td>0.00625</td>
<td>0.006336 (1.4)</td>
</tr>
<tr>
<td>3</td>
<td>0.007143</td>
<td>0.007250 (1.5)</td>
</tr>
<tr>
<td>4</td>
<td>0.008333</td>
<td>0.008469 (1.6)</td>
</tr>
<tr>
<td>5</td>
<td>0.01</td>
<td>0.010178 (1.8)</td>
</tr>
<tr>
<td>6</td>
<td>0.0125</td>
<td>0.012746 (2.0)</td>
</tr>
<tr>
<td>7</td>
<td>0.016667</td>
<td>0.017027 (2.2)</td>
</tr>
<tr>
<td>8</td>
<td>0.025</td>
<td>0.0255461 (2.2)</td>
</tr>
<tr>
<td>9</td>
<td>0.05</td>
<td>0.05</td>
</tr>
</tbody>
</table>
Table 2.4  Percentage of Improvement for Critical Values of Procedure A2 and Procedure B2 with \( n = 10, \beta = 0.5 \)

<table>
<thead>
<tr>
<th>( t )</th>
<th>Procedure A2</th>
<th>( \rho = 0.2 )</th>
<th>( \rho = 0.5 )</th>
<th>( \rho = 0.8 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.025024</td>
<td>0.025476 (1.8)</td>
<td>0.026997 (7.9)</td>
<td>0.032938 (31.6)</td>
</tr>
<tr>
<td>1</td>
<td>0.012512</td>
<td>0.012738 (1.8)</td>
<td>0.013498 (7.9)</td>
<td>0.016469 (31.6)</td>
</tr>
<tr>
<td>2</td>
<td>0.006256</td>
<td>0.006369 (1.8)</td>
<td>0.006749 (7.9)</td>
<td>0.008235 (31.6)</td>
</tr>
<tr>
<td>3</td>
<td>0.003128</td>
<td>0.003184 (1.8)</td>
<td>0.003375 (7.9)</td>
<td>0.004117 (31.6)</td>
</tr>
<tr>
<td>4</td>
<td>0.001564</td>
<td>0.001592 (1.8)</td>
<td>0.001687 (7.9)</td>
<td>0.002059 (31.6)</td>
</tr>
<tr>
<td>5</td>
<td>0.000782</td>
<td>0.000796 (1.8)</td>
<td>0.000844 (7.9)</td>
<td>0.001029 (31.6)</td>
</tr>
<tr>
<td>6</td>
<td>0.000391</td>
<td>0.000398 (1.8)</td>
<td>0.000422 (7.9)</td>
<td>0.000515 (31.7)</td>
</tr>
<tr>
<td>7</td>
<td>0.000195</td>
<td>0.000199 (2.1)</td>
<td>0.000211 (8.2)</td>
<td>0.000257 (31.8)</td>
</tr>
<tr>
<td>8</td>
<td>9.775e-05</td>
<td>9.951e-05 (1.8)</td>
<td>0.000105 (7.4)</td>
<td>0.000129 (32.0)</td>
</tr>
<tr>
<td>9</td>
<td>4.887e-05</td>
<td>4.976e-05 (1.8)</td>
<td>5.273e-05 (7.9)</td>
<td>6.433e-05 (31.6)</td>
</tr>
</tbody>
</table>
Table 2.5  Percentage of Improvement for Critical Values of Procedure A3 and Procedure B3 with \( n = 5 \)

<table>
<thead>
<tr>
<th>( s )</th>
<th>( t )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.018</td>
<td>0.014</td>
<td>0.01</td>
<td>0.006</td>
<td>0.002</td>
</tr>
<tr>
<td>1</td>
<td>0.0185</td>
<td>0.0145</td>
<td>0.0105</td>
<td>0.0065</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.0207</td>
<td>0.0167</td>
<td>0.0127</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>0.027</td>
<td>0.023</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.05</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Procedure B3 with \( \rho = 0.5 \)

<table>
<thead>
<tr>
<th>( s )</th>
<th>( t )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.02</td>
<td>0.016</td>
<td>0.012</td>
<td>0.008</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.02</td>
<td>0.016</td>
<td>0.012</td>
<td>0.008</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.0222</td>
<td>0.0182</td>
<td>0.0142</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.0289</td>
<td>0.0249</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.05</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Procedure B3 with \( \rho = 0.8 \)

<table>
<thead>
<tr>
<th>( s )</th>
<th>( t )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0219</td>
<td>0.0179</td>
<td>0.0139</td>
<td>0.0099</td>
<td>0.0059</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.0232</td>
<td>0.0192</td>
<td>0.0152</td>
<td>0.0112</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.0264</td>
<td>0.0223</td>
<td>0.0184</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.0333</td>
<td>0.0293</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.05</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
2.9 Concluding Remarks

The main focus of this chapter has been to advance the theory and methods of fixed sequence multiple testing for controlling the FWER. We introduced a generalized fixed sequence procedure and gave sufficient conditions for its FWER control. We proposed several new fixed sequence procedures by considering different critical value functions. Through extensive simulation studies, we have shown advantages of our proposed generalized fixed sequence procedures over the existing FWER controlling procedures. When the pairwise joint distributions of the true null $p$-values are known, we have improved the aforementioned procedures by incorporating the distributional information into the construction of these procedures while maintaining the control of the FWER. Specifically, in the case of bivariate normal distribution with common correlation, we have numerically shown improvements of the critical values of the improved procedures over the aforementioned procedures.

To use the fixed sequence methods, prior knowledge of the ordering of the tested hypotheses is required. When the ordering is completely correct, i.e., the false null hypotheses are ordered ahead of the true null hypotheses, even the conventional fixed sequence procedure, which does not allow any acceptance, has a natural advantage over the existing $p$-value based stepwise FWER controlling procedures such as the Holm and Hochberg procedures. However, when the ordering is not completely correct, the conventional fixed sequence procedure usually loses its edge over those stepwise procedures, whereas our proposed fixed sequence procedures can still perform well. Of course, when the ordering information is completely incorrect, our proposed fixed sequence procedures no longer have the advantage over those $p$-value based stepwise procedures. Therefore, a natural extension might be to use a combination of the a-priori ordering information and the $p$-values to order the hypotheses to be tested and then develop FWER controlling procedures based on such ordering.
CHAPTER 3

FAMILY-BASED GRAPHICAL APPROACH FOR SEQUENTIALLY TESTING HIERARCHICALLY ORDERED FAMILIES OF HYPOTHESES

3.1 Motivation

In Chapter 2, we deal with the hypotheses testing problem of single family. However, in complex clinical trial studies, the multiple research objectives often need to be grouped into several sets of objectives based on their hierarchical relationships and the formulated hypotheses are correspondingly grouped as several ordered families of hypotheses. For testing single family of hypotheses, Bretz et al. (2009) proposed a graphical approach by which most of the single family FWER controlling procedures can be elegantly represented by directed graphs. Figure 3.1 shows the graphical visualization of the Bonferroni-Holm procedure for three hypotheses and equal initial allocation of the critical values. As seen from Figure 3.1, each tested hypothesis is represented by a vertex with initial critical value $\alpha/3$. There is an edge with a weight $1/2$ between any two vertices. According to Figure 3.1, once a hypothesis $H_i, i = 1, 2, 3$ is rejected, its critical value $\alpha/3$ will split equally and passed on to the remaining two hypotheses as indicated by the directed edges. Thus, the critical values for the remaining two hypotheses will be updated as $\alpha/3 + \alpha/6 = \alpha/2$ and the weights on the edges between the remaining hypotheses also need to update by a specific algorithm. The whole procedure stops when no further hypothesis can be rejected. Bretz et al. (2009)’s approach is explicit and efficient for single family multiple testing procedure. However, in some cases involving multiple families of hypotheses, the graphical approach seems to show a lack of clarity. Consider an example that nine hypotheses are grouped into three families where each family
Figure 3.1 Hypothesis-based graphical visualization of the Bonferroni-Holm procedure with three hypotheses.

The procedure consists of three hypotheses, denoted as $F_i = \{H_{i1}, H_{i2}, H_{i3}\}$, for $i = 1, 2, 3$. The initial critical values for three families are set to be 0.04, 0.005 and 0.005. It is known in advance that these three families of hypotheses should be tested in a pre-specified order based on the hierarchical relationships among them and each one is locally tested by regular Bonferroni-Holm procedure with equal weights. Consider a testing strategy that if all three hypotheses in $F_1$ are rejected, then its critical value 0.04 will be passed on to $F_2$ and $F_3$ with weights $4/5$ and $1/5$, respectively. Otherwise, $F_2$ will be tested at its initially allocated critical value 0.005. Moreover, if all three hypotheses in $F_2$ are rejected, its whole local critical value will also be transferred to $F_3$. By using Bretz et al. (2009)s' hypotheses-based graphical approach, Figure 3.2 displays graphically the aforementioned testing strategy which involves edges with infinitesimally small weights $\epsilon$. An edge with an $\epsilon$ on it implies no critical value is passed. However, if a vertex only has infinitesimal outgoing edges, those edges are updated to non-infinitesimal edges such that critical value corresponding to this vertex can be passed to other hypotheses. It is easy to see that when dealing with multiple families of hypotheses, hypothesis-based graphical representation becomes too complicated. In fact, for multiple families of hypotheses problem, it is common to have hierarchical logical relationships among the families of hypotheses and thus, we focus more on these logical relationships between families.
rather than on the procedure locally used within each family. Therefore, we are motivated to develop a family-based graphical approach to illustrate the general hierarchical logical relationships among those families. And as for local procedure within each family, we can specifically illustrate it using hypothesis-based graphical approach alone without increasing complexity of resulting graphs. We assume that each vertex in our family-based graph corresponds to a family of hypotheses instead of a single hypothesis and between any two families, there is a directed edge with a pre-specified weight associated with it. Figure 3.2 can be equivalently illustrated as Figure 3.3. Within $F_i$, $i = 1, 2, 3$, we use regular Bonferroni-Holm procedure which can be particularly described by Figure 3.1. The weights on the edges means that after all three hypotheses in one family are rejected, the corresponding local critical value will be proportionally added to the subsequent families based on the edges with weights. Apparently, family-based graphical approach extracts the hierarchical logical relationships among tested families of hypotheses which is normally overlooked by hypothesis-based graphical approach. In particular, when dealing with large number of families of hypotheses, the family-based graphical approach, which does not involve small weights $\epsilon$, reduces the complexity of graph and make the testing strategy more
explicit and easy to explain. The rest of the chapter is organized as follows. In Section 3.2, we present some notations and basic assumption which are used through the whole chapter. The theoretical results are introduced in Section 3.3 and several cases to compare with the Bretz et al. (2009)s’ hypothesis-based graphical approach are discussed in Section 3.4. A clinical trial example is presented in Section 3.5 to show how the proposed approach can be used to better describe a procedure for testing multiple families of hypotheses procedure. Some concluding remarks and possible future research are discussed in Section 3.6.

3.2 Preliminary

In this section, we present some basic notations and assumptions. Suppose there are $N \geq 2$ hypotheses divided into $m \geq 2$ families based on their logical relationships. Furthermore, these $m$ families are grouped into $n$ levels $L_1, \cdots, L_n$ based on their hierarchical relationships, where $L_i$ consists of $l_i \geq 1$ families $F_{i1}, \cdots, F_{il_i}$ and $\sum_{i=1}^{l_i} l_i = m$. Moreover, each family $F_{ij}$ has $n_{ij} \geq 1$ null hypotheses, denoted as $F_{ij} = \{H_{ij1}, \cdots, H_{ijn_{ij}}\}$, for $i = 1, \cdots, n, j = 1, \cdots, l_i$ such that $\sum_{i=1}^{n} \sum_{j=1}^{l_i} n_{ij} = N$. Each hypothesis $H_{ijk}$ is tested based on its $p$-value $P_{ijk}, i = 1, \cdots, n, j = 1, \cdots, l_i, k = 1, \cdots, n_{ij}$. In this chapter, the true null $p$-values are always assumed to be stochastically greater than or equal to uniform distribution on $[0, 1]$. That is, for
given $i = 1, \cdots, n$ and $j = 1, \cdots, \ell_i$, if we let $T_{ij}$ denote the set of true null hypotheses in $F_{ij}$, then for $u \in [0, 1]$

$$\Pr \left\{ P_{ijk} \leq u \mid H_{ijk} \in T_{ij} \right\} \leq u, \quad (3.1)$$

where $k = 1, \cdots, n_{ij}$.

Given a pre-specified critical value $\alpha$, let $\alpha_1, \cdots, \alpha_n$ denote the initial allocation of $\alpha$ to the $n$ levels, such that $\sum_{i=1}^{n} \alpha_i = \alpha$. Moreover, we let the initial allocation of $\alpha_i$ to the $\ell_i$ families within level $L_i$ be $\alpha_{i1}, \cdots, \alpha_{il_i}$ with $\sum_{j=1}^{l_i} \alpha_{ij} = \alpha_i$. We need to strongly control the overall FWER at level $\alpha$ for the whole family $F = \bigcup_{i=1}^{n} \bigcup_{j=1}^{l_i} F_{ij}$ of all hypotheses, i.e., the probability of incorrectly rejecting at least one true null hypothesis will be controlled at $\alpha$ regardless of which and how many null hypotheses within each family are true.

Suppose each family of hypotheses are tested at its local critical value by any local FWER controlling procedure under arbitrary dependence. Due to the hierarchical relationships among the families, we consider an approach such that families of hypotheses in the different levels are tested sequentially. In other words, families in level $L_{i+1}$ can not be tested until all families within level $L_i$ are tested, $1 \leq i \leq n - 1$. However, families of hypotheses in the same level can be tested in any order. If rejections occurs in one family, part of its local critical value is distributed to the families in subsequent levels. This procedure stops testing when all the families of hypotheses in the last level $L_n$ are tested.

Let $g_{ijkl}$ denote the proportion of the local critical value of $F_{ij}$ transferred to $F_{kl}$. Let $G$ denote a set of transition coefficients with elements $g_{ijkl}$ satisfying the following conditions

$$\sum_{k=i+1}^{n} \sum_{l=1}^{l_k} g_{ijkl} \leq 1; g_{ijkl} = 0 \text{ if } i \geq k.$$
Figure 3.4 General graphical representation of family-based graphical approach.

Note that \( G \) includes \( m^2 \) elements. Also, the local critical values can only be transferred from the family in higher level to the one in lower level.

Based on the initial allocation of critical values \( \alpha_{ij}, i = 1, \cdots, n, j = 1, \cdots, l_i \) and transition coefficient set \( G \), we can construct a directed acyclic graph where the families \( F_{ij}, i = 1, \cdots, n, j = 1, \cdots, l_i \) are represented by vertices with associated initial critical value \( \alpha_{ij} \). The transition coefficient set \( G \) provides all the directed edges and its associated coefficients. Specifically, \( g_{ijkl} \) implies that there is a directed edge from \( F_{ij} \) to \( F_{kl} \). Since each vertex is associated with a family instead of a hypothesis. We term such graph as a family-based graph. Generally, the family-based graph can be described as a set of quintuples \( G_{ijkl} = (F_{ij}, F_{kl}, \alpha_{ij}, \alpha_{kl}, g_{ijkl}) \), where \( F_{ij} \) is head vertex with \( \alpha_{ij} \), \( F_{kl} \) is tail vertex with \( \alpha_{kl} \), and \( g_{ijkl} \) indicates the coefficient associated the directed edge between two vertices. The general family-based graph is illustrated in Figure 3.4.

In order to quantify the amount of local critical value of current family that can be passed on to families in subsequent levels, we need to use the definition of error rate function introduced in Section 1.3.1. Suppose for each family, a local FWER
controlling procedure is pre-specified. Thus, each family has its own local procedure with a particular error rate function. Let $\alpha_{ij}^*$ denote the local critical value for testing family $F_{ij}$. Let $A_{ij}$ be the set of accepted hypotheses associated to $F_{ij}$ with cardinality $|A_{ij}|$. Then based on $A_{ij}$, we can calculate $e^*(A_{ij})$ after testing $F_{ij}$ at level $\alpha_{ij}^*$ where $e^*(.)$ is the upper bound of error rate function corresponding to the local procedure. An amount of its local critical value $\alpha_{ij}^* - e^*(A_{ij})$ will be transferred to the families in the subsequent levels.

Remark 3.1. The error rate function was first introduced by Dmitrienko et al. (2008). They used it to develop a simple stepwise approach for gatekeeping strategies. In their discussion, the error rate function is required to be strictly less than $\alpha$ unless all of the hypotheses in one family are rejected. This is termed as separability condition. However, the error rate function used in this chapter is more general. The separability condition is unnecessary when choosing local procedure for family-based graphical approach.

### 3.3 Main Theoretical Results

In this section, we consider to construct a family-based graphical approach and prove that it strongly controls the overall FWER at level $\alpha$ under arbitrary dependence. We will begin with a simple case of two levels with two families of hypotheses in each level in Section 3.3.1. The general case of multiple levels with arbitrary number of families within each level will be introduced in Section 3.3.2.

#### 3.3.1 Two-Level Family-based Graphical Approach with Four Families

Consider $m = 4$ families of hypotheses being divided into two levels $L_1, L_2$ based on their hierarchal relationship with two families of hypotheses within each level. Based on the notation in Section 3.2, we can define a two-level family-based graphical approach though the following algorithm:
Algorithm 3.1.

**Step 1.** Set $L_1 = \{F_{11}, F_{12}\}, L_2 = \{F_{21}, F_{22}\}$. Test family $F_{1j}, j = 1, 2$, using any FWER controlling procedure at critical value $\alpha_{1j}$, calculate $e^*(A_{1j})$.

Update the graph:

$$L_1 \rightarrow L_1 \setminus F_{1j};$$

$$\alpha_{2k} \rightarrow \alpha_{2k} + (\alpha_{1j} - e^*(A_{1j}))g_{1j2k}, \ k = 1, 2;$$

$$g_{1l2k} \rightarrow g_{1l2k},\ l \neq j.$$  

If $L_1 \neq \emptyset$, go back to step 1; Otherwise, go to next step.

**Step 2.** Test $F_{2k}$ use any FWER controlling procedure at level $\alpha_{2k}$ and update the graph:

$$L_2 \rightarrow L_2 \setminus F_{2k}.$$  

If $L_2 \neq \emptyset$, go back to step 2; Otherwise stop.

Algorithm 3.1 starts the test from the families $F_{1j}, j = 1, 2$, in $L_1$. Once $F_{1j}$ has been tested, the critical values associated to the lower level families will be updated according to transition coefficient set $G$. And then, $G$ itself is updated by deleting all the elements associated with $F_{1j}$. This procedure can be fully described by a
graph given in Figure 3.5. For the aforementioned algorithm, we have the following theorem.

**Theorem 3.1.** The two-level multiple testing procedure described in Algorithm 3.1 strongly controls the overall FWER at level $\alpha$ under arbitrary dependence.

For the proof of theorem 3.1, see Appendix B.

### 3.3.2 General Multilevel Family-based Graphical Approach

The aforementioned two-level four-family case demonstrates the inherent nature of the family-based graphical approach. Now we generalize the approach from two levels to arbitrary $n, n \geq 2$ levels with arbitrary number of families of hypotheses within each level and prove that it also strongly controls the overall FWER at level $\alpha$ under arbitrary dependence. The general multilevel family-based graphical approach is defined through the following algorithm:

**Algorithm 3.2.**

**Step** $i$ $(1 \leq i \leq n-1)$. Test family $F_{ij}, j = 1, \cdots, l_j$ using any FWER controlling procedure at level $\alpha_{ij}$, and calculate $e^*(A_{ij})$.

Update the graph:

$$L_i \rightarrow L_i \setminus F_{ij};$$

$$\alpha_{kl} \rightarrow \alpha_{kl} + (\alpha_{ij} - e^*(A_{ij}))g_{ijkl}, \ k = i + 1, \cdots, n, \ l = 1, \cdots, l_k;$$

$$g_{iskl} \rightarrow g_{iskl}, \ s \neq j.$$

If $L_i \neq \emptyset$, go back to step $i$; Otherwise, go to next step.

**Step** $n$. Test $L_n = \{F_{n1}, \cdots, F_{nl_n}\}$. Use any FWER controlling procedure at level $\alpha_{nj}$ to test $F_{nj}$ and update $L_n \rightarrow L_n \setminus F_{nj}$. If $L_n \neq \emptyset$, go back to step $n$; Otherwise stop.
For this general multilevel family-based graphical approach, we have the following theorem.

**Theorem 3.2.** The general multilevel family-based graphical approach strongly controls the overall FWER at level $\alpha$ under arbitrary dependence.

For the proof of theorem 3.2, see Appendix B.

**Remark 3.2.** Consider the case that there is only one family in each level, i.e., there is a fixed sequence of families $F_{i1}, i = 1, \ldots, n$. Suppose initial critical values for all the families are $\alpha, 0, \ldots, 0$ and the elements of transition coefficient set $G$ are given by $g_{ik1} = 1$, if $1 \leq i \leq n - 1, k = i + 1$ and 0 otherwise. Then, the following remarks can be noted regarding to this approach.

1. If each family is tested using a local procedure which satisfies separability condition, i.e., the error rate function of the corresponding local procedure is strictly smaller than $\alpha$, then the multilevel family-based graphical approach can be regarded as a parallel gatekeeping procedure which is equivalent to the general multistage gatekeeping procedure proposed by Dmitrienko et al. (2008). The examples of such local procedures include Bonferroni, truncated Holm procedure, truncated fallback procedure, etc. For more details about truncated procedures, see Dmitrienko et al. (2008).

2. If each family is tested using a local procedure which do not satisfies separability condition, i.e., the upper bound on its error rate function is given by $e^*(I) = \alpha$ if $I \neq \emptyset$ or 0 if $I = \emptyset$, then the corresponding multilevel graphical approach is equivalent to a serial gatekeeping procedure. The examples of such local procedures including standard Holm’s procedure and conventional fixed sequence procedure, etc.

3. If each family has only one null hypothesis, then the multilevel graphical approach is equivalent to the conventional fixed sequence procedure.
4. If there are some dependence information regarding to the $p$-values of hypotheses in one family known in advance, more choices of local procedure can be made. For example, suppose it is known that the $p$-values corresponding to the hypotheses in one family are positive dependent or independent. Then, we can choose Hochberg or truncated Hochberg procedure as its local procedure.

3.4 Discussions

In Section 3.3, we introduced a family-based graphical approach for constructing the overall FWER controlling procedures when testing multiple hierarchically ordered families of hypotheses. Compared to Bretz et al. (2009)’s hypothesis-based graphical approach, it is more appropriate to deal with modern complex multiplicity issues arising in clinical trials with multiple research objectives grouped into several sets of objectives. In this section, we will use three cases in Bretz et al. (2009) to illustrate the efficiency and simplicity of our proposed approach comparing with hypothesis-based graphical approach in dealing with the problem of testing multiple families of hypotheses problem. In Figure 3.6, 3.7 and 3.8, the original hypothesis-based graphs in Bretz et al. (2009) are displayed in the left side of each figure, and its equivalent family-based graph we construct is displayed in the right side of the figure.

Case 1. Consider the case in Figure 3.6. The left side of Figure 3.6 displays the hypothesis-based procedure. We here consider an equivalent family-based procedure involving $m = 3$ families and $L = 2$ levels. $L_1 = \{F_{11}, F_{12}\}$ and $L_2 = \{F_{21}\}$. $F_{11} = \{H_1\}, F_{12} = \{H_2\}$ and $F_{21} = \{H_3, H_4\}$. Initially, the critical values allocated to the three families are $\alpha/2, \alpha/2, 0$, respectively, and the transition coefficient set $G$ is given by

$$g_{1121} = g_{1221} = 1;$$

$$g_{2111} = g_{2112} = g_{1112} = g_{1211} = 0.$$
Figure 3.6 Family-based graphical visualization of Case 1.

The procedure starts with testing \( F_{11}(F_{12}) \) using Bonferroni method. If \( H_1(H_2) \) is rejected, its critical value \( \alpha/2 \) is transferred to \( F_{21} \) as indicated by the directed edges with weight 1, such that the critical value of \( F_{21} \) is updated to \( \alpha_{21} = \alpha/2 \). Then, the procedure continues testing \( F_{12}(F_{11}) \) using Bonferroni method again. Once \( H_2(H_1) \) is rejected, its critical value \( \alpha/2 \) will be added to \( \alpha_{21} \). Otherwise, no critical value is transferred. After testing both families in \( L_1 \), if \( \alpha_{21} \neq 0 \), we continue testing \( H_3 \) and \( H_4 \) from \( F_{21} \) using Holm procedure with equal critical value \( \alpha_{21}/2 \) for each hypothesis. Thus, it specifies the equivalent sequentially multiple testing procedure displayed in the left side of Figure 3.6 using our family-based graphical procedure. It is easy to see that the right side of Figure 3.6 simply and clearly describes the hierarchical relationship among all hypotheses.

There are some situations where the hypotheses in one family can be tested only if all the hypotheses from another family are rejected. If one use the original hypothesis-based graphical approach to deal with the problem, then the generated graphs include edges with infinitesimally small weights, which are complex and difficult to communicate. However, if using the family-based graphical approach, the small weights can be removed in the graph.

Case 2. Consider a case involving three hypotheses \( H_1, H_2 \) and \( H_3 \). Suppose only if both \( H_1 \) and \( H_2 \) are rejected, \( H_3 \) has the chance to be tested. The graph using original graphical approach is shown in the left side of Figure 3.7 with a infinitesimally
small weight $\epsilon$. If using family-based graphical approach which is shown in the right graph of Figure 3.7, this procedure turns out to be a simple two levels, two families procedure with $L_1 = \{F_1\}$ and $L_2 = \{F_2\}$ where $F_1 = \{H_1, H_2\}$ and $F_2 = \{H_3\}$. The initial critical value for two families are, respectively, $\alpha, 0$. Thus, the specific procedure can be described as the following: test the two null hypotheses in $F_1$ using regular Holm method at level $\alpha$. If both hypotheses are rejected, then $\alpha$ can be passed on to $H_3$ in $F_2$ such that $H_3$ can be tested at critical value $\alpha$. Otherwise, the test stops.

**CASE 3.** Consider a more complicated case. Assume that there are four hypotheses $H_1, H_2, H_3$ and $H_4$. $H_3$ and $H_4$ are of interest only if both $H_1$ and $H_2$ were rejected. The graph using hypothesis-based graphical approach is shown in the left side of Figure 3.8 with an infinitesimally small weight $\epsilon$. Note that there involves pre-fixed weights $r_1$ and $r_2$, such that $r_1 + r_2 = 1$. The left graph in Figure 3.8 visualizes the procedure that if both hypotheses $H_1$ and $H_2$ are rejected, the critical value $\alpha$ is shuffled to $H_3$ and $H_4$ according to the weights $r_1$ and $r_2$ such that $H_3$ receives $r_1 \alpha$ and $H_4$ receives $r_2 \alpha$. If using family-based graphical approach which is shown in the right graph of Figure 3.8, this procedure turns out to be a simple two levels, two families procedure with $L_1 = \{F_1\}$ and $L_2 = \{F_2\}$ where $F_1 = \{H_1, H_2\}$ and $F_2 = \{H_3, H_4\}$. The initial critical values for two families are $\alpha$ and 0, respectively. Thus, the specific procedure can be described as the following: perform the Holm
procedure at level $\alpha$ for the two hypotheses in $F_1$. If both $H_1$ and $H_2$ can be rejected, then $\alpha$ can be passed on to $F_2$. But unlike case 2, weighted Holm procedure is used as local procedure to test the hypotheses in $F_2$ with weights $r_1$ and $r_2$.

Remark 3.3. Through case 2 and 3, it is easy to see that when dealing with complex multiplicity issues, our proposed family-based graphical approach usually makes the explanation of the whole procedure more clearly and easier to understand than the original hypothesis-based graphical approach with infinite small weight $\epsilon$.

### 3.5 A Clinical Trial Example

In this section, we apply our family-based graphical approach and Bretz et al. (2009)’s hypothesis-based graphical approach to illustrate the same multiple testing procedure for a real clinical trial example. The following discussion is meant to illustrate how the family-based graphical approach can be used to efficiently demonstrate a given multiple testing strategy. We revisit the Type II diabetes clinical trial example in Dmitrienko et al.(2007). The trial compares three doses of an experimental drug (Doses L, M and H) versus placebo (Plac) with respect to one primary endpoint (P: Haemoglobin A1c), and two secondary endpoint (S1: Fasting serum glucose, S2: HDL cholesterol). The three endpoints will be examined at each of the three doses. So a total of nine null hypotheses will be defined and grouped into three families. Family $F_1$ consists of three dose-placebo comparisons corresponding to the primary endpoint
(P): H vs Plac \((H_{11})\), M vs Plac \((H_{12})\) and L vs Plac \((H_{13})\). Family \(F_2\) and \(F_3\) consist of dose-placebo comparisons corresponding to the two secondary endpoints, respectively. We pre-specified \(\alpha = 0.05\). The raw \(p\)-values for nine null hypotheses are given in Table 3.1.

Based on the hierarchical relationship among the hypotheses known in advance, \(F_1\) must be tested before testing \(F_2\) and \(F_3\). In the following, we consider two testing multiple testing strategies and use the family-based and hypothesis-based graphical approaches to demonstrate them, respectively. Table 3.1 listed the testing results.

**Procedure 1.** We assume that \(F_2\) and \(F_3\) are equally important, thus are grouped into one level. Also, the dose-placebo comparisons within each family are ordered a priori (H vs. Plac through L vs. Plac). We choose the conventional fixed sequence procedure as local procedure for each family. The initial allocation of critical values for each family are 0.04, 0.005, 0.005. Once \(F_1\) is tested and all of its hypotheses are rejected, then its critical value are equally split to pass on to \(F_2\) and \(F_3\). Figure 3.9 (a) visualizes this strategy. Firstly, within \(F_1\), the three hypotheses are tested by the fixed sequence method at level 0.04, then all three hypotheses are rejected at this level. Thus all of its local critical value 0.04 will be equally assigned to \(F_2\) and \(F_3\). So the updated critical values for \(F_2\) and \(F_3\) will become 0.005 + 0.02 = 0.025. Therefore, we can arbitrarily pick \(F_2\) to be tested first since \(F_2\) and \(F_3\) are in the same level. We still use the fixed sequence procedure to test \(F_2\) and \(F_3\) at level 0.025, respectively. And the rejected hypotheses are \(H_{21}, H_{31}\) and \(H_{32}\).

Figure 3.9 (b) describes the same procedure by using hypothesis-based graphical approach.

It is worth noting that in the aforementioned two-level procedure, we treated the two secondary endpoints equally which means that there is no logical restriction between \(F_2\) and \(F_3\). Now, consider a procedure strictly follows the hierarchical relationship between \(F_2\) and \(F_3\) such that \(F_2\) needs to be tested before \(F_3\). Therefore,
we have the following procedure.

**PROCEDURE 2.** Suppose $F_1, F_2$ and $F_3$ have hierarchical relationships and thus, are tested in a pre-defined order. Consider using the procedure described in Figure 3.2 and 3.3 that Bonferroni-Holm procedure is used as local procedure for all three families. We initially assign non-zero critical values, $0.04, 0.005, 0.005$, to three families, respectively. If all of the three hypotheses in $F_1$ are rejected, its local critical value will be proportionally transferred to $F_2$ and $F_3$ with proportions 0.8 and 0.2, respectively. And if all of the three hypotheses in $F_2$ are rejected, its whole local critical value will be transferred to test $F_3$. Figure 3.3 visualizes the Procedure 2 by family-based graphical approach. According to this procedure, the three hypotheses in $F_1$ are still rejected using Bonferroni-Holm procedure at level 0.04. Then, the updated critical values for $F_2$ and $F_3$ are $0.04*0.8+0.005 = 0.037$ and $0.04*0.2+0.005 = 0.013$, respectively. Then test $F_2$ at level 0.037 using Bonferroni-Holm procedure. Still, all of the three hypotheses in $F_2$ are rejected. its whole local critical value will be transfered to $F_3$ such that the updated critical value of $F_3$ will be $0.013 + 0.037 = 0.05$. Lastly, use Bonferroni-Holm procedure to test $F_3$ at level 0.05. $H_{31}$ and $H_{32}$ are rejected finally.

Equivalently, Figure 3.2 shows the graphical representation by using hypothesis-based graphical approach.
Table 3.1 Two-level and Three-level Family based Procedures in the Type II Diabetes Clinical Trial

<table>
<thead>
<tr>
<th>Null hypothesis</th>
<th>Raw $p$-value</th>
<th>Procedure 1</th>
<th>Procedure 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_{11}$</td>
<td>0.005</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>$H_{12}$</td>
<td>0.011</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>$H_{13}$</td>
<td>0.018</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>$H_{21}$</td>
<td>0.009</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>$H_{22}$</td>
<td>0.026</td>
<td>NS</td>
<td>S</td>
</tr>
<tr>
<td>$H_{23}$</td>
<td>0.013</td>
<td>NS</td>
<td>S</td>
</tr>
<tr>
<td>$H_{31}$</td>
<td>0.010</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>$H_{32}$</td>
<td>0.006</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>$H_{33}$</td>
<td>0.051</td>
<td>NS</td>
<td>NS</td>
</tr>
</tbody>
</table>

Note: S=significant; NS=not significant; $\alpha = 0.05$.

3.6 Concluding Remarks

In this chapter, we have proposed a simple graphical tool to sequentially test hierarchically ordered families of hypotheses. We have described the algorithm and presented the theoretical results associating to the FWER control. Through some examples, we have shown the efficiency of our procedure comparing with the hypothesis-based graphical approach when deal with hierarchically related multiple families of hypotheses. In addition, when dealing with the problem of testing multiple families of hypotheses, our family-based graphs are easier to communicate to the non-statisticians than the original hypothesis-based graphs. By using family-based graphical approach, it is easy to illustrate the distinction of different testing strategies so that it is more convenient for us to pick the suitable one.
CHAPTER 4

BONFERRONI-BASED GATEKEEPING PROCEDURE WITH RETESTING OPTION

4.1 Introduction

In most of existing multiple testing procedures for testing multiple hierarchically ordered families of hypotheses, each family can only be tested once. In order to develop more desirable and more powerful procedures, some researchers have considered adding retesting steps to the existing procedures such that each family can be tested more than once. Guilbaud (2007), Dmitrienko, Kordzakhia and Tamhane (2011) and Dmitrienko, Soulakova and Millen (2011) constructed gatekeeping procedures with retesting option which test families of hypotheses in a pre-specified order. Guilbaud (2007) first considered involving retesting steps into bonferroni-based gatekeeping procedures. He presented that as long as all null hypotheses in the last family are rejected, the procedure allows retesting the families in a reverse order using more powerful procedure than the original Bonferroni procedure. Dmitrienko, Kordzakhia and Tamhane (2011) extended Guilbaud’s procedure by applying mixture procedure to each family instead of Bonferroni procedure. Later, restricted to two-family problem, Dmitrienko et al.(2011) improved the existing retesting procedure by considering the second family as a parallel gatekeeper instead of serial gatekeeper for the first family, i.e., as long as one hypothesis is rejected in second family, the first family can be retested using a more powerful procedure than the one used at previous time.

In contrast with the aforementioned sequential retesting procedures, Kordzakhia and Dmitrienko (2012) proposed a class of multiple testing procedures with retesting option on the basis of the simultaneous testing strategy, termed as superchain
procedures. Unlike those sequential retesting procedures, superchain procedures test all families concurrently at each step. Each family serves as a parallel gatekeeper for the other families. If at least one rejection occurs in either family, the rest of families are retested with updated procedures at updated critical values at the next step.

Comparing with superchain procedures, the sequential retesting procedures are usually simpler, easier to implement and more natural in a clinical sense. In practice, it’s common that there are hierarchical relationships among the tested families of hypotheses. For example, Dmitrienko et al. (2011) stated that ”the null hypotheses associated with superiority should not be tested if the corresponding null hypotheses of non-inferiority have failed to be rejected”. Therefore, it is natural to develop sequential retesting procedures. However, the existing sequential retesting procedures have certain limitations and are only restricted to some specific scenarios.

In this chapter, we introduce a Bonferroni-based gatekeeping procedure with retesting option which allows families of hypotheses to be repeatedly tested in a sequential manner. In this procedure, each family is repeatedly tested using Bonferroni procedure with different critical values. Initially, each family is assigned a fraction of $\alpha$ as its initial critical value. We term the critical value used to test one family local critical value. The local critical value for one family at each test depends on certain amount of local critical values passed down from higher - rank families and the initial critical values assigned to lower- rank families. In this class of procedures, each family can be iteratively retested with an increasingly updated local critical value. The proposed procedure exhibit several desirable features. First of all, we prove that the constructed procedure can strongly control the global familywise error rate (FWER), i.e., the probability of rejecting at least one hypothesis among all tested hypotheses across all families at a pre-specified level $\alpha$ under arbitrary dependence. Secondly, comparing with superchain procedure, the proposed procedure facilitates the implementation since it sticks to the simple Bonferroni method for testing each
family though the whole procedure and proceeds in a sequential manner. Additionally, comparing with the existing sequential retesting procedures, the proposed procedure is more general in the sense that it can be constructed under almost any scenarios. Finally, the procedure in this class can be described via a directed graph similar as the graphical approach (Bretz et al., 2009) except that the nodes of the graph here represent families instead of hypotheses, and it is easy to explain the underlying testing strategy with non-statisticians.

The rest of the chapter is organized as follows. In Section 4.2, we present some notations and definitions which are used through the whole chapter. The main theoretical results are introduced in Section 4.3 including the algorithms of the proposed procedures and the global FWER control. Several special cases are discussed in Section 4.4 to demonstrate the relationships among the proposed procedures and some existing procedures. Section 4.5 includes two clinical trial examples to illustrate the implementation of the proposed procedures and Section 4.6 presents the results of a simulation study. Section 4.7 extends the proposed procedures to a two-level structure with multiple families within each level while maintaining the control of the global FWER. Some concluding remarks are discussed in Section 4.8 and the Appendix C gives proofs of theoretical results.

4.2 Preliminary

In this section, we present some basic notations. Suppose $n \geq 2$ hypotheses are grouped into $m \geq 2$ families where the $i^{th}$ family consists of $n_i$ hypotheses, denoted as $F_i = \{H_{i1}, \cdots, H_{in_i}\}$, for $i = 1, \cdots, m$ with $\sum_{i=1}^{m} n_i = n$. Each hypothesis $H_{ij}$ is tested based on its $p$-value $p_{ij}, i = 1, \cdots, m, j = 1, \cdots, n_i$. In this chapter, the true null $p$-values are always assumed to be stochastically greater than or equal to uniform distribution on $[0, 1]$. That is, let $T_i$ denote the set of true null hypotheses in $F_i$, for
For any multiple testing procedure for testing a single family of hypotheses, it is necessary to control the familywise error rate (FWER) at a pre-defined level $\alpha$ where the FWER is defined as the probability of incorrectly rejecting at least one true null hypothesis within the family. For any multiple testing procedure for testing multiple families of hypotheses, it is required to control the overall FWER, that is, the probability of incorrectly rejecting at least one true null hypotheses among all families can not exceed a pre-defined level $\alpha$. In this chapter, we consider a procedure with retesting option, that is, each family of hypotheses has chance to be tested more than once. Correspondingly, we define the global FWER as the probability of incorrectly rejecting at least one true null hypothesis across all families of hypotheses regardless of which and how many null hypotheses within each family are true and how many times each family has been tested.

In this chapter, we consider to propose a procedure, termed as Bonferroni-based gatekeeping procedure with retesting option. In this procedure, $m, m \geq 2$ families of hypotheses are tested sequentially using Bonferroni procedure with different local critical values. Initially, each family is assigned a critical value and we let $\alpha_i, i = 1, \cdots, m$, be the initial critical value associated with $F_i$ with $\sum_{i=1}^{m} \alpha_i = \alpha$. After testing one family, a certain amount of its critical value can be distributed to test other families such that the local critical values of every other families can be updated. As the procedure goes on testing from $F_1$ to $F_m$, the local critical value for each family is constantly updated. Moreover, after finishing a round of tests of all $m$ families, the procedure can start over another round of tests from $F_1$ to $F_m$ based on the updated critical values using Bonferroni procedure. The whole procedure can stop only if there is no new rejection occurs in all $m$ families. The specific updating rule

\begin{align*}
\Pr \{ p_{ij} \leq u \mid H_{ij} \in T_i \} \leq u, i = 1, \cdots, m, j = 1, \cdots, n_i. \tag{4.1}
\end{align*}
Figure 4.1 Three-family Bonferroni-based gatekeeping procedure with retesting option.

for local critical values is described in Section 4.3. The distribution of the amount of critical value transferred among families can be pre-fixed by an $m \times m$ transition matrix which is defined as follows.

Denote $G = \{g_{ij}\}, i, j = 1, \cdots, m$ as a transition matrix which satisfies the following conditions:

$$0 \leq g_{ij} \leq 1; \quad g_{ij} = 0, \text{ if } i = j; \quad \sum_{j=1}^{m} g_{ij} = 1, \text{ for any } i = 1, \cdots, m.$$

Note that $g_{ij}$ is defined as the proportion of the critical value that can be transferred from $F_i$ to $F_j$. Figure 4.1 shows the graphical representation of a special case with $m = 3$.

Remark 4.1. Dmitrienko et al. (2008) quantified the amount of critical value of a tested family that can be transferred to test subsequent families of hypotheses. For a single family of hypotheses $F_i = \{H_{i1}, \cdots, H_{in_i}\} (i = 1, \cdots, m)$, suppose it is tested using Bonferroni procedure at level $\alpha$. Let $A_i$ and $R_i$ be the set of acceptances and rejections, respectively, with the corresponding cardinalities $|A_i|$ and $|R_i|$. It is known that the upper bound of FWER used when testing $F_i, i = 1, \cdots, m$, is $\frac{|A_i|}{n_i} \alpha$. In other
words, the remaining part of critical value \( \frac{|R_i|}{m_i} \alpha \) can be considered as unused critical value and used to test the subsequent families of hypotheses.

4.3 Main Results

In this section, we consider to construct a Bonferroni-based gatekeeping procedure with retesting option. It allows iteratively retesting all \( m, m \geq 2 \) families at non-decreasing critical values while controlling the global FWER for all families \( F = \bigcup_{i=1}^{m} F_i \) of all hypotheses at a pre-defined level \( \alpha \) in the strong sense. In the proposed procedure, every family is tested using Bonferroni procedure. We will begin with a simple case of two families of null hypotheses in Section 4.3.1. The general case of an arbitrary number of families will be introduced in section 4.3.2. Section 4.3.3 discusses the main property of the proposed procedure, that is, the global FWER control.

4.3.1 Two-Family Problem

Consider a multiple testing problem with two families of null hypotheses denoted as \( F_i, i = 1, 2 \). These two families are pre-ordered based on their hierarchical relationship. Initially, we assigned \( \alpha_1 \) and \( \alpha_2 \) with \( \alpha_1 + \alpha_2 = \alpha \) to \( F_1 \) and \( F_2 \), respectively. We let \( \alpha_1(j) \) and \( \alpha_2(j), j \geq 1 \) be the updated local critical values used for testing \( F_1 \) and \( F_2 \) the \( j^{th} \) time. The transition matrix is given by

\[
G = \begin{pmatrix}
0 & 1 \\
1 & 0
\end{pmatrix},
\]

i.e., \( g_{12} = g_{21} = 1 \), which implies the whole amount of critical value of one family that can be recycled is transferred to test the other family. The graphical representation of this case is shown in Figure 4.2. Denote \( R_{1(j)} \) and \( R_{2(j)} \), respectively, the sets of rejections when \( F_1 \) and \( F_2 \) are tested at the \( j^{th} \) time, while \( |R_{1(j)}| \) and \( |R_{2(j)}| \) are their
corresponding cardinalities. The proposed Bonferroni-based gatekeeping procedure with retesting option in the case of $m = 2$ is defined in the following.

Algorithm 4.1.

**Stage 1.** Test $F_1$ at local critical value $\alpha_{1(1)} = \alpha_1$ and then, test $F_2$ at local critical value $\alpha_{2(1)}$ where

$$\alpha_{2(1)} = \alpha_2 + \frac{|R_{1(1)}|}{n_1} \alpha_1$$

using Bonferroni procedure. If no null hypotheses are rejected in both families, the algorithm stops. Otherwise, proceed to next stage.

**Stage $k$ ($k \geq 2$).** Retest $F_1$ at local critical value $\alpha_{1(k)}$ where

$$\alpha_{1(k)} = \alpha_1 + \frac{|R_{2(k-1)}|}{n_2} \alpha_2. \tag{4.2}$$

Then, retest $F_2$ at local critical value $\alpha_{2(k)}$ where

$$\alpha_{2(k)} = \alpha_2 + \frac{|R_{1(k)}|}{n_1} \alpha_{1(k)}. \tag{4.3}$$

using Bonferroni procedure. If no new null hypotheses are rejected in both families, the algorithm stops. Otherwise, proceed to next stage.

**Remark 4.2.** Before the test, the transition matrix and initial allocation of critical values are pre-fixed for both families. Algorithm 4.1 allows iteratively retesting $F_1$.
and $F_2$ using Bonferroni procedure at increasingly updated local critical values. An amount of local critical value of $F_1$ can be accumulated to $F_2$ during each retesting stage, while the local critical value of $F_1$ can only be updated by using a certain amount of the initial critical value of $F_2$. Both families can be repeatedly tested as long as there is at least one new rejection occurs among two families of hypotheses at each retest stage.

Remark 4.3. If we initially assign critical values $\alpha_1 = \alpha$ and $\alpha_2 = 0$ to $F_1$ and $F_2$, respectively. There is no critical value transferred from $F_2$ to $F_1$ and thus, no retests involved. This case can be regarded as the original multistage parallel gatekeeping procedure (Dmitrienko et al., 2008)). For original multistage parallel gatekeeping procedure, although $F_1$ can be tested at full level $\alpha = 0.05$, if there is only a small number of rejections occurs in $F_1$, $F_2$ can only be tested at a small local critical value such that it can hardly have rejections. Consider an extreme case that if no rejection occurs in $F_1$, then $F_2$ has no chance to be tested. However, if we initially distribute $\alpha$ to both families. $F_2$ can be tested without completely depending on $F_1$. Furthermore, the local critical value for $F_1$ can still be increasingly updated during the retesting stages. If all hypotheses are rejected in $F_2$, $F_1$ can even be tested at full level $\alpha$ during the retesting stage.

4.3.2 Multi-Family Problem

In this section, we generalize the algorithm 4.1 to arbitrary $m \geq 2$ families. Based on the notations in Section 4.2, the algorithm for the general Bonferroni-based gatekeeping procedure with retesting option is defined as follows.

**Algorithm 4.2.**

**Stage 1.** Start with $F_1$. Let $\alpha_{1(1)} = \alpha_1$. Sequentially test the null hypotheses in
\( F_{i}, i = 1, \cdots, m, \) using Bonferroni procedure at its local critical value

\[
\alpha_{i(1)} = \alpha_{i} + \sum_{j=1}^{i-1} \frac{|R_{j(i)}|}{n_j} g_{ji}\alpha_{j(1)}.
\]

If for all \( i = 1, \cdots, m, \) \(|R_{i(1)}| = 0\), the algorithm stops. Otherwise, proceed to next stage.

**Stage \( k (k \geq 2) \).** Sequentially retest the null hypotheses in \( F_{i}, i = 1, \cdots, m, \) using Bonferroni procedure at its updated local critical value

\[
\alpha_{i(k)} = \alpha_{i} + \sum_{j=1}^{i-1} \frac{|R_{j(k)}|}{n_j} g_{ji}\alpha_{j(k)} + \sum_{l=i+1}^{m} \frac{|R_{l(k-1)}|}{n_l} g_{il}\alpha_{l}.
\] (4.4)

After retesting all families \( F_{i}, i = 1, \cdots, m \) at this stage, if no new hypotheses are rejected in any family, the algorithm stops. Otherwise, proceed to the next stage.

**Remark 4.4.** Algorithm 4.2 basically illustrates that the families of hypotheses are tested and retested in a sequential manner. The testing order of \( m \) families, the initial critical value for each family and the transition matrix are pre-specified. Each family is repeatedly tested using Bonferroni procedure at updated local critical value. After testing one family, an amount of its critical value, either local critical value at current stage or initial critical value, can be distributed to all other families according to the pre-specified transition matrix such that the local critical value of every family can be constantly updated when the procedure goes on testing from \( F_{1} \) to \( F_{m} \). After a round of tests of all \( m \) families, every family has an updated critical value. Then, the procedure can start over another round of tests from \( F_{1} \) by using the updated critical values. Note that if neither family has new null rejected hypotheses, the whole algorithm stops.

**Remark 4.5.** According to (4.4), it is easy to see that the updated local critical value of one family can be divided into three parts. One part is its own initial critical value. One part is transferred from the higher - rank families, and the other one part
is transferred from the lower-rank families. The amount of local critical values of higher-rank families can be accumulated to test lower-rank families. However, the amount of critical values transferred from lower-rank families to retest higher-rank families only depend on their initial critical values. The rejection number of either family is non-decreasing with respect to the number of tests, which implies that the updated local critical value of one family is also non-decreasing.

**Remark 4.6.** If we initially assigned \( \alpha_1 = \alpha \) and \( \alpha_2 = \cdots = \alpha_m = 0 \) to \( F_i, i = 1, \cdots, m \), respectively. Then, there is no critical values transferred from lower-rank families to higher-rank families since all the initial critical values are zero except \( F_1 \). Thus, no retesting stages can be involved. This procedure actually strictly follows parallel gatekeeping strategy, and it can be regarded as a general multistage parallel gatekeeping procedure (Dmitrienko et al., 2008). Moreover, if each family only has one hypothesis, i.e., \( F_1 = \{ H_{11} \}, \cdots, F_m = \{ H_{m1} \} \), then it can be regarded as the conventional fixed sequence procedure. (Maurer et al., 1995; Westfall and Krishen, 2001).

### 4.3.3 Global Familywise Error Rate Control

The following theorem presents that the Bonferroni-based gatekeeping procedure with retesting option proposed in Algorithm 4.2 controls the global FWER in the strong sense.

**Theorem 4.1.** The Bonferroni-based gatekeeping procedure with retesting option described in Algorithm 4.2 strongly controls the global FWER across the \( m, m \geq 2 \) families at level \( \alpha \) under arbitrary dependence.

For the proof of theorem 4.1, see Appendix C.

**Remark 4.7.** The two-family problem described in Algorithm 4.1 can be regarded as a special case of Algorithm 4.2 with \( m = 2 \). Hence, Theorem 4.1 is also true for
two-family Bonferroni-based gatekeeping procedure with retesting option described in Algorithm 4.1.

Remark 4.8. Theorem 4.1 illustrates that the proposed procedure controls the globe FWER without any assumption of dependence structure about test statistics. If there are some distribution information regarding to the test statistics known in advance, then it is possible to further improve the proposed procedure by exploiting the dependence information.

4.4 Discussions

This section illustrates two special cases based on Algorithm 4.2 and presents the relationships between the proposed procedure with the some existing multiple testing procedures.

Case 1. Suppose we assign $\alpha_i \neq 0$ to $F_i, i = 1, \cdots, m$, initially. Assume the transition matrix $G$ is an upper shift matrix which is given by

$$g_{ij} = \begin{cases} 1 & \text{if } j = i + 1, \text{ for } i = 1, \cdots, m, \\ 0 & \text{otherwise.} \end{cases}$$

This transition matrix implies that there is no retesting involved. This procedure can be considered as an extension of the general multistage gatekeeping procedure (Dmitrineko et al., 2008) in the sense that even no rejection occurs in the previous family, the current family still has chance to be tested. Figure 4.3 visualizes Case 1. Suppose each family only has one hypothesis, i.e., $F_1 = \{H_{11}\}, \cdots, F_m = \{H_{m1}\}$, then if the previous hypothesis is rejected, its critical value can be fully added to test the current hypothesis. However, if the previous one is an acceptance, the current one is tested at its initial critical value. It can be regarded as the fallback procedure (Wiens, 2003; Wiens and Dmitrienko, 2005).
Case 2. For Case 1, suppose retesting stages are involved and the new procedure is presented by Figure 4.4. Then, according to the aforementioned algorithm, after a round of tests for all \( m \) families, an amount of the initial critical value of \( F_m, \alpha_m \), can be used to update the critical value of \( F_1 \) such that all \( m \) families of hypotheses can have chances to be retested at the updated local critical values. Specially, if all hypotheses in \( F_m \) are rejected, then \( F_1 \) will be retested at updated critical value \( \alpha_1 + \alpha_m \).

Suppose each family only has one hypotheses, i.e., \( F_1 = \{H_{11}\}, \ldots, F_m = \{H_{m1}\} \). The new procedure firstly sequentially tests \( m \) hypotheses as fallback procedure mentioned in Case 1. Then, if the last hypothesis \( H_{m1} \) is rejected, the updated critical value of \( H_{11} \) is \( \alpha_1 + \alpha_m \) and thus, all \( m \) hypotheses can start over to be tested again. This procedure can be regarded as an improved version of fallback procedure described in Case 1.

4.5 Clinical Trial Examples

In this section, we illustrate our proposed Bonferroni-based gatekeeping procedures with retesting option through two clinical trial examples. We will also compare them with the proposed procedures without retesting option and the existing superchain
procedure. For notational convenience, the proposed procedure with and without retesting option are respectively named as Retest and No-retest procedure.

4.5.1 Two-Family Problem

This example is based on EPHESUS trial (Pitt et al., 2003) in which a balanced design clinical trial is to assess the effects of eplerenone on morbidity and mortality in patients with severe heart failure. There are two primary endpoints and two secondary endpoints grouped into two families:

- \( F_1 \): all-cause mortality (Endpoint P1) and cardiovascular mortality plus cardiovascular hospitalization (Endpoint P2).

- \( F_2 \): cardiovascular mortality (Endpoint S1) and all-cause mortality plus all-cause hospitalization (Endpoint S2).

The hypotheses of no treatment effect corresponding to four endpoints are \( H_{11}, H_{12}, H_{21} \) and \( H_{22} \). In this trial, we applied the proposed procedure with and without retesting option, the existing superchain procedure to this trial. The overall critical value is set as \( \alpha = 0.05 \) and the initial critical values for two families are set as 0.04 and 0.01 for all three procedures. For proposed Retest procedure and Superchain procedure, we used the same graphical representation which is shown in Figure 4.5. For Superchain procedure, we applied the Holm-based superchain procedure to this example. At first step, we used Bonferroni procedure to test both families simultaneously at the initial critical values. According to the testing results of first step, we proceeded to test both families using truncated Holm procedure at updated critical values during the subsequent steps. Due to the complexity of updating rules for local critical values and truncation parameters for truncated Holm procedure, we omit the detail steps here. For more information about updating rules of Superchain procedure, see Kordzakhia and Dmitrienko, 2012. The raw \( p \)-values for
four null hypotheses and the test results using the aforementioned three procedures are given in Table 4.1. The Retest procedure is implemented as follows.

**Stage 1.** Test null hypotheses of $F_1$ at level $\alpha_{1(1)} = 0.04$. Since $p_{11} < \frac{0.04}{2}$, $p_{12} > \frac{0.04}{2}$, $R_{1(1)} = \{H_{11}\}$. Thus, the updated local critical value for $F_2$ is

$$\alpha_{2(1)} = \alpha_2 + \frac{1}{2} \alpha_{1(1)} = 0.01 + 0.02 = 0.03.$$  

Test $F_2$ at level $\alpha_{2(1)}$. Since $p_{21} < \frac{0.03}{2}$, $p_{22} > \frac{0.03}{2}$, $R_{2(1)} = \{H_{21}\}$. So far, the No-retest procedure stops. To proceed the Retest procedure, the updated local critical value for $F_1$ is

$$\alpha_{1(2)} = \alpha_1 + \frac{1}{2} \alpha_2 = 0.045.$$  

**Stage 2.** Retest $F_1$ using Bonferroni method at level $\alpha_{1(2)}$. $p_{11} < \frac{0.045}{2}$, $p_{12} > \frac{0.045}{2}$, $R_{1(2)} = \{H_{11}\} = R_{1(1)}$. Thus, the updated local critical value for $F_2$ is

$$\alpha_{2(2)} = \alpha_2 + \frac{1}{2} \alpha_{1(2)} = 0.0325.$$  

Retest $F_2$ using Bonferroni method at level $\alpha_{2(2)}$. Since $p_{21} < \frac{0.0325}{2}$, $p_{22} < \frac{0.0325}{2}$, $R_{2(1)} = \{H_{21}, H_{22}\}$. Thus, the updated local critical value for $F_1$ is

$$\alpha_{1(3)} = \alpha_1 + \frac{2}{2} \alpha_2 = 0.05.$$  

71
### Table 4.1 Comparison Results of Three Procedures in the EPHESUS Trial Example

<table>
<thead>
<tr>
<th>Family</th>
<th>Null hypothesis</th>
<th>Raw p-value</th>
<th>Retest</th>
<th>No-retest</th>
<th>Superchain</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>$H_{11}$</td>
<td>0.0121</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>$H_{12}$</td>
<td>0.0337</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>$F_2$</td>
<td>$H_{21}$</td>
<td>0.0084</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>$H_{22}$</td>
<td>0.0160</td>
<td>S</td>
<td>NS</td>
<td>S</td>
</tr>
</tbody>
</table>

Note: the initial allocation of critical values for $F_1$ and $F_2$ are 0.04 and 0.01, respectively. The globe Type I error rate is $\alpha = 0.05$. S=significant; NS=not significant.

**Stage 3.** Retest $F_1$ using Bonferroni method at level $\alpha_{1(3)}$. Since for both $F_1$ and $F_2$, there is no new rejections. The whole Retest procedure stops.

As seen from Table 4.1, the No-retest procedure only rejects two null hypotheses and the proposed Retest procedure and Superchain procedure reject more hypotheses which is three.

### 4.5.2 Three-Family Problem

In this section, we consider the example introduced in Kordzahia and Dmitrienko (2012). It is a balanced design clinical trial in which two doses (D1, D2) treatment are compared with a placebo (P) in the general population of patients as well as two pre-specified subpopulations of patients. The subpopulations are defined by phenotype or genotype markers. The three populations are labeled Group 1 (General population), Group 2 (First subpopulation) and Group 3 (Second subpopulation). There are six null hypotheses grouped into three families:

- $F_1$: $H_{11}$ (D1 vs P in Group 1) and $H_{12}$ (D2 vs P in Group 1).
- $F_2$: $H_{21}$ (D1 vs P in Group 2) and $H_{22}$ (D2 vs P in Group 2).
- $F_3$: $H_{31}$ (D1 vs P in Group 3) and $H_{32}$ (D2 vs P in Group 3).
Figure 4.6 Graphical visualization of three-family clinical trial problem.

We applied the proposed Retest, No-retest and Superchain procedures to this example. The global FWER needs to be controlled at a one-sided $\alpha = 0.025$ and the initial critical value for each family was set to be $\alpha_1 = \frac{1}{2} \alpha = 0.0125$, $\alpha_2 = \frac{1}{3} \alpha = 0.00833$ and $\alpha_3 = \frac{1}{6} \alpha = 0.00417$. For Retest procedure and Superchain procedure, the transition matrix is pre-defined as

\[
G = \begin{pmatrix}
0 & 0.5 & 0.5 \\
0.5 & 0 & 0.5 \\
0.5 & 0.5 & 0
\end{pmatrix}.
\]

Figure 4.6 shows the graphical representation. Similar as aforementioned two-family problem, we also applied Holm-based superchain procedure to this example. As said in Kordzakhia and Dmitrienko (2013), they assumed the three families of null hypotheses are interchangeable and can be tested in any order. At first step, we used Bonferroni procedure to test three families simultaneously at the initial critical values. According to the testing results of first step, we proceeded to test three families using truncated Holm procedure at updated local critical values during the subsequent steps. Again, we omit the detail steps here due to the complexity of updating rules for local critical values and truncation parameters for truncated Holm procedure. For more information about updating rules of superchain procedure, see Kordzakhia and Dmitrienko (2012). The raw $p$-values for six null hypotheses and the test results using the aforementioned three procedures are shown in Table 4.2. The
proposed Retest procedure is implemented as follows.

**Stage 1.** Test $F_1$ using Bonferroni method at level $\alpha_{1(1)} = 0.0125$. Since $p_{11} > \frac{0.0125}{2}, p_{12} > \frac{0.0125}{2}, R_{1(1)} = \emptyset$. Then test $F_2$ at level $\alpha_{2(1)} = \alpha_2 = 0.00833$. Since $p_{21} > \frac{0.00833}{2}, p_{22} > \frac{0.00833}{2}, R_{2(1)} = \emptyset$. Test $F_3$ at level $\alpha_{3(1)} = \alpha_3 = 0.00417$. Since $p_{31} > \frac{0.00417}{2}, p_{32} < \frac{0.00417}{2}, R_{3(1)} = \{H_{32}\}$. The No-retest procedure stops here. The proposed Retest procedure proceeds to next stage.

**Stage 2.** Retest $F_1$ using Bonferroni method at level $\alpha_{1(2)}$ where

\[
\alpha_1 + \frac{1}{2} g_{32} \alpha_3 = 0.0125 + \frac{1}{2} \cdot 0.5 \cdot 0.00417 = 0.0135.
\]

Since $p_{11} > \frac{0.0135}{2}, p_{12} > \frac{0.0135}{2}, R_{2(1)} = \emptyset$. Then retest $F_2$ at level $\alpha_{2(2)}$ where

\[
\alpha_2 + \frac{1}{2} g_{32} \alpha_3 = 0.00833 + \frac{1}{2} \cdot 0.5 \cdot 0.00417 = 0.00937.
\]

Since $p_{21} > \frac{0.00937}{2}, p_{22} < \frac{0.00937}{2}, R_{2(2)} = \{H_{22}\}$. Retest $F_3$ at level $\alpha_{3(2)}$ where

\[
\alpha_3 + g_{23} \alpha_{2(2)} = 0.00417 + \frac{1}{2} \cdot 0.5 \cdot 0.00937 = 0.0065.
\]

Since $p_{31} > \frac{0.0065}{2}, p_{32} < \frac{0.0065}{2}, R_{3(2)} = \{H_{32}\}$. $F_3$ has the same rejection as the one rejected in Stage 1 and because there are no new rejections, the testing algorithm stops. The final set of rejected null hypotheses are $\{H_{22}, H_{32}\}$.

As seen from Table 4.2, the No-retest procedure performs worst which only rejects $H_{32}$. By contrast, the proposed Retest procedure and Holm-based superchain procedure reject $H_{32}$ as well as $H_{22}$.

Remark 4.9. From aforementioned two examples, it is easy to see that our proposed Bonferroni-based gatekeeping procedure with retesting option increases the power comparing with the procedure without retesting stages. Also, it is not only comparable with Holm-based Superchain procedure with respect to power, but also much simpler and easier to implement. The implementation of Superchain procedure is complicated due to the updating rules of critical values and truncation parameters.
Table 4.2 Comparison Results of Three Procedures in the Dose-Response Trial Example

<table>
<thead>
<tr>
<th>Family</th>
<th>Null hypothesis</th>
<th>Raw $p$-value</th>
<th>Retest</th>
<th>No-retest</th>
<th>Superchain</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>$H_{11}$</td>
<td>0.0092</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>$H_{12}$</td>
<td>0.0105</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>$F_2$</td>
<td>$H_{21}$</td>
<td>0.0059</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>$H_{22}$</td>
<td>0.0044</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>$F_3$</td>
<td>$H_{31}$</td>
<td>0.0271</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>$H_{32}$</td>
<td>0.0013</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
</tbody>
</table>

Note: initial allocation of critical values for $F_1$, $F_2$ and $F_3$ are 0.0125, 0.00833 and 0.00417, respectively. The globe FWER is $\alpha = 0.025$. S=significant; NS=not significant.

of the testing procedure at each stage. The computational complexity is even more when the number of families are large. In practice, it is difficult to explain to non-statistician. Moreover, the sequential manner makes our procedure more explicit in a clinical trial sense.

### 4.6 Numerical Findings

In this section, we only concern the two-family problem. Simulation studies were performed to investigate the performances of the proposed procedure (Retest) in terms of the global FWER control and power compared to the existing Bonferroni parallel gatekeeping procedure (Parallel), superchain procedure (Superchain) and the proposed procedure without retesting stage (No-retest) with respect to the correlation $\rho$ among test statistics, the proportion $d_i, i = 1, 2$ of false null hypotheses with large effect size among all false null hypotheses within $F_i$.

To simulate the values of global FWER and average power, the expected proportion of false nulls that are rejected, for each $F_i, i = 1, 2$, we first generated
dependent normal random variables $T_{ij} \sim N(\mu_{ij}, 1), j = 1, \ldots, n_i,$ with $n_{i0}$ of the $\mu_{ij}$’s being equal to 0 and the rest being equal to $\mu_{ij} \neq 0,$ and an equicorrelation matrix with correlation $\rho.$ We then applied each aforementioned procedure to the generated data to test $H_{ij} : \mu_{ij} = 0$ against $K_{ij} : \mu_{ij} \neq 0$ simultaneously for $i = 1, 2, j = 1, \ldots, n_i,$ at level $\alpha = 0.05.$ The above steps were repeated for 2,000 times.

In the simulation, the $p$-value $P_{ij}$ corresponding to the hypothesis $H_{ij}$ was calculated by $P_{ij} = 2(1 - \Phi(T_{ij})), i = 1, 2, j = 1, \ldots, n_i,$ where $\Phi(\cdot)$ is the cdf of $N(0, 1).$ For Retest, No-retest and superchain procedure, the initial allocation of critical values for two families are set as 0.03 and 0.02. For Retest and Superchain procedure, the graphical representation is shown in Figure 4.2. For Superchain procedure, truncated Holm procedures were applied within each family and the initial truncation parameters were set as 0 for both families.

In this simulation, for $i = 1, 2,$ we set $n_i = 20$ with $n_{i0} = 10$ for all aforementioned procedures. For $F_i, i = 1, 2$ with $n_i - n_{i0} = 10$ false null hypotheses, we set 10$d_i$ of the $\mu_{ij}$’s being equal to 4 indicating large effect sizes and the rest $\mu_{ij}$ being equal to 2. $\rho$ is set to be 0, 0.2, 0.5 and 0.8.

Tables 4.3 - 4.7 present comparisons of the simulated global FWERs and average powers of the four aforementioned procedures under the scenarios that two families have equal number of false null hypotheses but different proportion of effect sizes. All five tables show that the global FWERs of all these procedures are controlled at level $\alpha$ under dependence. Apparently, Retest procedure has larger power compared with No-retest procedure in any scenario. And as seen from Tables 4.3, 4.4 and 4.5, when the proportion of false null hypotheses with large effect sizes in $F_1$ is higher than $F_2$ indicating the order of two families is strictly follow the hierarchical structure, all four procedures are comparable. The power of Bonferroni parallel gatekeeping procedure is slightly larger than other procedures. But when $F_1$ has less number of false null
Table 4.3  Simulated Global FWERs and Average Powers for Four Procedures under Equal Correlation $\rho$ for $d_1 = 1, d_2 = 0$

<table>
<thead>
<tr>
<th></th>
<th>$\rho = 0$</th>
<th></th>
<th>$\rho = 0.2$</th>
<th></th>
<th>$\rho = 0.5$</th>
<th></th>
<th>$\rho = 0.8$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FWER</td>
<td>Power</td>
<td>FWER</td>
<td>Power</td>
<td>FWER</td>
<td>Power</td>
<td>FWER</td>
<td>Power</td>
</tr>
<tr>
<td>Retest</td>
<td>0.0275</td>
<td>0.4619</td>
<td>0.0230</td>
<td>0.4610</td>
<td>0.0255</td>
<td>0.4603</td>
<td>0.0185</td>
<td>0.4575</td>
</tr>
<tr>
<td>No-Retest</td>
<td>0.0270</td>
<td>0.4605</td>
<td>0.0220</td>
<td>0.4590</td>
<td>0.0245</td>
<td>0.4586</td>
<td>0.0180</td>
<td>0.4559</td>
</tr>
<tr>
<td>Parallel</td>
<td>0.0310</td>
<td>0.4685</td>
<td>0.0265</td>
<td>0.4680</td>
<td>0.0295</td>
<td>0.4653</td>
<td>0.0185</td>
<td>0.4619</td>
</tr>
<tr>
<td>Superchain</td>
<td>0.0280</td>
<td>0.4632</td>
<td>0.0240</td>
<td>0.4626</td>
<td>0.0275</td>
<td>0.4624</td>
<td>0.0195</td>
<td>0.4602</td>
</tr>
</tbody>
</table>

hypotheses with large effect size than $F_2$, as seen in Tables 4.6, 4.7, the Bonferroni parallel gatekeeping procedure loses its edge. Our proposed retest procedure has almost the same power as superchain procedure.

Summarizing the above observations, the aforementioned four procedures control the FWER well in each setting. However, there is not an uniformly powerful procedure. Apparently, adding retesting stages increases the power comparing with the procedure without retesting option. Compared to Superchain procedure, our proposed procedure is slightly less powerful since superchain uses truncated Holm procedure as its local method which is more powerful than Bonferroni procedure. However, our proposed procedure is based on simple Bonferroni procedure but its power is almost comparable to superchain procedure under different scenarios. If the order of families of hypotheses violates the intrinsic hierarchical structure, Bonferroni parallel gatekeeping procedure loses its power. Therefore, when there is doubt about the testing order, proposed procedure and superchain procedure are better choices.

4.7 An Extension

In the preceding sections, we only consider testing one family at a time and testing $m, m \geq 2$ families in a sequential manner. However, the structure of multiple
Table 4.4  Simulated Global FWERs and Average Powers for Four Procedures under Equal Correlation $\rho$ for $d_1 = 0.8, d_2 = 0.2$

<table>
<thead>
<tr>
<th></th>
<th>$\rho = 0$</th>
<th>$\rho = 0.2$</th>
<th>$\rho = 0.5$</th>
<th>$\rho = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retest</td>
<td>0.0270</td>
<td>0.4626</td>
<td>0.0220</td>
<td>0.4603</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0265</td>
<td>0.4580</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0170</td>
</tr>
<tr>
<td>No-Retest</td>
<td>0.0260</td>
<td>0.4587</td>
<td>0.0215</td>
<td>0.4570</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0240</td>
<td>0.4565</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0165</td>
</tr>
<tr>
<td>Parallel</td>
<td>0.0275</td>
<td>0.4616</td>
<td>0.0255</td>
<td>0.4604</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0265</td>
<td>0.4598</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0185</td>
</tr>
<tr>
<td>Superchain</td>
<td>0.0290</td>
<td>0.4642</td>
<td>0.0240</td>
<td>0.4633</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0270</td>
<td>0.4631</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0180</td>
</tr>
</tbody>
</table>

Table 4.5  Simulated Global FWERs and Average Powers for Four Procedures under Equal Correlation $\rho$ for $d_1 = 0.5, d_2 = 0.5$

<table>
<thead>
<tr>
<th></th>
<th>$\rho = 0$</th>
<th>$\rho = 0.2$</th>
<th>$\rho = 0.5$</th>
<th>$\rho = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retest</td>
<td>0.0250</td>
<td>0.4560</td>
<td>0.0210</td>
<td>0.4588</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0265</td>
<td>0.4586</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0175</td>
</tr>
<tr>
<td>No-Retest</td>
<td>0.0220</td>
<td>0.4546</td>
<td>0.0200</td>
<td>0.4540</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0230</td>
<td>0.4536</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0165</td>
</tr>
<tr>
<td>Parallel</td>
<td>0.0250</td>
<td>0.4473</td>
<td>0.0240</td>
<td>0.4451</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0255</td>
<td>0.4400</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0180</td>
</tr>
<tr>
<td>Superchain</td>
<td>0.0285</td>
<td>0.4631</td>
<td>0.0220</td>
<td>0.4624</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0275</td>
<td>0.4621</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0185</td>
</tr>
</tbody>
</table>
Table 4.6  Simulated Global FWERs and Average Powers for Four Procedures under Equal Correlation $\rho$ for $d_1 = 0.2, d_2 = 0.8$

<table>
<thead>
<tr>
<th></th>
<th>$\rho = 0$</th>
<th>$\rho = 0.2$</th>
<th>$\rho = 0.5$</th>
<th>$\rho = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FWER</td>
<td>Power</td>
<td>FWER</td>
<td>Power</td>
</tr>
<tr>
<td>Retest</td>
<td>0.0250</td>
<td>0.4575</td>
<td>0.0205</td>
<td>0.4571</td>
</tr>
<tr>
<td>No-Restest</td>
<td>0.0215</td>
<td>0.4506</td>
<td>0.0190</td>
<td>0.4507</td>
</tr>
<tr>
<td>Parallel</td>
<td>0.0242</td>
<td>0.4158</td>
<td>0.0230</td>
<td>0.4090</td>
</tr>
<tr>
<td>Superchain</td>
<td>0.0270</td>
<td>0.4607</td>
<td>0.0230</td>
<td>0.4601</td>
</tr>
</tbody>
</table>

Table 4.7  Simulated Global FWERs and Average Powers for Four Procedures under Equal Correlation $\rho$ for $d_1 = 0, d_2 = 1$

<table>
<thead>
<tr>
<th></th>
<th>$\rho = 0$</th>
<th>$\rho = 0.2$</th>
<th>$\rho = 0.5$</th>
<th>$\rho = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FWER</td>
<td>Power</td>
<td>FWER</td>
<td>Power</td>
</tr>
<tr>
<td>Retest</td>
<td>0.0240</td>
<td>0.4512</td>
<td>0.0205</td>
<td>0.4517</td>
</tr>
<tr>
<td>No-Restest</td>
<td>0.0202</td>
<td>0.4432</td>
<td>0.0175</td>
<td>0.4450</td>
</tr>
<tr>
<td>Parallel</td>
<td>0.0220</td>
<td>0.3269</td>
<td>0.0220</td>
<td>0.2897</td>
</tr>
<tr>
<td>Superchain</td>
<td>0.0255</td>
<td>0.4535</td>
<td>0.0210</td>
<td>0.4537</td>
</tr>
</tbody>
</table>
objectives in clinical trial can be more complex. For example, multiple primary and secondary endpoints are evaluated in several patient populations. For each particular endpoint, the constructed hypotheses corresponding to different patient populations are grouped into a family. Therefore, multiple primary and secondary endpoints indicate that there are multiple primary and secondary families of hypotheses. In the following, we group these primary families into a primary level. Similarly, multiple secondary families of hypotheses are associated with multiple secondary endpoints and are grouped into a secondary level.

In this section, we consider a case in which families of hypotheses are divided into two levels. By using the similar idea as in developing aforementioned algorithm, we develop a procedure with retesting option under which the families between levels are tested sequentially but the families within level are tested simultaneously. Each family is still allowed to be iteratively retested using Bonferroni procedure with repeatedly updated local critical values. We need to strongly control the global FWER at level $\alpha$ for all the families of hypotheses among two levels, i.e., the probability of incorrectly rejecting at least one true null hypothesis across tested families will be controlled at $\alpha$ regardless of which and how many null hypotheses within each family are true and how many times the family has been tested.

Notationally, suppose $n \geq 2$ hypotheses are grouped into $m \geq 2$ families. Moreover, these $m$ families of hypotheses are divided into two levels based on prior knowledge, where each level $L_i, i = 1, 2$ consists of families $F_{i1}, \ldots, F_{im_i}$ with $\sum_{i=1}^{2} m_i = m$. Each family $F_{ij}$ has $n_{ij} \geq 1$ null hypotheses with $\sum_{i=1}^{2} \sum_{j=1}^{m_i} n_{ij} = n$. For $i = 1, 2, j = 1, \ldots, m_i$, let $\alpha_{ij}$ denote the initial allocation of the critical value to $F_{ij}$ such that $\sum_{i=1}^{2} \sum_{j=1}^{m_i} \alpha_{ij} = \alpha$. The distribution of critical values transferred among families can be pre-fixed by a transition coefficient set which is defined as follows.
Figure 4.7  Two levels of four families generalized multistage Bonferroni procedure with retesting.

Denote $G = \{g_{ijkl}\}$, $i, k = 1, 2, j = 1, \cdots, m_i$, $l = 1, \cdots, m_k$ as a set of transition coefficient satisfying the following conditions:

\begin{align*}
0 \leq g_{ijkl} \leq 1; & \quad g_{ijkl} = 0, \text{ if } i = k; \\
\sum_{l=1}^{m_2} g_{1jl} = 1, & \text{ for any } j = 1, \cdots, m_1; \quad \sum_{j=1}^{m_1} g_{2lj} = 1, \text{ for any } l = 1, \cdots, m_2.
\end{align*}

Note that $g_{ijkl}$ is defined as the proportion of critical value of $F_{ij}$ that can be transferred to $F_{kl}$. Figure 4.7 shows the graphical representation of the case with two levels of four families.

The proposed procedure allows all $m$ families of hypotheses to be tested more than once. For notational conveniences, we let the local critical value used to test $F_{ij}$ the $t^{th}$ time be $\alpha_{ij(t)}$. Let $R_{ij(t)}$ be the set of rejected hypotheses when $F_{ij}$ is tested at the $t^{th}$ time with cardinality $|R_{ij(t)}|$. The algorithm for the two-level Bonferroni-based gatekeeping procedure with retesting option is as follows.

Algorithm 4.3.

Stage 1. Test $F_{1j}, j = 1, \cdots, m_1$ simultaneously using Bonferroni multiple testing method at level $\alpha_{1j(1)} = \alpha_{1j}$. Update the local critical values for $F_{2l}, l = 1, \cdots, m_2$:  

$$
\alpha_{2l(t)} = \alpha_{2l} + \sum_{j=1}^{m_1} \frac{|R_{1j(t)}|}{n_{1j}} g_{1jl} \alpha_{1j}.
$$
Test $F_{2l}, l = 1, \cdots, m_2$ simultaneously at level $\alpha_{2l(1)}$ using Bonferroni procedure. If no hypotheses are rejected among all $m$ families, the algorithm stops. Otherwise, continue to the next stage.

**Stage $k (k \geq 2)$**. For $j = 1, \cdots, m_1$, set

$$
\alpha_{1j(k)} = \alpha_{1j} + \sum_{l=1}^{m_2} \frac{|R_{2l(k-1)}|}{n_{2l}} g_{2l1j} \alpha_{2l}.
$$

(4.5)

Retest $F_{1j}, j = 1, \cdots, m_1$ simultaneously at level $\alpha_{ij(k)}$ and update the local critical values for $F_{2l}, l = 1, \cdots, m_2$:

$$
\alpha_{2l(k)} = \alpha_{2l} + \sum_{j=1}^{m_1} \frac{|R_{1j(k)}|}{n_{1j}} g_{1j2l} \alpha_{1j(k)}.
$$

Retest $F_{2l}, l = 1, \cdots, m_2$, simultaneously at level $\alpha_{2l(k)}$. If no new null hypotheses are rejected among all $m$ families, the algorithm stops. Otherwise, continue to the next stage.

**Remark 4.10**. Algorithm 4.3 tests the families between levels in a sequential manner. However, within each level, the families of null hypotheses are tested concurrently. The updated local critical value for one family depends on the results of most recent tests from the families at the other level. It is easy to see that the rejection number of every family is non-decreasing with respect to the times of stages which implies that the local critical value of the corresponding family is also non-decreasing. Besides, the more rejections occur in one family, the more critical value will be transferred to test the families in another level. When all families within one level have no new hypotheses rejected, the whole algorithm stops.

**Remark 4.11**. The following remarks can be noted regarding some special cases of this procedure. For notational convenience, we only consider four-family problem with two families within each level as described in Figure 4.7.
1. Suppose $\alpha_{12} = \alpha_{22} = 0$ and $g_{1121} = g_{2111} = 1$. Each level only has one family of hypotheses. Then, this case reduces to the two-family Bonferroni-based gatekeeping procedure proposed in Algorithm 4.1.

2. Suppose $g_{1122} = g_{2211} = g_{1221} = g_{2112} = 0$, each family in $L_2$ only relates to one specific family in $L_1$. The hierarchical logical relationships among families are pre-specified and this particular case takes into account the special logical relationships among families.

3. Suppose $\alpha_{12} = 0$ and $g_{1221} = g_{1222} = g_{2112} = g_{2122} = 0$, both families in $L_2$ only rely on the testing results of one particular family in $L_1$. We can consider it as the case that both “child” families share one “parent” family. This case is similar to tree structure but with retesting option. Moreover, if $g_{2112} \neq 0$ and $g_{2212} \neq 0$, then $F_{12}$ with $\alpha_{12} = 0$ still have chance to be tested during retesting stage due to the contributions from the “child” families in $L_2$.

4. Suppose $\alpha_{22} = 0$, then the testing results of both families in $L_1$ can contribute to a single family in $L_2$. In other words, we can consider it as the case that one “child” family has two “parent” families.

For algorithm 4.3, we have the following theorem.

**Theorem 4.2.** The two-level Bonferroni-based gatekeeping procedure with retesting option described in Algorithm 4.3 strongly controls the global FWER at level $\alpha$ under arbitrary dependence.

For the proof of theorem 4.2, see Appendix C.

### 4.8 Concluding Remarks

The main focus of this paper has been to develop simple and powerful procedures for testing ordered families of hypotheses. We introduced a new multiple testing
procedure, termed as Bonferroni-based gatekeeping procedure with retesting option, under which the families of hypotheses are repeatedly tested by Bonferroni procedure at updated local critical values in a sequential manner. We show that the proposed procedure strongly control the global FWER at level $\alpha$ under arbitrary dependence. Through two real clinical trial examples, we have illustrated the straightforward testing algorithms of our proposed procedures.

Both the superchain procedure and our proposed procedure allow iteratively retest families of hypotheses and both of them have power improvement compared with the procedure without retesting option. Apparently, by choosing the optimized initial parameters and initial multiple testing procedure for each family, superchain procedure has its advantage over our procedure with respect to power. However, the proposed procedure is comparable. In practice, to solve multiple ordered families of hypotheses problem in real life, it is also important to consider the simplicity of the testing procedure. To deal with several ordered families of hypotheses, the sequential testing strategy of our proposed procedure are more nature than the superchain procedure which tests all families concurrently. From the graphical point of view, given the tested families, transition matrix as well as the initial critical values, our procedure can be implemented based on simple Bonferroni procedure. No matter how many iterations each family has been through, the testing procedure used at each stage for each family never changes. On the contrary, for superchain procedure, even the graph is given, the specific algorithm can not be defined. One graph may have different superchain algorithms which lead to completely different testing results. It has been mentioned by Kordzakhia and Dmitrienko (2012) that the performance of superchain procedure is heavily depends on the choices of initial parameters, i.e., the initial truncation parameters, the initial local method. They have shown that by choosing different initial values of truncated parameters, the power of the resulting superchain procedure can have huge difference. Therefore, before starting
to implement superchain procedure, we first need to make some efforts to decide the optimal values of initial procedure parameters which adds more complexities into the implementation.

From the computational point of view, it is easy to see that our procedure is simple and easy to implement. The implementation of superchain procedure is complicated due to the updating rules of critical values and truncation parameters of the testing procedure for each family at each stage. The computational complexity is even more when the number of families are large. Due to that, it is difficult to communicate with non-statistician about the algorithm.

Of course, to use Bonferroni procedure as basic multiple testing procedure for each family make our proposed procedure slightly less powerful than superchain procedure in some cases. Therefore, a natural extension might be to use more powerful multiple testing procedure as basic procedure and then develop global FWER controlling sequential procedures with retesting option based on such basic procedure. Moreover, the proposed procedure controls the globe FWER without any assumption of dependence structure about test statistics. If there are some distribution information regarding to the test statistics known in advance, then it is possible to further improve the proposed procedure by exploiting the dependence information.
CHAPTER 5

PROCEDURES FOR TESTING MULTIPLE FAMILIES WITH
PARTIALLY ORDERED HIERARCHICAL STRUCTURE

5.1 Motivation

Multiple testing problems in confirmatory clinical trials usually involves hypotheses with complex structure which is beyond the simple hierarchical structures associated with serial and parallel gatekeeping methods. For example, the treatment comparison corresponding to a secondary endpoint is generally of interest if the treatment comparisons corresponding to some primary endpoints associated to this secondary endpoint have been rejected. Hung and Wang (2010) mentioned that “it does not make common sense to condition the rejection of a secondary endpoint on the rejection of the unrelated primary endpoints”. The following are two examples of some complex multiple comparison problems which the procedures discussed in the preceding chapters can either not work out or can not fully explain the logical relationships among the hypotheses.

Example 1. Consider a clinical trial for comparing a new treatment to an active control on two primary endpoints (P1 and P2) and one secondary endpoint (S1). There are two tests, non-inferiority (N) and superiority (S), for each endpoint such that there are totally six hypotheses. The null hypothesis for superiority can be tested only if the hypothesis for non-inferiority corresponding to the same endpoint could be rejected. Besides, the non-inferiority (superiority) hypothesis for secondary endpoint can be tested only if at least one of the non-inferiority (superiority) hypotheses for two primary endpoints are rejected. Based on such hierarchical logical relationships, these six hypotheses can be grouped into the following four families:

- $F_1 : H_{11}$ (N for P1), $H_{12}$ (N for P2).
Figure 5.1  Graphical visualization of logical structure for multiple testing problem in Example 1.

- $F_2 : H_{21}$ (S for P1), $H_{22}$ (S for P2).

- $F_3 : H_{31}$ (N for S1).

- $F_4 : H_{41}$ (S for S1).

Figure 5.1 visualizes the logical relationships among the six hypotheses. $H_{31}$ can only be tested when $H_{11}$ or $H_{12}$ is significant. $H_{41}$ can only be tested when $H_{21}$ or $H_{22}$ is significant.

Example 2. Consider a trial testing for treatment effects for three primary endpoints (P1, P2 and P3) and two secondary endpoints (S1, S2) at low (L), medium (M) and high (H) doses of a new treatment compared to placebo (P). Therefore, for primary endpoints, there are nine hypotheses total and for secondary endpoints, there are six hypotheses. For each endpoint, the high dose must be significant in order to test medium dose, then low dose. S1 is closely related to P1 and P2 in the sense that the hypotheses of S1 for a specific dose level can be carried out only when at least one hypotheses of P1 and P2 show meaningful treatment efficacy for that dose level.
Similarly, the hypotheses of S2 is of interested if only at least one hypotheses of P2 and P3 is significant. Theses fifteen hypotheses can be grouped into six families.

- $F_1 : H_{11}$ (H vs P for P1), $H_{12}$ (H vs P for P2), $H_{13}$ (H vs P for P3).
- $F_2 : H_{21}$ (M vs P for P1), $H_{22}$ (M vs P for P2), $H_{23}$ (M vs P for P3).
- $F_3 : H_{31}$ (L vs P for P1), $H_{32}$ (L vs P for P2), $H_{33}$ (L vs P for P3).
- $F_4 : H_{41}$ (H vs P for S1), $H_{42}$ (H vs P for S2).
- $F_5 : H_{51}$ (M vs P for S1), $H_{52}$ (M vs P for S2).
- $F_6 : H_{61}$ (L vs P for S1), $H_{62}$ (L vs P for S2).

Based on the logical relationships, we have, for example, $H_{61}$ can be tested only if at least one of $H_{31}$ and $H_{32}$ is rejected, and both $H_{41}$, $H_{51}$ are rejected, $H_{62}$ can be tested only if at least one of $H_{32}$ and $H_{33}$ is rejected, and both $H_{42}$, $H_{52}$ are rejected.

From the above two examples, we can see that besides the hierarchical structure between families of hypotheses, there are also intrinsic logical relationships among specific hypotheses between different families. The regular multiple testing procedures for testing a fixed sequence of families overlook the intrinsic partial hierarchical structure information and thus, lose testing power. Hence, the problem about partial hierarchical testing has been arising and the related research has been investigated by many authors, such as Dmitrienko et al. (2006), Rosenbaum (2008), Goeman and Mansmann (2008), Meinshausen (2008), Quan, Capizzi and Zhang (2009), Goeman and Finos (2009), as well as others. However, the existing methods have certain limitations and restrictions.

Dmitrienko, Wiens, Tamhance and Wang (2006) proposed a tree-structured gatekeeping procedure using the closure principle for testing logically related hypotheses in hierarchically ordered families. Dmitrienko, Tamhane, Liu and Wiens (2008) further discussed the statistical properties of this procedure. Since the
Figure 5.2  Graphical visualization of logical structure for multiple testing problem in Example 2.
tree-structured procedure is based on the closure principle, the computation is enormously complicated when the number of null hypotheses are large. Later, Maurer et al. (2011) proposed a partial hierarchically ordered procedure. In the paper, they suggested a “parent-decedent” relationship between primary and secondary hypotheses, i.e., inference of one decedent hypothesis in secondary family might only related to part of parent hypotheses in primary family. But their procedure is only suitable for testing two families of hypotheses. Meinshausen (2008) proposed a hierarchical testing procedure which can strongly control the FWER at level $\alpha$ only under the condition that if the “parent” hypothesis is true, then its corresponding “decedent” hypotheses are true with probability 1.

Therefore, we are motivated to propose a new multiple testing procedure for testing multiple families with partially ordered hierarchical structure. The idea is that the decision for testing one hypothesis in lower-ranked families may only depend on the testing results of several hypotheses, referred to as “parent hypotheses”, in higher-ranked families. For each family, the proposed procedure only tests those hypotheses whose corresponding “parent hypotheses” satisfy some pre-specified criteria instead of the whole family. This property reduces the number of testing hypotheses, enhances the testing efficiency. It is shown that the proposed procedure can also control the FWER at a pre-specified level $\alpha$ under the assumption that test statistics of null hypotheses from different families are independent.

The rest of the chapter is organized as follows. In Section 5.2, we present some notations and basic assumption which are used through the whole chapter. The main theoretical results are introduced in Section 5.3 and two clinical trial examples are illustrated in Section 5.4 to show the implementation of the proposed procedure. Some conclusion remarks are discussed in Section 5.5.
5.2 Preliminary

To define the multilevel partial hierarchical procedure, consider testing \( n \) hypotheses that can be grouped into \( m \geq 2 \) families based on their hierarchical logical relationships, labeled as \( F_i = H_{i1}, \ldots, H_{im_i} (1 \leq i \leq m) \). Note that \( n_i \) corresponds to the number of hypotheses in \( F_i \) and \( \sum_{i=1}^{m} n_i = n \). We need to strongly control the overall FWER at level \( \alpha \) for the whole family \( F = \bigcup_{i=1}^{m} F_i \) of all hypotheses, i.e., the probability of incorrectly rejecting at least one true null hypothesis will be controlled at \( \alpha \) regardless of which and how many null hypotheses in the \( m \) families are true. For each family \( F_i, i = 1, \ldots, m \), we let \( T_i \) and \( U_i \) be the true null set and false null set and the corresponding cardinalities be defined as \( |T_i| \) and \( |U_i| \), respectively. Also, let the \( A_i \) and \( R_i \) be the corresponding sets of accepted and rejected hypotheses respectively with cardinality \( |A_i| \) and \( |R_i| \). The general strategy of a partial hierarchical testing procedure is that a hypothesis can be tested or not only depends on testing results of some in stead of all of hypotheses in previous families.

In general, the procedure sequentially tests \( m \) hierarchically ordered families of hypotheses. For each hypothesis \( H_{ij} \) in \( F_i, i = 2, \ldots, m \), there is a corresponding subset of hypotheses in \( F_1, \ldots, F_{i-1} \), referred to as “parent hypotheses”. \( H_{ij}, i = 2, \ldots, m \), is only of interest if the testing results of its respective “parent hypotheses” satisfy some pre-specified conditions. Otherwise, it will be directly attained. Denote \( sF_i \) the set of testable hypotheses in \( F_i \) with critical value \( \alpha_i \). Note that \( F_1 = sF_1 \) and \( \alpha_1 = \alpha \). Especially, if \( sF_i = \emptyset \), the whole procedure stops. Before introducing the procedure, we need to make the following assumption:

**Assumption 5.1.** Assume that the p-values corresponding to the null hypotheses in \( F_i \) are independent of the p-values corresponding to the null hypotheses in \( F_j, i \neq j \).
5.3 Main Theoretical Results

5.3.1 General Multilevel Partial Hierarchical Procedure

In this section, we need to use the error rate function introduced in Section 1.3.1 in developing the general multilevel partial hierarchical procedure. The algorithm for the procedure is as follows.

Algorithm 5.1.

Step 1. Test $F_1$ with a separable multiple testing procedure at level $\alpha$. Let $A_1$ be the set of acceptances in $F_1$. Calculate $e^*_1(A_1)$ where $e^*_i(.)$ is the upper bound of error rate function of the corresponding separable procedure.

Step $k$ ($2 \leq k \leq m - 1$). Test the testable hypotheses in $sF_k$ with a separable multiple testing procedure at level $\alpha_k$ where

$$\alpha_k = \alpha_{k-1} - e^*_{k-1}(A_{k-1}).$$ (5.1)

In (5.1), $A_{k-1}$ is the set of acceptances in $sF_{k-1}$ and $e^*_{k-1}(.)$ is the upper bound of error rate function of the separable procedure used in testing $sF_{k-1}$.

Retain all hypotheses in $F_k \setminus sF_k$. If $A_k = F_k$, i.e., all hypotheses are accepted in $F_k$, then stop the algorithm and accept all hypotheses in $F_{k+1}, \cdots, F_m$. Otherwise, continue to next step.

Step $m$. Use any multiple testing procedure to test $F_m$ at level $\alpha_m$.

Remark 5.1. Regarding to Algorithm 5.1, we know that if all hypotheses in $F_i$ are rejected at step $i, i = 1, \cdots, m - 1$, then the whole critical value of $F_i$ will be transferred to $F_{i+1}$, i.e., $\alpha_{i+1} = \alpha_i$. The essence of Algorithm 5.1 is the same as the general multistage gatekeeping procedure. Both of the procedures are stepwise and the critical value for each family depends on the testing result of the previous family. However, the general multistage gatekeeping procedure tests all hypotheses in one family without considering the partial hierarchical logical relationships. On the
contrary, our proposed procedure only tests the subfamily $sF_i$ which is the testable hypotheses in $F_i$. Thus, those tested hypotheses will have more chances to be rejected.

Furthermore, we have the following theorem.

**Theorem 5.1.** Under assumption 5.1, the general multilevel partial hierarchical procedure described in algorithm 5.1 can strongly control the FWER at level $\alpha$.

**Proof.** Within $sF_i$, let $T_i^*$ be the subset of true null hypotheses with cardinality $|T_i^*|$. And we denote the number of testable hypotheses in $sF_i$ as $n_i^*$. Let $E_i(x)$ denote the event that at least one true null hypothesis in $F_i$ is rejected at level $x$, $i = 1, \cdots, m$. Thus the FWER can be represented as

$$FWER = \Pr \left\{ \bigcup_{i=1}^{m} E_i(\alpha_i) \right\}$$

$$= \Pr \{E_1(\alpha_1)\} + \Pr \{\overline{E}_1(\alpha_1) \cap E_2(\alpha_2)\}$$

$$+ \cdots + \Pr \{\cap_{i=1}^{m-1} \overline{E}_i(\alpha_i) \cap E_m(\alpha_m)\},$$

(5.2)

where $\overline{E}_i(\alpha_i)$ is the complement of $E_i(\alpha_i)$.

Firstly, we know that

$$\Pr \{E_1(\alpha_1)\} \leq e_1^*(T_1)$$

(5.3)

which is due to the definition of error rate function.

Secondly, after testing $F_1$, for any $k, k = 2, \cdots, m - 1$, $F_k$ will be tested at level $\alpha_k = \alpha_{k-1} - e_{k-1}^*(A_{k-1})$. When $\cap_{i=1}^{k-1} \overline{E}_i(\alpha_i)$ is true, there is no true null hypotheses being rejected in $F_1, \cdots, F_{k-1}$.

$$\Pr \left\{ \cap_{i=1}^{k-1} \overline{E}_i(\alpha_i) \cap E_k(\alpha_k) \right\}_{R_{k-1}}$$

$$\leq \Pr \{E_k(\alpha_k) \mid R_{k-1}\}$$

$$\leq E \{e_k^*(T_k^*)\}$$

93
The first inequality holds due to assumption 5.1 and true null hypotheses follow $U(0, 1)$. The second inequality holds according to the definition of error rate function. Therefore,

$$\Pr \left\{ \bigcap_{i=1}^{k-1} E_i(\alpha_i) \cap E_k(\alpha_k) \right\} = E \left\{ \Pr \left\{ \bigcap_{i=1}^{k-1} E_i(\alpha_i) \cap E_k(\alpha_k) \bigg| R_{k-1} \right\} \right\} \leq E \left\{ E \left\{ e^*_k(T^*_k) \right\} \right\} \quad (5.4)$$

Lastly, consider the event $\bigcap_{i=1}^{m-1} E_i(\alpha_i) \cap E_m(\alpha_m)$. If $\bigcap_{i=1}^{m-1} E_i(\alpha_i)$ is true, i.e., there is no true null hypotheses rejected in the first $m - 1$ families. It implies that $T^*_i \subset A_i$ for any $i = 2, \ldots, m - 1$ and $T_1 \subset A_1$. Therefore, according to the monotonicity of error rate function we have the following recursive relation:

$$e^*_t(A_t) \geq e^*_t(T^*_t),$$

$$\alpha_t - e^*_t(A_t) \leq \alpha_t - e^*_t(T^*_t) \leq \alpha_{t-1} - e^*_t(T^*_t) - e^*_t(T^*_t),$$

which implies that

$$\alpha_m \leq \alpha_{m-1} - e^*_m(T^*_m)$$

$$\leq \alpha_{m-2} - e^*_{m-2}(T^*_{m-2}) - e^*_m(T^*_m)$$

$$\leq \alpha_t - e^*_t(T_1) - \sum_{i=2}^{m-1} e^*_i(T^*_i). \quad (5.5)$$

Also, since $F_m$ is tested using any FWER controlling multiple testing procedure at level $\alpha_m$, we have

$$\Pr \left\{ \bigcap_{i=1}^{m-1} E_i(\alpha_i) \cap E_m(\alpha_m) \bigg| R_{m-1} \right\}$$

$$\leq \Pr \left\{ \bigcap_{i=1}^{m-1} E_i(\alpha_i) \cap E_m(\alpha_{m-1} - e^*_m(T^*_m)) \bigg| R_{m-1} \right\}$$

$$\leq \Pr \left\{ E_m(\alpha_{m-1} - e^*_m(T^*_m)) \bigg| R_{m-1} \right\}$$

$$\leq E \left\{ \alpha_{m-1} - e^*_m(T^*_m) \right\}$$

$$\leq E \left\{ \alpha - e^*_1(T_1) - \sum_{i=2}^{m-1} e^*_i(T^*_i) \right\}. \quad 94$$
Therefore,

\[
\Pr \left\{ \cap_{i=1}^{m-1}\mathcal{E}_i(\alpha_i) \cap \mathcal{E}_m(\alpha_m) \right\} = E \left\{ \Pr \left\{ \cap_{i=1}^{m-1}\mathcal{E}_i(\alpha_i) \cap \mathcal{E}_m(\alpha_m) \mid R_{m-1} \right\} \right\} 
\leq E \left\{ E \left\{ \alpha - e_1^*(T_1) - \sum_{i=2}^{m-1} e_i^*(T_i^*) \right\} \right\}. \tag{5.6}
\]

Thus, based on (5.3), (5.4) and (5.6), we have

\[
\text{FWER} \leq e_1^*(T_1) + E \left\{ E \left\{ e_2^*(T_2^*) \right\} \right\} 
+ \cdots + E \left\{ E \left\{ e_{m-1}^*(T_{m-1}^*) \right\} \right\} 
+E \left\{ E \left\{ \alpha - e_1^*(T_1) - \sum_{i=2}^{m-1} e_i^*(T_i^*) \right\} \right\} 
= \alpha \tag{5.7}
\]

The last inequality holds due to that it is the expectation of a constant. The proof is complete. \hfill \square

Remark 5.2. The proposed procedure specially deals with multiple families of hypotheses with partially ordered hierarchical structure. Both serial and parallel gatekeeping procedures are special cases of our procedure. For both serial and parallel gatekeeping, all hypotheses in one family are the “parent hypotheses” for each hypothesis in the subsequent family. For serial gatekeeping, one hypothesis is testable only if all its “parent hypotheses” are rejected. And for parallel gatekeeping, one hypothesis is testable as long as at least one of its “parent hypotheses” is rejected.

5.4 Illustration Analysis

This section illustrates our proposed procedures using the two clinical trial examples presented in Section 5.1 and compares them with two existing procedures, tree-structure gatekeeping procedure in Dmitrienko et al. (2006) and general multistage gatekeeping procedure in Dmitrienko et al. (2008) introduced in Section 1.3. Each procedure is based on the regular Bonferroni method and the global FWER is to be
Table 5.1 Comparison Results of Proposed Procedure, Multistage Gatekeeping Procedure and Tree-Structure Gatekeeping Procedure in the Motivation Example 1

<table>
<thead>
<tr>
<th>Family</th>
<th>Null hypothesis</th>
<th>Raw p-value</th>
<th>Partial hierarchical</th>
<th>Multistage gatekeeping</th>
<th>Tree structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>$H_{11}$</td>
<td>0.001</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>$H_{12}$</td>
<td>0.026</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>$F_2$</td>
<td>$H_{21}$</td>
<td>0.015</td>
<td>S</td>
<td>NS</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>$H_{22}$</td>
<td>0.208</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>$F_3$</td>
<td>$H_{31}$</td>
<td>0.020</td>
<td>S</td>
<td>NS</td>
<td>S</td>
</tr>
<tr>
<td>$F_4$</td>
<td>$H_{41}$</td>
<td>0.578</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
</tbody>
</table>

Note: S = Significant, NS = Not significant.

controlled at level $\alpha = 0.05$.

**Example 1.** The hypothesis testing problem illustrated in the Example 1 in Section 5.1 includes four families of null hypotheses. Suppose that the raw $p$-values for the six hypotheses are given in Table 5.1. The proposed partial hierarchical procedure is implemented as follows.

*Step 1.* Test $F_1$ using Bonferroni method at level $\alpha = 0.05$, only $H_{11}$ is rejected in $F_1$.

*Step 2.* Based on the logical restriction, $sF_2 = \{H_{21}\}$. Test $sF_2$ at level $\alpha_2 = \frac{\alpha}{2}$ using Bonferroni method. $H_{21}$ is rejected and $H_{22}$ is directly accepted.

*Step 3.* Since one of hypotheses in $F_1$ is rejected, $sF_3 = \{H_{31}\}$. Test $sF_3$ at level $\frac{\alpha}{2}$ using Bonferroni method, and thus $H_{31}$ is rejected.

*Step 4.* Since one of hypotheses in $F_2$ are rejected and $H_{31}$ is rejected, $sF_4 = \{H_{41}\}$. Test $sF_4$ at level $\frac{\alpha}{2}$ using Bonferroni method and $H_{41}$ is accepted.

Table 5.1 gives the testing results using the aforementioned three procedures. It is easy to see that by considering the logical relationships among particular
null hypotheses, our proposed procedure and tree-structure gatekeeping procedure rejected more hypotheses than general multistage gatekeeping procedure which only consider the general hierarchical relationships among families. Although the tree-structure gatekeeping procedure has the same performance as our proposed procedure, it is based on closure principle so that we need to calculate the Bonferroni $p$-values for the $2^6 - 1 = 127$ intersection hypotheses. By contrast, our proposed procedure follows a simple stepwise form such that computational complexity is greatly reduced.

**Example 2.** The hypothesis testing problem illustrated in the Example 2 in Section 5.1 includes six families of null hypotheses and the intrinsic logical relationships are more complicated. Suppose that the raw $p$-values for the fifteen hypotheses are given in Table 5.2. The proposed partial hierarchical procedure is implemented as follows.

1. **Step 1.** Test $F_1$ using Bonferroni method at level $\alpha = 0.05$, $H_{11}$ and $H_{12}$ are rejected in $F_1$.

2. **Step 2.** Based on the logical restriction, $sF_2 = \{H_{21}, H_{22}\}$. Test $sF_2$ at level $\alpha_2 = \frac{2\alpha}{3}$ using Bonferroni method. Only $H_{21}$ is rejected and $H_{22}$ is accepted. $H_{23}$ is directly accepted.

3. **Step 3.** Since only $H_{21}$ is rejected in $F_2$, $sF_3 = \{H_{31}\}$. Test $sF_3$ at level $\alpha_3 = \frac{\alpha}{3}$ using Bonferroni method. $H_{31}$ is rejected. Both $H_{32}$ and $H_{33}$ are directly accepted.

4. **Step 4.** Since $H_{11}$ and $H_{12}$ are rejected, $sF_4 = \{H_{41}, H_{42}\}$. Test $sF_4$ at level $\frac{\alpha}{3}$ using Bonferroni method, and thus $H_{41}$ is rejected.

5. **Step 5.** Since $H_{21}$ is rejected and $H_{41}$ is rejected, $sF_5 = \{H_{51}\}$. Test $sF_5$ at level $\frac{\alpha}{6}$ using Bonferroni method and $H_{51}$ is rejected and $H_{52}$ is directly accepted.

6. **Step 6.** Since $H_{31}$ is rejected and $H_{51}$ is rejected, $sF_6 = \{H_{61}\}$. Test $sF_6$ at level $\frac{\alpha}{6}$ using Bonferroni method and $H_{61}$ is not rejected and $H_{62}$ is directly accepted.

Table 5.2 gives the testing results using the aforementioned three procedures. In this example, our proposed procedure rejects the most number of hypotheses.
Table 5.2 Comparison Results of Proposed Procedure, Multistage Gatekeeping Procedure and Tree-Structure Gatekeeping Procedure in the Motivation Example 2

<table>
<thead>
<tr>
<th>Family</th>
<th>Null hypothesis</th>
<th>Raw p-value</th>
<th>Partial hierarchical</th>
<th>Multistage gatekeeping</th>
<th>Tree structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>$H_{11}$</td>
<td>0.001</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>$H_{12}$</td>
<td>0.008</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>$H_{13}$</td>
<td>0.026</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>$F_2$</td>
<td>$H_{21}$</td>
<td>0.015</td>
<td>S</td>
<td>NS</td>
<td>S</td>
</tr>
<tr>
<td></td>
<td>$H_{22}$</td>
<td>0.018</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>$H_{23}$</td>
<td>0.208</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>$F_3$</td>
<td>$H_{31}$</td>
<td>0.010</td>
<td>S</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>$H_{32}$</td>
<td>0.030</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>$H_{33}$</td>
<td>0.302</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>$F_4$</td>
<td>$H_{41}$</td>
<td>0.008</td>
<td>S</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>$H_{42}$</td>
<td>0.200</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>$F_5$</td>
<td>$H_{51}$</td>
<td>0.005</td>
<td>S</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>$H_{52}$</td>
<td>0.100</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>$F_6$</td>
<td>$H_{61}$</td>
<td>0.013</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td></td>
<td>$H_{62}$</td>
<td>0.578</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
</tbody>
</table>

Note: S = Significant, NS = Not significant.
Tree-structure gatekeeping procedure rejects one more hypothesis than general multistage gatekeeping procedure but less than our proposed procedure.

5.5 Concluding Remarks

In this chapter, we have proposed a multilevel partial hierarchical procedure under which the decision for testing a hypothesis in current families only depend on the testing results of its respective “parent hypotheses” from previous families. For each family except the first one, the procedure only tests those testable hypotheses instead of the whole family. We have also proved that this procedure can control the FWER at level $\alpha$ under the assumption that $p$-values of hypotheses from different families are independent. Clinical trial examples illustrate the advantages of the proposed procedure that it is powerful, simple to implement, and can manage those multiple comparison problems with more complex structured hypotheses.

A possible future work might be to generalize our proposed partial hierarchical procedure from independence to arbitrary dependence.
CHAPTER 6

CONCLUSION AND FUTURE WORK

In applications of clinical trials, the hypotheses to be tested often exhibit a hierarchical structure. They are usually hierarchically ordered based on their importance, clinical relevance, or dose concentration, etc., and thus are tested in a pre-defined fixed sequence. In some more complex cases, the hypotheses to be tested are hierarchically grouped into several families, and thus the families are tested in a sequential order. Although such problems of structured multiple testing have received much attention in the last decades and several popular FWER controlling procedures, such as conventional fixed sequence procedure, fallback procedure and gatekeeping procedure have been introduced, not much progress has been made yet advancing their theory and methods. This thesis contributes to the development of theory and methods for structured multiple testing problems.

In Chapter 2, we proposed a generalized fixed sequence procedure for testing a single family of hypotheses and gave sufficient conditions for its FWER control under arbitrary dependence. Through extensive simulation studies, we illustrate the advantages of our proposed procedures over the existing FWER controlling procedures in terms of FWER control and power. When the pairwise joint distribution of the true null \( p \)-values are known, we improved the proposed procedure by incorporating the distributional information into the construction of the procedure while maintaining the control of the FWER. To use the fixed sequence procedures, prior knowledge of the ordering of the tested hypotheses is required. However, when the ordering is not completely correct, the fixed sequence procedures usually lose their edge over the conventional \( p \)-value based stepwise procedures. Therefore, a natural extension of work in Chapter 2 is to use a combination of the \( p \)-values and the a-priori ordering
information to order the hypotheses to be tested and then develop FWER controlling procedures based on such ordering.

In Chapter 3, we have proposed a simple graphical tool to sequentially test hierarchically ordered families of hypotheses. We have described the algorithm and presented the theoretical results associating to the FWER control. Through some examples, we have shown the efficiency of our procedure comparing with the hypothesis-based graphical approach (Bretz et al., 2008) when deal with hierarchically related multiple families of hypotheses.

In Chapter 4, we developed a Bonferroni-based gatekeeping procedure with retesting option for testing ordered families of hypotheses. We proved that it can strongly control the global FWER under arbitrary dependence. By this procedure, each family of hypotheses is repeatedly tested using Bonferroni procedure with updated local critical values. Using Bonferroni procedure as basic multiple testing procedure for each family makes our proposed procedure slightly conservative comparing to superchain procedure (Dmitrienko et al., 2013) in some cases. A possible future work is to use more powerful multiple testing procedures, such as Holm procedure, to replace Bonferroni procedure as the local procedure. Another possible future work is that if there are some distribution information regarding to the test statistics known in advance, then it is possible to further improve the proposed procedure by exploiting the dependence information.

In Chapter 5, we introduced a general multilevel partial hierarchical procedure for testing multiple families of hypotheses with partial hierarchical structures, such as tree structure. By this procedure, the decision for testing a hypotheses in current families only depend on the testing results of its respective “parent hypotheses” in previous families. We have also proved that this procedure can control the FWER at level $\alpha$ under the assumption that $p$-values of hypotheses from different families are independent. Clinical trial examples illustrated the advantages of the proposed
procedure over the general multistage procedure (Dmitrienko et al., 2008) and the tree-structure procedure (Dmitrienko et al., 2007). A possible future work is to generalize our proposed partial hierarchical procedure from independence assumption to arbitrary dependence.
This appendix contains the proofs of the proposition and theorems stated but not proved in Chapter 2.

A.1 Proof of Proposition 2.1.

Proof. Let $V$ denote the set of falsely rejected hypotheses which can be expressed as a set function in terms of the $p$-values

$$V(P_1, \cdots, P_n) = \bigcup_{i=1}^{n} \{H_i \text{ is true : } P_i \leq \alpha(s_i, t_i)\} = \bigcup_{j=1}^{n_0} \{\hat{P}_j \leq \alpha(s_j, t_j)\}. \quad (A.1)$$

And for any given $1 \leq i \leq n$, the number of rejections before testing $H_i$ satisfies

$$s_i = \sum_{j=1}^{t_i} I(1 \leq i \leq n, P_j \leq \alpha(s_j, t_j)), \quad (A.2)$$

where $I(.)$ is an indicator function.

For any $0 \leq t_i \leq n - 1$, $\alpha(s_i, t_i)$ is non-decreasing in $s_i$ due to the property of critical value function. Also from (A.2), we can see that $s_i$ will decrease as $p$-values increase. Thus, $\alpha(s_i, t_i)$ is a non-increasing function with respect to $P_1, \cdots, P_n$. Hence, it is easy to see that $V(P_1, \cdots, P_n)$ is a decreasing set function with respect to $P_1, \cdots, P_n$, i.e., for any given $P_i' \leq P_i, i = 1, \cdots, n$, we have $V(P_1, \cdots, P_n) \subseteq V(P_1', \cdots, P_n')$. Therefore, the set of rejections $V(P_1, \cdots, P_n)$ will become larger if we let all false null $p$-values equal to zero.

After replacing all of the false null $p$-values by 0, it is necessary to show that all false null hypotheses listed in front of true null hypotheses is the worst case in the sense that the FWER based on this structure attains the maximum. From equation

\[
\text{\textit{APPENDIX A}}
\]

\text{\textbf{PROOFS FOR FIXED SEQUENCE PROCEDURE}}

This appendix contains the proofs of the proposition and theorems stated but not proved in Chapter 2.

A.1 Proof of Proposition 2.1.

\textit{Proof.} Let $V$ denote the set of falsely rejected hypotheses which can be expressed as a set function in terms of the $p$-values

$$V(P_1, \cdots, P_n) = \bigcup_{i=1}^{n} \{H_i \text{ is true : } P_i \leq \alpha(s_i, t_i)\} = \bigcup_{j=1}^{n_0} \{\hat{P}_j \leq \alpha(s_j, t_j)\}. \quad (A.1)$$

And for any given $1 \leq i \leq n$, the number of rejections before testing $H_i$ satisfies

$$s_i = \sum_{j=1}^{t_i} I(1 \leq i \leq n, P_j \leq \alpha(s_j, t_j)), \quad (A.2)$$

where $I(.)$ is an indicator function.

For any $0 \leq t_i \leq n - 1$, $\alpha(s_i, t_i)$ is non-decreasing in $s_i$ due to the property of critical value function. Also from (A.2), we can see that $s_i$ will decrease as $p$-values increase. Thus, $\alpha(s_i, t_i)$ is a non-increasing function with respect to $P_1, \cdots, P_n$. Hence, it is easy to see that $V(P_1, \cdots, P_n)$ is a decreasing set function with respect to $P_1, \cdots, P_n$, i.e., for any given $P_i' \leq P_i, i = 1, \cdots, n$, we have $V(P_1, \cdots, P_n) \subseteq V(P_1', \cdots, P_n')$. Therefore, the set of rejections $V(P_1, \cdots, P_n)$ will become larger if we let all false null $p$-values equal to zero.

After replacing all of the false null $p$-values by 0, it is necessary to show that all false null hypotheses listed in front of true null hypotheses is the worst case in the sense that the FWER based on this structure attains the maximum. From equation
\( V(P_1, \ldots, P_n) \) is empty \[= \bigcap_{j=1}^{n_0} \{ \hat{P}_j > \alpha(s_j, t_j) \}. \] \tag{A.3}

For any \( j = 1, \ldots, n_0 \), let \( i_j \) be the smallest index in the original sequence of hypotheses satisfying \( P_{i_j} = \hat{P}_j \), and \( \tilde{s}_j \) and \( \tilde{t}_j \), respectively, be the number of rejections and acceptances before testing the \( j^{th} \) true null hypothesis under the DO configuration. Assume that before testing the \( j^{th} \) true null, all of the false null hypotheses are directly rejected and true null hypotheses are accepted. Then, it implies that for any \( j = 1, \ldots, n_0 \), \( s_j = i_j - j \leq \tilde{s}_j = n_1 \) and \( t_j = \tilde{t}_j = j - 1 \). Since the function \( \alpha(s, t) \) is non-decreasing in \( s \), \( \alpha(s_j, t_j) \leq \alpha(\tilde{s}_j, \tilde{t}_j) \) for each \( j \). Hence we have

\[
\bigcap_{j=1}^{n_0} \{ \hat{P}_j > \alpha(s_j, t_j) \} \supseteq \bigcap_{j=1}^{n_0} \{ \hat{P}_j > \alpha(\tilde{s}_j, \tilde{t}_j) \},
\]

which implies

\[
\Pr \left\{ \bigcap_{j=1}^{n_0} \{ \hat{P}_j > \alpha(s_j, t_j) \} \right\} \geq \Pr \left\{ \bigcap_{j=1}^{n_0} \{ \hat{P}_j > \alpha(\tilde{s}_j, \tilde{t}_j) \} \right\},
\]

\[
1 - \text{FWER} \geq 1 - \text{FWER}_{\text{DO}}.
\]

Thus, the desired result is proved. \( \square \)

### A.2 Proof of Theorem 2.1.(ii)

**Proof.** We will construct a joint distribution under the DO configuration such that the true null \( p \)-values are \( U(0, 1) \) and the FWER is exactly \( \alpha \). Consider the following
construction:

\[\hat{P}_1 \sim U(0, 1),\]

\[\hat{P}_2 \sim \begin{cases} U(0, 1 - \alpha(n_1, 0)), & \text{if } \hat{P}_1 > \alpha(n_1, 0), \\ U(1 - \alpha(n_1, 0), 1), & \text{if } \hat{P}_1 \leq \alpha(n_1, 0), \end{cases}\]

\[\vdots\]

\[\hat{P}_{n_0} \sim \begin{cases} U(0, 1 - \sum_{i=1}^{n_0-1} \alpha(n_1, i - 1)), & \text{if } \bigcap_{i=1}^{n_0-1} \{\hat{P}_i > \alpha(n_1, i - 1)\}, \\ U(1 - \sum_{i=1}^{n_0-1} \alpha(n_1, i - 1), 1), & \text{otherwise.} \end{cases}\]

Since \(\alpha(n_1, j - 1) \leq 1 - \sum_{i=1}^{j-1} \alpha(n_1, i - 1), t = 1, \ldots, n_0\), when the event \(\{\hat{P}_j \leq \alpha(n_1, j - 1)\}\) occurs, \(\bigcap_{i=1}^{j-1} \{\hat{P}_i > \alpha(n_1, i - 1)\}\) occurs with probability 1. That is, for \(i = 0, \ldots, m_0 - 1\),

\[\Pr\{\hat{P}_1 > \alpha(n_1, 0), \ldots, \hat{P}_{i-1} > \alpha(n_1, i - 2), \hat{P}_i \leq \alpha(n_1, i - 1)\}\]

\[= \Pr\{\hat{P}_i \leq \alpha(n_1, i - 1)\}. \quad (A.4)\]

We will use induction to show that the true null hypotheses are \(U(0, 1)\). Trivially, \(\hat{P}_1\) is \(U(0, 1)\). Assume the first \(j - 1, j = 2, \ldots, n_0\) true null hypotheses are \(U(0, 1)\). Then,

\[\Pr\{\bigcap_{i=1}^{j-1} \hat{P}_i > \alpha(n_1, i - 1)\} = 1 - \Pr\{\bigcup_{i=1}^{j-1} \hat{P}_i \leq \alpha(n_1, i - 1)\}\]

\[= 1 - \sum_{i=1}^{j-1} \Pr\{\hat{P}_1 > \alpha(n_1, 0), \ldots, \hat{P}_{i-1} > \alpha(n_1, i - 2), \hat{P}_i \leq \alpha(n_1, i - 1)\}\]

\[= 1 - \sum_{i=1}^{j-1} \Pr\{\hat{P}_i \leq \alpha(n_1, i - 1)\} = 1 - \sum_{i=1}^{j-1} \alpha(n_1, i - 1). \quad (A.5)\]

The third inequality follows from (A.4).
Let $0 \leq u \leq 1 - \sum_{i=1}^{j-1} \alpha(n_1, j - 1)$,

$$\Pr\{\hat{P}_j \leq u\} = \Pr\left\{\hat{P}_j \leq u \mid \bigcap_{i=1}^{j-1} \hat{P}_i > \alpha(n_1, i - 1)\right\} \Pr\left\{\bigcap_{i=1}^{j-1} \hat{P}_i > \alpha(n_1, i - 1)\right\}$$

$$= \frac{u}{1 - \sum_{i=1}^{j-1} \alpha(n_1, i - 1)} \Pr\left(\bigcap_{i=1}^{j-1} \{\hat{P}_i > \alpha(n_1, i - 1)\}\right)$$

$$= u.$$

Now, let $1 - \sum_{i=1}^{j-1} \alpha(n_1, i - 1) < u \leq 1$,

$$\Pr\{\hat{P}_j \leq u\} = \Pr\left\{\hat{P}_j \leq u \mid \bigcap_{i=1}^{j-1} \hat{P}_i > \alpha(n_1, i - 1)\right\} \Pr\left\{\bigcap_{i=1}^{j-1} \hat{P}_i > \alpha(n_1, i - 1)\right\}$$

$$+ \Pr\left\{\hat{P}_j \leq u \mid \bigcup_{i=1}^{j-1} \hat{P}_i \leq \alpha(n_1, i - 1)\right\} \left(1 - \Pr\left\{\bigcap_{i=1}^{j-1} \hat{P}_i > \alpha(n_1, i - 1)\right\}\right)$$

$$= \Pr\left\{\bigcap_{i=1}^{j-1} \hat{P}_i > \alpha(n_1, i - 1)\right\}$$

$$+ \frac{u - (1 - \sum_{i=1}^{j-1} \alpha(n_1, i - 1))}{\sum_{i=1}^{j-1} \alpha(n_1, i - 1)} \left(1 - \Pr\left(\bigcap_{i=1}^{j-1} \hat{P}_i > \alpha(n_1, i - 1)\right)\right)$$

$$= u.$$

Finally,

$$\text{FWER}_{\text{DO}} = \sum_{t=1}^{n_0} \Pr\left\{\hat{P}_1 > \alpha(n_1, 0), \ldots, \hat{P}_{t-1} > \alpha(n_1, t - 2), \hat{P}_t \leq \alpha(n_1, t - 1)\right\}$$

$$= \sum_{t=1}^{n_0} \Pr\left\{\hat{P}_t \leq \alpha(n_1, t - 1)\right\} = \sum_{i=1}^{n_0} \alpha(n_1, t - 1) = \alpha.$$

The second equality follows from (A.4).

A.3 Proof of Theorem 2.3.

Proof. To prove these two procedures are equivalent, we only need to prove the following two results for any individual hypothesis $H_i, i = 1, \ldots, n:

Result 1. If $H_i$ is accepted by the generalized fixed sequence procedure, then it is also accepted by the closed testing procedure.
**Result 2.** If $H_i$ is accepted by the closed testing procedure, then it is also accepted by the generalized fixed sequence procedure.

*Proof of Result 1.* Consider an intersection hypothesis, $\tilde{H}_i = \bigcap_{j=1}^{i} \{ H_j : P_j > \alpha(s_{j-1}, t_{j-1}) \}$. Since $H_i$ is accepted by the generalized fixed sequence procedure, $P_i > \alpha(s_{i-1}, t_{i-1})$ and thus $H_i$ is contained in $\tilde{H}_i$. Note that for all $H_j$ contained in $\tilde{H}_i$, we have $s_{j-1}^* = s_{j-1}$ and $t_{j-1}^* = t_{j-1}$. Thus, $\alpha(s_{j-1}^*, t_{j-1}^*) = \alpha(s_{j-1}, t_{j-1})$ and $P_j > \alpha(s_{j-1}^*, t_{j-1}^*)$. Therefore, $\tilde{H}_i$ is not rejected by the local test and thus $H_i$ is accepted by the closed testing procedure.

*Proof of Result 2.* Since $H_i$ is accepted by the closed testing procedure, thus there exists an index set $I$ with $i \in I$ such that the intersection hypothesis, $\bigcap_{j \in I} H_j$ is not rejected by the corresponding local test. Thus, for any $j \in I$, we have $P_j > \alpha(s_{j-1}^*, t_{j-1}^*)$. By using the mathematical induction method, it is easy to show that $s_{j-1}^* \geq s_{j-1}$ and $t_{j-1}^* \leq t_{j-1}$ for all $j \in I$. Thus, $P_i > \alpha(s_{i-1}^*, t_{i-1}^*) \geq \alpha(s_{i-1}, t_{i-1})$ follows due to monotonicity of the critical value function. Therefore, $H_i$ is accepted by the generalized fixed sequence procedure. 

$\Box$
This appendix contains the proofs of the theorems stated but not proved in Chapter 3.

**B.1 Proof of Theorem 3.1**

*Proof.* Suppose the family $F_{ij}$ is tested at level $\alpha_{ij}^*$, then we know

$$\alpha_{1j}^* = \alpha_{1j},$$

$$\alpha_{2i}^* = \alpha_{2i} + \sum_{j=1}^{2} (\alpha_{1j}^* - e^*(A_{1j}))g_{1j2i}. \quad (B.1)$$

For $i, j = 1, 2$, define event $E_{ij}(x) = \{\text{at least one true null hypothesis being rejected in } F_{ij} \text{ at significant level } x\}$. Thus,

$$\text{FWER} = \Pr\{\bigcup_{j=1}^{2} E_{ij}(\alpha_{ij}^*)\} \quad (B.2)$$

$$= \Pr\{\bigcup_{j=1}^{2} E_{1j}(\alpha_{1j}^*)\} + \Pr\{\bigcap_{j=1}^{2} \overline{E}_{1j}(\alpha_{1j}^*) \cap \bigcup_{j=1}^{2} E_{2j}(\alpha_{2j}^*)\},$$

where $\bigcap_{j=1}^{2} \overline{E}_{1j}(\alpha_{1j}^*)$ is the complement set of $\bigcup_{j=1}^{2} E_{1j}(\alpha_{1j}^*)$.

Let $T_{ij}$ denote the sets of true null hypotheses in $F_{ij}$ and the rejection set and acceptance set are denoted as $R_{ij}$ and $A_{ij}$, respectively.

First of all, let us consider $\bigcup_{j=1}^{2} E_{1j}(\alpha_{1j}^*)$. Note that

$$\Pr\{\bigcup_{j=1}^{2} E_{1j}(\alpha_{1j}^*)\} \leq \sum_{j=1}^{2} \Pr\{E_{1j}(\alpha_{1j}^*)\} \leq \sum_{j=1}^{2} e_{ij}^*(T_{ij}). \quad (B.3)$$

The inequality in (B.3) follows from the definition of the error rate function.

Next, we consider $\bigcap_{j=1}^{2} \overline{E}_{1j}(\alpha_{1j}^*) \cap \bigcup_{j=1}^{2} E_{2j}(\alpha_{ij}^*)$. If $\cap_{j=1}^{2} \overline{E}_{1j}(\alpha_{1j}^*)$ is true, that is, all of the rejected hypotheses in the families of level 1 are false, then $A_{11} \supseteq$
\( T_{11}, A_{12} \supseteq T_{12} \) which implies \( e^*(A_{11}) \geq e^*(T_{11}), e^*(A_{12}) \geq e^*(T_{12}) \). And based on (B.1), we have

\[
\alpha_{2i}^* = \alpha_{2i} + \sum_{j=1}^{2} (\alpha_{1j}^* - e^*(A_{1j})) g_{1j2i} \
\leq \alpha_{2i} + \sum_{j=1}^{2} (\alpha_{1j}^* - e^*(T_{1j})) g_{1j2i}.
\]

Therefore,

\[
\left( \bigcap_{j=1}^{2} E_{1j}(\alpha_{1j}^*) \right) \cap \left( \bigcup_{j=1}^{2} E_{2j}(\alpha_{2j}^*) \right)
\subseteq \bigcup_{j=1}^{2} E_{2j} \left( \alpha_{2i} + \sum_{j=1}^{2} (\alpha_{1j}^* - e^*(T_{1j})) g_{1j2i} \right),
\]

and

\[
\Pr \left\{ \left( \bigcap_{j=1}^{2} E_{1j}(\alpha_{1j}^*) \right) \cap \left( \bigcup_{j=1}^{2} E_{2j}(\alpha_{2j}^*) \right) \right\} \leq \Pr \left\{ \bigcup_{j=1}^{2} E_{2j}(\alpha_{2j}^*) \right\} \\
\leq \Pr \left\{ \bigcup_{i=1}^{2} E_{2i} \left( \alpha_{2i} + \sum_{j=1}^{2} (\alpha_{1j}^* - e^*(T_{1j})) g_{1j2i} \right) \right\} \\
\leq \sum_{i=1}^{2} \Pr \left\{ E_{2i} \left( \alpha_{2i} + \sum_{j=1}^{2} (\alpha_{1j}^* - e^*(T_{1j})) g_{1j2i} \right) \right\} \\
\leq \sum_{i=1}^{2} \left( \alpha_{2i} + \sum_{j=1}^{2} (\alpha_{1j}^* - e^*(T_{1j})) g_{1j2i} \right) \\
= \sum_{i=1}^{2} \alpha_{2i} + \sum_{j=1}^{2} ((\alpha_{1j} - e^*(T_{1j})) \sum_{i=1}^{2} g_{1j2i} \\
\leq \sum_{i=1}^{2} \alpha_{2i} + \sum_{j=1}^{2} (\alpha_{1j} - e^*(T_{1j})) \\
= \sum_{i=1}^{2} \alpha_{2i} + \sum_{j=1}^{2} \alpha_{1j} - \sum_{j=1}^{2} e^*(T_{1j}) \\
= \alpha - \sum_{j=1}^{2} e^*(T_{1j}). \hspace{1cm} (B.4)
\]
The fourth inequality of (B.4) follows from the assumption of the $p$-values corresponding to the null hypotheses in families of $L_1$ are independent of the $p$-values corresponding to the null hypotheses in families of $L_2$. The fifth equality holds because that for any $j = 1, 2$, $\sum_{i=1}^{2} g_{1j2i} \leq 1$.

Therefore, using (B.3) and (B.4) in (B.2), we have

$$FWER \leq \sum_{j=1}^{2} e^*(T_{1j}) + \alpha - \sum_{j=1}^{2} e^*(T_{1j}) = \alpha.$$ 

Thus, the desire result is proved. \hfill \Box

### B.2 Proof of Theorem 3.2

*Proof.* Denote $FWER_n(\alpha_1, \cdots, \alpha_n)$ be the overall FWER for the $n$ levels family-based procedure with initial allocation of the critical values to the $n$ levels $\alpha_1, \cdots, \alpha_n$. We use mathematical induction method to prove that

$$FWER_n(\alpha_1, \cdots, \alpha_n) \leq \sum_{i=1}^{n} \sum_{j=1}^{l_i} \alpha_{ij} \leq \alpha. \quad (B.5)$$

If $n = 2$, through the proof of Theorem 3.1, we can get that $FWER_2(\alpha_1, \alpha_2) \leq \sum_{i=1}^{2} \sum_{j=1}^{l_i} \alpha_{ij} \leq \alpha$.

Assume that (B.5) is true when $n=k$, which is

$$FWER_k(\alpha_1, \cdots, \alpha_k) \leq \sum_{i=1}^{k} \sum_{j=1}^{l_i} \alpha_{ij} \leq \alpha.$$ 

We must prove that (B.5) is also true for $n=k+1$, i.e.,

$$FWER_{k+1}(\alpha_1, \cdots, \alpha_{k+1}) \leq \sum_{i=1}^{k+1} \alpha_i \leq \alpha.$$ 

Define the events $B_1 = \{\text{at least one true null being rejected among all the families in level 1}\}$ and $B_2 = \{\text{at least one true null being rejected among the families}}$
in all the levels except level 1}. Then we have

\[
\text{FWER}_{k+1}(\alpha_1, \ldots, \alpha_{k+1}) = \Pr \{B_1\} + \Pr \{\overline{B}_1 \cap B_2\}.
\]

Note that \(\Pr \{B_1\} \leq \sum_{j=1}^{l_1} e_{1j}(T_{1j})\) due to the definition of error rate function. Let’s consider the event \(\overline{B}_1 \cap B_2\).

After testing all families in \(L_1\), we know that the total significant level \(\sum_{j=1}^{l_1} \alpha_{1j} - \sum_{j=1}^{l_1} e^*_1(A_{1j})\) will be transferred to the families from \(L_2\) to \(L_n\). Once the families in \(L_1\) are tested, for \(i = 2, \ldots, k + 1, j = 1, \ldots, l_i\), we let the updated significant level for \(F_{ij}\) be \(\alpha^*_ij\) which is

\[
\alpha^*_ij = \alpha_{ij} + \sum_{l=1}^{l_1} (\alpha_{il} - e^*_l(A_{il}))g_{1lij},
\]

and denote \(\alpha^*_i\) the updated critical value for the \(i^{th}\) level and \(\alpha^*_1, \ldots, \alpha^*_l\) the updated local critical value for each family within level.

If \(\overline{B}_1\) is true, which means that no true null hypotheses in any families in \(L_1\) are rejected, then it implies type I error can only occurs in the \(L_2\) to \(L_{k+1}\). So we can get \(\Pr \{\overline{B}_1 \cap B_2\} = \text{FWER}_k(\alpha^*_1, \ldots, \alpha^*_{k+1})\). And also for any \(F_{ij}, j = 1, \ldots, l_1, T_{1j} \subseteq A_{1j}\) which means \(e^*(T_{1j}) \leq e^*(A_{1j})\) due to the monotonicity condition of error rate function. Thus

\[
\begin{align*}
\sum_{i=2}^{k+1} \sum_{j=1}^{l_i} \alpha^*_ij & = \sum_{i=2}^{k+1} \sum_{j=1}^{l_i} \left[ \alpha_{ij} + \sum_{l=1}^{l_1} (\alpha_{il} - e^*_l(A_{il}))g_{1lij} \right] \\
& \leq \sum_{i=2}^{k+1} \sum_{j=1}^{l_i} \alpha_{ij} + \sum_{l=1}^{l_1} \alpha_{il} - \sum_{l=1}^{l_1} e^*_l(A_{il}) \\
& \leq \sum_{i=2}^{k+1} \sum_{j=1}^{l_i} \alpha_{ij} + \sum_{l=1}^{l_1} \alpha_{il} - \sum_{l=1}^{l_1} e^*_l(T_{il}) \\
& = \sum_{i=1}^{k+1} \sum_{j=1}^{l_i} \alpha_{ij} - \sum_{j=1}^{l_1} e^*_1j(T_{1j}).
\end{align*}
\]

(B.6)
The first inequality of (B.6) holds due to the condition of transition matrix that for any fixed $k = 1, \cdots, l_1$, $\sum_{i=2}^{l_1} \sum_{j=1}^{l_i} g_{ij} \leq 1$. Therefore,

\[
\text{FWER}_{k+1}(\alpha_1, \cdots, \alpha_{k+1}) = \Pr \{B_1\} + \Pr \{\overline{B}_1 \cap B_2\} \\
\leq \sum_{j=1}^{l_1} e^*_i(T_{1j}) + \text{FWER}_k(\alpha^*_2, \cdots, \alpha^*_{k+1}) \\
\leq \sum_{j=1}^{l_1} e^*_i(T_{1j}) + \sum_{i=2}^{k+1} \sum_{j=1}^{l_i} \alpha^*_{ij} \\
\leq \sum_{j=1}^{l_1} e^*_i(T_{1j}) + \sum_{j=1}^{l_1} \alpha_{1j} - \sum_{j=1}^{l_1} e^*_i(T_{1j}) + \sum_{i=2}^{k+1} \sum_{j=1}^{l_i} \alpha_{ij} \\
= \sum_{i=1}^{k+1} \sum_{j=1}^{l_i} \alpha_{ij} \leq \alpha.
\]

This completes the induction, and show that the property is true for any positive $n$. \qed
This appendix contains the proofs of the theorems stated but not proved in Chapter 4.

C.1 Proof of Theorem 4.1

**Proof.** Let $V_k$ be the total number of false rejections among all $m$ families of hypotheses in the first $k$ stages by using Bonferroni-based gatekeeping procedure with retesting option. Denote $\text{FWER}_k$ as the FWER of this procedure in the first $k$ stages such that $\text{FWER}_k = \Pr(V_k \geq 1)$. Therefore, the FWER of this procedure is $\text{FWER} = \Pr(\bigcup_{k=1}^\infty \{V_k \geq 1\})$. Define the event $E_{i(j)}=\{$at least one true null hypothesis being rejected in $F_i$ at stage $j\}$ and $E_{i(j)}$ being the complement of $E_{i(j)}$, $i = 1, \cdots, m$. For notational convenience, denote $D_j$ as the event that at least one true null hypotheses is rejected among all $m$ families at stage $j$, that is, $D_j = \bigcup_{i=1}^m E_{i(j)}, j = 1, \cdots, k$. Then, we have

$$\text{FWER}_k = \Pr(V_k \geq 1)$$

$$= \Pr\{\bigcup_{j=1}^k D_j\}.$$ 

Therefore,

$$1 - \text{FWER}_k = \Pr(\overline{D}_k \cap \{\bigcap_{j=1}^{k-1} \overline{D}_j\})$$

$$= \Pr(\bigcap_{j=1}^{k-1} \overline{D}_j | \overline{D}_k) \Pr(\overline{D}_k)$$

$$= \Pr(\overline{D}_k).$$
where the second equality holds due to that if no true null hypotheses are rejected in any family at stage $k$, then no true nulls are rejected during previous $k - 1$ stages with probability 1.

In order to show $\text{FWER}_k \leq \alpha$, it is sufficient to show

$$1 - \Pr(\overline{D}_k) \leq \alpha.$$ 

Let $p_{ij}, i = 1, \cdots, m, j = 1, \cdots, n_i$ denote the $p$-value corresponding to the null hypothesis $H_{ij}$ in family $F_i$. Define $T_i$ the set of true null hypotheses within $F_i$ with cardinality $|T_i|, i = 1, \cdots, m$. We have

$$\Pr(\overline{D}_k) = \Pr\left(\bigcap_{i=1}^{m} E_{i(k)}\right) \geq \Pr\left\{\bigcap_{i=1}^{m} \bigcap_{H_{ij} \in T_i} \{p_{ij} \geq \frac{\alpha_{i(k)}^{*}}{n_i}\}\right\}$$

where

$$\alpha_{1(k)}^{*} = \alpha_1 + \sum_{l=2}^{m} \left(1 - \frac{|T_l|}{n_l}\right) g_{l1} \alpha_l$$ (C.1)

and for $i = 2, \cdots, m$,

$$\alpha_{i(k)}^{*} = \alpha_i + \sum_{j=1}^{i-1} \left(1 - \frac{|T_j|}{n_j}\right) g_{ji} \alpha_{j(k)} + \sum_{l=i+1}^{m} \left(1 - \frac{|T_l|}{n_l}\right) g_{li} \alpha_l.$$ (C.2)

Note that the event $\overline{D}_k$ means that no true null hypotheses are rejected among all $m$ families at stage $k$. It implies that no true null hypotheses are rejected in the first $k - 1$ stages. Thus, for any $F_j, j = 1, \cdots, m$, $|R_{j(k-1)}| \leq |R_{j(k)}| \leq n_j - |T_j|$. Therefore, comparing (C.2) with (4.4), it is easy to see that the critical value used for testing $F_i, i = 1, \cdots, m$, are $\alpha_{i(k)} \leq \alpha_{i(k)}^{*}$, respectively. Besides, by combining (C.1) and (C.2), it can be proved that $\alpha_{i(k)}^{*}$ is a constant for any $i = 1, \cdots, m$. 

114
In order to prove the FWER control, we need to use the following lemma.

**Lemma 1.** Consider a function \( f \) defined by

\[
\begin{align*}
    f(j) &= \sum_{i=1}^{j} \left[ \frac{|T_i|}{n_i} + \left(1 - \frac{|T_i|}{n_i}\right) \left(\sum_{l=j+1}^{m} g_{il}\right) \right] \alpha^*_{i(k)} \\
    &\quad + \sum_{l=j+1}^{m} \left[ \frac{|T_l|}{n_l} + \left(1 - \frac{|T_l|}{n_l}\right) \left(\sum_{i=j+1}^{m} g_{li}\right) \right] \alpha_l 
\end{align*}
\]

(C.3)

on the integers \( 2, \ldots, m-1 \). The function \( f(j) \) is non-increasing in terms of \( j \).

**Proof of Lemma 1.** To show \( f(j) \) is a non-increasing function on \( j = 2, \ldots, m-1 \), it is sufficient to show that \( f(j) \leq f(j-1) \) for any \( j = 2, \ldots, m-1 \). Therefore, we have

\[
\begin{align*}
    f(j) &= \sum_{i=1}^{j} \left[ \frac{|T_i|}{n_i} + \left(1 - \frac{|T_i|}{n_i}\right) \left(\sum_{l=j+1}^{m} g_{il}\right) \right] \alpha^*_{i(k)} \\
    &\quad + \sum_{l=j+1}^{m} \left[ \frac{|T_l|}{n_l} + \left(1 - \frac{|T_l|}{n_l}\right) \left(\sum_{i=j+1}^{m} g_{li}\right) \right] \alpha_l \\
    &= \sum_{i=1}^{j-1} \left[ \frac{|T_i|}{n_i} + \left(1 - \frac{|T_i|}{n_i}\right) \left(\sum_{l=j}^{m} g_{il}\right) \right] \alpha^*_{i(k)} \\
    &\quad + \sum_{l=j+1}^{m} \left[ \frac{|T_l|}{n_l} + \left(1 - \frac{|T_l|}{n_l}\right) \left(\sum_{i=j+1}^{m} g_{li}\right) \right] \alpha_l \\
    &\quad + \left[ \frac{|T_j|}{n_j} + \left(1 - \frac{|T_j|}{n_j}\right) \left(\sum_{i=j}^{m} g_{ji}\right) \right] \alpha_j + \sum_{l=j+1}^{m} \left(1 - \frac{|T_l|}{n_l}\right) g_{jl} \alpha_l \\
    &= \sum_{i=1}^{j-1} \left[ \frac{|T_i|}{n_i} + \left(1 - \frac{|T_i|}{n_i}\right) \left(\sum_{l=j}^{m} g_{il}\right) \right] \alpha^*_{i(k)} \\
    &\quad + \sum_{l=j}^{m} \left[ \frac{|T_l|}{n_l} + \left(1 - \frac{|T_l|}{n_l}\right) \left(\sum_{i=j}^{m} g_{li}\right) \right] \alpha_l \\
    &\quad + \left[ \frac{|T_j|}{n_j} + \left(1 - \frac{|T_j|}{n_j}\right) \left(\sum_{i=j}^{m} g_{ji}\right) \right] \alpha_j + \sum_{l=j}^{m} \left(1 - \frac{|T_l|}{n_l}\right) g_{jl} \alpha_l \\
    &= f(j-1). 
\end{align*}
\]

(C.4)
The proof of Lemma 1 is complete. 

Because of Lemma 1, we have

\[ 1 - \Pr \left( \bar{D}_k \right) \]

\[ \leq \sum_{i=1}^{m} \Pr \left\{ \bigcup_{H_{ij} \in T_i} \left\{ \hat{p}_{ij} \leq \frac{\alpha^*_{i(k)}}{n_i} \right\} \right\} \leq \sum_{i=1}^{m} \frac{|T_i|}{n_i} \alpha^*_{i(k)} \leq \sum_{i=1}^{m} \frac{|T_i|}{n_i} \alpha^*_{i(k)} \]

\[ = \sum_{i=1}^{m-1} \left( \frac{|T_i|}{n_i} + \left( \frac{1 - |T_i|}{n_i} \right) g_i \alpha^*_{i(k)} \right) \]

\[ + \sum_{l=1}^{m-1} \left( \frac{|T_l|}{n_l} + \left( \frac{1 - |T_l|}{n_l} \right) \left( \sum_{j=2}^{m} g_{ij} \right) \right) \alpha^*_{i(l)} \]

\[ = \alpha^*_{1(k)} + \sum_{l=2}^{m} \left( \frac{|T_l|}{n_l} + \left( \frac{1 - |T_l|}{n_l} \right) \left( \sum_{j=2}^{m} g_{lj} \right) \right) \alpha^*_{l} \]

\[ = \alpha_{1} + \sum_{l=2}^{m} \left( \frac{|T_l|}{n_l} + \left( \frac{1 - |T_l|}{n_l} \right) \left( \sum_{j=1}^{m} g_{ij} \right) \right) \alpha^*_{l} \]

\[ = \sum_{l=1}^{m} \alpha_{l} = \alpha \]

where (C.5) holds due to the fact that \( \alpha^*_{i(k)} \) is a constant for any \( i = 1, \ldots, m \) and the assumption that true null \( p \)-values follow \( U(0, 1) \). The inequality (C.6) holds due to Lemma 1 and the equalities (C.7) and (C.8) hold due to the transition matrix condition that for any \( i = 1, \ldots, m \), \( \sum_{j=1}^{m} g_{ij} = 1 \) and \( g_{ii} = 0 \). Therefore, we can get

\[ \text{FWER}_k = \Pr(V_k \geq 1) \leq \alpha. \]  

(C.9)

The above argument shows that for arbitrary integer \( k \), \( \Pr (V_k \geq 1) \leq \alpha \). To complete the proof, we need to show that \( \text{FWER}_\infty \leq \alpha \), i.e., when \( k \to \infty \), \( \lim_{k \to \infty} \Pr(V_k \geq 1) \leq \alpha \). Define an event \( A_k = \{ V_k \geq 1 \} \). Since \( V_k \) is non-decreasing in \( k \), \( (A_k)_{k \geq 1} \) is
an increasing sequence of events. Let $A = \lim_{k \to \infty} A_k = \bigcup_{k \geq 1} A_k$. Then we have

$$
FWER_\infty = \Pr(A) = \Pr(\lim_{k \to \infty} A_k) = \lim_{k \to \infty} \Pr(A_k) \leq \alpha. \quad \text{(C.10)}
$$

Based on (C.9) and (C.10), we have

$$
FWER = \Pr(\bigcup_{k=1}^\infty \{V_k \geq 1\}) \leq \alpha.
$$

The proof is complete.

\[ \Box \]

### C.2 Proof of Theorem 4.2

**Proof.** Let $V_k$ be the total number of false rejections among all $m$ families of hypotheses in the first $k$ stages by using two-level Bonferroni-based gatekeeping procedure with retesting option. Denote $FWER_k$ as the FWER of this procedure in the first $k$ stages so that $FWER_k = \Pr(V_k \geq 1)$. Therefore, the FWER of this procedure is $FWER = \Pr(\bigcup_{k=1}^\infty \{V_k \geq 1\})$. Define the event $E_{ij(k)} = \{\text{at least one true null hypothesis being rejected in } F_{ij} \text{ at stage } k\}$ and $\overline{E}_{ij(k)}$ being the complement of $E_{ij(k)}$, $i = 1, 2, j = 1, \ldots, m_i$. For notational convenience, let $D_k$ denote as the event that at least one true null hypotheses among $F = \bigcup_{i=1}^2 \bigcup_{j=1}^{m_i} F_{ij}$ is rejected at stage $k$, that is, $D_k = \bigcup_{i=1}^2 \bigcup_{j=1}^{m_i} E_{ij(k)}$. Then, we have

$$
FWER_k = \Pr(V_k \geq 1)
= \Pr\left(\bigcup_{t=1}^k D_t\right).
$$

Therefore,

$$
1 - FWER_k = \Pr(\overline{D}_k \cap \{\bigcap_{t=1}^{k-1} \overline{D}_t\})
= \Pr(\bigcap_{t=1}^{k-1} \overline{D}_t | \overline{D}_k) \Pr(\overline{D}_k)
= \Pr(\overline{D}_k).
$$
where the second equality holds due to the fact that if no true null hypotheses are rejected in each family at stage $k$, then no true nulls are rejected in the first $k-1$ stages with probability 1.

In order to show $\text{FWER}_k \leq \alpha$, it is sufficient to show

$$1 - \Pr(\overline{D}_k) \leq \alpha.$$

Let $p_{ij}$ represent the $p$-value corresponding to the null hypothesis $H_{ijs}$ in family $F_{ij}$, $i = 1, 2, j = 1, \cdots, m_i$ and $s = 1, \cdots, n_{ij}$. Define $T_{ij}$ be the set of true null hypotheses within $F_{ij}$ with cardinality $|T_{ij}|$. We have

$$\Pr(\overline{D}_k) = \Pr\left(\bigcap_{i=1}^{m_i} \bigcap_{j=1}^{n_{ij}} E_{ij(k)}\right) \geq \Pr\left(\left\{\bigcap_{j=1}^{m_i} \bigcap_{H_{1js} \in T_{1j}} \left\{p_{1js} \geq \frac{\alpha_{1j(k)}}{n_{1j}}\right\}\right\} \bigcap \left\{\bigcap_{l=1}^{m_2} \bigcap_{H_{2ls} \in T_{2l}} \left\{p_{2ls} \geq \frac{\alpha_{2l(k)}}{n_{2l}}\right\}\right\}\right)$$

where

$$\alpha_{1j(k)}^* = \alpha_{1j} + \sum_{l=1}^{m_2} \left(1 - \frac{|T_{2l}|}{n_{2l}}\right) g_{2l} \alpha_{2l};$$

$$\alpha_{2l(k)}^* = \alpha_{2l} + \sum_{j=1}^{m_1} \left(1 - \frac{|T_{1j}|}{n_{1j}}\right) g_{1j} \alpha_{1j(k)}^*.$$}

Note that the event $\overline{D}_k$ means that no true null hypotheses are rejected at stage $k$. It implies that no true null hypotheses are rejected at stage $k-1$. Thus, $|R_{ij(k-1)}| \leq n_{ij} - |T_{ij}|$ for any $F_{ij}$. Also, if $E_{ij(k)}$ occurs which means that no true null hypotheses are rejected in $F_{ij}$ at stage $k$. Then, it implies that $|R_{ij(k)}| \leq n_{ij} - |T_{ij}|$. It is easy to
see that the critical value used for testing \( F_{ij}, i = 1, 2 \) is \( \alpha_{ij(k)} \leq \alpha_{ij}^\ast \). Therefore,

\[
1 - \Pr (\overline{D}_k) \leq \sum_{j=1}^{m_1} \Pr \left\{ \bigcup_{H_{1js} \in T_{1j}} \left\{ p_{1js} \leq \frac{\alpha_{1j}^*}{n_{1j}} \right\} \right\} + \sum_{l=1}^{m_2} \Pr \left\{ \bigcup_{H_{2ls} \in T_{2l}} \left\{ p_{2ls} \leq \frac{\alpha_{2l}^*}{n_{2l}} \right\} \right\}
\]

\[
\leq \sum_{j=1}^{m_1} \frac{|T_{1j}|}{n_{1j}} \alpha_{1j}^* + \sum_{l=1}^{m_2} \frac{|T_{2l}|}{n_{2l}} \left[ \alpha_{2l} + \sum_{j=1}^{m_1} \left( 1 - \frac{|T_{1j}|}{n_{1j}} \right) g_{ij2l} \alpha_{1j}^* \right]
\]

\[
= \sum_{j=1}^{m_1} \alpha_{1j} + \sum_{l=1}^{m_2} \left( 1 - \frac{|T_{2l}|}{n_{2l}} \right) g_{2l1j} \alpha_{2l} + \sum_{l=1}^{m_2} \frac{|T_{2l}|}{n_{2l}} \alpha_{2l}
\]

\[
= \sum_{j=1}^{m_1} \alpha_{1j} + \sum_{l=1}^{m_2} \alpha_{2l} = \alpha.
\]

The second inequality holds due to the assumption that true null p-values follow \( U(0, 1) \) and the third and fourth equalities hold due to the conditions of transition coefficient set. Therefore, we can get

\[
\text{FWER}_k = \Pr (V_k \geq 1) \leq \alpha.
\]  \hspace{1cm} (C.11)

The above argument shows that for arbitrary integer \( k \), \( \Pr (V_k \geq 1) \leq \alpha \). To complete the proof, we need to show that \( \text{FWER}_\infty \leq \alpha \), i.e., when \( k \to \infty \), \( \lim_{k \to \infty} \Pr (V_k \geq 1) \leq \alpha \). Define an event \( A_k = \{ V_k \geq 1 \} \). Since \( V_k \) is non-decreasing in \( k \), \( (A_k)_{k \geq 1} \) is an increasing sequence of events. Let \( A = \lim_{k \to \infty} A_k = \bigcup_{k \geq 1} A_k \). Then we have

\[
\text{FWER}_\infty = \Pr (A) = \Pr (\lim_{k \to \infty} A_k) = \lim_{k \to \infty} \Pr (A_k) \leq \alpha.
\]  \hspace{1cm} (C.12)

Based on (C.11) and (C.12), we have

\[
\text{FWER} = \Pr (\bigcup_{k=1}^{\infty} \{ V_k \geq 1 \}) \leq \alpha.
\]

The proof is complete.


