Fall 2017

Topics on multiple hypotheses testing and generalized linear model

Yalin Zhu
New Jersey Institute of Technology

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

ABSTRACT

TOPICS ON MULTIPLE HYPOTHESES TESTING AND GENERALIZED LINEAR MODEL

by
Yalin Zhu

In applications such as studying drug adverse events (AE) in clinical trials and identifying differentially expressed genes in microarray experiments, the data of the experiments usually consists of frequency counts. In the analysis of such data, researchers often face multiple hypotheses testing based on discrete test statistics. Incorporating this discrete property of the data, several stepwise procedures, which allow to use the CDF of p-values to determine the testing threshold, are proposed for controlling familiwise error rate (FWER). It is shown that the proposed procedures strongly control the FWER and are more powerful than the existing ones for discrete data. Through some simulation studies and real data examples, the proposed procedures are shown to outperform the existing procedures in terms of the FWER control and power. An R package "MHTdiscrete" and a web application are developed for implementing the proposed procedures for discrete data.

Many complex biomedical studies, such as clinical safety studies and genome-wide association studies, often involve testing multiple families of hypotheses. Most existing multiple testing methods cannot guarantee strong control of appropriate type 1 error rates suitable for such increasingly complex research questions. A novel two-stage procedure based on the recently developed idea of selective inference for clinical safety studies is introduced. In the first stage, some significant families are selected by using some family-level global test, which guarantees control of generalized familywise error rate (k-FWER) among the selected families. In the second stage, individual hypotheses are tested for each selected families by using some multiple testing procedure, which controls conditional false discovery rate (cFDR) based on the fact that the family is selected. By applying the proposed procedure to clinical safety studies, one can not only

efficiently flag the significant clinical adverse events (AEs) but also select body systems of interest (BSoI) as extra information for further research. The simulation studies show that the proposed procedure can be more reliable than alternative methods such as Mehrotra and Heyse's double FDR procedure in the setting of clinical safety. The proposed procedure for multiple families structure is implemented in the R package "MHTmult".

Categorical data arises in biomedical and healthcare experiments naturally. In many of these cases, the outcome variables of interest are the numbers of special events. At least one distinct special event category is observed, when the negative multinomial and extended negative multinomial or generalized inverse sampling scheme-based regression models are used. The new model, based on generalized inverse sampling scheme for several special events, is developed in this dissertation. This research is an adaption to the widely used multinomial logistic regression model. The resulting equations of the proposed model, corresponding to the natural log of the ratio of the expected responses, appears similar to the multinomial logistic regression. Using this expected response ratio of a category to that of the special category, the maximum likelihood estimator of the regression parameters can be computed by creating score equations and the Hessian matrix of the likelihood. The covariance matrix of estimators of the regression parameters for the new model can be estimated by inverting the Hessian matrix to develop the inference. This research also develops model diagnostics such as normality check with deviance and Pearson residuals, and likelihood based computations. The proposed model is implemented in the R package "mvlogit".

TOPICS ON MULTIPLE HYPOTHESES TESTING AND GENERALIZED LINEAR MODEL

by
Yalin Zhu

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
and Rutgers, The State University of New Jersey – Newark
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Mathematical Sciences

Department of Mathematical Sciences, NJIT
Department of Mathematics and Computer Science, Rutgers – Newark

August 2017

APPROVAL PAGE

TOPICS ON MULTIPLE HYPOTHESES TESTING AND GENERALIZED LINEAR MODEL

Yalin Zhu

Dr. Sunil Dhar, Dissertation Co-Advisor — Date
Professor, Department of Mathematical Sciences, NJIT

Dr. Wenge Guo, Dissertation Co-Advisor — Date
Associate Professor, Department of Mathematical Sciences, NJIT

Dr. Ji Meng Loh, Committee Member — Date
Associate Professor, Department of Mathematical Sciences, NJIT

Dr. Sundarraman Subramanian, Committee Member — Date
Associate Professor, Department of Mathematical Sciences, NJIT

Dr. Satrajit Roychoudhury, Committee Member — Date
Senior Director, Statistical Research and Data Science Center, Pfizer, NY

BIOGRAPHICAL SKETCH

Author: Yalin Zhu

Degree: Doctor of Philosophy

Date: August 2017

Undergraduate and Graduate Education:

- Doctor of Philosophy in Mathematical Sciences,
  New Jersey Institute of Technology, Newark, NJ, 2017
- Master of Science in Applied Statistics,
  New Jersey Institute of Technology, Newark, NJ, 2017
- Bachelor of Science in Statistics,
  Tianjin University of Finance and Economics, Tianjin, P. R. China, 2012

Major: Applied Statistics

Presentations:

Y. Zhu, "A Selective Inference-based Two-stage Procedure for Clinical Safety Studies," Joint Statistical Meeting (JSM'2017), Baltimore, MD, July 29-August 3, 2017.

Y. Zhu, "Multivariate Logistic Type Models Based on Inverse Sampling Scheme," 14th Annual Conference on Frontiers in Applied and Computational Mathematics (FACM'2017), NJIT, Newark, NJ, June 24-25, 2017.

Y. Zhu, "A Selective Inference-based Two-stage Procedure for Clinical Safety Studies," ASA Biopharmaceutical Section Nonclinical Biostatistics Conference (NCB'2017), Rutgers University, Piscataway, NJ, June 12-14, 2017.

Y. Zhu, "FWER Controlling Procedures for Discrete Data in Clinical Safety Analysis," ASA Biopharmaceutical Section Nonclinical Biostatistics Conference (NCB'2017), Rutgers University, Piscataway, NJ, June 12-14, 2017.

Y. Zhu, "Generalized Inverse Sampling Scheme-Based GLM Package," NJIT President's Forum and 2017 Innovation Day, NJIT, Newark, NJ, April 10, 2017.

Y. Zhu, "FWER Controlling Procedures for Discrete Data in Clinical Safety Analysis," 4th Annual ASA NJ Chapter/Bayer Statistics Workshop, Bayer Pharmaceuticals, Whippany, NJ, November 11, 2016.

Y. Zhu, "Statistical Designs for Phase II Oncology Clinical Trials," Biostatistics and Data Management Seminar, Regeneron Pharmaceuticals, Basking Ridge, NJ, August 23, 2016.

Y. Zhu, "Metrics and Performance Response Functions for Assessment of Resilience of Urban Infrastructure Systems," 9th GSA Annual Graduate Students Research Day, NJIT, Newark, NJ, October 31, 2013.

Publications:

Y. Zhu and W. Guo, "A Selective Inference-based Two-Stage Multiple Testing Procedure in Clinical Safety Studies," in preparation, 2017.

Y. Zhu and W. Guo, "FWER Controlling Multiple Testing Procedures for Discrete Data," in preparation, 2017.

Y. Zhu and S. Dhar, "Multivariate Logistic-Type Models Based on an Inverse Sampling Scheme," in preparation, 2017.

E. Inde, S. Zamudio, J. M. Loh, Y. Zhu, J. Woytanowski, L. Rosen, M. Liu and B. Buckley, "Exposure to Bisphenol A and Common Substitutes Among Fasting Mothers: Analysis of Fetal Cord Blood and Maternal Urine," submitted for publication, 2017.

Y. Zhu and W. Guo, "R package MHTmult: Multiple Hypotheses Testing for Multiple Families/Groups Structure," The Comprehensive R Archive Network (CRAN), 2017.

Y. Zhu and W. Guo, "R package FixSeqMTP: Fixed Sequence Multiple Testing Procedures," The Comprehensive R Archive Network (CRAN), 2017.

Y. Zhu and W. Guo, "R package MHTdiscrete: Multiple Hypotheses Testing for Discrete Data," The Comprehensive R Archive Network (CRAN), 2016.

Y. Zhu and R. Qin, "R package ph2bye: Phase II Clinical Trial Design Using Bayesian Methods," The Comprehensive R Archive Network (CRAN), 2016.

Y. Zhu, "Web Application: Multiple Testing Procedures Controlling FWER/FDR," https://allen.shinyapps.io/MTPs/, 2016. [Accessed August 25, 2017]

Y. Zhu, "Web Application: Bayesian Design for Binary Outcomes," https://allen.shinyapps.io/BayesDesign/, 2016. [Accessed August 25, 2017]

To my beloved God, for He being with me every second and giving me wisdom and guidance throughout my life. To my beloved parents, Yuzhang Zhu and Yanfen Chen, for supporting me all the way.

ACKNOWLEDGMENT

I am enormously grateful to my Co-advisors, Dr. Sunil Dhar and Dr. Wenge Guo, who guided me in the right direction not only as academic advisors but also as true well wishers and friends. It would not have been possible to complete this dissertation without their help, guidance, support and constant encouragement.

I would like to extend my gratitude to the other members of my dissertation committee, Dr. Ji Meng Loh, Dr. Sundarraman Subramanian and Dr. Satrajit Roychoudhury for their encouragement and support throughout my PhD.

I am grateful to Dr. Rui Qin, Dr. Anjana Grandhi, Dr. Zhiying Qiu and Dr. Gavin Lynch for their guidance and extremely important insights throughout my research. I would also like to thank all the other faculty, staff and PhD students in the Department of Mathematical Sciences for all the help they have been giving me during my studies.

Last, but not the least, I take this opportunity to mention the names of my parents, Mr. Yuzhang Zhu and Mrs. Yanfen Chen, in an effort to convey my gratitude to them for their unconditional love and continuous guidance, which made me achieve this goal.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

Table **Page**

LIST OF TABLES
(Continued)

Table **Page**

LIST OF TABLES
(Continued)

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Introduction

In the applications of clinical trials and genealogy study, multiple hypotheses testing is a very useful statistical tool to analyze experimental data. Simultaneously testing multiple hypotheses is often required in such applications. In single hypothesis testing, a typical error measure which needs to be controlled is called type I error rate, the probability of falsely rejecting the hypothesis while the hypothesis is true. There are several possible measures for the overall type I error rate while testing multiple hypotheses. A popular error rate is the familywise error rate (FWER), which is the probability of making at least one false rejection. It is appropriate to control the FWER when the number of hypotheses tested is small or moderate, but it is too conservative when a large number of hypotheses are tested simultaneously which is typically the case in large scale experiments like microarray or fMRI study. Benjamini and Hochberg (1995) introduced false discovery rate (FDR) as an appropriate measure to be controlled while simultaneously testing a large number of hypotheses. FDR is defined as the expected proportion of false rejections among all rejections. To control the FDR, it allows more hypotheses to be rejected while controlling the proportion of false rejections, thus opening an opportunity for the development of more powerful procedures than those using FWER as an error measure to control. For a review of multiple testing procedures controlling the FWER, see Dmitrienko et al. (2009). For a review of FDR controlling procedures, refer to Benjamini (2010). Most existing procedures are constructed for continuous data, but these procedures may be highly conservative when testing discrete data. Tarone (1990) proposed a modified Bonferroni procedure to make it more powerful for discrete data. The modification is to reduce the number of significance tests by eliminating those tests with relative large minimal

attainable p-values. But the Tarone procedure lacks α-consistency, that is, a hypothesis which is accepted at a given α level may be rejected at a lower α level (Roth, 1999). To overcome this issue, two modified Tarone procedures were developed by Hommel and Krummenauer (1998) and Roth (1999), which not only control FWER, but also have the property of α-consistency. Hommel and Krummenauer (1998) improved the Holm procedure to develop a step-down procedure for discrete data by using Tarone's idea. Roth (1999) presented a two-stage step-up procedure for improving Hochberg procedure for discrete data based on the similar Tarone's idea, but this step-up procedure lacks α-consistency. Westfall and Wolfinger (1997) suggested a resampling based approach to simulate the null distribution of minimal p-value, which uses the full set of attainable p-values for each p-value. But this method is computationally complicated and only ensure asymptotic control of the FWER. Gutman and Hochberg (2005) proposed new stepwise procedures which use the Westfall and Wolfinger's resampling algorithm and the idea of Tarone procedure.

There are many adverse events (AE) classified by body systems (BS) in clinical safety studies In clinical safety studies, there are many adverse events (AE) recorded in one clinical trial. The goal for assessing the safety of an experimental drug is to flag "reasonable" or "correct" AEs among these AE types. Most AE detecting or flagging methods do not control for overall type 1 error rates, such as FWER or FDR. Thus, similar as dealing with multiple endpoints in drug efficacy analysis, multiplicity effect should be also considered in drug safety analysis. However, the number of AEs in safety analysis is much larger than the number of endpoints in efficacy analysis for the experimental drugs. Simply applying FWER controlling procedures such as Bonferroni procedure may fail to flag more important AEs. Therefore, some FDR controlling procedures such as BH procedure can be applied to detect the signals of the AEs, since the number of AEs in clinical safety studies is usually large. Moreover, searching for significant AEs, the AE types are often classified by several body systems (BS). So the multiple-family structure should be considered for the drug safety data

analysis. Recently, some structured BH-type procedures are developed for multiple families of hypotheses (Mehrotra and Heyes, 2004; Mehrotra and Adewale, 2012; Hu et al., 2010; Benjamini and Bogomolov, 2014). However, most existing methods do not clearly separate the selection effect and multiplicity effect. To overcome this problem, selective inference by using conditional inference such as conditional type 1 error rate control, selection adjusted confidence interval is developed recently (Fithian et al., 2015; Weinstein et al., 2013; Heller et al., 2016). By using the selective inference idea, the second part of this dissertation introduces a multiple testing procedure for multiple families structure in clinical safety studies.

In the last part of this dissertation, we highlight logistic-type model while introducing generalized linear models for multi-level data. Logistic model is the most important model for categorical response data. It is used increasingly in a wide variety of applications, such as biomedical studies, social science researches, marketing, etc. An area of increasing application of logistic model is genetics. For instance, Henshall and Goddard (1999) used logistic regression to estimate quantitative trait loci effects, modeling the probability that an offspring inherits an allele of one type instead of another type as a function of phenotypic values on various traits for that offspring. Levinson et al. (2000) used logistic regression for analysis of the genotype data of affected sibling pairs (ASPs) and their parents from several research centers. The model studied the probability that ASPs have identity-by-descent allele sharing and tested its heterogeneity among the centers. In clinical trial experiments, the data are usually collected as count data for several group/categories, where some categories are classified as severe and rare diseases (or called "stages"), such as the data in Desmet et al. (1994). We use this information in building our model, for more details, please refer to Chapter 4.

1.2 Basic Concepts of Multiple Hypotheses Testing

Consider simultaneously testing m hypotheses $H_1, \ldots, H_m$, based on the corresponding p-values $P_1, \ldots, P_m$. Let m_0 denote the number of true null hypotheses and $m_1 = m - m_0$ denote the number of false null hypotheses. Let I_0 denote the set of indices of true null hypotheses.

Let V denote the number of falsely rejected hypotheses, S denote the number of correctly rejected hypotheses and R denote the total number of hypotheses rejected, thus $R = S + V$. Table 1.1 summarizes the notations for all possible outcomes.

Table 1.1 Summary of the Outcomes while Simultaneously Testing m Hypotheses

	Number of Hypotheses Not Rejected	**Number of Hypotheses rejected**	**Total Number**
True Null Hypotheses	$m_0 - V$	V	m_0
False Null Hypotheses	$m_1 - S$	S	m_1
Total	$m - R$	R	m

Note that m and m_0 are fixed but m_0 is usually unknown, R, V and S are random but only R is observable, and V and S are unobservable.

When dealing with multiple testing problems, it is essential to choose an appropriate overall measure of error rate and power measure.

1.2.1 Error Rate Definition

The overall error rate measure for multiple testing is not unique. Several commonly used error rates are defined as follows.

- *Per family error rate* (PFER) is the expected number of incorrectly rejected hypotheses, which is given by

$$PFER = E\{V\}.$$

- *Comparisonwise error rate* (CWER) is the proportion of falsely rejected hypotheses among all tested hypotheses, which is given by

$$CWER = \frac{E\{V\}}{m}.$$

- *Familywise error rate* (FWER) is the probability of making at least one false rejection, which is given by

$$FWER = P\{V > 0\}.$$

- *False discovery rate* (FDR) is the proportion of falsely rejected hypotheses among all rejected hypotheses and is formally defined by Benjamini and Hochberg (1995) as

$$FDR = E\left\{\frac{V}{R \vee 1}\right\} = E\left\{\frac{V}{R} I(R > 0)\right\},$$

where $R \vee 1 = \max\{R, 1\}$ and $I(\cdot)$ is indicator function. Note that when all null hypotheses are true, that is $m_0 = m$, FDR reduces to FWER. The relationship among the above four error rate measures is $CWER \leq FDR \leq FWER \leq PFER$.

1.2.2 Definition of Power

The power of a single test is defined as the probability of rejecting a false null hypothesis. There are several types of power measure when testing multiple hypotheses simultaneously, so it is important to use an appropriate power measure to evaluate performance of a MTP. Several commonly used concepts of power are described below:

- *Minimal power* is the probability of rejecting at least one false null hypothesis, which is given by

$$\text{Minimal Power} = Pr(S > 0).$$

- *Complete power* is the probability of rejecting all false null hypotheses, which is given by

$$\text{Complete Power} = Pr(S = m_1).$$

- *Average power* is the expected proportion of rejected false null hypotheses among all false null hypotheses, which is given by

$$\text{Average Power} = \frac{E\{S\}}{m_1}.$$

- Another concept of power is from the *false non-discovery rate (FNR)*, which is given by

$$1 - FNR = 1 - E\left\{\frac{m_1 - S}{(m - R) \vee 1}\right\}.$$

Theoretically, the definition of "universally more powerful" can be used to compare two procedures.

Definition 1.1. *If procedure A rejects all hypotheses rejected by procedure B for every possible configuration, then we can say procedure A is universally more powerful than procedure B.*

1.2.3 Strong Control and Weak Control

Strong control is to control type I error rate under any combination of true and false null hypotheses, while weak control is to control type I error rate only when all null hypotheses are true. Generally, strong control of type I error rate is desired, since the combination of true and false hypotheses in the actual setting is unknown.

In applications of clinical trials, strong control of the FWER for the primary objects is mandated by regulators. For example, in drug safety studies, the FWER control is needed for the adverse events related to the trial drugs. For all other adverse events, it is reasonable to control the FWER or the FDR.

1.2.4 Assumptions of p-values

- The p-value is calculated from given test statistic, thus the distribution of p-value could be continuous or discrete based on the distribution of the test statistic.

  Assumption 1.1. *True null p-values are stochastically greater than or equal to the* $U[0,1]$ *distribution, that is,*

  $$Pr(P_i \leq u) \leq u, \quad \text{for } i \in I_0 \text{ and } u \in [0,1]. \tag{1.2.1}$$

  We should note that the equality does not always hold in the assumption. For instance, for finite discrete p-values, when u takes the values except attainable p-values, "$\leq$" in the above assumption becomes "$<$". Only when u takes the value of the attainable p-values, $Pr(P_i \leq u) = u$. In Chapter 2, we will use this property of discrete null p-values to develop some more powerful MTPs.

- Another assumption is the joint dependence structures of the p-values for multiple hypotheses. Several dependence structures while developing MTPs are used including: independence, block dependence (Storey, 2003; Guo and Sarkar, 2012), positive regression dependence on subset (PRDS) (Benjamini and Yekutieli, 2001; Sarkar 2002), arbitrary dependence, which allows any dependence structure including previous ones. The PRDS property is defined as follows.

  Assumption 1.2. *A set of p-values* $\{P_1 \dots P_m\}$ *is said to be PRDS, if for any non-decreasing function of the p-values* ϕ, $E\{\phi(P_1, \dots, P_m) | P_i \leq p\}$ *is non-decreasing in p for each true null hypothesis* H_i.

1.2.5 Distributions of the p-values for Discrete Data

We will now look into several typical discrete p-value distributions, which include the Binomial Distribution, Hypergeometric Distribution.

Binomial distribution Consider testing a single hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta < \theta_0$. Suppose that under H_0, the test statistic $X \sim Bin(n, \theta_0)$. Then the attainable p-value for this test can be calculated by $p_i = Pr(X \leq i|H_0)$ for $i = 0, 1, \ldots, n$. If the probability mass function (PMF) of P_i is $Pr(P = p_i)$, then the cumulative distribution function (CDF) of the P_i is $F(p) = Pr(P \leq p) = \sum_{p_i \leq p} Pr(P = p_i)$. For example, suppose $X \sim Bin(5, 0.3)$ under H_0. Then the set of the attainable p-values are $\{0.16807, 0.52822, 0.83692, 0.96922, 0.99757, 1\}$. The PMF and CDF of the p-value can be given as

$$
\begin{aligned}
Pr(P = 0.16807) = Pr(X = 0) &= 0.16807, \\
Pr(P = 0.52822) = Pr(X = 1) &= 0.36015, \\
Pr(P = 0.83692) = Pr(X = 2) &= 0.3087, \\
Pr(P = 0.96922) = Pr(X = 3) &= 0.1323, \\
Pr(P = 0.99757) = Pr(X = 4) &= 0.02835, \\
Pr(P = 1) = Pr(X = 5) &= 0.00243
\end{aligned}
$$

and

$$F(p) = \begin{cases} 0, & 0 \leq p < 0.16807 \\ 0.16807, & 0.16807 \leq p < 0.52822 \\ 0.52822, & 0.52822 \leq p < 0.83692 \\ 0.83692, & 0.83692 \leq p < 0.96922 \\ 0.96922, & 0.96922 \leq p < 0.99757 \\ 0.99757, & 0.99757 \leq p < 1 \\ 1, & p = 1. \end{cases} \tag{1.2.2}$$

The test statistic in binomial exact test (BET) follows binomial distribution.

Hypergeometric distribution Another popular discrete distribution is hypergeometric distribution $T_i \sim Hypergeometric(x_{1i}, n_{1i}, x_{\cdot i}, n_{\cdot i})$, which describes the probability of x_{1i} successes in n_{1i} draws, without replacement, from a finite population of size $n_{\cdot i}$ that contains exactly $x_{\cdot i}$ successes, wherein each draw is either a success or a failure.

For instance, let us consider the test statistics of Westfall and Wolfinger (1997) in Table 1. In that example,

$$T_1 \sim Hypergeometric(5, 48, 5, 98),$$
$$T_2 \sim Hypergeometric(3, 48, 7, 98),$$
$$T_3 \sim Hypergeometric(4, 48, 4, 98),$$
$$T_4 \sim Hypergeometric(4, 48, 10, 98).$$

Note that each p-value is matched with corresponding test statistic, then we can find PMF of the corresponding p-value P_i for each given T_i. For example,

$$Pr(P_1 = 0.02521) = Pr(T_1 = 5) = \frac{\binom{48}{5}\binom{50}{0}}{\binom{98}{5}} = 0.02521,$$

$$Pr(P_1 = 0.16848) = Pr(T_1 = 4) = \frac{\binom{48}{4}\binom{50}{1}}{\binom{98}{5}} = 0.14327.$$

Therefore, we can get CDF's F_i of P_i, $i = 1, \ldots, 4$, as follows:

$$F_1(p) = \begin{cases} 0, & 0 \le p < 0.02521 \\ 0.02521, & 0.02521 \le p < 0.16848 \\ 0.16848, & 0.16848 \le p < 0.48047 \\ 0.48047, & 0.48047 \le p < 0.80602 \\ 0.80602, & 0.80602 \le p < 0.96880 \\ 0.96880, & 0.96880 \le p < 1 \\ 1, & p = 1, \end{cases} \tag{1.2.3}$$

$$F_2(p) = \begin{cases} 0, & 0 \le c < 0.00532 \\ 0.00532, & 0.00532 \le p < 0.04967 \\ 0.04967, & 0.04967 \le p < 0.20129 \\ 0.20129, & 0.20129 \le p < 0.47697 \\ \ldots, & \ldots \\ 0.99278, & 0.99278 \le p < 1 \\ 1, & p = 1, \end{cases} \tag{1.2.4}$$

$$F_3(p) = \begin{cases} 0, & 0 \leq p < 0.05387 \\ 0.05387, & 0.05387 \leq p < 0.29327 \\ 0.29327, & 0.29327 \leq p < 0.67580 \\ 0.67580, & 0.67580 \leq p < 0.93625 \\ 0.93625, & 0.93625 \leq p < 1 \\ 1, & p = 1, \end{cases} \tag{1.2.5}$$

and

$$F_4(p) = \begin{cases} 0, & 0 \leq p < 0.00047 \\ 0.00047, & 0.00047 \leq p < 0.00645 \\ 0.00645, & 0.00645 \leq p < 0.03946 \\ 0.03946, & 0.03946 \leq p < 0.14250 \\ \ldots, & \ldots \\ 0.99927, & 0.99927 \leq p < 1 \\ 1. & c = 1. \end{cases} \tag{1.2.6}$$

Figure 1.1 shows the CDF's of true null p-values for these four tests.

Fisher's exact test (FET) is usually used to test association between two variables of interest for a 2×2 contingency table. The test statistic in FET follows hypergeometric distribution.

Remark 1.1. *χ^2 test and Fisher exact test are two popular approaches used for analyzing Adverse Events (AEs) data. Fisher's exact test is desired when the expected draws (x_{1i}) is small.*

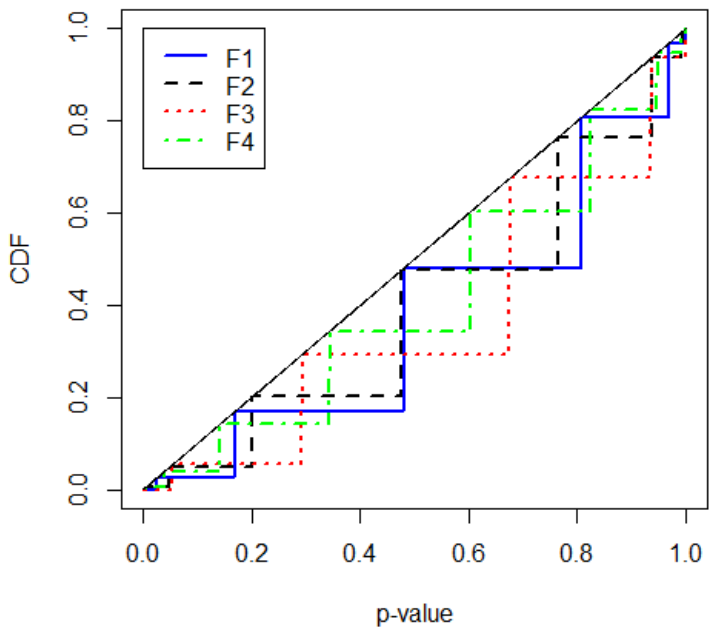

Figure 1.1 The values of the p-value P_i versus the values of the CDF F_i for $i = 1, \ldots, 4$.

1.2.6 α-consistency

In hypothesis testing, α-consistency is a type of desired statistical property in terms of the significant level α, which is defined as follow:

Definition 1.2. *A hypothesis that is rejected at a given α level must be rejected at a higher α level. This property is called α-consistency.*

For single hypothesis testing, it is trivial that this property is satisfied. For multiple hypotheses testing, not all procedures have this property. Here α-consistency means when α becomes larger, the set of rejections determined by the MTP will not become smaller.

1.2.7 p-value Monotonicity

Another favorable property of a multiple testing procedure is monotonicity in terms of p-values, which can be defined as follow:

Definition 1.3. *If one or more p-values are made smaller, then at least the same or even more hypotheses would be rejected for the same procedure. We say such a procedure is p-value monotone.*

It is easy to see that the property of p-value monotone is always satisfied by stepwise procedures. It is an essential requirement for multiple testing procedures.

1.2.8 Adjusted p-value

It is very useful to calculate *adjusted p-value* in multiple testing, since adjusted p-values capture the degree of multiplicity adjustment, so that we can make decisions of rejection and acceptance as in single hypothesis by comparing the adjusted p-values with the given significance level. Decision rules based on adjusted p-values are equivalent to ones based on original procedures.

A general definition of an adjusted p-value is given in Westfall and Young (1993): the adjusted p-value for a hypothesis is the smallest significance level at which one would reject the hypothesis using the given multiple testing procedure. Let $\tilde{P}_i$ denote the adjusted p-value corresponding to H_i, which is given by

$$\tilde{P}_i = \inf\{\alpha : H_i \text{ is rejected using the given procedure at level } \alpha\}. \tag{1.2.7}$$

If this procedure controls FWER/FDR at level α, then

$$\tilde{P}_i = \inf\{\alpha : H_i \text{ is rejected when FWER/FDR is controlled at level } \alpha\}. \tag{1.2.8}$$

So we can make the decision based on adjusted p-values: reject H_i if $\tilde{P}_i \leq \alpha$. This calculation can help researchers make decisions easily and fast.

1.3 Multiple Testing Procedures (MTPs)

Several multiple testing procedures have been developed in the literature for various scenario. They can be broadly classified as three main types: p-value based MTP, parametric MTP and resampling based MTP.

- *p-value based MTP*: These procedures do not make any assumptions about the joint distribution of the test statistics and only rely on the univariate p-values. Typical examples are Bonferroni (1936) procedure, Holm (1979) procedure, BH (1995) procedure.

- *parametric MTP*: These procedures make specific assumptions about the distribution of the test statistics. For instance, the joint distribution might be a multivariate normal or a multivariate t-distribution. A typical procedure is Dunnett procedure (1955).

- *resampling based MTP*: These procedures use resampling techniques like bootstrap, permutation, etc., that make fewer assumptions about the data-generating process while still exploiting the dependence structure of the underlying test statistics in multiple testing procedures, see Westfall and Young (1993).

1.3.1 Multiple Testing Procedures based on p-values

The first two parts of this dissertation mainly focuses on p-value based MTPs. Commonly, the p-value based MTPs test hypotheses step by step. According to the order of p-values, the MTPs can be classified as single-step, step-down, and step-up procedures.

Let $P_{(1)} \leq \cdots \leq P_{(m)}$ be the ordered p-values and $H_{(1)}, \ldots, H_{(m)}$ be the corresponding hypotheses. By using a sequence of non-decreasing critical constants $\alpha_1 \leq \cdots \leq \alpha_m$, the stepwise MTPs are described as follows:

- *Single-step* procedure compares p-values with the same critical constant c, that is, reject H_i if $P_i \leq c$ for $i = 1, \ldots, m$. A widely used single-step procedure is Bonferroni procedure, for which the critical constant is defined by $c = \dfrac{\alpha}{m}$.

- *Step-down* procedure starts with the most significant hypothesis $H_{(1)}$ corresponding to the smallest p-value $P_{(1)}$, If $P_{(1)} > \alpha_1$, accept all m hypotheses; otherwise, keep rejecting hypotheses until an acceptance is observed. The

rest hypotheses are accepted automatically. That is, the procedure rejects $H_{(1)}, \ldots, H_{(r)}$, accept $H_{(r+1)}, \ldots, H_{(m)}$, where r is the largest index satisfying $P_{(1)} \leq \alpha_1, \ldots, P_{(r)} \leq \alpha_r$. A typical example is Holm procedure (Holm, 1979), which is a step-down procedure with the critical constant $\alpha_i = \dfrac{\alpha}{m - i + 1}$.

- *Step-up* procedure starts with the least significant hypothesis $H_{(m)}$ corresponding to the largest p-value $P_{(m)}$. If $P_{(m)} \leq \alpha_m$, reject all m hypotheses; otherwise, keep accepting hypotheses until a rejection is observed. The rest hypotheses are rejected automatically. That is, the procedure rejects $H_{(1)}, \ldots, H_{(r)}$, accept $H_{(r+1)}, \ldots, H_{(m)}$, where r is the largest index satisfying $P_{(r)} \leq \alpha_r$.

  A typical example is Hochberg procedure (1988), which is a step-up procedure with the same critical constants as Holm procedure.

1.3.2 FWER Controlling Procedures

A well known FWER controlling procedure is the Bonferroni procedure (1936), which is a single-step procedure with the critical constant $c = \dfrac{\alpha}{m}$. A more powerful single-step procedure is the Sidak procedure (1967), which has the critical constant $1 - (1 - \alpha)^{\frac{1}{m}}$. The critical constant is slightly larger than that of the Bonferroni procedure. Holm (1979) developed a step-down procedure we have defined in last subsection with the critical constant $\alpha_i = \dfrac{\alpha}{m - i + 1}$, which strongly controls the FWER under arbitrary dependence. Hochberg (1988) proposed a step-up procedure with the same critical value as the Holm procedure, which controls the FWER under the PRDS property.

Taking discrete property of statistics into account, Tarone (1990) introduced a modified Bonferroni procedure by using the smallest attainable p-values to eliminate the non-significant tests. The critical constant of this modified procedure is larger than that of Bonferroni procedure, which implies this procedure is more powerful than Bonferroni procedure.

Procedure 1.1 (Tarone procedure)**.** *Let p_i^* be the smallest attainable p-values for H_i, $M_\alpha(k) = \sum_{i=1}^{m} I\{p_i^* < \alpha/k\} \leq m$. Define $K_\alpha = \min\{1 \leq k \leq m : M_\alpha(k) \leq k\}$ and $R_{K_\alpha} = \{i : p_i^* < \frac{\alpha}{K_\alpha}\}$. Then reject H_i if $i \in R_{K_\alpha}$ and $P_i < \alpha/K_\alpha$.*

Remark 1.2. *Note that in the definition of Procedure 1.1, since $P_i \geq p_i^*$, $P_i < \alpha/K_\alpha$ implies $i \in R_{K_\alpha} = \{i : p_i^* < \frac{\alpha}{K_\alpha}\}$. That means rejecting H_i only needs $P_i < \alpha/K_\alpha$.*

Note that for continuous p-values $M_\alpha(k) = m$ for $k = 1, \ldots, m$, so $K_\alpha = m$ and Procedure 1.1 reduces to Bonferroni procedure.

Unfortunately, the Tarone procedure lacks α-consistency. For example, suppose we simultaneously test two hypotheses H_1 and H_2 with the corresponding actual p-values being $P_1 = 0.07$ and $P_2 = 0.1$. The smallest attainable p-values for H_1 and H_2 are $p_1^* = 0.06$ and $p_2^* = 0.08$, respectively. When $\alpha = 0.1$, $M_\alpha(1) = 2$ and $M_\alpha(2) = 0$, so $K_\alpha = 2$. Thus, no any hypothesis is rejected since P_1 and P_2 are larger than $\alpha/K_\alpha = 0.05$. However, when $\alpha = 0.075$, $M_\alpha(1) = 1$ and $M_\alpha(2) = 0$, so $K_\alpha = 1$ and $\alpha/K_\alpha = 0.075$. Therefore H_1 is rejected since $P_1 = 0.07 < 0.075$. Thus, when α becomes smaller, more hypotheses can be rejected. That is, the Tarone procedure does not satisfy the property of α-consistency.

To overcome the issue of lacking α-consistency, Hommel and Krummenauer (1998) developed a modified Tarone procedure, which is proved satisfying the property of α-consistency.

Procedure 1.2 (T^*)**.** *Suppose p_i^* is the smallest attainable p-value for H_i and let $\gamma \in (0, \alpha]$ and $M_\gamma(k) = \sum_{i=1}^{m} I\{p_i^* < \gamma/k\}$. Define $K_\gamma = \min\{1 \leq k \leq m : M_\gamma(k) \leq k\}$, then reject H_i if there exists a γ, such that $P_i \leq \gamma/K_\gamma$*

In Procedure 1.1, k and K_α can only take integer values from 1 to m, Roth (1999) suggested another modified Tarone procedure allowing k and K_α to take fractional values. The critical constant α/X_α introduced in the following procedure will be continuous and monotone. Thus it retains α-consistency.

Procedure 1.3. *Suppose p_i^* be the smallest attainable p-values for H_i. For $1 \leq x \leq m$, let $M_\alpha(x) = \sum_{i=1}^{m} I\{p_i^* < \frac{\alpha}{x}\}$. Define $X_\alpha = \inf\{1 \leq x \leq m : M_\alpha(x) \leq x\}$, then reject H_i if $P_i < \alpha/X_\alpha$.*

Roth (1999) showed that Procedure 1.3 is equivalent to Procedure 1.2 and these two procedures are universally more powerful than Tarone procedure.

Hommel and Krummenauer (1998) also proposed a step-down procedure for discrete data by incorporating Tarone's idea into Holm procedure.

Procedure 1.4 (TH^*)**.**

1. *Set $I = \{1, \ldots, m\}$.*
2. *For $k = 1, \ldots, |I|$, define $M_I(k) = \#\{i \in I : p_i^* \leq \alpha/k\}$ as the number of hypotheses with indices in I that can be rejected at level α/k. Let $K_I(\alpha) = \min\{k = 1, \ldots, |I| : M_I(k) \leq k\}$ and define $b_I(\gamma) = \dfrac{\gamma}{K_I(\gamma)}$.*
3. *For $i \in I$, reject H_i if and only if $P_i \leq b_I(\gamma)$ for some $0 < \gamma \leq \alpha$.*
4. *Let J be the index set of hypotheses rejected in step 3.*
5. *If J is empty then stop, otherwise set $I = I - J$ and return to step 2.*

Note that if the test statistics is continuous, then Procedure 1.4 reduces to Holm procedure.

Roth (1999) also introduced a two stage step-up procedure based on Hochberg procedure (1998).

Procedure 1.5. *Roth's step-up procedure is consisted of the following two stages:*

- *Stage 1:*
  1. *Accept all hypotheses outside of $R_1 = \{H_i : p_i^* < \alpha\}$.*
  2. *For the $M_\alpha(1)$ hypotheses in the set R_1, order their available p-values from highest to lowest as $P_{1,(1)} \geq \cdots \geq P_{1,(M_\alpha(1))}$ with corresponding hypotheses $H_{1,(1)}, \ldots, H_{1,(M_\alpha(1))}$.*
  3. *Define $r = \min\{j : P_{1,(j)} < \frac{\alpha}{j}$ and $H_{1,(j)} \in R_1\}$.*

4. *Reject all of the* $H_i \in R_1$, *such that* $P_i < \frac{\alpha}{r}$.

- *Stage 2:*

1. *Consider only the hypotheses in* R_K,, *order their available p-values from highest to lowest by* $P_{K,(1)} \geq \cdots \geq P_{K,(M_\alpha(K))}$. *If* $M_\alpha(K) < K$, *then let* $P_{K,(i)} = 0$ *for* $i = M_\alpha(K)+1, \ldots, K$, *where* K *is the same as* K_α *defined in Procedure 1.1.*
2. *For* $j = 1, \ldots, K$, *define* $P_j^* = \max\{\{P_{j,(j)}\} \bigcup \{P_i : H_i \in R_j - R_K\}\}$.
3. *Define* $r' = \min\{j : P_j^* < \frac{\alpha}{j}\}$.
4. *Reject* H_i *if* $P_i < \frac{\alpha}{r'}$.

Then this procedure rejects H_i *if it was rejected in stage 1 or 2.*

Adjusted p-value for several existing procedures

Based on (1.2.8), adjusted p-value for the Bonferroni procedure for each hypothesis H_i can be obtained by

$$\tilde{P}_{i,Bonf} = \min\{1, mP_i\}, \quad \text{where } i = 1, \ldots, m.$$

The adjusted p-value for the Sidak procedure is given by

$$\tilde{P}_{i,Sidak} = 1 - (1 - P_i)^m, \quad \text{where } i = 1, \ldots, m.$$

The adjusted p-values for Procedure 1.2 is defined by Hommel and Krummernauer (1998) as follow.

Proposition 1.1. *Adjusted p-value for Procedure 1.2* (T^*) *Order the minimal attainable p-values* $p_1^* \leq \cdots \leq p_m^*$. *For each* P_i, *determine* $q(P_i)$, *such that* $p_{q(P_i)}^* \leq P_i < p_{q(P_i)+1}^*$, *the adjusted p-value is*

$$\tilde{P}_{i,T^*} = \min\{1, q(P_i) \cdot P_i\}, \quad \textit{for } i = 1, \ldots, m. \tag{1.3.1}$$

The adjusted p-value for Holm procedure (1979) can be obtained as follow:

Proposition 1.2. *The adjusted p-value of Holm procedure for each hypothesis $H_{(i)}$ is defined by*

$$\tilde{P}_{(i),Holm} = \begin{cases} \min\left\{1,\ mP_{(1)}\right\}, & i = 1 \\ \max\left\{\tilde{P}_{(i-1),Holm}, \min\left\{1,\ (m-i+1)P_{(i)}\right\}\right\}, & i = 2,\ldots,m \end{cases}$$

Hommel and Krummenauer (1998) obtained an algorithm for computing the adjusted p-values for Tarone-Holm Procedure 1.4 as follow:

Proposition 1.3. *Adjusted p-value for Procedure 1.4 (TH^*)*

1. *Set $j = 1$. Let indices $i_1, \ldots, i_m$ according to ordered p-values $P_{(1)} < P_{(2)} < \cdots < P_{(m)}$. Determine $q = q(P_{(1)})$ such that $p^*_{i_q} < P_{(1)} < p^*_{i_{q+1}}$. Set*

$$\tilde{P}_{(1),TH^*} = \min\{1, q(P_{(1)}) \cdot P_{(1)}\}.$$

2. *Set $j = j + 1$,*

3. *Set $I = \{(j), (j+1), \ldots, (m)\} = \{i_1, ..., i_t\}$ with $t = m - j + 1$ and $i_1 < ... < i_t$. Determine $q = q(P_{(j)})$ such that $p^*_{i_q} < P_{(j)} < p^*_{i_{q+1}}$ If $P_{(j)} > \alpha_{i^*_t}$, choose $q(P_{(j)}) = t$.*

4. *Compute*

$$\tilde{P}_{(j),TH^*} = \max\{\tilde{P}_{(j-1),TH^*}, \min\{1, q(P_{(j)}) \cdot P_{(j)}\}\}$$

5. *If $j = m$, stop; otherwise go back to step 2.*

The adjusted p-value of Hochberg procedure (1988) is derived as follow:

Proposition 1.4. *The adjusted p-value of Hochberg procedure for each hypothesis $H_{(i)}$ is defined by*

$$\tilde{P}_{(i),Hochberg} = \begin{cases} P_{(m)}, & i = m \\ \min\left\{\tilde{P}_{(i+1),Hochberg}, (m-i+1)P_{(i)}\right\}, & i = m-1,\ldots,1 \end{cases}$$

In addition, Gutman and Hochberg (2007) proposed single-step and stepwise procedures by using the Westfall and Wolfinger procedure on the set R_K defined in Procedure 1.1. Kulinskaya and Lewin (2009) proposed a fuzzy Bonferroni procedure

based on the idea of randomized test, but the interpretation of the results is not very straightforward.

1.3.3 FDR Controlling Procedures

In some experimental settings such as DNA microarray experiments, genome-wide association studies (GWAS), functional Magnetic Resonance Imaging (fMRI) experiments and adverse events detection in clinical trials, there are a large number of hypotheses. Familywise Error Rate (FWER) controlling procedures are quite conservative for such testing problems. Benjamini and Hochberg (1995) introduced the False Discovery Rate (FDR) as an alternative error measure to the FWER. They also introduced BH procedure for controlling FDR, which is a simple step-up procedure with the critical constant of $\alpha_i = \frac{i}{m}\alpha$. Benjamini and Yekutieli (2001) show that BH procedure controls the FDR under PRDS condition. They also introduced BY procedure, which is another step-up procedure with the critical constant $\alpha_i = \frac{i}{mC_m}\alpha$, where $C_m = \sum_{i=1}^{m} 1/i$. BY procedure controls FDR under arbitrary dependence. Benjamini and Liu (1999) and Romano and Shaikh (2006) proposed two different step-down procedures which can control FDR under certain conditions. Storey (2002, 2004) introduced an estimation approach to FDR that is the opposite of stepwise methods. In the stepwise methods, the rejection region (critical constants) is determined based on the fixed FDR level, but Storey's approach is to fix the rejection region and estimate the FDR of the rejection region. For some other methods, see Sarkar (2008) and Benjamini (2010).

Adjusted p-value for FDR controlling procedure

The adjusted p-value of BH procedure is derived as follow:

Proposition 1.5. *The adjusted p-value of BH procedure for each hypothesis $H_{(i)}$ is defined by*

$$\tilde{P}_{(i),BH} = \begin{cases} P_{(m)}, & i = m \\ \min\left\{\tilde{P}_{(i+1),BH}, \frac{m}{i}P_{(i)}\right\}, & i = m-1, \ldots, 1 \end{cases}$$

The BH procedure can be used for developing conditional FDR controlling procedure in Chapter 3.

1.4 Research Motivation and Dissertation Outline

In this dissertation, we focus on developing some new methods for analyzing biomedical or clinical data. In Chapter 2, several stepwise multiple testing procedures are proposed for real data applications, that take the discreteness of the test statistics into account, and control FWER as required by the problem of interest. In Chapter 3, by exploiting selective inference idea, one class of two-stage multiple testing procedures are developed for controlling type 1 error rates for different levels based on the multiple families structure. The proposed procedure can efficiently select body system of interest and flag adverse events in clinical safety studies. In Chapter 4, a logistic-type model considering an inverse sampling scheme is established for modeling categorical data, which shares common covariates in each sample. In the following, we discuss the motivation behind the research.

In clinical trials, discrete data often arise and FWER control is commonly required while testing multiple hypotheses. In the literature, most FWER controlling procedures are developed for continuous data. By fully exploiting the discrete information, we can generally develop more powerful procedures than the usual ones. Previous researches on FWER control procedures for discrete data are either based on the partial information of p-values (minimal attainable p-value), or resampling and randomization methods which needs intensive computation but only ensured asymptotic control of the FWER. For example, Tarone procedure only uses the minimal attainable p-value to reduce the number of tested hypotheses. In practice, the CDF of the null p-values are often known

for discrete data. By using the distributional information of the null p-values, we can develop more powerful FWER controlling procedures for discrete data.

In many modern applications, it may be more appropriate to apply a multiple testing procedure that controls FDR. One such application is to study clinical safety for drug development, which collect and monitor spontaneous reports of suspected adverse events from health care providers. In order to detect new adverse drug reactions after marketing approval, we can use multiple testing methods to test the association between drugs and adverse events while controlling the FDR. Since the adverse events are naturally classified by different body system, a two-stage multiple testing procedure is considered, where the first stage is to select body systems for further research discovery, and the second stage is to flag AE in selected body systems. It is desired to control type 1 error rates for both screening and testing stages. Since the selection and testing stages use the same data, the selective (conditional) inference should be taken into account.

This application also motivates to develop a GLM model in Chapter 4. When the response is categorical data, with several level as rare ones, but the covariates are the same in each sample, the traditional logistic model or ENMn model is not suitable any more. We suggest a logistic-type model with the response following ENMn distributions. We can make corresponding inference, such as parameter estimation, confidence interval, hypotheses testing, etc. Real data analysis and comparisons are performed as well.

This dissertation is outlined as follows: Chapter 1 provides some basic concepts on multiple testing and background on generalized linear model. In Chapter 2, several stepwise FWER controlling procedures for discrete data are proposed. We also compare our proposed procedures with some existing MTPs through real data analysis and simulation studies. In Chapter 3, we develop a two-stage selective inference based multiple testing procedures for multiple families structure, which can be well applied in clinical safety data analysis. Simulation studies through which we compare the proposed procedure with other procedures for multiple families are also presented. In Chapter

4, a multivariate logistic-type model based on an inverse sampling scheme is developed for modeling categorical data including several special and non-special event groups, statistical inference and model diagnostics for the proposed model in comparison with conventional logistic regression are also provided.

CHAPTER 2

FWER CONTROLLING PROCEDURES FOR DISCRETE DATA

2.1 Introduction

In this chapter, we consider to develop several FWER controlling procedures for discrete data. In the existing literature, most FWER controlling procedures are developed for continuous data, such as Bonferroni procedure (1936), Holm procedure (1979) and Hochberg procedure (1988), etc. These procedures control FWER under various dependence condition. However, they might be highly conservative when they are used to analyze discrete data. A few of researches have been devoted to develop FWER controlling procedures for discrete data. Tarone (1990) improved Bonferroni procedure by using the minimal attainable p-value, which reduces the actual number of hypotheses by removing the non-significant ones. The modified Bonferroni procedure controls the FWER under arbitrary dependence and is more powerful than the original Bonferroni procedure for discrete test statistics. But Tarone's procedure lacks α-consistency (Roth, 1999). To overcome this problem, two types of improved Tarone procedures were developed by Hommel and Krummenauer (1998) and Roth (1999), which not only control FWER, but also maintain α-consistency. Furthermore, Hommel and Krummenauer (1998) incorporated Tarone's idea to improve the Holm procedure for discrete data. By using the similar idea, Roth (1999) introduced a two stage step-up procedure based on Hochberg procedure (1988). However, this procedure lacks α-consistency.

In this chapter, we introduce several FWER controlling procedures that exploits the discrete nature of test statistics. We first consider a single-step modified Bonferroni procedure using CDF's of p-values, which exploits enough information of discrete p-values. Compared with existing single-step methods, the proposed procedure has several good properties. It is more powerful than the existing single-step procedures for discrete

data. By using similar idea, we also develop step-down and step-up procedures for discrete data. The proposed procedures not only control the FWER, but also have α-consistency and p-value monotonicity, which are desired properties in multiple testing. Adjusted p-value of the proposed procedures can also be easily calculated, while closed-forms of adjusted p-values are very difficult to obtain for resampling based methods or randomized tests. We illustrate an application for detecting differentially expressed cDNA transcripts among multiple nucleotides, where the experiments are conducted by using discrete data. Through real data analysis and simulation studies, we compare the performances of the proposed methods with those of the available procedures.

The rest of the chapter is organized as follows. Section 2.2 introduces some basic notations, concepts and existing procedures for discrete data. In Section 2.3, the new single-step procedure is proposed and some desired statistical properties of this procedure are discussed. Section 2.4 and 2.5 respectively introduce new step-down and step-up procedures for discrete data to control the FWER. Section 2.6 summarizes and discusses some future work. Statistical computing tools such as R package and web application are also developed.

2.2 Preliminary

Consider the problem of simultaneously testing m hypotheses $H_1, \ldots, H_m$, suppose there are m_0 true null hypotheses and m_1 false null hypotheses. Assume the test statistics are discrete. Let P_i denote the p-value for testing H_i and $\mathbb{P}_i$ denote the full set of all attainable p-values for H_i such that $P_i \in \mathbb{P}_i$. Suppose F_i denote the cumulative distribution function (CDF) of P_i when H_i is true, that is $F_i(u) = Pr(P_i \leq u | H_i \text{ is true})$. For any $u \in \mathbb{P}_i$, $F_i(u) = u$; otherwise, $F_i(u) < u$.

Typically, the hypotheses are ordered based on their p-values and are tested using a single-step or stepwise procedure. Let $P_{(1)} \leq \cdots \leq P_{(m)}$ denote the ordered p-values and $H_{(1)}, \ldots, H_{(m)}$ denote the corresponding hypotheses. Let $F_{(i)}$ denote the CDF of $P_{(i)}$ when $H_{(i)}$ is true, and $\mathbb{P}_{(i)}$ denote the set of all attainable p-values of $P_{(i)}$.

2.3 A Single-step Procedure for Discrete Data

A simple and commonly used single-step procedure is Bonferroni procedure, which rejects H_i if $P_i \leq \frac{\alpha}{m}$. The Bonferroni procedure controls FWER under arbitrary dependence. Taking discreteness of data into account, Tarone (1990) proposed a novel single-step procedure (Procedure 1.1) controlling FWER, which is more powerful than Bonferroni procedure. However, the procedure only use partial information of true null distributions, so it might be conservative. In this section, we consider using full sets of discrete p-values to develop a more powerful single-step procedure.

2.3.1 A New Single-step Procedure

In this subsection, we present a new single-step procedure for discrete data. The proposed procedure fully exploits the marginal distribution of the true null p-values, and is defined as follow:

Procedure 2.1 (Modified Bonferroni)**.** *Let* $t = \max\{p \in \bigcup_{i=1}^{m} \mathbb{P}_i : \sum_{i=1}^{m} F_i(p) \leq \alpha\}$ *and set* $t = \frac{\alpha}{m}$ *if the maximum does not exist. Then reject* H_i *if its corresponding p-value* $P_i \leq t$.

Remark 2.1. It should be noted that the proposed modified Bonferroni procedure 2.1 for discrete data is a natural extension of the usual Bonferroni method for continuous data. When all true null test statistics have continuous distributions, which implies $F_i \sim U[0,1]$, $F_i(p) = p$, then $t = \max\{p \in (0,1] : mp \leq \alpha\} = \frac{\alpha}{m}$ is exactly the critical value of Bonferroni procedure. Thus, the above procedure reduces to Bonferroni procedure.

In the following, we prove that Procedure 2.1 strongly controls the FWER under arbitrary dependence.

Theorem 2.1. *Procedure 2.1 strongly controls the FWER at level* α *under arbitrary dependence.*

Proof. Let V denote the number of falsely rejected hypotheses, I_0 denote the index set of true null hypotheses with $|I_0| = m_0$, then

$$\begin{aligned} FWER &= Pr\{V \geq 1\} = Pr\left\{\bigcup_{i\in I_0}\{P_i \leq t\}\right\} \\ &\leq \sum_{i\in I_0} Pr\{P_i \leq t\} = \sum_{i\in I_0} F_i(t) \\ &\leq \sum_{i=1}^{m} F_i(t) \leq \alpha. \end{aligned} \tag{2.3.1}$$

The first inequality follows from Bonferroni inequality.

If the maximum does not exist, by using the property of the CDF for discrete p-value $F_i(t) \leq t$, $\sum_{i=1}^{m} F_i(t) \leq \sum_{i=1}^{m} t = m \cdot \frac{\alpha}{m} = \alpha$. The proof is complete. □

In the following, we compare the proposed method with several existing single-step procedures, and prove that the proposed procedure is more powerful than these procedures. First of all, we want to show the proposed procedure is more powerful than Tarone's procedure, which is described in supplementary materials.

Proposition 2.1. *Procedure 2.1 is universally more powerful than Procedure 1.1.*

Proof. We firstly show that $\sum_{i=1}^{m} F_i(\frac{\alpha}{K_\alpha}) \leq \alpha$.

Since $R_{K_\alpha} = \{i : p_i^* < \frac{\alpha}{K_\alpha}\}$, $|R_{K_\alpha}| = M_\alpha(K_\alpha) \leq K_\alpha$. Therefore,

$$\sum_{i=1}^{m} F_i(\frac{\alpha}{K_\alpha}) = \sum_{i\in R_{K_\alpha}} F_i(\frac{\alpha}{K_\alpha}) \leq |R_{K_\alpha}| \cdot \frac{\alpha}{K_\alpha} \leq \alpha.$$

Let $t = \max\{p \in \bigcup_{i=1}^{m} \mathbb{P}_i : \sum_{i=1}^{m} F_i(p) \leq \alpha\}$, and let t^* be the smallest attainable p-value greater than t, that is, $t^* = \min\left\{p \in \bigcup_{i=1}^{m} \mathbb{P}_i : p > t\right\}$, then $\sum_{i=1}^{m} F_i(t^*) > \alpha$. We have shown $\sum_{i=1}^{m} F_i(\frac{\alpha}{K_\alpha}) \leq \alpha$, so $\frac{\alpha}{K_\alpha} < t^*$. Then there are two cases: (1) when $\frac{\alpha}{K_\alpha} \leq t$, it is trivial the set of rejections using Tarone's procedure is no more than the Procedure 2.1; (2) when $t < \frac{\alpha}{K_\alpha} < t^*$, by the property of discreteness, $\{H_i : P_i \leq t\} = \{H_i : P_i < \frac{\alpha}{K_\alpha}\} = \{H_i : P_i < t^*\}$. So based on (1) and (2) rejection set using Tarone's procedure is less than or equal to the one using Procedure 2.1.

Therefore, Procedure 2.1 always rejects as many hypotheses as Tarone's procedure. That is, Procedure 2.1 is universally more powerful than Tarone's procedure. □

We can also show that the proposed procedure is more powerful than modified Tarone's Procedure.

Proposition 2.2. *Procedure 2.1 is universally more powerful than Procedure 1.2.*

Proof. We need to show that for $\forall\gamma \leq \alpha$, $\sum_{i=1}^{m} F_i(\frac{\gamma}{K_\gamma}) \leq \alpha$.

Let $R_{K_\gamma} = \{i : p_i^* < \frac{\gamma}{K_\gamma}\}$, then $|R_{K_\gamma}| = M_\gamma(K_\gamma) \leq K_\gamma$. Therefore,

$$\begin{aligned} \sum_{i=1}^{m} F_i(\frac{\gamma}{K_\gamma}) &= \sum_{i \in R_{K_\gamma}} F_i(\frac{\gamma}{K_\gamma}) \\ &\leq |R_{K_\gamma}| \cdot \frac{\gamma}{K_\gamma} \\ &= M_\gamma(K_\gamma) \cdot \frac{\gamma}{K_\gamma} \\ &\leq \gamma \leq \alpha. \end{aligned} \tag{2.3.2}$$

The rest of argument is similar as the proof of Proposition 2.1 and the conclusion follows.

□

So far, we have shown the new single-step procedure is more powerful than the Tarone's procedure and its modified versions for discrete data. Next, we look into some other good properties of this procedure.

α-consistency

Proposition 2.3. *Procedure 2.1 is an α-consistent procedure.*

Proof. Since $t = \max\{p \in \bigcup_{i=1}^{m} \mathbb{P}_i : \sum_{i=1}^{m} F_i(p) \leq \alpha\}$. It is equivalent to show that the threshold t is a non-decreasing function in α. It is trivial.

□

p-value monotonicity Based on (2.3.3), it is easy to show Procedure 2.1 has p-value monotonicity. Since for each i, true null CDF $F_i(\cdot)$ is a non-decreasing function, $\sum_{i=1}^{m} F_i(\cdot)$ is also a non-decreasing function. When some p-values become smaller, the corresponding adjusted p-values will not become larger, thus the procedure will reject the same hypotheses and possibly more. So we have the following proposition.

Proposition 2.4. *Procedure 2.1 is p-value monotone.*

Adjusted p-value Now, we can derive the adjusted p-value for our proposed procedure as follow:

Proposition 2.5 (Modified Bonferroni Procedure 2.1)**.**
If P_i is the available p-value for H_i, then the adjusted p-value for corresponding hypothesis is

$$\tilde{P}_{i,MBonf} = \min\left\{1,\ \sum_{j=1}^{m} F_j(P_i)\right\}, \quad for\ i = 1, \ldots, m. \tag{2.3.3}$$

It is easy to see that the adjusted p-value of the proposed modified Bonferroni procedure is smaller than or equal to that of original Bonferroni procedure, since for each fixed i and any $j = 1, \ldots, m$, $F_j(P_i) \leq P_i$, then $\sum_{j=1}^{m} F_j(P_i) \leq mP_i$. Therefore, the Procedure 2.1 could have more rejections than Bonferroni procedure for the same available p-values.

In the following, we compare the adjusted p-values of the proposed Procedure 2.1 with those of Bonferroni procedure and Procedure 1.2 through a simple example.

Example 2.1. Suppose there are $m = 2$ hypotheses H_1 and H_2, the attainable p-values for H_1 is $\mathbb{P}_1 = \{0.05, 1\}$; for H_2 is $\mathbb{P}_2 = \{0.1, 1\}$. The actual p-values are $P_1 = 0.05,\ P_2 = 0.1$. Thus the minimal attainable p-values are $p_1^* = 0.05,\ p_2^* = 0.1$. Now we can calculate the adjusted p-values for Bonferroni procedure are $\tilde{P}_{1,Bonf} = 2 \times P_1 = 0.1$ and $\tilde{P}_{2,Bonf} = 2 \times P_2 = 0.2$.

To calculate the adjusted p-value of Procedure 1.2, firstly we need to determine q. For $P_1 = 0.05$, $p_1^* \leq P_1 < p_2^*$, so $q = 1$, $\tilde{P}_{1,T^*} = 1 \times P_1 = 0.05$. For $P_2 = 0.1$, since $P_2 \geq p_2^*$, then $q = m = 2$, $\tilde{P}_{2,T^*} = 2 \times P_2 = 0.2$. The CDF of p-values for

the two hypotheses can be expressed by $F_1(c) = 0.05 \times I\{0.05 \leq c < 1\} + I\{c = 1\}$, $F_2(c) = 0.1 \times I\{0.1 \leq c < 1\} + I\{c = 1\}$, where I is an indicator function. So $\tilde{P}_{1,MBonf} = F_1(0.05) + F_2(0.05) = 0.05 + 0 = 0.05$, $\tilde{P}_{2,MBonf} = F_1(0.1) + F_2(0.1) = 0.05 + 0.1 = 0.15$, which are smaller than those of the Bonferroni procedure and Procedure 1.2.

Now, suppose we set the significant level $\alpha = 0.06$, then by comparing the adjusted p-values of the above procedures with α, we can conclude the Bonferroni procedure reject no hypothesis, Procedure 1.2 and Procedure 2.1 reject H_1. But if set $\alpha = 0.16$, the Bonferroni procedure and Procedure 1.2 only reject H_1, while the proposed Procedure 2.1 rejects H_1 and H_2.

2.3.2 Applications for Single-step Procedures

cDNA transcripts data Tarone (1990) analyzed an experiment in which complementary DNA (cDNA) transcripts were produced from transcribed RNA obtained from cells grown under normal conditions and from cells grown under an unusual study condition. The cDNA transcripts from a gene of interest were sequenced and compared to the known nucleotide sequence to determine the number of individual nucleotide changes in the transcripts. The frequencies of the changes were compared from the control and study cells to evaluate differences in the transcribed RNA.

The data in Table 2.1 is from Hommel and Krummenauer (1998, Table 1), which reports the frequencies of nucleotide changes observed at nine sites. The DNA sequences examined in the experiment were 200 nucleotides in length. Our analysis includes nine changed nucleotides, which are those with a sufficient number of changes to possibly detect statistical significance at the significant level $\alpha = 0.05$ using the Fisher's Exact Test (FET), conditional on the fixed marginal totals, and assuming independence between sites. In the data, N_{ji} is the number of transcripts at nucleotide i in group j and X_{ji} is the observed number of change in transcripts, which is the events of interest, where $i = 1, ..., 9$ and $j = 0, 1$ (0 is control group and 1 is study group). The first column shows the index of the ordered nucleotide p-values reported in Hommel and Krummenauer (1998). The second and third columns are the frequencies of the observed change in the control and study groups. The nucleotides available p-values P_i

in Table 2.1 are calculated by using one-sided FET:

$$P_i = \sum_{k=X_{1i}}^{X_{\cdot i}} \frac{\binom{N_{1i}}{k}\binom{N_{0i}}{X_{\cdot i}-k}}{\binom{N_{\cdot i}}{k}}, \tag{2.3.4}$$

where $X_{\cdot i} = X_{0i} + X_{1i}, N_{\cdot i} = N_{0i} + N_{1i}$. For Tarone's procedure, the minimal attainable significance level at site i is given by

$$p_i^* = \frac{\binom{N_{1i}}{X_{\cdot i}}}{\binom{N_{\cdot i}}{X_{\cdot i}}}. \tag{2.3.5}$$

Table 2.1 A Comparison of Adjusted p-values for the Bonferroni Procedure, Sidak Procedure, Modified Tarone Procedure and Procedure 2.1 when Testing the Hypotheses in the cDNA Example from Hommel and Krummenauer (1998)

i	X_{0i}/N_{0i}	X_{1i}/N_{1i}	P_i	$\tilde{P}_{i,Bonf}$	$\tilde{P}_{i,Sidak}$	$\tilde{P}_{i,T^*}$	$\tilde{P}_{i,MBonf}$
1	1/10	8/11	0.0058	0.0522	0.0510	0.0116	0.0097
2	0/8	5/7	0.0070	0.0629	0.0612	0.0210	0.0167
3	0/11	4/10	0.0351	0.3158	0.2749	0.2100	0.1072
4	1/11	3/9	0.2167	1.0000	0.8890	1.0000	0.6184
5	2/11	4/10	0.2678	1.0000	0.9395	1.0000	1.0000
6	1/10	3/10	0.2910	1.0000	0.9547	1.0000	1.0000
7	2/9	2/8	0.6647	1.0000	1.0000	1.0000	1.0000
8	2/9	2/9	0.7118	1.0000	1.0000	1.0000	1.0000
9	2/9	2/9	0.7118	1.0000	1.0000	1.0000	1.0000

From Table 2.1, we can see that for nucleotide $i = 1, \ldots, 4$, the adjusted p-values of Procedure 2.1 are smaller than those of other traditional procedures, which implies these hypotheses are more likely to be rejected by the Procedure 2.1 than others.

Clinical safety data We can also apply the proposed single-step procedure for clinical safety studies, since clinical safety data is usually based on the count of patients to illustrate the adverse events exposures. The data in Table 2.2 (first three columns) are from Mehrotra and Heyse (2004, Table 1), which reports the AE types for two groups

of toddlers for Body System 10. For illustration purpose, we reorder the data based on the corresponding p-values. The goal of this clinical safety study is to detect significant AEs (so-called "flagging"). Our analysis includes nine AE types of No. 10 body system (skin), which are those with a sufficient number of AE types to possibly detect statistical significance at the significant level $\alpha = 0.05$ using the Fisher's Exact Test (FET), conditional on the fixed marginal totals, and assuming independence between sites. In the data, N_j is the total number of toddlers at group j, and X_{ji} is the observed number of the j-th group toddlers experiencing the i-th AE, which is the events of interest, where $i = 1, ..., 9$ and $j = 1, 2$ (1 is control group receiving MMR and 2 is study group receiving the candidate vaccine MMRV). Here $N_1 = 148$ and $N_2 = 132$. The first column shows the index of the AE types after reordering the data. The second and third columns are the number of toddlers experiencing the corresponding AE in the control and study groups. The available p-values P_i for i-the AE type in Table 2.2 are calculated by using two-sided FET.

Table 2.2 A Comparison of Adjusted p-values for the Bonferroni Procedure, Sidak Procedure, Procedure 1.2 and Procedure 2.1 when Testing the Hypotheses for Nine AE types of Body System 10 in the Clinical Safety Data Example from Mehrotra and Heyse (2004), where the Numbers of Patients for Two Groups Are $N_1 = 148$ and $N_2 = 132$

i	X_{1i}	X_{2i}	P_i	$\tilde{P}_{i,Bonf}$	$\tilde{P}_{i,Sidak}$	$\tilde{P}_{i,T^*}$	$\tilde{P}_{i,MBonf}$
1	13	3	0.0209	0.1880	0.1731	0.0836	0.0534
2	8	1	0.0388	0.3490	0.2995	0.1551	0.1343
3	4	0	0.1248	1.0000	0.6986	0.8734	0.7134
4	0	2	0.2214	1.0000	0.8948	1.0000	1.0000
5	6	2	0.2885	1.0000	0.9533	1.0000	1.0000
6	2	0	0.4998	1.0000	0.9980	1.0000	1.0000
7	1	2	0.6033	1.0000	0.9998	1.0000	1.0000
8	4	2	0.6872	1.0000	1.0000	1.0000	1.0000
9	2	1	1.0000	1.0000	1.0000	1.0000	1.0000

From Table 2.2, we can see that for the first three AE p-values $P_1, \ldots, P_3$, the adjusted p-values of Procedure 2.1 are smaller than those of other traditional procedures, which implies these hypotheses are more likely to be rejected by the Procedure 2.1 than others, that is, those AE are more easily flagged by using the Procedure 2.1.

2.3.3 Simulation Studies for Single-step Procedures Comparisons

In the following, simulation studies were performed to investigate the performances of the proposed Procedures 2.1 in terms of the FWER control and minimal power compared with some existing single-step FWER controlling procedures.

Basic settings of the simulation The simulations are conducted based on two typical discrete tests: Fisher's Exact Test (FET) and Binomial Exact Test (BET).

1. Fisher's Exact Test: Suppose we have two groups, study (1) and control (2) group. There are m independent binomial responses X_{ij} observed for each of N individuals in each group i, such as $X_{i1} \sim Bin(N, p_{i1})$, $X_{i2} \sim Bin(N, p_{i2})$ for $i = 1, \ldots, m$. The goal is to simultaneously test the m hypotheses $H_i : p_{i1} = p_{i2}$, where p_{ij} is the success probability for the i-th response in group j, and $i = 1, \ldots, m$, $j = 1, 2$. So there are m of 2×2 contingency tables in each simulation as described in Chapter 1. We conduct the experiment using one-sided FET under $\alpha = 0.05$, then the test statistic $T_i \sim Hypergeometric(X_{i1}, N, X_{i1} + X_{i2}, 2N)$. Set the number of hypotheses $m = \{5, 10, 15\}$, with true null proportion $\pi_0 = \{0.2, 0.4, 0.6, 0.8\}$ respectively. The sample size for the binomial response per group used are $N = \{25, 50, 75, 100, 125, 150\}$. For true null hypotheses, set the success probability parameter of binomial response in each group as 0.1, and for false null hypotheses set the success probability for study group as 0.1, and for control group as 0.2. The observed individuals in the two groups are chosen randomly from the Binomial distributions.

2. Binomial Exact Test: Suppose we have two groups: study (1) and control (2) group. There are m Poisson responses observed in each group, such as $X_{i1} \sim Poi(\lambda_{i1})$, $X_{i2} \sim Poi(\lambda_{i2})$ for $i = 1, \ldots, m$. The goal is to simultaneously test the m hypotheses $H_i : \lambda_{1i} = \lambda_{2i}$, where λ_{ij} is the mean parameter for the i-th response in group j, and $i = 1, \ldots, m$, $j = 1, 2$. We conduct the experiment using one-sided BT under $\alpha = 0.05$ and $\alpha = 0.1$ respectively, then the test statistic for reference group follow binomial distribution. Here we assume group 1 as reference group, then $T_i \sim Bin(X_{i1} + X_{i2}, p_i)$, where $p_i = \dfrac{\lambda_{i1}}{\lambda_{i1} + \lambda_{i2}}$. Set the number of hypotheses $m = \{5, 10, 15\}$, with true null proportion $\pi_0 = \{0.2, 0.4, 0.6, 0.8\}$ respectively. For true null hypotheses set the mean parameter of Poisson response in each group as $\lambda_{1i} = \lambda_{2i} = 2$, and for false null hypotheses set the mean parameter for group 1 as $\lambda_{1i} = 2$, and for group 2 as $\lambda_{2i} = 10$. The study and control group observed individual are chosen randomly from the Binomial distributions.

Using the FET or BET we can calculate the available p-value P_i and all attainable p-values in the set $\mathbb{P}_i$. Then we compute the simulated FWER, minimal power, number of rejections by taking average of $B = 2000$ iterations.

$$\text{Power} = \Pr\{\text{correctly rejecting at least one null hypotheses}\}.$$

Results of the simulation under independence Tables A.1 and A.2 in the Appendix A show the simulated FWER levels and minimal powers of the compared four procedures using the FET statistics. First, the proposed Modified Bonferroni procedure (Procedure 2.1) always has higher FWER level, and more powerful than the other three procedures. The simulation results also verify that two discrete FWER controlling procedures (Modified Bonferroni and Tarone) have higher FWER levels and provide more power than the other two classic procedures (Bonferroni and Sidak). Second, the FWER levels are less conservative, and the power advantages are larger for smaller size N, since the data was more discrete for smaller N, then the improvement is more obvious. For example, when testing $m = 10$ hypotheses, $\pi_0 = 0.2$, which

implies there are 2 true nulls and 8 false nulls, the simulation result shows that the FWER improvement of Procedure 2.1 (0.0020) is 300% higher than Tarone procedure (0.0005) when the simulated data is from binomial with $N = 5$. But when sample size $N = 125$, the improvement is only 35.7% (0.0095 versus 0.0070). Third, as the true null proportion becomes bigger, the proposed procedures FWER is closer to nominal significant level 0.05, but power becomes smaller. The power of Procedure 2.1 becomes larger when testing more hypotheses or using larger sample size N. We also plot the simulation results in Figures 2.1 and 2.2 for the FWER and minimal power comparisons.

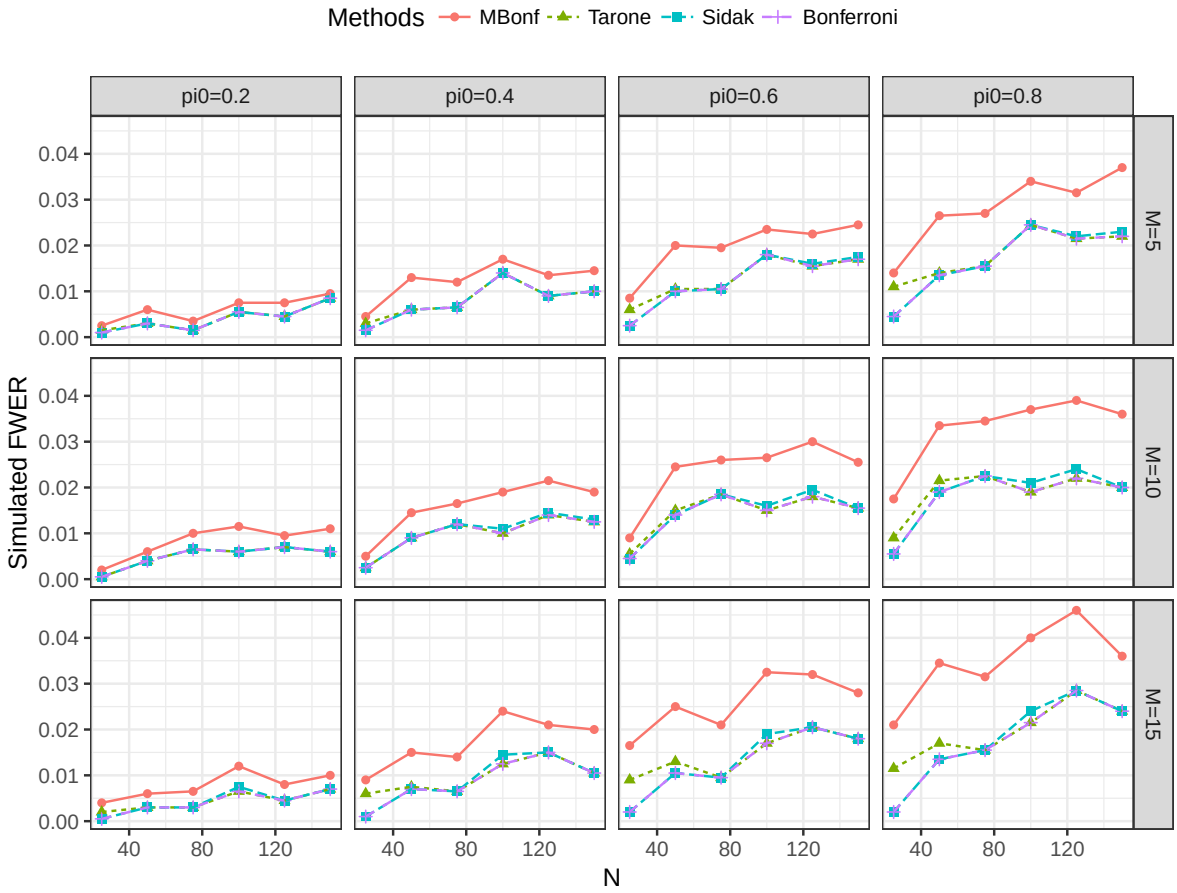

Figure 2.1 Simulated FWER comparisons for different single-step procedures based on FET.

Tables A.3 and A.4 in the appendix show the simulated FWER levels and minimal powers comparisons using the BET statistics. The results show the proposed Procedure

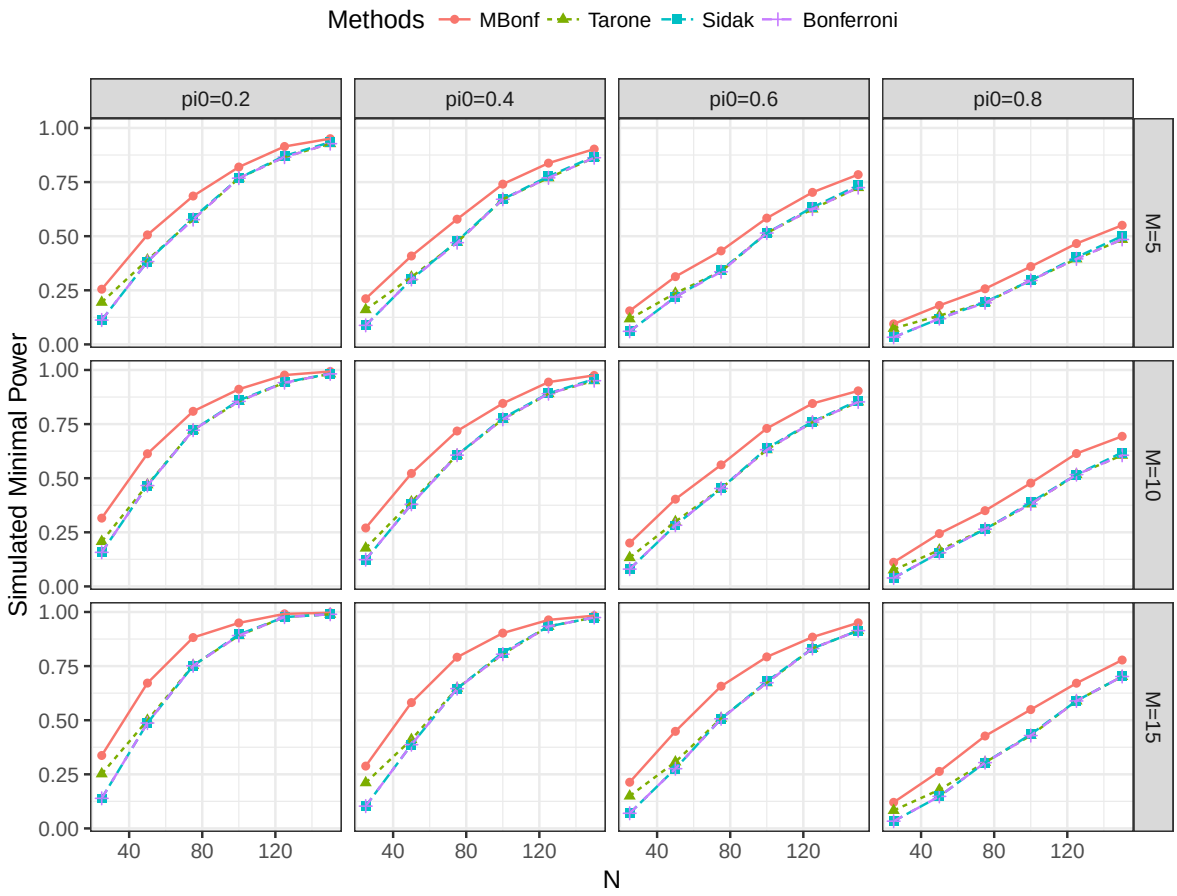

Figure 2.2 Simulated minimal power comparisons for different single-step procedures based on FET.

2.1 controls FWER and are more powerful than other three procedures under such settings. For other findings, they are similar to the simulation results based on FET.

2.3.4 Extension for the Proposed Procedures for the Mixed Data Structure of the Hypotheses

In practice, the hypotheses are commonly involved in a mixed structure for both discrete and continuous data. then the proposed procedures can be naturally extended as follows.

Suppose there are m hypotheses $H_1, \ldots, H_m$. Let I_c denote the index set of the hypotheses for continuous data, so the number of continuous test statistics is $m_c = |I_c|$. The corresponding available p-values under the null one uniformly distributed in $[0, 1]$,

that is, $F_i(p) = p$ for $i \in I_c$ and any $p \in [0,1]$. Let I_d denote the index set of the hypotheses for discrete data, the number of discrete test statistics is $m_d = |I_d| = m - m_c$. The corresponding available p-values are stochastically greater than or uniformly distributed in $[0,1]$, that is, $F_i(p) = p$ for $i \in I_d$ and $p \in \mathbb{P}_i$; $F_i(p) < p$ for $i \in I_d$ and $p \notin \mathbb{P}_i$. The mixed CDF can be expressed by

$$F_i(p) = pI\{i \in I_c\} + F_i(p)I\{i \in I_d\}.$$

Procedure 2.2 (Mixed Bonferroni procedure)**.** *Let the critical constant be* $t = \max\{0 \le p \le 1 : m_c p + \sum_{i \in I_d} F_i(p) \le \alpha\}$*, then reject* H_i *if* $P_i \le t$*.*

Example 2.2. If there are only two hypotheses H_1 and H_2. The test statistics T_1 of H_1 is continuous, then the corresponding p-value $P_1 \sim Unif(0,1)$. The test statistics T_2 of H_2 is discrete, let P_2 be the corresponding p-value with atom at $0.01, 0.19, 1$ (eg: $T_2 \sim Binomial(2, 0.1)$, then the CDF of P_2 is

$$F_2(p) = \begin{cases} 0 & 0 \le p < 0.01 \\ 0.01 & 0.01 \le p < 0.19 \\ 0.19 & 0.19 \le p < 1 \\ 1 & p = 1 \end{cases} \tag{2.3.6}$$

If the significant level $\alpha = 0.05$, then the critical constant of mixed Bonferroni procedure is $t = \max\{0 \le p \le 1 : p + F_2(p) \le 0.05\} = 0.04$. While the critical constant for traditional Bonferroni procedure is $t' = 0.05/2 = 0.025$, which is more conservative. If we observe $P_1 = 0.03, P_2 = 0.01$, then only H_2 is rejected for Bonferroni procedure, but H_1 and H_2 are rejected for mixed Bonferroni procedure.

So the mixed Bonferroni procedure could less conservative than Bonferroni procedure.

Proposition 2.6 (Adjusted p-value for mixed Bonferroni procedure)**.**
If P_i *is the available p-value for* H_i*, then the adjusted p-value for corresponding hypothesis is*

$$\tilde{P}_{i,MixBonf} = \min\left\{1,\ m_c P_i + \sum_{j \in I_d} F_j(P_i)\right\},\ \textit{for } i = 1, \ldots, m. \tag{2.3.7}$$

Since the mixed Bonferroni procedure is just a special case of Procedure 2.1 only if some test statistics are continuous, the mixed procedure strongly controls the FWER under arbitrary dependence and holds all desired properties such as α-consistency, p-value monotonicity, etc.

2.4 A Step-down Procedure for Discrete Data

By exploiting the discreteness of data, Hommel and Krummenauer (1998) improved Holm procedure as Procedure 1.4. This procedure could be further improved by utilizing full sets of p-values.

2.4.1 A New Step-down Procedure

In the last section, we have proposed a new single-step procedure based on the marginal CDF's of p-values, and now we develop a more powerful step-down procedure by exploiting the the same distributional information of null p-values.

Procedure 2.3 (Modified Holm)**.** *Let* $\alpha_i = \max\{p \in \bigcup_{j=i}^{m} \mathbb{P}_{(j)} : \sum_{j=i}^{m} F_{(j)}(p) \leq \alpha\}$ *with* $\alpha_0 = 0$. *Set* $\alpha_i = \max\left\{\alpha_{i-1}, \frac{\alpha}{m-i+1}\right\}$ *if the maximum does not exist. Then reject no null hypotheses if* $P_{(1)} > \alpha_1$*; otherwise, reject* $H_{(1)}, \ldots, H_{(r)}$ *and retain* $H_{(r+1)}, \ldots, H_{(m)}$*, where* r *is the largest index satisfying* $P_{(1)} \leq \alpha_1, \ldots, P_{(r)} \leq \alpha_r$.

Remark 2.2. It should be noted that when the test statistics have continuous distributions, which implies $F_i \sim U[0,1]$, $F_i(p) = p$, then

$$\alpha_i = \max\{p \in (0,1] : (m-i+1)p \leq \alpha\} = \frac{\alpha}{m-i+1}.$$

Thus, the procedure reduces to Holm procedure.

Theorem 2.2. *Procedure 2.3 strongly controls the FWER at level* α *under arbitrary dependence.*

Proof. Let I_0 be the indices of the true null hypotheses and V denote the number of falsely rejected hypotheses. If $|I_0| = 0$, then $V = 0$, $FWER = 0 \leq \alpha$ is trivial. When $|I_0| = m_0 \geq 1$, let $\hat{P}_{(1)} \leq \cdots \leq \hat{P}_{(m_0)}$ denote the m_0 true null p-values, and $\check{P}_{(1)} \leq \cdots \leq \check{P}_{(m_1)}$ denote the m_1 false null p-values.

Let k be the smallest random index of whole p-values satisfying $P_{(k)} = \hat{P}_{(1)}$, that is $P_{(k)} = \min_{i \in I_0} P_i$. It implies $P_{(k)}, \ldots, P_{(m)}$ include all true null p-values, that is,

$$\left\{\hat{P}_{(1)}, \ldots, \hat{P}_{(m_0)}\right\} \subseteq \left\{P_{(k)}, \ldots, P_{(m)}\right\}.$$

Therefore,

$$\begin{aligned} FWER = Pr\{V \geq 1\} &= Pr\{\min_{i \in I_0} P_i \leq \alpha_k\} \\ &\leq \sum_{i \in I_0} Pr\{P_i \leq \alpha_k\} \leq \sum_{j=k}^{m} F_{(j)}(\alpha_k) \leq \alpha \end{aligned} \tag{2.4.1}$$

The last inequality follows by the definition of $\alpha_k = \max\{p \in \bigcup_{j=k}^{m} \mathbb{P}_{(j)} : \sum_{j=k}^{m} F_{(j)}(p) \leq \alpha\}$.

If the maximum for k does not exist, then $\alpha_k = \max\left\{\alpha_{k-1}, \ \dfrac{\alpha}{m-k+1}\right\}$.

If $\alpha_{k-1} \leq \dfrac{\alpha}{m-k+1}$, that is, $\alpha_k = \dfrac{\alpha}{m-k+1}$, by the property of CDF for discrete p-value, the last inequality will become $\sum_{j=k}^{m} F_{(j)}(\alpha_k) \leq \sum_{j=k}^{m} \alpha_k = \sum_{j=k}^{m} \dfrac{\alpha}{m-k+1} = \alpha$.

If $\alpha_{k-1} > \dfrac{\alpha}{m-k+1}$, and if the maximum as definition exists for $k-1$, that is, $\alpha_{k-1} = \max\{p \in \bigcup_{j=k-1}^{m} \mathbb{P}_{(j)} : \sum_{j=k-1}^{m} F_{(j)}(p) \leq \alpha\}$, then the last inequality will become $\sum_{j=k}^{m} F_{(j)}(\alpha_k) = \sum_{j=k}^{m} F_{(j)}(\alpha_{k-1}) \leq \sum_{j=k-1}^{m} F_{(j)}(\alpha_{k-1}) \leq \alpha$. If the maximum as definition does not exist for $k-1$, then $\alpha_k = \alpha_{k-1} = \max\left\{\alpha_{k-2}, \ \dfrac{\alpha}{m-k+2}\right\}$. By the similar argument, $FWER \leq \alpha$ when $\alpha_k = \alpha_{k-1} = \dfrac{\alpha}{m-k+2}$.

By iteration, we can prove until for some $l-1$, the maximum exists, then

$$\alpha_k = \alpha_{k-1} = \cdots = \alpha_l = \alpha_{l-1}, \text{ then}$$

$$\sum_{j=k}^{m} F_{(j)}(\alpha_k) = \sum_{j=k}^{m} F_{(j)}(\alpha_{l-1}) \leq \sum_{j=l-1}^{m} F_{(j)}(\alpha_{l-1}) \leq \alpha,$$

which completes the proof. □

α-consistency Based on Definition 1.2, we can also explore this desired property for Procedure 2.3. This property can be proved by using the similar argument in the proof of Proposition 2.3.

Proposition 2.7. *Procedure 2.3 is an α-consistent procedure.*

Adjusted p-value We can directly calculate the adjusted p-value of Procedure 2.3 based on the Definition 1.2.8.

Proposition 2.8 (Adjusted p-value for Procedure 2.3)**.**
If $P_{(1)} \leq \cdots \leq P_{(m)}$ are the available p-value for $H_{(1)}, \ldots, H_{(m)}$, then the adjusted p-value of Procedure 2.3 for corresponding hypothesis $H_{(i)}$ is

$$\tilde{P}_{(i),MHolm} = \begin{cases} \min\left\{1,\ \sum_{j=1}^{m} F_{(j)}(P_{(1)})\right\}, & i = 1 \\ \max\left\{\tilde{P}_{(i-1),MHolm}, \min\left\{1,\ \sum_{j=i}^{m} F_{(j)}(P_{(i)})\right\}\right\}. & i = 2, \ldots, m \end{cases} \tag{2.4.2}$$

p-value monotonicity According to the calculation of the adjusted p-value Eq. (2.4.2), we can show that Procedure 2.3 is also p-value monotone using similar argument of Proposition 2.4.

Proposition 2.9. *Procedure 2.3 is p-value monotone.*

2.4.2 Applications for Step-down Procedures

cDNA transcripts data We compare the proposed Procedure 2.3 with Holm procedure and Tarone-Holm Procedure 1.4 using the previous cDNA transcripts example. We also use their adjusted p-values to make decisions of rejection and acceptance.

Table 2.3 shows for hypotheses $H_{(1)}, \ldots, H_{(5)}$, the adjusted p-values of Procedure 2.3 are smaller than those of Holm and Tarone-Holm procedures. That means Procedure 2.3 has more chances to reject $H_{(1)}, \ldots, H_{(5)}$ than the other two procedures, which implies our proposed Procedure 2.3 could be more powerful than other two.

Table 2.3 A Comparison of Adjusted p-values for the Holm Procedure, Tarone-Holm Procedure and Procedure 2.3 when Testing the Hypotheses in the cDNA Transcript Example from Hommel and Krummenauer (1998)

(i)	X_{0i}/N_{0i}	X_{1i}/N_{1i}	$P_{(i)}$	$\tilde{P}_{(i),Holm}$	$\tilde{P}_{(i),TH^*}$	$\tilde{P}_{(i),MHolm}$
(1)	1/10	8/11	0.0058	0.0552	0.0116	0.0097
(2)	0/8	5/7	0.0070	0.0559	0.0140	0.0109
(3)	0/11	4/10	0.0351	0.2456	0.1404	0.1072
(4)	1/11	3/9	0.2167	1.0000	1.0000	0.4268
(5)	2/11	4/10	0.2678	1.0000	1.0000	0.6347
(6)	1/10	3/10	0.2910	1.0000	1.0000	1.0000
(7)	2/9	2/8	0.6647	1.0000	1.0000	1.0000
(8)	2/9	2/9	0.7118	1.0000	1.0000	1.0000
(9)	2/9	2/9	0.7118	1.0000	1.0000	1.0000

Clinical safety data We also compare these step-down procedures using the previous clinical safety data example.

Table 2.4 A Comparison of Adjusted p-values for the Holm Procedure, Procedure 1.4 and Procedure 2.3 when Testing the Hypotheses for AE Types of Body System 10 in the Clinical Safety Data Example from Mehrotra and Heyse (2004), where the Numbers of Patients for Two Groups are $N_1 = 148$ and $N_2 = 132$

(i)	X_{1i}	X_{2i}	$P_{(i)}$	$\tilde{P}_{(i),Holm}$	$\tilde{P}_{(i),TH^*}$	$\tilde{P}_{(i),MHolm}$
(1)	13	3	0.0209	0.1880	0.0836	0.0534
(2)	8	1	0.0388	0.3103	0.1163	0.0982
(3)	4	0	0.1248	0.8734	0.6238	0.5050
(4)	0	2	0.2214	1.0000	1.0000	1.0000
(5)	6	2	0.2885	1.0000	1.0000	1.0000
(6)	2	0	0.4998	1.0000	1.0000	1.0000
(7)	1	2	0.6033	1.0000	1.0000	1.0000
(8)	4	2	0.6872	1.0000	1.0000	1.0000
(9)	2	1	1.0000	1.0000	1.0000	1.0000

Table 2.4 shows for hypotheses $H_{(1)}, \ldots, H_{(3)}$, the adjusted p-values of Procedure 2.3 are smaller than those of Holm and Tarone-Holm procedures. It means Procedure 2.3 has more chances to reject $H_{(1)}, \ldots, H_{(3)}$ than the other two procedures, which implies our proposed Procedure 2.3 could be more powerful than other two.

2.4.3 Simulation Study for Step-down Procedures Comparisons

In this section, simulation studies were performed to investigate the performances of the proposed Procedure 2.3 in terms of the FWER level and minimal power compared with two existing step-down procedures: Holm procedure and Tarone-Holm procedure in Hommel and Krummenauer (1998). The step-down procedures simulations are conducted by using Fisher's Exact Test only, since using binomial exact test produces similar patterns. The same simulation settings in Section 2.3.3 are used for this comparison. The simulations results are shown in the Tables A.5 and A.6 in the Appendix A.

The results show that Procedure 2.3 always controls FWER and are more powerful than other procedures. Moreover, by comparing the results in the Tables A.1 and A.2, the proposed step-down Procedure 2.3 is more powerful than proposed single-step Procedure 2.1. We also plot the simulation results in Figures 2.3 and 2.4 for the FWER and minimal power comparisons.

2.5 A Step-up Procedure for Discrete Data

2.5.1 A New Step-up Procedure

By using the same critical constants of Procedure 2.3, we can develop a new step-up procedure for discrete data.

Procedure 2.4 (Modified Hochberg)**.** *Let* $\alpha_i = \max\{p \in \bigcup_{j=i}^{m} \mathbb{P}_{(j)} : \sum_{j=i}^{m} F_{(j)}(p) \leq \alpha\}$ *with* $\alpha_0 = 0$. *Set* $\alpha_i = \max\left\{\alpha_{i-1}, \ \frac{\alpha}{m-i+1}\right\}$ *if the maximum does not exist. Then reject all hypotheses* $H_{(1)}, \ldots, H_{(m)}$ *if* $P_{(m)} \leq \alpha_m$*; otherwise, reject* $H_{(1)}, \ldots, H_{(r)}$ *and retain* $H_{(r+1)}, \ldots, H_{(m)}$*, where* r *is the largest index satisfying* $P_{(r)} \leq \alpha_r$.

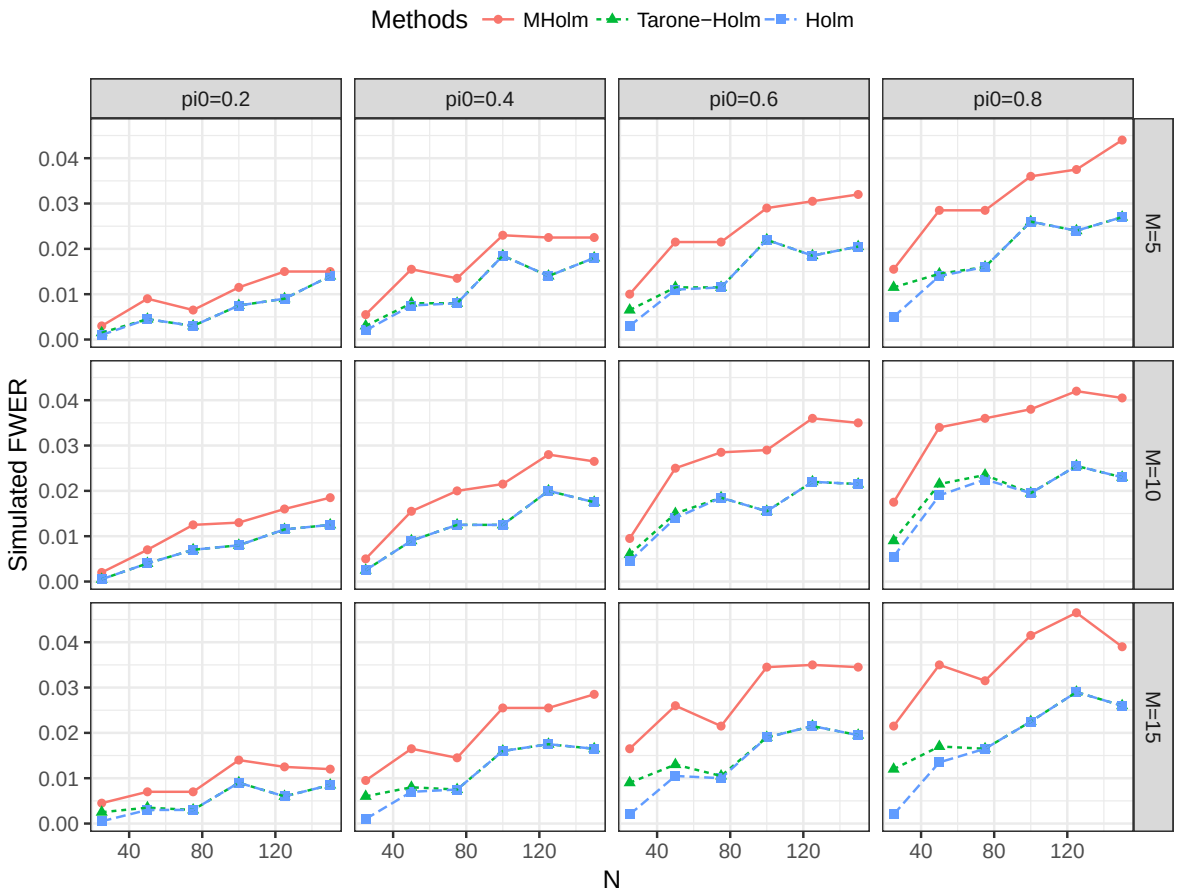

Figure 2.3 Simulated FWER comparisons for different step-down procedures based on FET.

Remark 2.3. It should be noted that when the true null test statistics have continuous distributions, which implies that all true null p-values are uniformly distributed in $[0, 1]$, the above procedure will reduce to conventional Hochberg procedure.

Theorem 2.3. *If the true null test statistics are identically distributed, (i) then Procedure 2.4 strongly controls the FWER at level α under the Assumption 1.2. (ii) Moreover, the Procedure 2.4 rejects the same number of hypotheses as Hochberg procedure.*

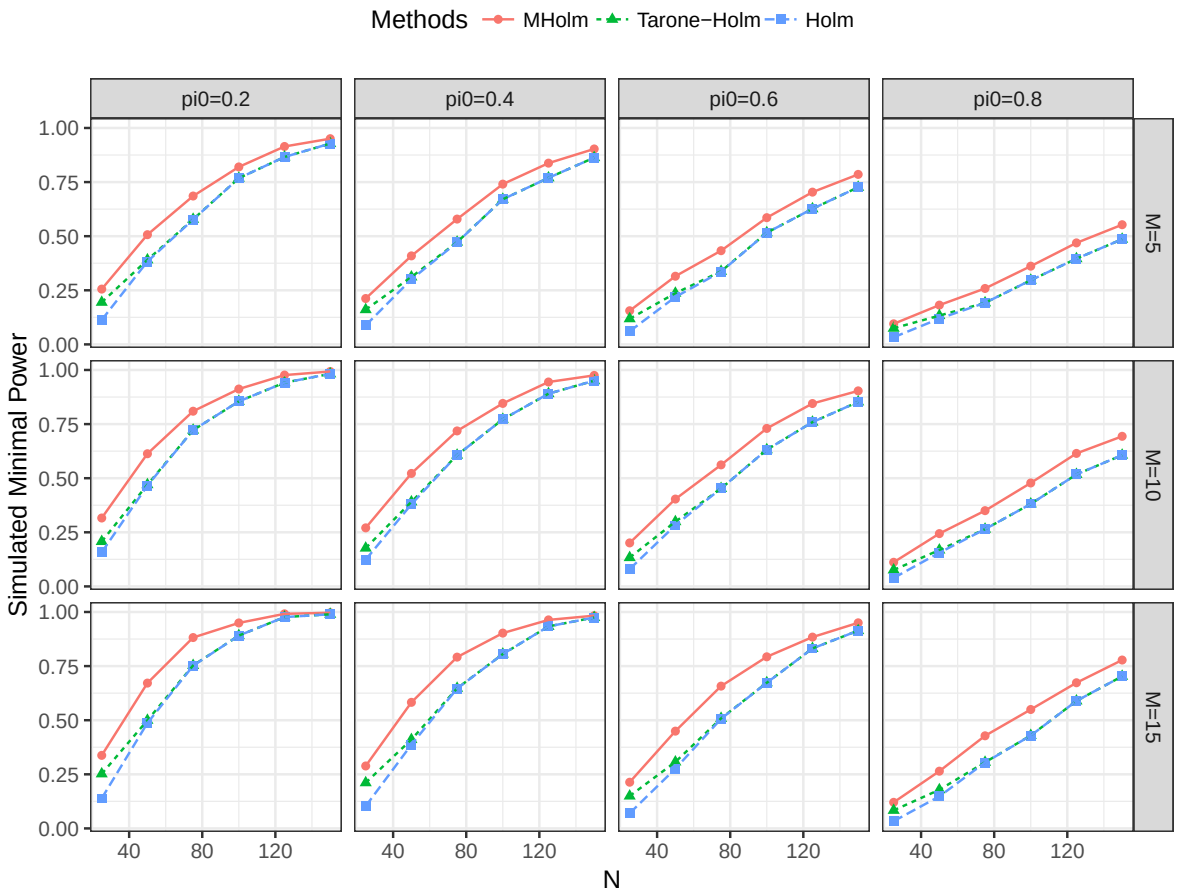

Figure 2.4 Simulated minimal power comparisons for different step-down procedures based on FET.

Proof. Since the test statistics have the same distribution, then true null p-values have the same domain $\mathbb{P}$ and CDF $F(\cdot)$, then for each i,

$$\begin{aligned}\alpha_i &= \max\{p \in \bigcup_{j=i}^{m} \mathbb{P}_{(j)} : \sum_{j=i}^{m} F_{(j)}(p) \leq \alpha\} \\ &= \max\{p \in \mathbb{P} : (m-i+1)F(p) \leq \alpha\} \\ &= \max\left\{p \in \mathbb{P} : p \leq \frac{\alpha}{m-i+1}\right\}.\end{aligned} \tag{2.5.1}$$

The last equation follows from the Assumption 1.1. Obviously, $\alpha_i \leq \dfrac{\alpha}{m-i+1}$. Then Procedure 2.4 also controls the FWER since Hochberg procedure controls the FWER.

To prove (ii), let $R = \max\{i : P_{(i)} \leq \dfrac{\alpha}{m-i+1}\}$ be the number of rejections using Hochberg procedure, then the critical p-value $P_{(R)}$ of the Hochberg procedure can be

written as

$$P_{(R)} = \max\{P_{(i)} : P_{(i)} \leq \frac{\alpha}{m-i+1}\} = \max\{P_i : P_i \leq \frac{\alpha}{m-R+1}\},$$

which is the critical value α_R of Procedure 2.4. That is, Procedure 2.4 has the same number of rejections as Hochberg procedure. □

Theorem 2.4. *If the true null p-values only take two attainable values between 0 and 1, then Procedure 2.4 controls the FWER under arbitrary dependence.*

Proof. Since the true null p-values only take two attainable values between 0 and 1, the domain of each p-value is

$$\mathbb{P}_i = \{p_i, 1\}, \text{ where } 0 < p_i < 1.$$

Suppose there are only two hypotheses H_1 and H_2, the corresponding available p-values are P_1 and P_2, where $P_1 \in \{p_1, 1\}$ and $P_2 \in \{p_2, 1\}$. Without loss of generality, assume $p_1 \leq p_2$. Then the critical values of Procedure 2.4 based on the definition are computed as:

$$\alpha_1 = \begin{cases} \alpha/2, & \alpha < p_1 \\ p_1, & p_1 \leq \alpha < p_1 + p_2 \\ p_2, & p_1 + p_2 \leq \alpha \leq 1 \end{cases}$$

and

$$\alpha_2 = \begin{cases} \alpha, & \alpha < p_2 \\ p_2, & p_2 \leq \alpha \geq 1. \end{cases}$$

There are three cases containing rejections based on the values of attainable p-values.

Case 1: If $P_1 = p_1, P_2 = 1$, since $P_2 = 1 > \alpha_2$, accept H_2. To reject H_1, one need to check $P_1 \leq \alpha_1$ if and only if $p_1 \leq \alpha$. *Case 2:* If $P_1 = 1, P_2 = p_2$, similarly, accept H_1. To reject H_2, one need to check $P_2 \leq \alpha_1$ if and only if $p_1 + p_2 \leq \alpha$. *Case 3:* If $P_1 = p_1, P_2 = p_2$, only need to check whether $P_1 \leq \alpha_1$ to ensure at least one rejection.

Then the maximum FWER is

$$\begin{aligned} FWER_{max} &= \Pr(P_1 = p_1, P_2 = 1) + \Pr(P_1 = p_1, P_2 = p_2) + \Pr(P_1 = 1, P_2 = p_2) \\ &\leq Pr(P_1 = p_1) + Pr(P_2 = p_2) \\ &= p_1 + p_2 \leq \alpha, \end{aligned}$$

which completes the proof. □

This step-up procedure has similar properties as those of Procedure 2.1 and Procedure 2.3 such as α-consistency and p-value monotonicity. But it provides a higher power, since a step-up procedure is uniformly more powerful than the corresponding step-down procedure using the same threshold. Real data analysis and simulation studies also show this procedure outperforms the others.

α-consistency Procedure 2.4 is an α-consistent procedure, since the critical values of these procedures are non-decreasing in α. So we have the following propositions:

Proposition 2.10. *Procedure 2.4 is an α-consistent procedure.*

Adjusted p-value The adjusted p-value of Procedure 2.4 can be obtained based on Definition 1.2.8.

Proposition 2.11 (Adjusted p-value for Procedure 2.4)**.**
If $P_{(1)} \leq \cdots \leq P_{(m)}$ are the available p-value for $H_{(1)}, \ldots, H_{(m)}$, then the adjusted p-value $\tilde{P}_{(i),MHoch}$ for corresponding hypothesis $H_{(i)}$ is

$$\tilde{P}_{(i),MHoch} = \begin{cases} F_{(m)}(P_{(m)}), & i = m \\ \min\left\{ \tilde{P}_{(i+1),MHoch}, \sum_{j=i}^{m} F_{(j)}(P_{(i)}) \right\}. & i = m-1, \ldots, 1 \end{cases}$$

p-value monotonicity Similar as previous argument in Proposition 2.4, we can show that Procedure 2.4 are p-value monotone.

2.5.2 Applications for Step-up Procedures

cDNA transcripts We also compare the proposed step-up Procedure 2.4 with traditional Hochberg procedure using the previous cDNA transcripts example. We also use their adjusted p-values to make decisions. The results are shown in Table 2.5.

Table 2.5 A Comparison of Adjusted p-values for the Hochberg Procedure, Procedure 1.5 and Procedure 2.4 when Testing the Hypotheses in the cDNA Transcript Example from Hommel and Krummenauer (1998)

(i)	X_{0i}/N_{0i}	X_{1i}/N_{1i}	$P_{(i)}$	$\tilde{P}_{(i),Hochberg}$	$\tilde{P}_{(i),Roth}$	$\tilde{P}_{(i),MHoch}$
(1)	1/10	8/11	0.0058	0.0552	0.0117	0.0097
(2)	0/8	5/7	0.0070	0.0559	0.0140	0.0109
(3)	0/11	4/10	0.0351	0.2456	0.1765	0.0944
(4)	1/11	3/9	0.2167	0.7118	0.7118	0.4268
(5)	2/11	4/10	0.2678	0.7118	0.7118	0.6347
(6)	1/10	3/10	0.2910	0.7118	0.7118	0.7118
(7)	2/9	2/8	0.6647	0.7118	0.7118	0.7118
(8)	2/9	2/9	0.7118	0.7118	0.7118	0.7118
(9)	2/9	2/9	0.7118	0.7118	0.7118	0.7118

The results in Table 2.5 show that for hypotheses $H_{(1)}, \ldots, H_{(5)}$, the adjusted p-values of Procedure 2.4 are smaller than those of Hochberg and Roth procedure. It means Procedure 2.4 has more chances to reject $H_{(1)}, \ldots, H_{(5)}$ than Hochberg and Roth procedure, which implies our proposed Procedure 2.4 could be more powerful than Hochberg procedure.

Clinical safety data We also compare these step-up procedures using the previous clinical safety data example.

Table 2.6 shows for each AE with the hypotheses $H_{(1)}, \ldots, H_{(3)}$, the adjusted p-values of Procedure 2.3 are smaller than those of Holm and Tarone-Holm procedures. It means Procedure 2.3 has more chances to reject $H_{(1)}, \ldots, H_{(3)}$ than the other two

Table 2.6 A Comparison of Adjusted p-values for the Hochberg Procedure, Procedure 1.5 and Procedure 2.4 when Testing the Hypotheses for AE types of Body System 10 in the Clinical Safety Data Example from Mehrotra and Heyse (2004), where the Numbers of Patients for Two Groups Are $N_1 = 148$ and $N_2 = 132$

(i)	X_{1i}	X_{2i}	$P_{(i)}$	$\tilde{P}_{(i),Hochberg}$	$\tilde{P}_{(i),Roth}$	$\tilde{P}_{(i),MHoch}$
(1)	13	3	0.0209	0.1880	0.0836	0.0534
(2)	8	1	0.0388	0.3103	0.1552	0.0982
(3)	4	0	0.1248	0.8734	0.7246	0.5050
(4)	0	2	0.2214	1.0000	1.0000	1.0000
(5)	6	2	0.2885	1.0000	1.0000	1.0000
(6)	2	0	0.4998	1.0000	1.0000	1.0000
(7)	1	2	0.6033	1.0000	1.0000	1.0000
(8)	4	2	0.6872	1.0000	1.0000	1.0000
(9)	2	1	1.0000	1.0000	1.0000	1.0000

procedures, which implies our proposed Procedure 2.3 could be more powerful than other two.

2.5.3 Simulation Studies for Step-up Procedures Comparisons

We now present simulation studies comparing the new step-up FWER controlling procedure (Proc 2.4) with Hochberg procedure. We only show the simulations based on Fisher's Exact Test, since simulations using Binomial Exact Test have similar patterns. The results of comparisons are shown in Tables A.7 and A.8. The results show that Procedure 2.4 controls FWER and is universally more powerful than Hochberg Procedure. We also plot the simulation results in Figures 2.5 and 2.6 for the FWER and minimal power comparisons

2.5.4 Simulation Studies for the Dependence Settings

We have performed the simulation studies for stepwise FWER controlling procedures when p-values are independent. Now we focus on the dependence case. Since it is

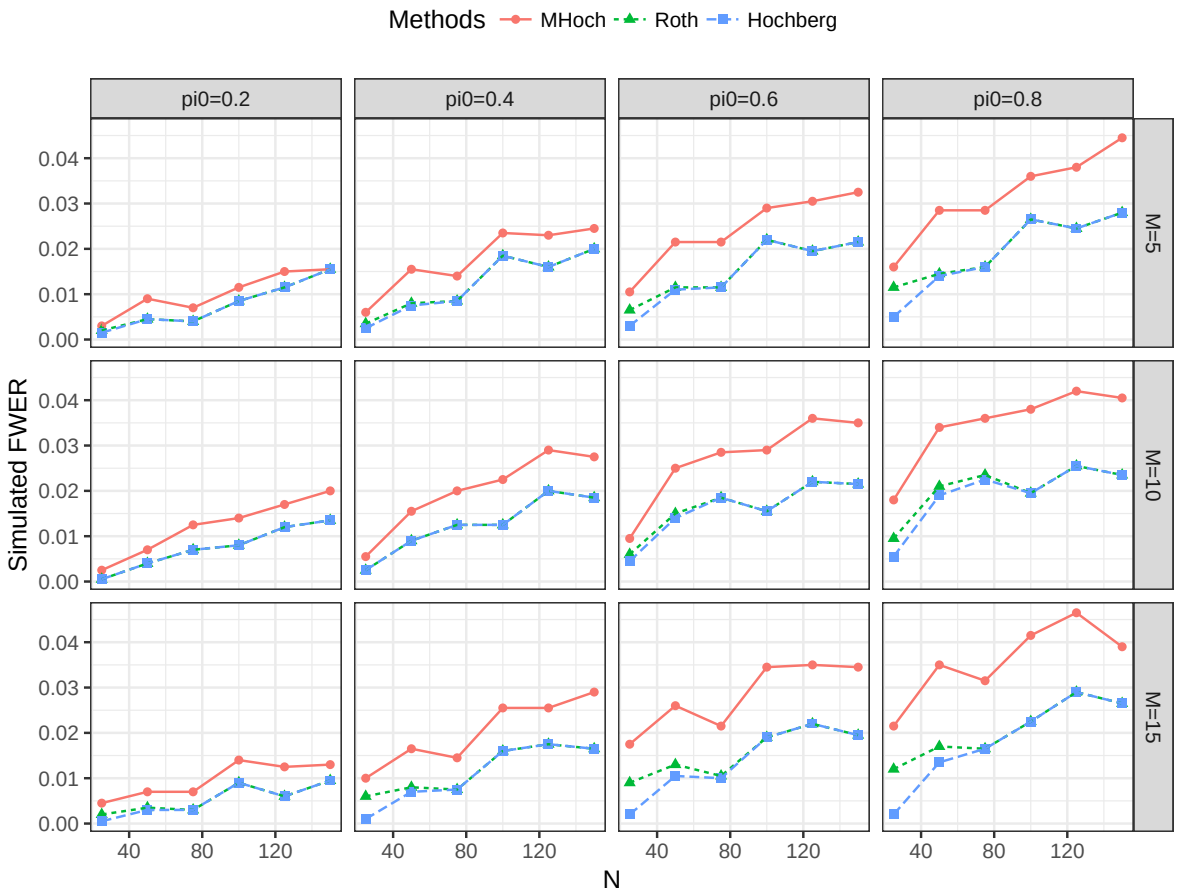

Figure 2.5 Simulated FWER comparisons for different step-up procedures based on FET.

not easy to construct an arbitrary dependence structure for discrete test statistics, we construct a special blocking dependence for the simulation.

Let $Poi(\lambda)$ denote the Poisson distribution with mean λ, $Bin(n, p)$ denote the binomial distribution with probability of success p and number of trials n. For the BET, let x_{1i} and x_{2i} be the observed counts from two independent Poisson distributions with mean λ_{1i} and λ_{2i}, where $i = 1, \ldots, m$. Then test $H_i : \lambda_{1i} = \lambda_{2i}$ versus $H_i' : \lambda_{1i} < \lambda_{2i}$, where $i = 1, \ldots, m$. The test statistic T_i for each i is based on the total $c_i = x_{1i} + x_{2i}$ is $Bin(c_i, \theta_i)$, where $\theta_i = \lambda_{1i}/(\lambda_{1i} + \lambda_{2i})$. (Lehman and Romano, 2005) Then, under the null hypothesis $H_i : \lambda_{1i} = \lambda_{2i}$, T_i follows binomial distribution $Bin(c_i, 0.5)$ and its distribution only depends on c_i. Then the following simulations are conducted based

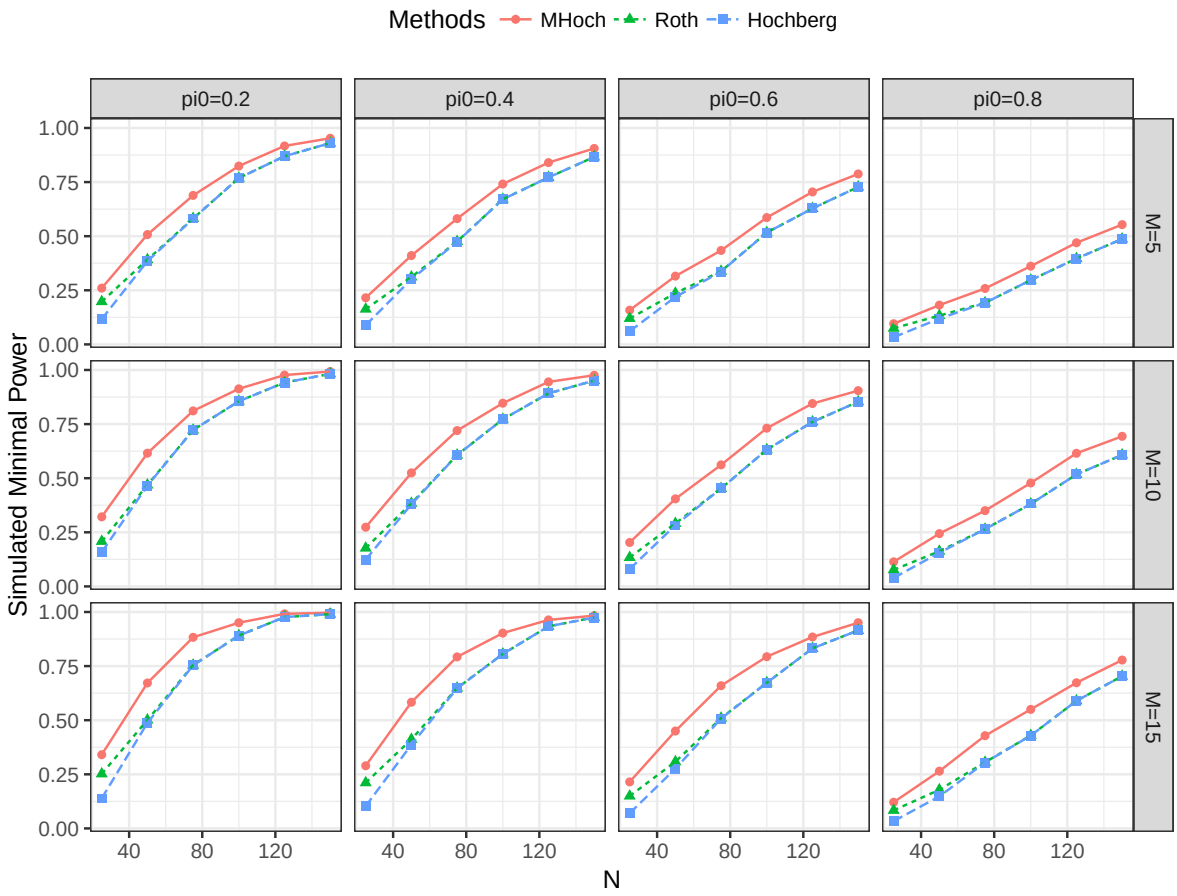

Figure 2.6 Simulated minimal power comparisons for different step-up procedures based on FET.

on dependent binomial exact test (BET) statistics. The details for generating the dependent simulation data can be found in Appendix A.

In the simulation, set the number of hypotheses $m = \{5, 10\}$, with true null proportion $\pi_0 = \{0.4, 0.6, 0.8\}$ respectively. Set the mean parameter of Poisson response in each group as $\lambda_{1i} = \lambda_{2i} = 2$ for $i = 1, \ldots, m_0$, and set the mean parameter for group 1 as $\lambda_{1i} = 2$, and for group 2 as $\lambda_{2i} = 10$ for $i = m_0 + 1, \ldots, m$, where $m_0 = \pi_0 m$. Then compute the simulated FWER, minimal power to compare the different procedures by taking average of $B = 2000$ iterations.

Remark 2.4. Note that the above simulated p-values have joint dependence within the group of true null hypotheses, but the p-values corresponding to true null hypotheses

are independent of the false ones, which is the same as condition (3.5) in Romano and Shaikh [58]. This setting also satisfies PRDS condition.

The simulation results comparisons for stepwise procedures (single-step, step-down, and step-up) are displayed in Figures 2.7 to 2.12. More simulation results for different scenarios can be found in Appendix A.

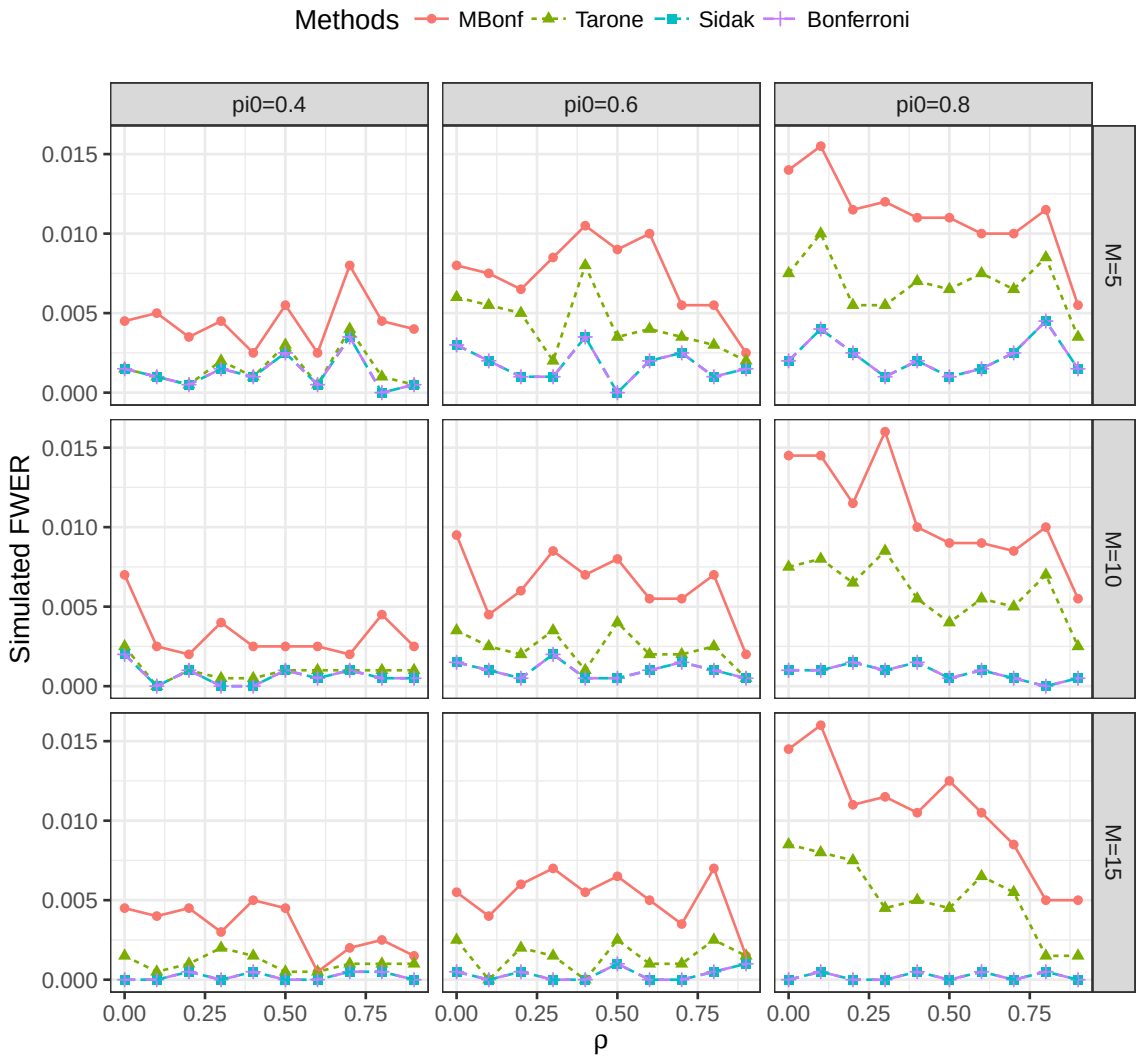

Figure 2.7 Simulated FWER comparisons for different single-step procedures based on the blocking dependent BET.

From the simulation results, we can see Procedure 2.1, 2.3 and 2.4 control FWER under the significant level $\alpha = 0.05$. Under different settings of the correlation among the p-values, Modified Bonferroni procedure is more powerful than Tarone's and

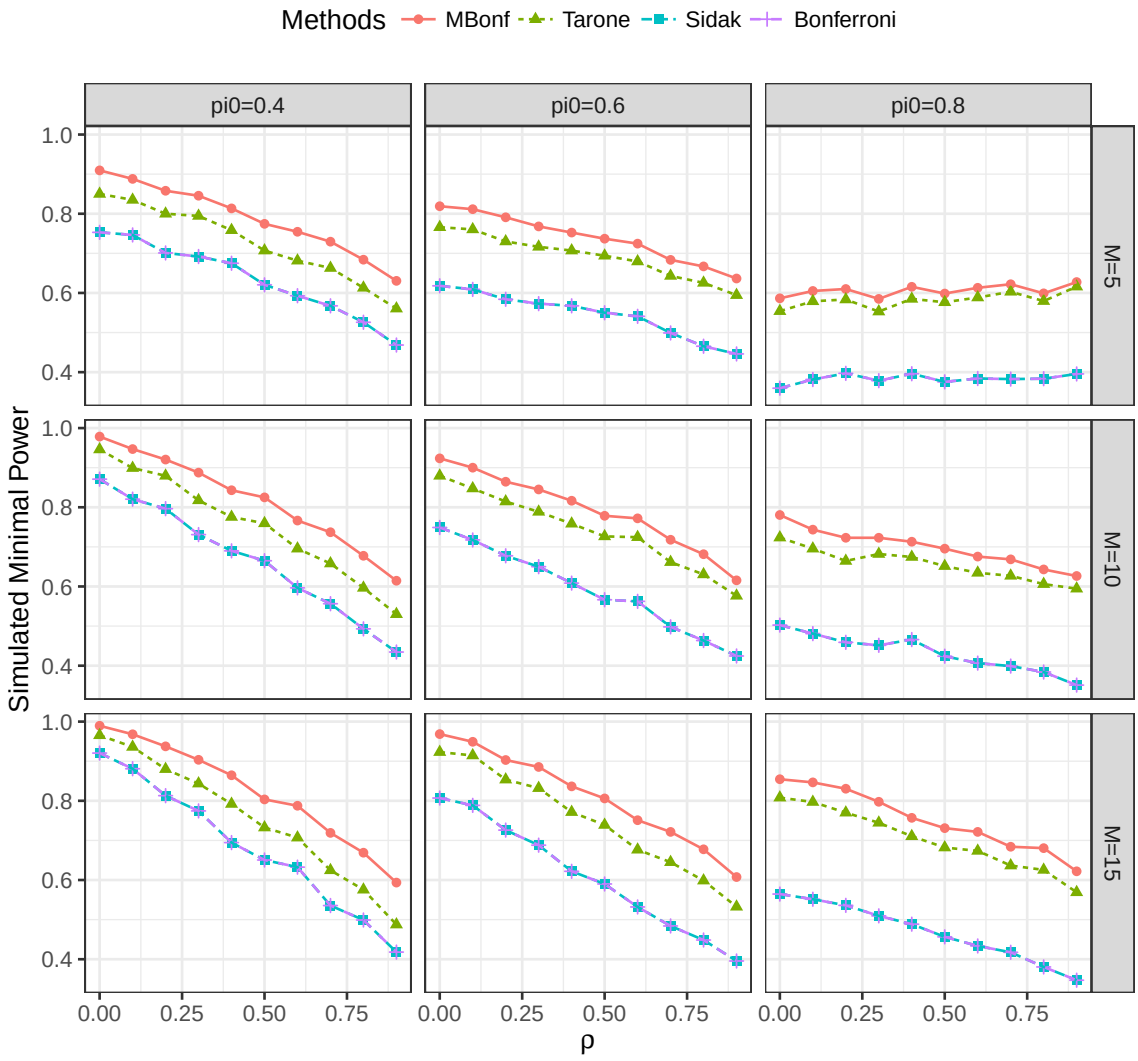

Figure 2.8 Simulated minimal power comparisons for different single-step procedures based on the blocking dependent BET.

Bonferroni procedures, Modified Holm procedure is more powerful than Tarone-Holm and Holm procedures, and Modified Hochberg procedure is more powerful than Roth's and Hochberg procedures.

2.6 Conclusions and Discussion

In this chapter, we have developed several FWER controlling procedures for discrete data by exploiting the information of discreteness for test statistics. The proposed Procedure 2.1 and Procedure 2.3 control FWER under arbitrary dependence, Procedure

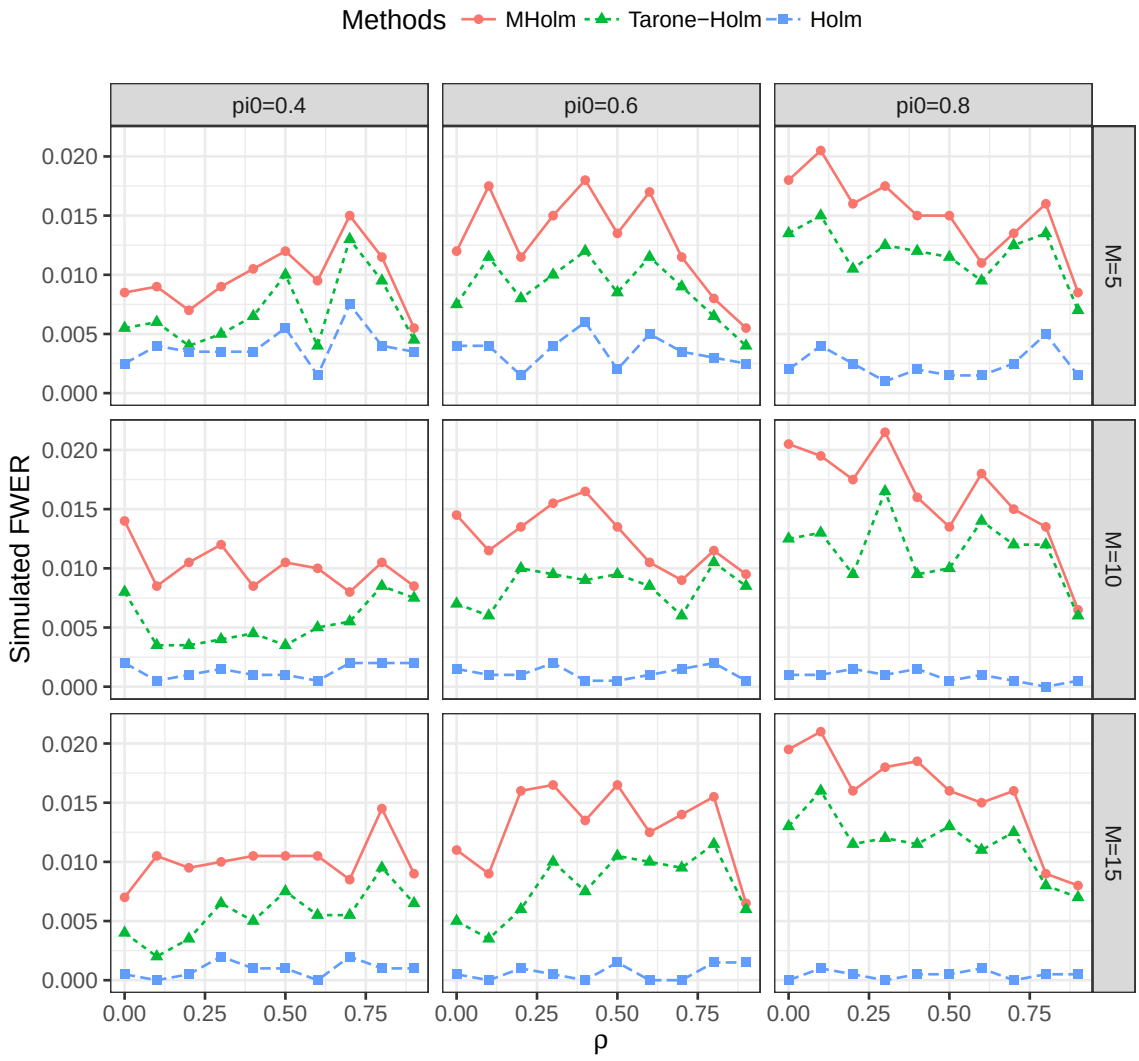

Figure 2.9 Simulated FWER comparisons for different step-down procedures based on the blocking dependent BET.

2.4 controls FWER under PRDS condition for some specific distributions settings. Real data analysis in both clinical safety studies and cDNA transcript data reveals that the proposed procedures have more rejections than conventional procedures. The simulation studies show that when the proportion of true null hypotheses is large, which is usually the case in practical applications, the proposed stepwise procedures can outperform the corresponding Bonferroni, Holm and Hochberg procedures and even better than some existing discrete procedures, such as Tarone and Tarone-Holm procedures in terms of minimal power.

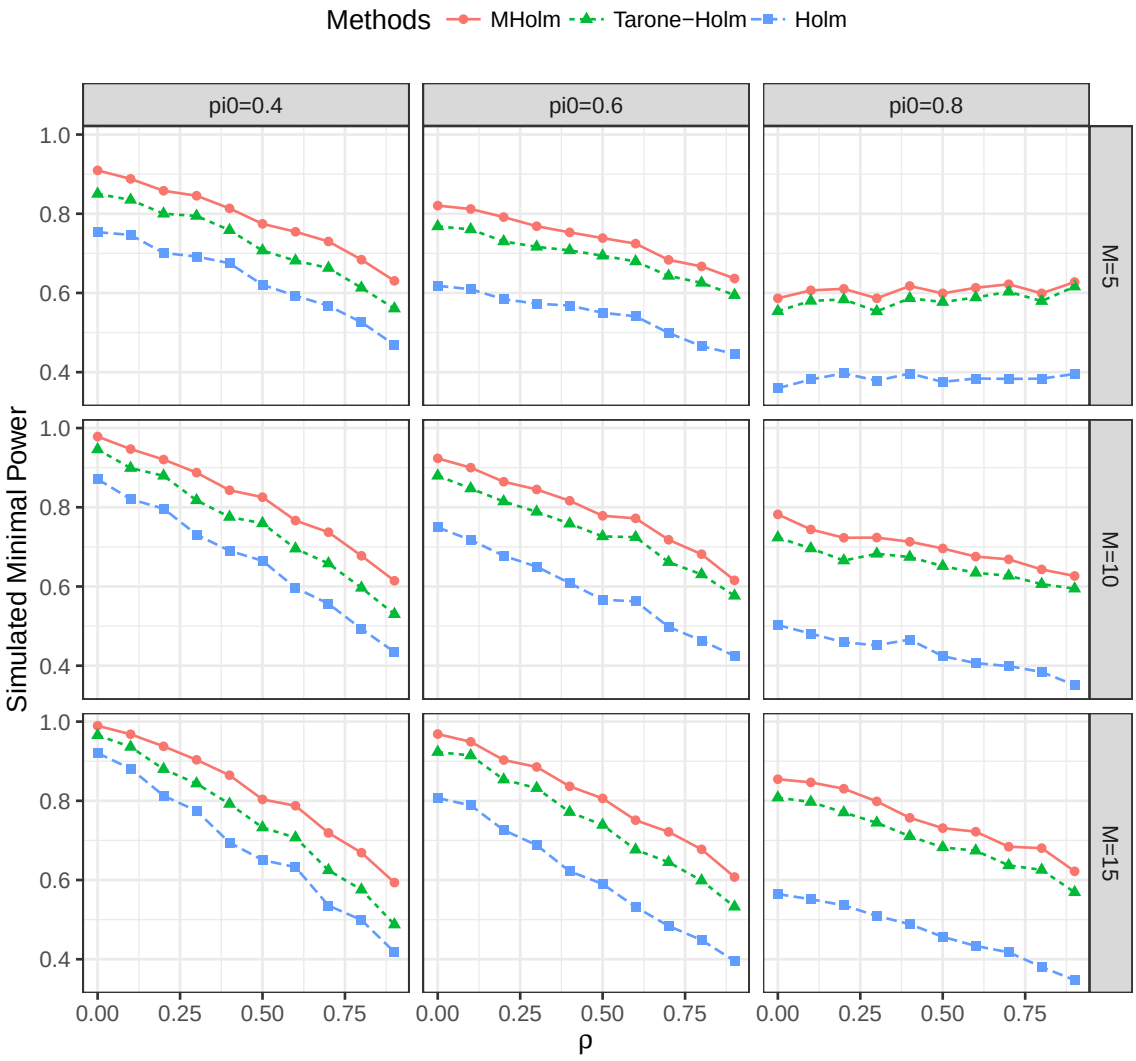

Figure 2.10 Simulated minimal power comparisons for different step-down procedures based on the blocking dependent BET.

A possible future work is to explore optimality of the suggested Procedure 2.1 and 2.3 under arbitrary dependence, which means for some joint distribution of the discrete p-values, one cannot increase even one of the critical constants while keeping the remaining fixed without losing control of the FWER. Another possible future work is to incorporate some data driven weights into the proposed procedures to develop more powerful FWER controlling procedures for discrete data.

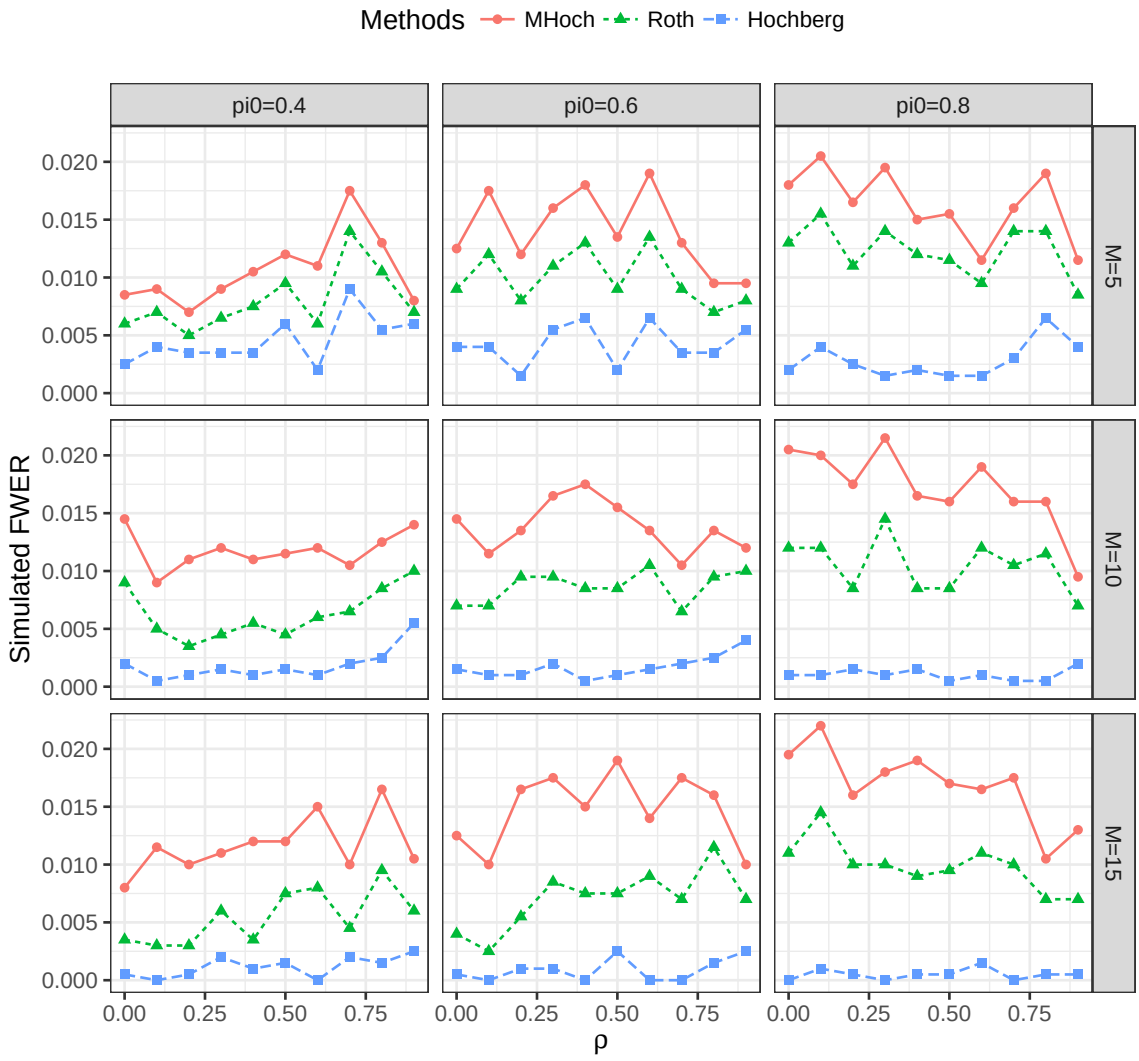

Figure 2.11 Simulated FWER comparisons for different step-up procedures based on the blocking dependent BET.

2.7 Software

The FWER controlling procedures for discrete data described in this chapter have been implemented as a part of the MHTdiscrete R package [Zhu and Guo, 2017], which is available online at https://cran.r-project.org/web/packages/MHTdiscrete. A web application for the proposed procedures and most existing FWER and FDR controlling procedures is developed at https://allen.shinyapps.io/MTPs.

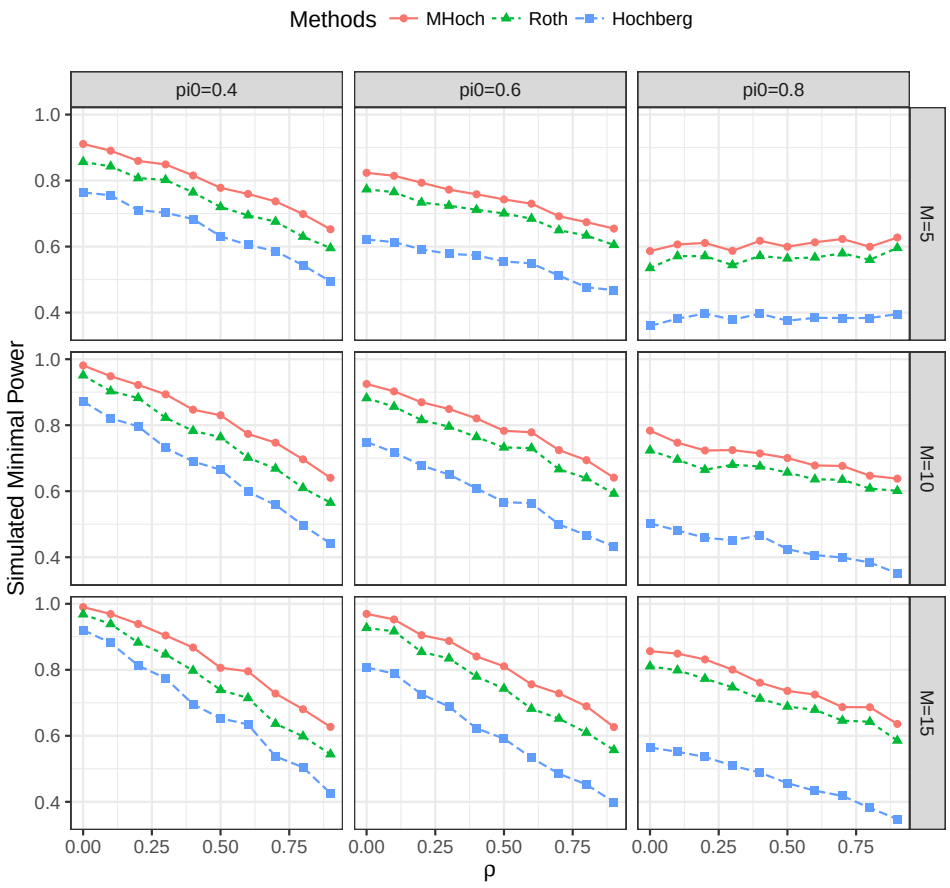

Figure 2.12 Simulated minimal power comparisons for different step-up procedures based on the blocking dependent BET.

CHAPTER 3

SELECTIVE INFERENCE IN CLINICAL SAFETY STUDIES

3.1 Introduction

In clinical safety studies, there are many adverse events (AE) recorded in one clinical trial. The goal for assessing the safety of an experimental drug is to flag "reasonable" or "correct" AEs among these AE types. Chen et al. (2015) [13] summarized most existing signal detection methods, such as proportional reporting ratios [19], reporting odds ratios [60], and the maximum likelihood ratio test [41], Bayesian confidence propagation neural network method [2] and multi-item Gamma Poisson shrinker [66] for spontaneous reporting data, and Pearsons chi-square test, Fishers exact test, and the chi-square test for rates comparison for flagging safety signal in clinical trials. The above detecting or flagging methods do not control for overall type 1 error rates, such as FWER or FDR. In fact, the number of possible AEs is usually very large (see TCTAE v4.03 and MedDRA® v19.1). Thus, similar as dealing with multiple endpoints in drug efficacy analysis, multiplicity effect should be also considered in drug safety analysis. However, the number of AEs in safety analysis is much larger than the number of endpoints in efficacy analysis for the experimental drugs. Therefore, FWER controlling procedures such as Bonferroni procedure may fail to flag more important AEs. Benjamini and Hochberg (1995) introduced the concepts of false discovery rate (FDR), which is defined as the expected proportion of the false rejections among all discoveries. Their proposed BH step-up procedure becomes the most popular multiple testing procedure (MTP) for the large-scale multiple hypotheses testing in the last two decades. The BH procedure can be applied to detect the signals of the AEs, since the number of AEs in clinical safety studies is usually large.

Searching for significant AEs, the AE types (Preferred Term (PT) in MedDRA®) are often classified by several body systems (BS) (System Organ Classes (SOC) in

MedDRA®). Each AE can be regarded as a hypothesis, and the hypotheses of AEs from the same body system naturally forms a family. So the multiple-family structure should be considered for the drug safety data analysis. Recently, some structured BH-type procedures are developed for multiple families of hypotheses (Mehrotra and Heyes, 2004; Mehrotra and Adewale, 2012; Hu et al., 2010; Benjamini and Bogomolov, 2014). Mehrotra and Heyes (2004) proposed a two-stage double FDR (DFDR) procedure, which firstly uses BH procedure on the minimum original p-values of each family under level α_1, then applies the BH procedure on the hypotheses in the selected families under level α. The problem is this procedure cannot guarantee FDR control. It is also not clear for how to choose the significant level α_1 in the first step. Mehrotra and Adewale (2012) modified the DFDR procedure by using BH procedure in the first step on the minimum BH-adjusted p-value under the same significant level α as the second step. The DFDR2 procedure still cannot guarantee FDR control. The main reason is the procedure does not consider the selection effect for the first step. Other recent references also examined related questions for multiple-family multiple hypotheses testing. Hu et al. (2010) introduced a p-value weighting group BH procedure (GBH) by estimating the true null proportion for each group, the method asymptotically controls global FDR. Benjamini and Bogomolov (2014) provides a general framework for multiple families multiple testing considering selection effect, and defined average FDR and average FWER over the selected families as desired type 1 error rates. Their proposed procedure (BB) can guarantee the average FDR control. Actually, the DFDR2 procedure and original BB procedure select the same families by using the same selection rule in the first step, the main difference between these two procedures is original BB procedure uses $R\alpha/m$ to conduct individual test in each selected family. Barber and Ramdas (2016) proposed a multilayer FDR controlling procedure, which can guarantee FDR control on family level and control global FDR, but they do not consider any error rate control within selected families. The above methods do not clearly separate the selection effect and multiplicity effect. To overcome this problem, selective inference

by using conditional inference such as conditional type 1 error rate control, selection adjusted confidence interval is developed recently (Fithian et al., 2015; Weinstein et al., 2013; Heller et al., 2016).

In practice, especially in clinical safety studies, we may not only need to flag the AE, but also want to investigate the body systems for the further research (Berry and Berry, 2004). For example, if some body systems are selected, but there is no AE flagged in these selected body systems, we can also use the body system information to conduct follow-up studies. Thus controlling some type 1 error in family level is also necessary. Since the number of body systems is not very large (commonly 5-50), and sometimes we may allow more than one type 1 errors made when doing selection. Thus, the generalized familywise error rate (k-FWER) is a suitable error measure in practice, which is the probability of making at least k false rejections. Some existing k-FWER controlling procedure include Lehmann and Romano (2005), Guo and Romano (2007), Sarkar (2006, 2007). And this selection rule using k-FWER controlling procedure is also a simple selection rule. We can conclude some similar conclusions for selective inference within the selected families as Benjamini and Bogomolov (2014) and Heller et al. (2016).

So far, there are growing literature of approaches for testing multiple hypotheses with multiple families structure. But very few valid methods were used in clinical safety adverse events. So this chapter aims to develop a selective inference-based two-stage procedure in the settings of clinical safety studies. We consider using combining methods on the conditional p-values, as described in the preliminaries in Section 3.2. Some existing multiple testing procedures for multiple families are introduced for clinical safety settings as well. The valid two-stage procedure is proposed in Section 3.3. Theoretical results of our proposed procedures are provided in Section 3.4. Then we discuss the selection rules effect in Section 3.5. In Section 3.6, simulation studies are also conducted to compare the proposed procedure with other existing procedures and compared the proposed procedure for different selection rules regarding choices for

combining methods and generalized FWER fold k. Section 3.7 explores several clinical safety examples to illustrate the proposed procedure and compare the outcomes with other existing multiple families multiple testing procedures. Some concluding remarks are made in Section 3.8 and one R package "MHTmult" is implemented in Section 3.9. Proofs of some results are given in the Appendix B.

3.2 Preliminaries

In this section, some necessary notations and basic concepts are introduced. In clinical safety studies, several body systems contain amount of AE types, the body systems are regarded as multiple families, AE types classified by body systems forms individual hypotheses. Therefore, flagging significant AE in some body systems can be formulated as a multiple testing problem with hierarchical multiple families structure.

3.2.1 Notations

Suppose that there are $n = \sum_{i=1}^{m} n_i$ AE types, denoted by AE_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, appearing in m body systems denoted by $BS_1, \ldots, BS_m$. The hypotheses H_{ij} to flag AE_{ij} are to be simultaneously tested based on their corresponding p-values P_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. Let $\tilde{H}_i$ be the global null hypothesis for i-th body system, that is, $\tilde{H}_i = \bigcap_{j=1}^{n_i} H_{ij}$. The global hypothesis is true only if all H_{ij} in this body system are true. Let $\tilde{P}_i$ denote the p-value of the corresponding global hypothesis $\tilde{H}_i$.We use the global test to select body systems. Let $\mathcal{S}$ denote the index set of selected hypotheses, that is, $\mathcal{S} = \{i : \tilde{H}_i \text{ is rejected}\}$

Assume individual p-values are stochastically greater than Uniform(0,1), that is, P_{ij} satisfies

$$\Pr\{P_{ij} \leq t\} \leq t, \text{ for } t \in (0, 1), \tag{3.2.1}$$

then the global p-value $\tilde{P}_i$ satisfies

$$\Pr\{\tilde{P}_i \leq \tilde{t}\} \leq \tilde{t}, \text{ for } \tilde{t} \in (0, 1). \tag{3.2.2}$$

3.2.2 Several Type 1 Error Rates

The most commonly used type 1 error rates are familywise error rate (FWER) and false discovery rate (FDR), which are defined in Chapter 1. Here we specifically introduce some type 1 error rates for multiple families structure.

Type 1 error rate for family level In the clinical safety studies, the number of body systems is often moderate (5-50). Based on the requirement of future research, the error measurement is usually not as strict as allowing to make only one type I error (falsely selecting one body system). Sometimes, the clinical studies would like to control the probability of falsely selecting at least k body systems under 5%-10%, where $1 \leq k \leq m$. Thus when selecting body systems of interest (BSoI), controlling the generalized FWER for body system level is desired. Let $\tilde{V}$ denote the number of false selections when selecting BSoI using global p-values, then generalized FWER is defined as

$$k\text{-FWER} = \Pr(\tilde{V} \geq k)$$

If $k = 1$, k-FWER will become FWER. The commonly used k-FWER procedures are generalized Bonferroni and generalized Sidak procedure.

Procedure 3.1 (Generalized Sidak procedure)**.** *Reject H_i if $p_i \leq \tilde{t}$, where $\tilde{t}$ satisfies* $\sum_{l=k}^{m} \binom{m}{l} \tilde{t}^l (1-\tilde{t})^{m-l} = \alpha_1$.

Type 1 error rate for individual level Considering the selection effect, it is valid to use conditional inference for flagging AEs in testing step. Then define the FWER and FDR measures conditional on the selected BSoI as follows.

Let V_i denote the number of false rejections for i-th body system, and R_i denote the number of total rejections for i-th family. Then the conditional FWER selected i-th family is defined as

$$\text{cFWER}_i = \Pr(V_i > 0 | i \in \mathcal{S}) = E\{I(V_i > 0) | i \in \mathcal{S}\},$$

and the conditional FDR for a selected body system is defined as

$$\text{cFDR}_i = E\left\{\frac{V_i}{R_i \vee 1}\Big| i \in \mathcal{S}\right\}.$$

The advantage of conditional error measure is it can distinguish the selection effect introduced in the first stage from the multiplicity effect in the second stage. We can use this information to develop more powerful procedure than existing procedure. In practice, t is preferred to control cFDR for flagging AE within the selected body systems in clinical safety studies, since the number of AEs is very large, cFWER control is too strict.

Overall type 1 error rate An overall error measure addressing selective inference is the average FDR over selected families, which is defined as

$$\text{average-FDR} = E\left\{\frac{\sum_{i\in\mathcal{S}} \frac{V_i}{R_i \vee 1}}{|\mathcal{S}| \vee 1}\right\},$$

which is expected average of false rejection proportions across selected families.

Another overall error measure is global FDR, which is used in several references such as Hu et al. (2010), Guo and Sarkar (2016), Barber and Ramdas (2016).

$$\text{global-FDR} = E\left\{\frac{\sum_{i\in\mathcal{S}} V_i}{\sum_{i\in\mathcal{S}} R_i \vee 1}\right\}$$

Remark 3.1. We can observe if the numbers of total rejections are the same, the average-FDR and global-FDR are equivalent. Otherwise, these two overall FDR measures are different. The procedures considering average FDR control often give up the global FDR control (Benjamini and Bogomolov (2014)). It makes more sense to consider average FDR control if the selected families are heterogeneous, such as the body systems in clinical safety studies are functionally different regarding the AE response for the experimental drugs.

3.2.3 Several Existing Two-stage Multiple Testing Procedures

The original double FDR (DFDR) procedure and modified double FDR (DFDR2) procedure are proposed in clinical safety settings, but original Benjamini and Bogomolov (BB) procedure and its modification are proposed in GWAS settings, which are much larger scale settings.

The original double FDR (DFDR) procedure proposed by Mehrotra and Heyse (2004) is described as follows.

Procedure 3.2 (DFDR)**.**

Step 1: *(a) For each body system, find the minimum p-value* $\tilde{p}_i = \min_{1\le j\le n_i}\{p_{ij}\}$.
(b) Apply BH procedure on $\tilde{p}_1, \ldots, \tilde{p}_m$ *at level* α_1 *to select BSoI.*

Step 2: *In the i-th selected body system, apply BH procedure on* $p_{i1}, \ldots, p_{in_i}$ *at level* α *to flag AEs.*

Note that the minimum p-values $\tilde{p}_i$ are not uniformly distributed any more, so the BH procedure applying on those minimum p-values cannot strongly control FDR. The modified double FDR (DFDR2) procedure developed by Mehrotra and Adewale (2012) replaces the minimum p-value by minimum BH-adjusted p-value for each body system.

Procedure 3.3 (DFDR2)**.**

Step 1: *(a) For each body system, compute the minimum BH-adjusted p-value as* $\tilde{p}_i = \min_{1\le j\le n_i}\{p_{ij}^{BH\text{-}adj}\}$.
(b) Apply BH procedure on $\tilde{p}_1, \ldots, \tilde{p}_m$ *at level* α *to select BSoI.*

Step 2: *In the i-th selected body system, apply BH procedure on* $p_{i1}, \ldots, p_{in_i}$ *at level* α *to flag AEs.*

However, since it still uses α as in Step 2, which brings selection bias in testing step. Then the DFDR2 still cannot control average FDR or global FDR under α.

In R software, use p.adjust() function to compute the BH-adjusted p-values within each body system, and find the minimum one. Note that the minimum BH-adjusted

p-value is equivalent to Simes global null p-value $\min_{1\leq j\leq n_i}\{\frac{n_i}{j}p_{i(j)}\}$, which is a valid global p-value uniformly distributed in [0,1].

Benjamini and Bogomolov (2014) recommended to use the BH procedure on the set of minimum BH-adjusted p-values as a simple selection rule, then in each selected family apply BH procedure at an selection-adjusted level which is less than the nominal significant level, and determined by number of selected families.

Procedure 3.4 (Original BB: BB-α-BH-α)**.**

Step 1: (a) *For each body system, compute the minimum BH-adjusted p-value as* $\tilde{p}_i = \min_{1\leq j\leq n_i}\{p_{ij}^{BH\text{-}adj}\}$.

(b) *Apply BH procedure on* $\tilde{p}_1,\ldots,\tilde{p}_m$ *at level* α *to select BSoI.*

Step 2: (a) *Count the number of selected body systems as* $|\mathcal{S}|$.

(b) *In the i-th selected body system, apply BH procedure on* $p_{i1},\ldots,p_{in_i}$ *at level* $\frac{|\mathcal{S}|}{m}\alpha$ *to flag AEs.*

It can be seen that the original BB procedure proceeds the same selecting step (Step 1) as DFDR2 procedure, the only difference is in testing step (Step 2), original BB procedure uses a adjusted significant level $\frac{|\mathcal{S}|}{m}\alpha$, which is smaller than α.

Peterson et al. (2016) modified the original BB procedure (BB-α-BH-α_1), which applies BH procedure at level α_1 to select BSoI. The selection level α_1 is not necessarily equal to testing level α.

Although original BB procedure and its modification guarantee average FDR control, they both select body systems using BH procedure, which is too liberal for moderate number of body systems in clinical safety studies.

Procedure 3.5 (Group BH-TST estimation)**.**

Step 1: (a) *For each body system, using Benjamini et al. (2006) two-stage estimator, compute the estimated true null proportion* $\hat{\pi}_{0i} = \hat{n}_{0i}/n_i$ *for i-th body system.*

(b) *Compute weighted p-value for each AE as* $p_{ij}^w = \min\left\{\frac{\hat{\pi}_{0i}}{1-\hat{\pi}_{0i}}p_{ij}, 1\right\}$.

Step 2: *(a) Compute the pooled estimated true null proportion* $\hat{\pi}_0 = \frac{\sum_{i=1}^{m} \hat{n}_{0i}}{n}$.

(b) Pool the weighted p-value together as $p_1^w, \ldots, p_n^w$*, then apply BH procedure on the pooled p-values at level* $\dfrac{\alpha}{(1+\alpha)(1-\hat{\pi}_0)}$ *to flag AEs.*

3.2.4 Combining Functions and Conditional p-values

Since we are dealing with a selected family of hypotheses based on the global test, we will omit the index i for family to simplify the notation. Also use n replace n_i to denote the number of hypothesis within a family.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a combining function for testing the global null on each family. Here we require f is a monotone function of each p_j. Let t denote the fixed hypothesis/family selection threshold, i.e. $f(P_1, \ldots, P_n) \geq t$ for non-increasing f or $f(P_1, \ldots, P_n) \leq t$ for non-decreasing f. Let b_j be an inflation factor to determine the conditional p-value p'_j satisfying $f(\mathbf{p}^{(-j)}, b_j) = t$ for $b_j \in (0, 1]$, where $\mathbf{p}^{(-j)} = p_1, \ldots, p_{j-1}, p_{j+1}, \ldots, p_n$. Then we can define conditional p-value as

$$P'_j := P_j | f(\mathbf{p}^{(-\mathbf{j})}, P_j) \leq t = \frac{P_j}{b_j}$$

The following are three typical examples for calculating inflation factor and conditional p-values using Fisher's, Stouffer's and Tippet's combining methods.

For Fisher's combining method, the family will be selected if and only if the p-values of this family satisfy $-2\sum_{j=1}^{n} \log p_j \geq t$, which implies $p_j \prod_{l=1(\neq j)}^{n} p_l \leq e^{-t/2}$, then within the family, the inflation factor of p_j will be

$$b_j = \begin{cases} \dfrac{e^{-t/2}}{\prod_{l=1(\neq j)}^{n} p_l} & \text{if } \prod_{l=1(\neq j)}^{n} p_l > e^{-t/2} \\ 1 & \text{otherwise.} \end{cases}$$

So the conditional p-value can be calculated as follow,

$$p'_j = \begin{cases} \dfrac{\prod_{l=1}^{n} p_l}{e^{-t/2}} & \text{if } \prod_{l=1(\neq j)}^{n} p_l > e^{-t/2} \\ p_j & \text{otherwise.} \end{cases}$$

For Stouffer's combining method, the family will be selected if and only if the p-values of this family satisfy $\dfrac{\sum_{j=1}^{n} z_j}{\sqrt{n}} = \dfrac{\sum_{j=1}^{n} \Phi^{-1}(1-p_j)}{\sqrt{n}} \geq t$, which implies $\Phi^{-1}(1-p_j) + \sum_{\substack{l=1 \\ (\neq j)}}^{n} \Phi^{-1}(1-p_l) \geq \sqrt{n}t$. Then within the selected family, the inflation factor of p_j will be $b_j = 1 - \Phi(\sqrt{n}t - \sum_{\substack{l=1 \\ (\neq j)}}^{n} \Phi^{-1}(1-p_l))$. Therefore, the conditional p-value can be calculated as

$$p'_j = \frac{p_j}{1 - \Phi(\sqrt{n}t - \sum_{\substack{l=1 \\ (\neq j)}}^{n} \Phi^{-1}(1-p_l))}.$$

For Tippet's (minP) combining method, the family will be selected if and only if the p-values of this family satisfy $\min\{p_1, \ldots, p_n\} \leq t$, that is, $\min\{p_j, \mathbf{p}^{(-j)}\} \leq t$. So we call this method as minP combining method. Then within the selected family, the inflation factor of p_j will be

$$b_j = \begin{cases} t & \text{if } \min\{\mathbf{p}^{(-j)}\} > t \\ 1 & \text{otherwise.} \end{cases}$$

The conditional p-value can be calculated as

$$p'_j = \begin{cases} \dfrac{p_j}{t} & \text{if } \min\{\mathbf{p}^{(-j)}\} > t \\ p_j & \text{otherwise.} \end{cases}$$

Remark 3.2. If there is only one hypothesis with the p-value P in selected family, the conditional p-value follows $\Pr\{P \leq p | P \leq t\} = p/t$ for any $t > p$. According to the inflation factor b satisfies $b = t$, the conditional p-value will be $p' = p/b = p/t$. Similarly, for Fisher's combining method, the inflation factor b satisfies $-2\log b = t$, then $b = e^{-t/2}$, so the conditional p-value is $p' = p/e^{-t/2}$. For Stouffer's combining

method, the inflation factor b has to satisfy $\frac{\Phi^{-1}(1-b)}{\sqrt{n}} = t$, then $b = 1 - \Phi(t)$, so the conditional p-value is $p' = \frac{p}{1-\Phi(t)}$.

3.3 A Valid cFDR Controlling Procedure Using k-FWER Controlling Selection Rule

When we conduct the two-stage procedure, the first stage (selecting step) selects some significant body systems based on the global null p-values $\tilde{P}_1, \ldots, \tilde{P}_m$. Let $\mathcal{S} \subseteq \{1, \ldots, m\}$ denote the set of selected body systems. Selecting body systems could be controlled for generalized familywise error rate (denoted by k-FWER) with $k \geq 1$ folds under α level. A simple example is to consider generalized Bonferroni procedure selection rule, then $\mathcal{S} = \{i : \tilde{P}_i \leq \frac{k\alpha}{m}\}$.

The second stage (testing step) simultaneously tests the individual hypotheses in the selected body systems based on the individual p-values $P_{ij}, i \in \mathcal{S}, j = 1, \ldots, n_i$, to flag significant AEs, which is the final goal for the safety studies. Since testing individual hypothesis is conditional on their body systems being selected, the procedure could guarantee conditional FDR control. The two-stage selective inference-based procedure is described as follows.

Procedure 3.6 (cFDR-α-minP-k-Sidak-α_1)**.**

Step 1: (a) *For each body system, compute the global p-value* $\tilde{p}_i = n_i \min_{1 \leq j \leq n_i} \{p_{ij}\}$.

(b) *Apply generalized Sidak procedure on* $\tilde{p}_1, \ldots, \tilde{p}_m$ *at level* α_1 *to select body systems of interest (BSoI), that is, select the i-th body system if* $\tilde{p}_i \leq \tilde{t}$, *where* $\tilde{t}$ *satisfies* $\sum_{l=k}^{m} \binom{m}{l} \tilde{t}^l (1-\tilde{t})^{m-l} = \alpha_1$.

Step 2: (a) *In the i-th selected body system, calculate the conditional p-value for* H_{ij}: $p'_{ij} = \frac{p_{ij}}{t_i}$ *if* $\min_{1 \leq s \leq n_i, s \neq j} \{p_{is}\} > t_i$; *otherwise,* $p'_{ij} = p_{ij}$, *where* $t_i = \frac{\tilde{t}}{n_i}$.

(b) *Apply BH procedure on* $p'_{i1}, \ldots, p'_{in_i}$ *at level* α *to flag AEs.*

Remark 3.3. *The selection rule:* $\tilde{p}_i \leq \tilde{t}$ *is equivalent to minP combining function* $f(p_{i1}, \ldots, p_{in_i}) = \min_{1 \leq j \leq n_i} \{p_{ij}\} \leq t_i$.

Remark 3.4. A modified BB procedure using the same minP combining k-FWER controlling selection rule as Procedure 3.6 can be naturally developed. That is, the procedure uses Step 1 of the Procedure 3.6 to select body systems, and Step 2 of the Procedure 3.4 to flag AEs (BB-α-minP-k-Sidak-α_1). Such a procedure still uses a simple selection rule, so it strongly controls average FDR as the Theorem 1 in Benjamini and Bogomolov (2014) stated. We compare this modified BB procedure using the same selection rule as the proposed procedure in the simulation studies.

3.4 Theoretical Results

Since conditional p-values can be easily obtained using the above combination methods, it is natural to conduct MTPs for individual hypothesis (AE type) in the selected families (BSoI). The followings are some theoretical results for the two-stage procedure controlling conditional FDR within the selected families. Theorem 3.1, Lemma 3.1 and Theorem 3.2 still omit the index i for family to simplify the notation. Also use identical number of hypothesis within each family n replace n_i.

Theorem 3.1 (Fisher's combining selection rule)**.** *Let $P_1, \ldots, P_n$ be independent p-values with $U(0,1)$ under true null. If $f(p_1, \ldots, p_n) = -2\sum_{j=1}^{n} \log p_j \geq t$, then the BH procedure on conditional p-values $p'_1, \ldots, p'_n$ controls the cFDR at level $\frac{n_0}{n}\alpha \leq \alpha$.*

Some desired properties for conditional p-values can be found.

Note that for Fisher's combing method, we have the following desired statistical property of the conditional p-values.

Lemma 3.1 (conditional p-value monotonicity)**.** *For Fisher's combining function, if the unconditional p-value $p_1 \leq \cdots \leq p_n$, then the conditional p-value $p'_1 \leq \cdots \leq p'_n$.*

Remark 3.5. Note that the conditional p-values using minP selection rule does not always satisfy monotone property. Here is a counterexample: suppose use minP selection rule with $t = 0.05$ to select two hypotheses $p_1 = 0.04 \leq p_2 = 0.06$, Since $p_2 > t$, $p'_1 = p_1/t = 0.8$. Since $p_1 \leq t$, $p'_2 = p_2 = 0.06$. So $p'_1 > p'_2$.

Remark 3.6. For minP combining method, based on the assumption $\min\{p_1, \ldots, p_n\} \leq t$, at least one inflation factor should be 1, which means at least one hypothesis will not

be inflated. The fewer inflated p-values, the smaller selective effect the MTP produces. Actually, for minP combing method, except for the minimum p-value, other p-values do not need to be inflated.

For minP combining method, the conditional FDR can also be controlled under α level.

Theorem 3.2 (minP combining selection rule)**.** *Let $P_1, \ldots, P_n$ be independent p-values with $U(0,1)$ under true null. If $f(p_1, \ldots, p_n) = \min\{p_1, \ldots, p_n\} \leq t$ and $\alpha \leq t \leq 1$, then the BH procedure on conditional p-values $p'_1, \ldots, p'_n$ controls the cFDR at level $\frac{n_0}{n}\alpha \leq \alpha$.*

Note that the threshold t is the same as the $t_i = \frac{\tilde{t}}{n_i}$ in Procedure 3.6. In Procedure 3.6, $\tilde{p}_i = n_i \min_{1 \leq j \leq n_i} \{p_{ij}\} \leq \tilde{t}$ is equivalent to $\min_{1 \leq j \leq n_i} \{p_{ij}\} \leq 1-(1-t)^{1/n_i}$, where t satisfies $\sum_{l=k}^{m} \binom{m}{l} t^l (1-t)^{m-l} = \alpha_1$. Since $1-(1-t)^{1/n_i}$ can be regarded as a fixed threshold for each body system i, thus we have the following result.

Theorem 3.3 (cFDR control)**.** *For each selected body system $\mathcal{F}_i$, if individual p-values $P_{i1}, \ldots, P_{in_i}$ are mutually independent with $U(0,1)$ under true null, then Procedure 3.6 strongly controls the conditional FDR at level α for flagging AEs within the selected body systems.*

Now we have the following result for selecting families.

Theorem 3.4 (k-FWER control)**.** *If global p-values $\tilde{P}_1, \ldots, \tilde{P}_m$ are mutually independent with $U(0,1)$ under true null, then Procedure 3.6 strongly controls the k-FWER at level α_1 across body systems.*

This theorem follows from the proof of Theorem 2.2 in Guo and Romano (2007).

Corollary 3.1 (average FDR control)**.** *Under the independence assumption of Theorem 3.3 and 3.4, Procedure 3.6 strongly controls the average FDR over selected body systems at level α for flagging AEs.*

Theorem 3.3, Theorem 3.4 and Corollary 3.1 imply the proposed procedure can strongly control various type 1 error rates for selecting body systems (k-FWER),

flagging AE within each selected body system (condition FDR) and overall flagging AEs average on all selected body systems (average FDR). Next we will look into how the selections rules affect the proposed procedure.

3.5 Selection Rule Comparisons

As illustrated in Procedure 3.6, using a multiple testing procedure (MTP) on the combined p-values (global p-values), and selecting the families accordingly, is a natural approach in selective inference. In the following, we will compare some different selection rules regarding *p-value combining methods* and *selection procedures*.

We start from a simple example. Suppose we have selected a family of $n = 2$ hypotheses, since conditional p-value is based on inflation factor, which is relevant to other p-values and selection cutoff, we want to look into how the other p-values and selection cutoff affect inflation factor under the examples iin the Section 3.2.4.

3.5.1 Inflation Factor

Firstly, for Fisher's combining function, suppose the family will be selected if and only if $-2(\log p_1 + \log p_2) > t_F$, that is, $p_1 p_2 \leq e^{-t_F/2}$. Within the selected family, the inflation factor of p_1 is

$$b_1 = \begin{cases} \dfrac{e^{-t/2}}{p_2} & \text{if } p_2 > e^{-t_F/2} \\ 1 & \text{otherwise.} \end{cases}$$

Figure 3.1 is to show b_1 versus p_2, set $t = \{0, 1, \ldots, 10\}$; Figure 3.2 is to show b_1 versus t, set $p_2 = \{0, 0.02, 0.04, 0.06, 0.1, \ldots, 1\}$.

For Stouffer's combining function, suppose the family will be selected if and only if $\Phi^{-1}(1-p_j) + \sum_{\substack{l=1 \\ (\neq j)}}^{n} \Phi^{-1}(1-p_l) \geq \sqrt{n}t$, then $b_1 = 1 - \Phi(\sqrt{n}t - \sum_{\substack{l=1 \\ (\neq j)}}^{n} \Phi^{-1}(1-p_l))$.

Figure 3.3 is to show b_1 versus p_2, set $t = \{-3, -2.5, \ldots, 3\}$; the second plot of Figure 3.4 is to show b_1 versus t, set $p_2 = \{0, 0.1, 0.2, \ldots, 1\}$.

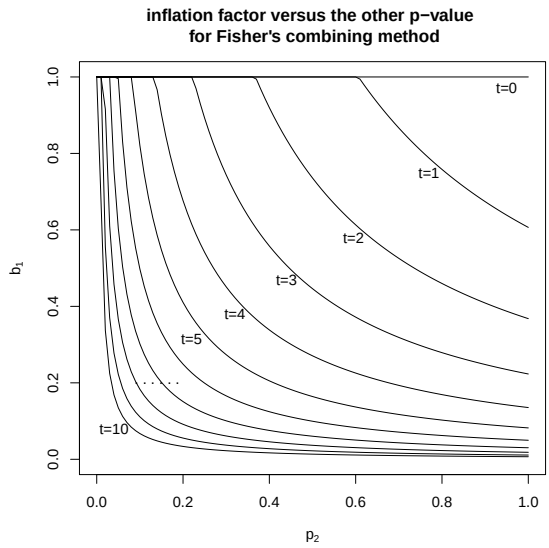

Figure 3.1 Comparison of inflation factor b_1 with respect to p_2 for different values of threshold t using Fisher's combining method with $n = 2$ hypotheses.

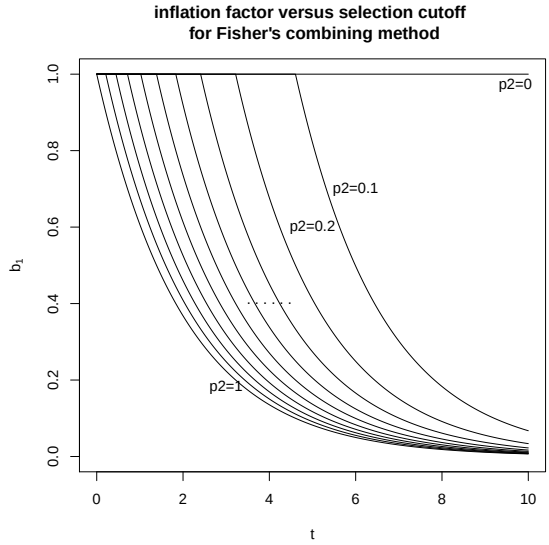

Figure 3.2 Comparison of inflation factor b_1 with respect to threshold t for different values of p_2 using Fisher's combining method with $n = 2$ hypotheses.

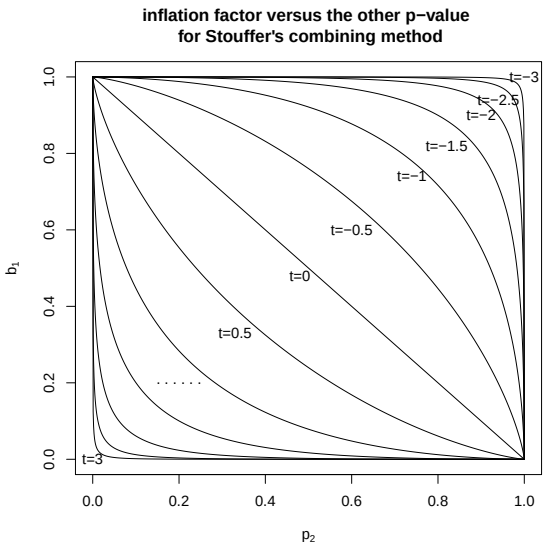

Figure 3.3 Comparison of inflation factor b_1 with respect to p_2 for different values of threshold t using using Stouffer's combining method with $n = 2$ hypotheses.

For minP combining function, suppose the family will be selected if and only if $\min\{p_1, p_2\} \leq t_M$. Within the selected family, the inflation factor of p_1 is

$$b_1 = \begin{cases} t & \text{if } p_2 > t \\ 1 & \text{otherwise.} \end{cases}$$

Figure 3.5 is to show b_1 versus p_2, set $t = \{0, 0.1, \ldots, 1\}$; the Figure 3.6 is to show b_1 versus t, set $p_2 = \{0, 0.1, 0.2, \ldots, 1\}$.

Remark 3.7. Since the p-values are exchangeable, the inflation factor b_2 has the same tendency as b_1.

Now, we want to compare the different selection rules using Fisher's, minP and Stouffer's combining methods. Suppose given the threshold $t \in [\alpha, 1]$ for minP selection rule such that the hypotheses/family is selected if $\min\{\mathbf{p} \leq t\}$. A simple way to make them comparable is to find equivalent cutoffs t_F for Fisher's method or t_S for Stouffer's method satisfying

$$\Pr\left\{-2\sum_{j=1}^{n}\log p_j \geq t_F\right\} = \Pr\left\{\min\{p_1, \ldots, p_n\} \leq t\right\},$$

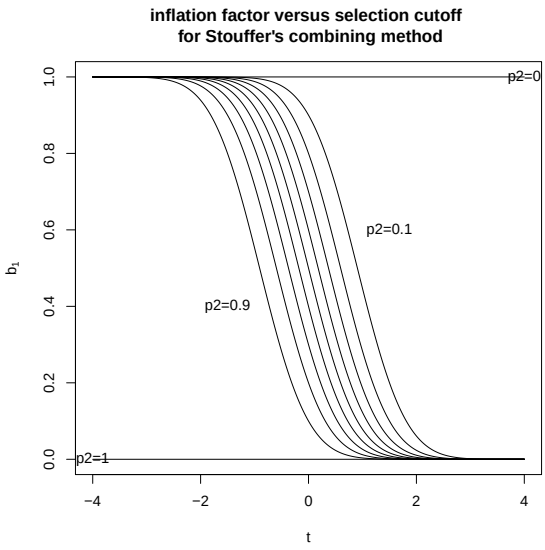

Figure 3.4 Comparison of inflation factor b_1 with respect to threshold t for different values of p_2 using Stouffer's combining method with $n = 2$ hypotheses.

or

$$\Pr\left\{\frac{\sum_{j=1}^{n}\Phi^{-1}(1-p_j)}{\sqrt{n}} \geq t_S\right\} = \Pr\left\{\min\{p_1,\ldots,p_n\} \leq t\right\}.$$

It means the chances of the family is selected are the same for these methods.

We use the following calculations find equivalent selection threshold for Fisher's and minP combining methods. Here, assume the p-values are identical and independent. In order to compare Fisher's method with minP method, let

$$\Pr\left\{-2\sum_{j=1}^{n}\log p_j > t_F\right\} = \Pr\left\{\min\{p_1,\ldots,p_n\} \leq t\right\},$$

which is equivalent to

$$\Pr(T \geq t_F) = 1-(1-t)^n,$$

where $T \sim \chi^2_{2n}$. Then

$$\Pr(T \leq t_F) = (1-t)^n.$$

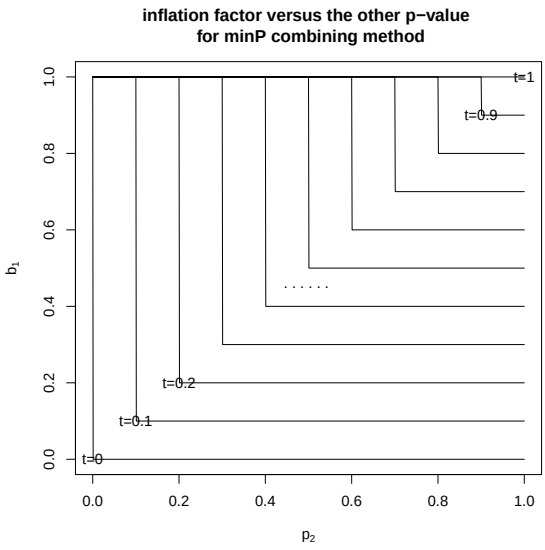

Figure 3.5 Comparison of inflation factor b_1 with respect to p_2 for different values of threshold t using minP combining method with $n = 2$ hypotheses.

We can also compare Stouffer's method with minP method. Let

$$\Pr\left\{\frac{\sum_{j=1}^{n}\Phi^{-1}(1-p_j)}{\sqrt{n}} \geq t_S\right\} = \Pr\left\{\min\{p_1,\ldots,p_n\} \leq t\right\},$$

which is equivalent to

$$\Pr(Z \geq t_S) = 1-(1-t)^n,$$

where $Z \sim N(0,1)$. Then

$$\Pr(Z \leq t_S) = (1-t)^n.$$

It means for the minP threshold t, we can always find somehow equivalent threshold t_F and t_S for Fisher's and Stouffer's combining method.

We can also plot the comparison for three combining methods when combining two hypotheses. From Figure 3.7, we can observe the cross points for different two lines is the equivalent point for the inflation factor and the other p-value. We also compare different combining methods by conducting the simulations studies in Section 3.6.1.

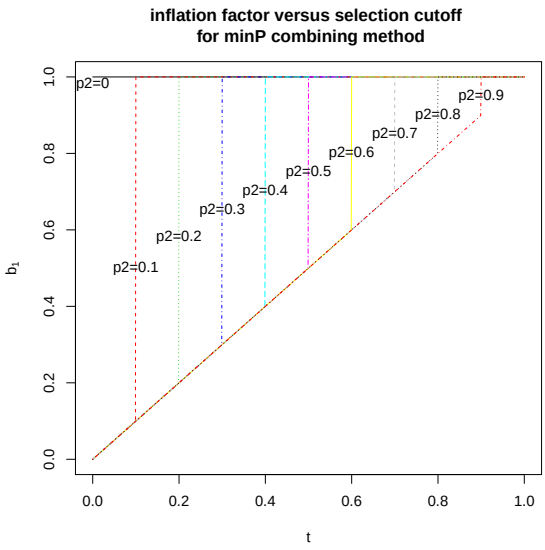

Figure 3.6 Comparison of inflation factor b_1 with respect to threshold t for different values of p_2 using minP combining method with $n = 2$ hypotheses.

3.5.2 Selection Rules Using MTP Controlling k-FWER

It has been discussed that controlling generalized FWER is more suitable than controlling either FWER or FDR for selecting body systems in clinical safety studies. Since FWER is too strict and FDR is too liberal in such setting, k-FWER is an error measure between these two conventional error measures. Using k-FWER controlling procedure to select BSoI brings two questions. Which procedure should be used for selecting more signals? How to choose a suitable k?

For multiple hypotheses testing without family structure, it has been shown that the generalized Sidak procedure is much more powerful than generalized Bonferroni procedure (Guo and Romano , 2007) under independence. In Section 3.6.1, we will also compare generalized Sidak and generalized Bonferroni selection rules in the simulation studies.

Another factor which affects the selection rule is how many fold we need to choose for k-FWER control for family level. On one hand, choosing k depends on the requirement of practice, such as clinicians' experience or clinical research interest. For example, if there are many body systems in the trial, and we are interested in

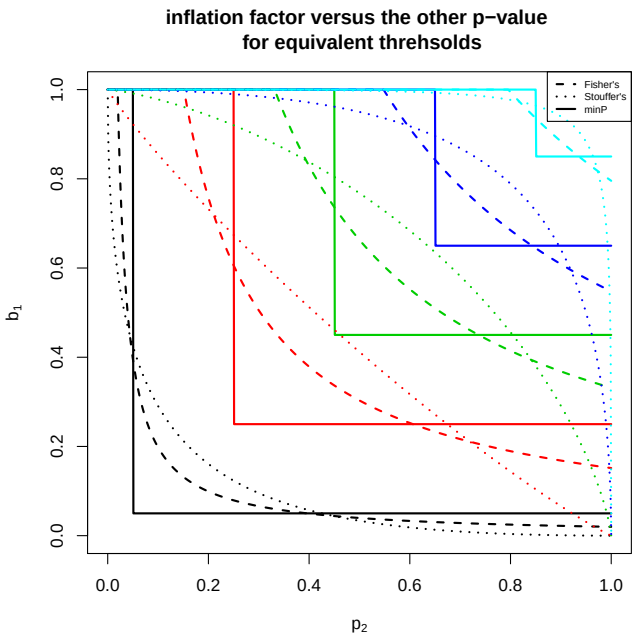

Figure 3.7 Comparisons of inflation factor b_1 with respect to p_2 for equivalent thresholds of Fisher's, Stouffer's and minP combining methods when $n = 2$.

selecting more body systems, then we can choose $k = 2$ or $k = 3$ to allow to make two or three type 1 errors for the selection. On the other hand, we can also optimize the k by doing simulations, which will be discussed in Section 3.6.

3.6 Simulation Studies

This section investigates the performance of the proposed two-stage procedure using simulations under various dependence settings: (i) the p-values are independent both within body system and across body system; (ii) the p-values are dependent within each body system and independent of the p-values in other body systems; (iii) the p-values are independent within each body system and dependent of the p-values in other body systems.

Since the families contains true nulls and false nulls, it is more reasonable to use average FDR over selected families (average-FDR) and conditional FDR (cFDR) for all true null and non-null families respectively to evaluate the performance of the

procedures, rather than use global FDR or simple FDR of one family. We use analogous average power over selected families as average power.

3.6.1 Simulations for the Independence Settings

Numerical comparisons for various number of true null families (m_0) and true null individual hypotheses (n_0) First of all, we conduct simulations to compare the proposed procedure using Sidak procedure (similar as Bonferroni procedure) for global test of minP combination to select families (cFDR-minP-Sidak) with (I) average FDR controlling procedure using the same selection rule (BB-minP-Sidak), (II) original double FDR procedure (DFDR) and (III) modified double FDR procedure (DFDR2). All the simulations are conducted for 2000 times. Set level $\alpha = 0.05$ for both selection for family level and testing for individual hypotheses level. Each simulated data set is based on one-sided one sample Z-test for testing $H_{ij} : \mu = 0$ versus $H'_{ij} : \mu = \mu_1 > 0$.

Set the number of families m to be 10, in each family set the number of hypotheses n to be 20. We vary the number of all true null families m_0 as 2, 4, 6 and 8, and the number of true null hypotheses in each non-true null family n_0 as 5, 10 and 15. Set $\mu_1 = 3$, $\alpha = \alpha_1 = 0.05$

From Figures 3.8 to 3.11, we can see the cFDR-minP and BB procedures using Sidak selection rule can control FWER for family level, and control average FDR and conditional FDR under all scenarios. But DFDR and DFDR2 procedure fail to control average FDR when n_0 and m_0 become bigger. For example, in Figure 3.10, when $n_0 = 15$, both DFDR and DFDR2 lines are above 0.05 when the proportion of null families is greater than 0.4. DFDR and DFDR2 also fail to control conditional FDR as well.

Numerical comparisons for various combining methods In this section, we compare the procedures using different combining methods, here we mainly compare Fisher's and minP methods. The selection threshold here we used generalized Sidak

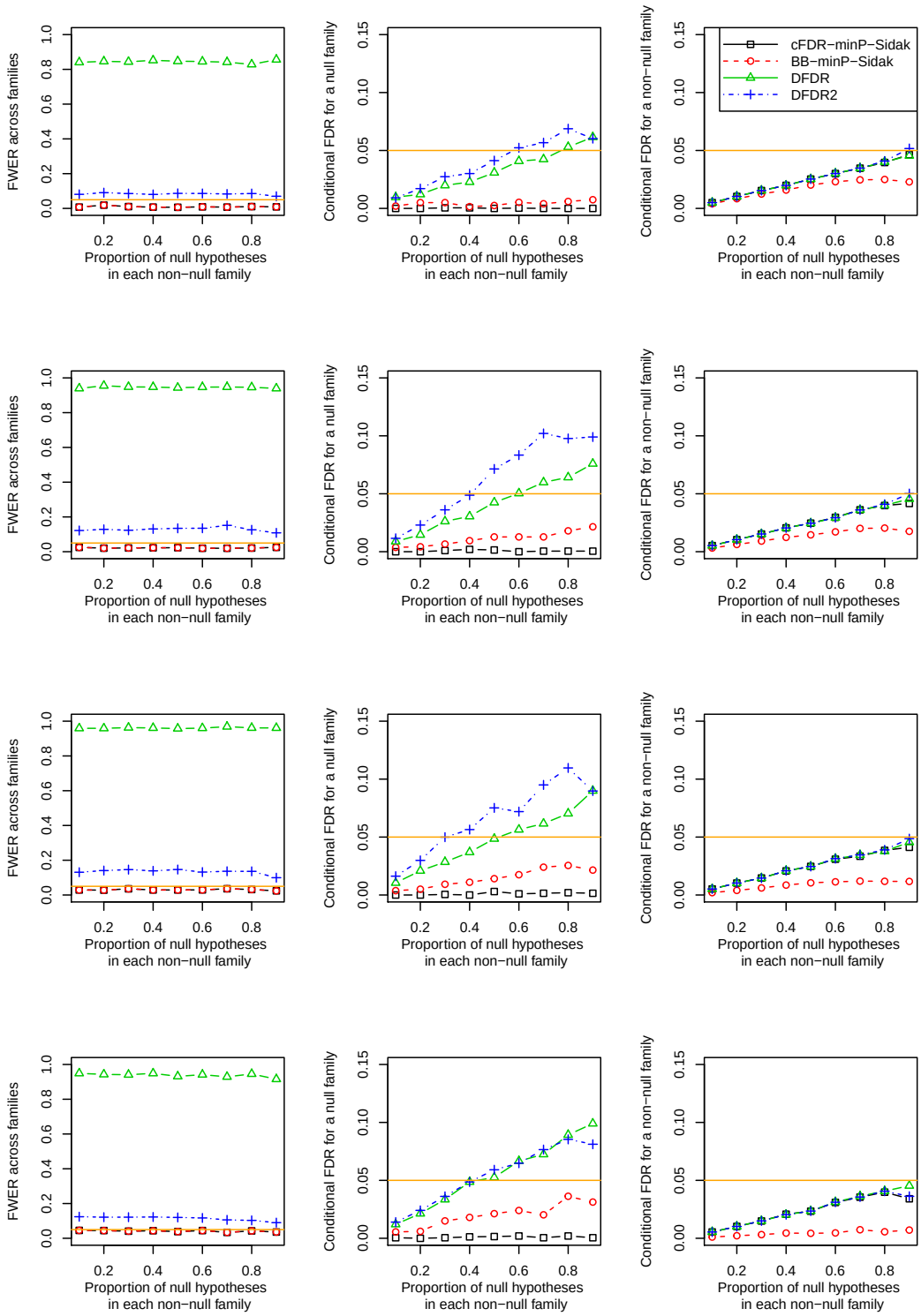

Figure 3.8 From the left to right panels are simulated FWER across families, conditional FDR for a null family and conditional FDR for a non-null family versus proportion of null hypotheses in each non-null family (n_0/n). From the top to bottom panels, the numbers of true null families are $m_0 = 2, 4, 6, 8$ out of $m = 10$ families, there are $n = 20$ hypotheses in each family, $\mu_1 = 3$, $\alpha = \alpha_1 = 0.05$.

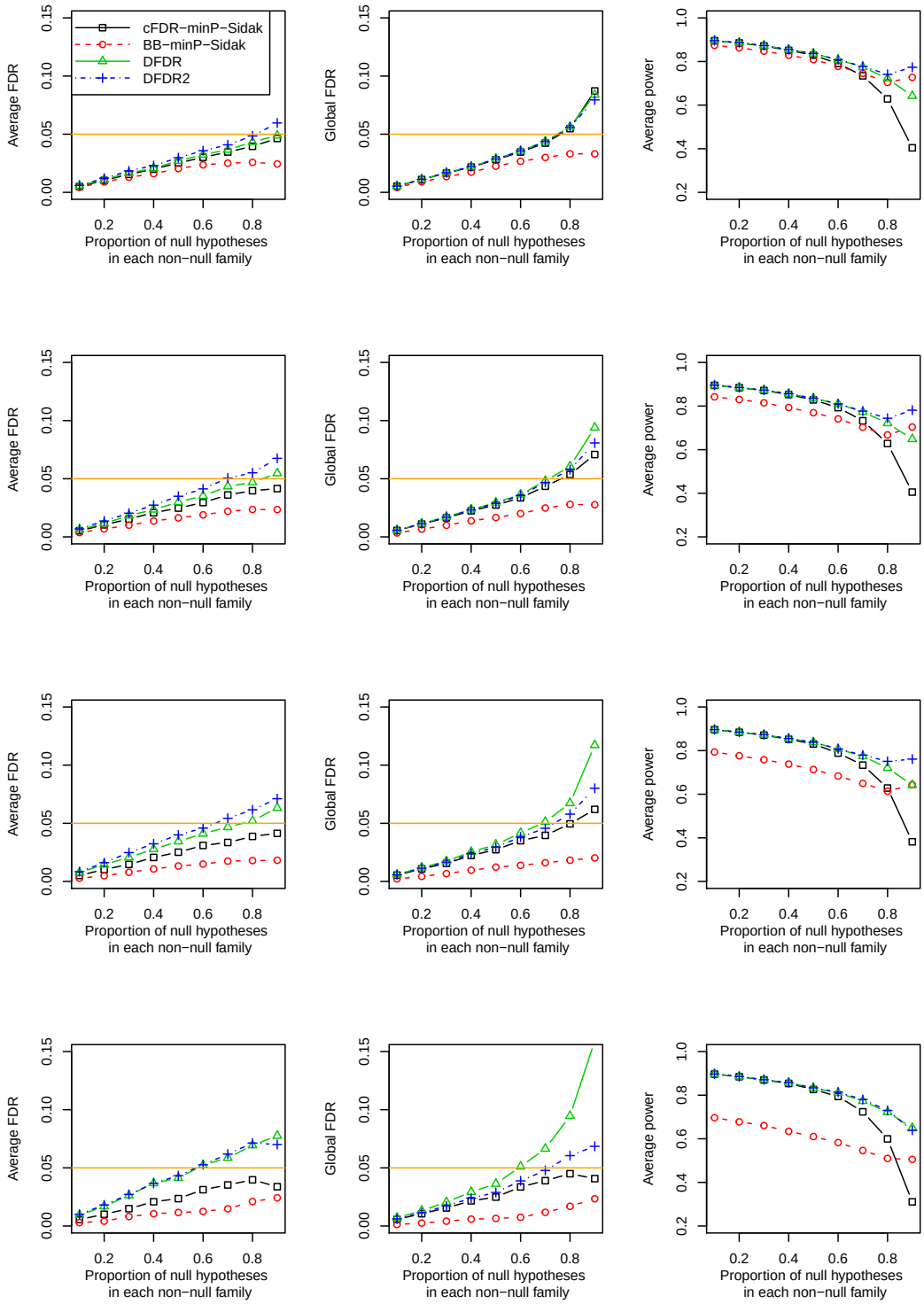

Figure 3.9 From the left to right panels are simulated average FDR over selected families, global FDR and average power versus proportion of null hypotheses in each non-null family (n_0/n). From the top to bottom panels, the numbers of true null families are $m_0 = 2, 4, 6, 8$ out of $m = 10$ families, there are $n = 20$ hypotheses in each family, $\mu_1 = 3$, $\alpha = \alpha_1 = 0.05$.

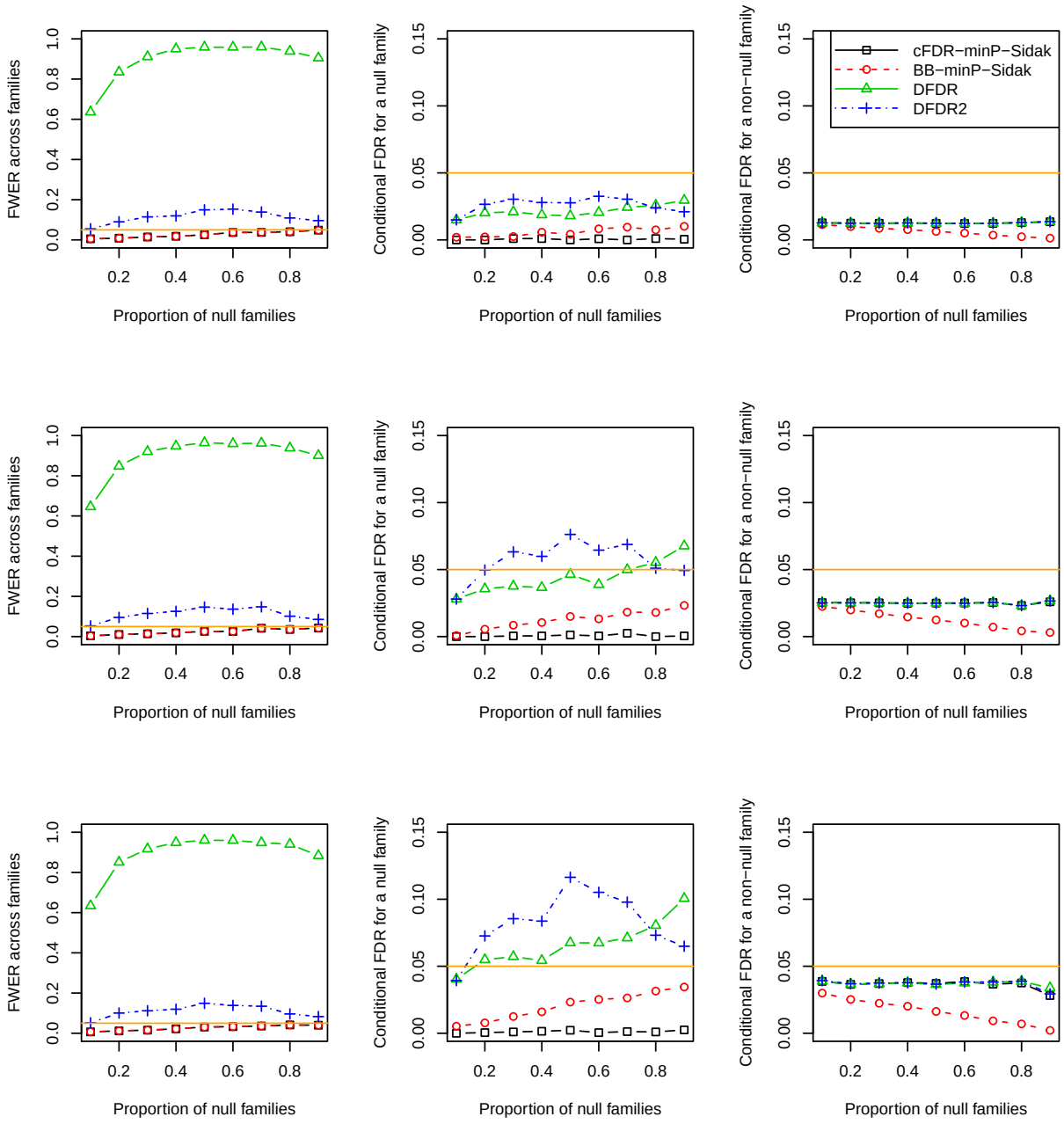

Figure 3.10 From the left to right panels are simulated FWER across families, conditional FDR for a null family and conditional FDR for a non-null family versus proportion of null families (m_0/m). From the top to bottom panels, the numbers of true null hypotheses in each non-null family are $n_0 = 5, 10, 15$ out of $n = 20$ hypotheses, there are $m = 10$ families, $\mu_1 = 3$, $\alpha = \alpha_1 = 0.05$.

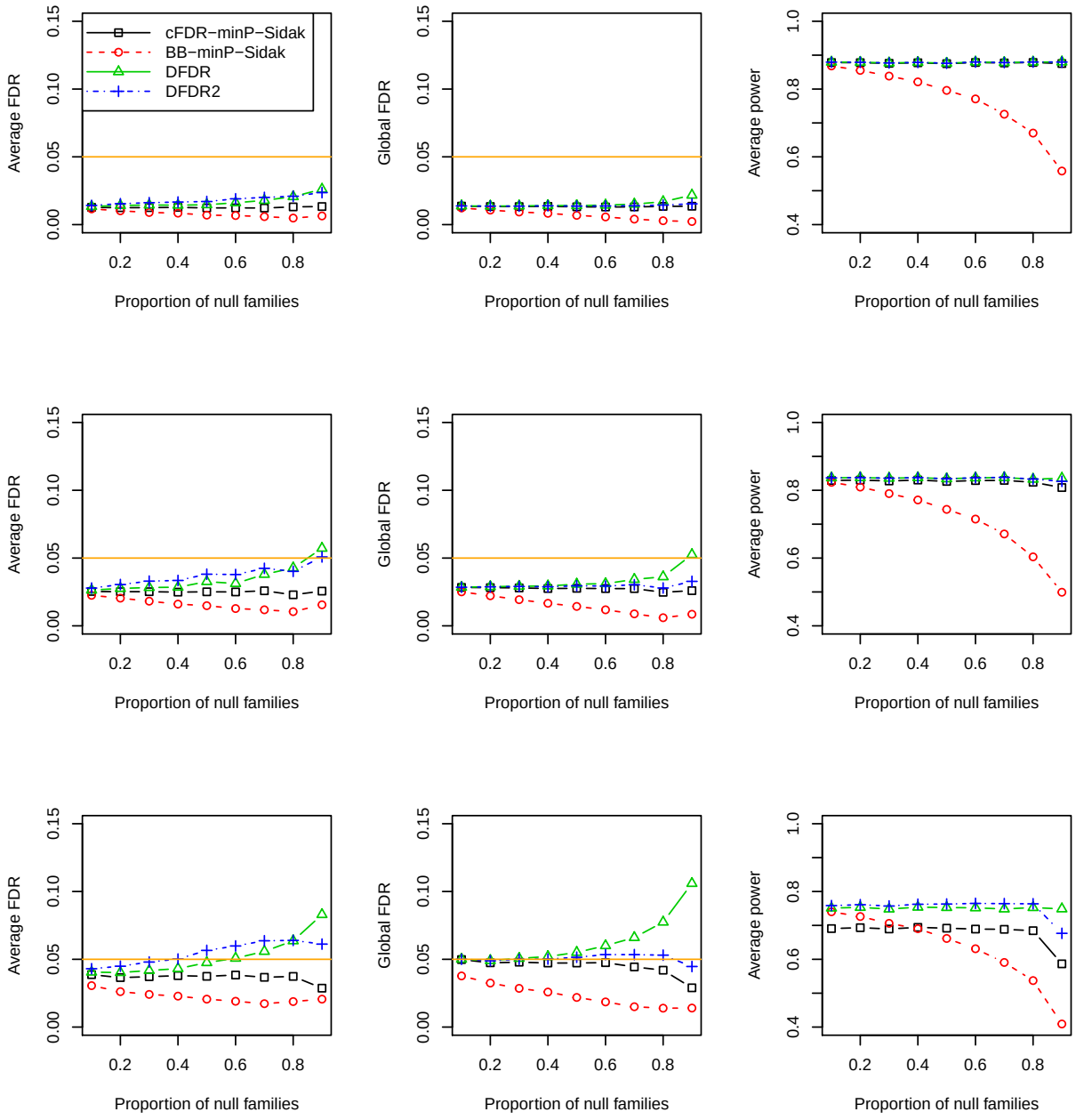

Figure 3.11 From the left to right panels are simulated average FDR over selected families, global FDR and average power versus proportion of null families (m_0/m). From the top to bottom panels, the numbers of true null hypotheses in each non-null family are $n_0 = 5, 10, 15$ out of $n = 20$ hypotheses, there are $m = 10$ families, $\mu_1 = 3$, $\alpha = \alpha_1 = 0.05$.

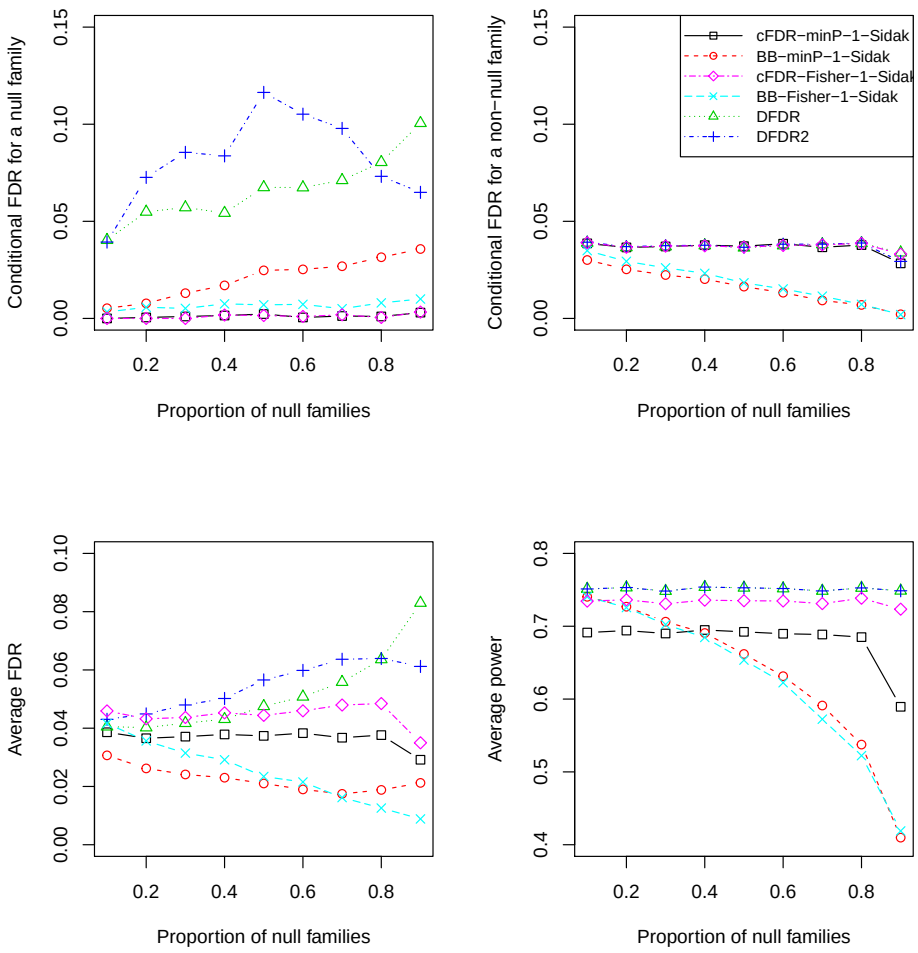

Figure 3.12 Comparisons for different combining methods using Sidak selection rules ($k = 1$) for independent structure, $m = 10$, $n = 20$, $n_0 = 15$, $\alpha = \alpha_1 = 0.05$.

threshold. In Figure 3.12, we can see the proposed procedure using Fisher's combining method as FWER selection rule is more powerful than using minP combining method. But BB procedure using Fisher's combining method is slightly less powerful than the one using minP combining method.

However, when considering generalized Sidak procedure with a slightly larger k, the proposed procedure using minP combining method is still less powerful than the one using Fisher's combining method, but the two lines are very close. The BB procedure using these two combining methods are also almost the same regarding the power performance, see Figure 3.13.

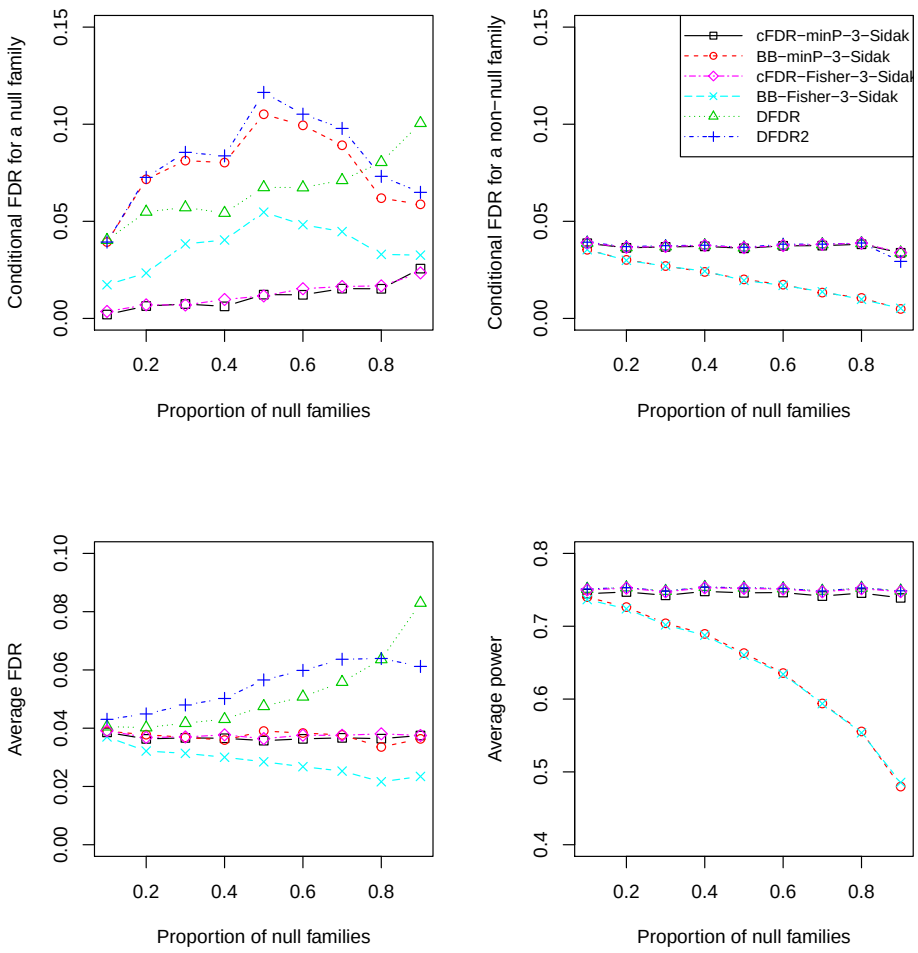

Figure 3.13 Comparisons for different combining methods using generalized Sidak with $k = 3$ selection rules for independent structure, $m = 10$, $n = 20$, $n_0 = 15$, $\alpha = \alpha_1 = 0.05$.

Now we also consider the simulated average power, average FDR, conditional FDR for a null family and a non-null family versus different significant level α_1 for selecting families, set $\alpha_1 = 0.05, 0.1, 0.15, 0.2, 0.25, 0.3$, simulate the proportion of containing signals among the selected families and proportion of detecting signals among selected families containing signals. From Figure 3.14, we can observe similar results as Figure 3.13. However, from the real data analysis in the later section, in some cases using minP combing method can find more signals in selecting step. And minP combining is more convenient to calculate in practice.

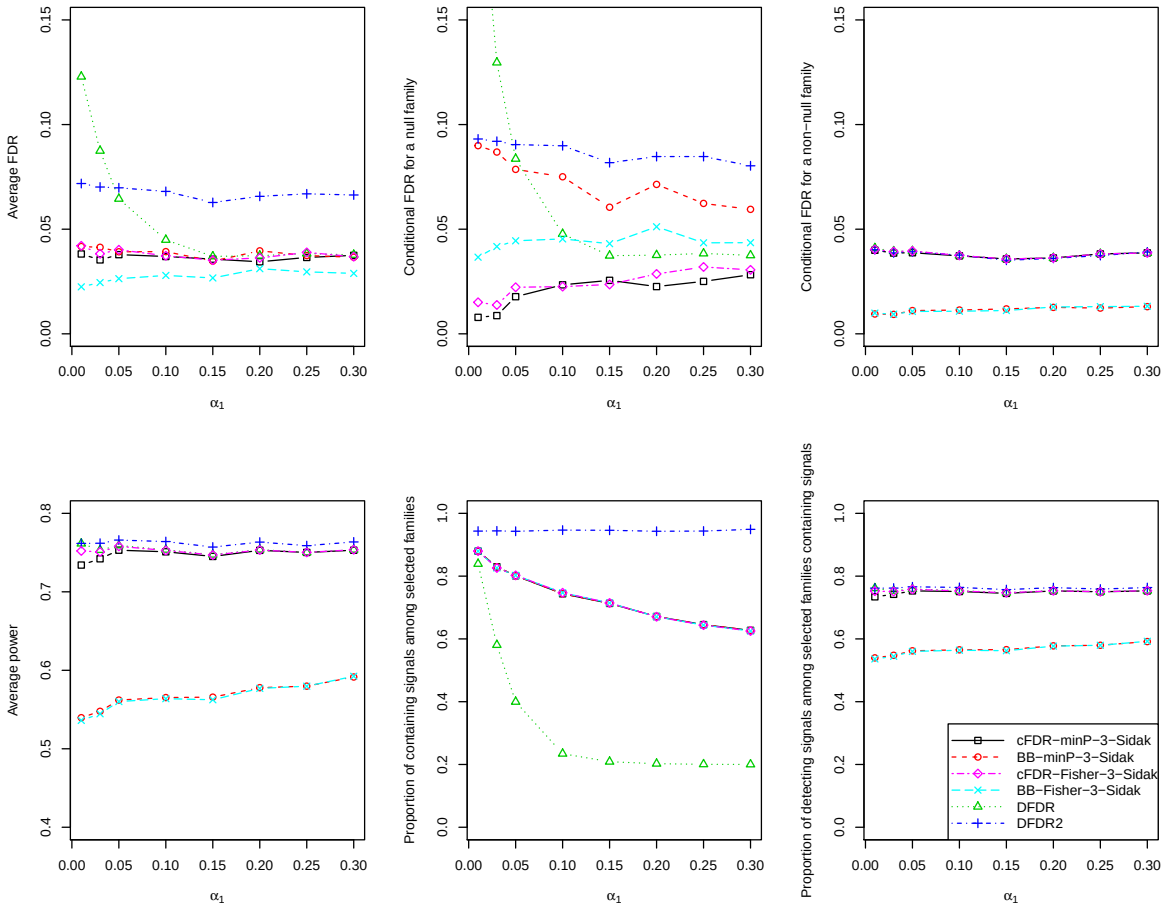

Figure 3.14 Comparisons for different combining method for independent structure using generalized Sidak with $k = 3$ selection rules versus different selection significant level α_1.

Numerical comparisons for various k-FWER selection procedures For different k-FWER selection rule, such as generalized Bonferroni versus generalized Sidak; choice of k, we also perform some simulation to investigate the differences.

In Figure 3.16, when k becomes larger, the power for cFDR or BB will become higher, since the procedure allows to make more than one type 1 error in selecting stage, more families containing signals could be selected. Moreover, the proposed procedures (cFDR-minP-k-Sidak and cFDR-minP-k-Bonf) are more powerful than BB procedures (BB-minP-k-Sidak and BB-minP-k-Bonf) when proportion of null families is greater

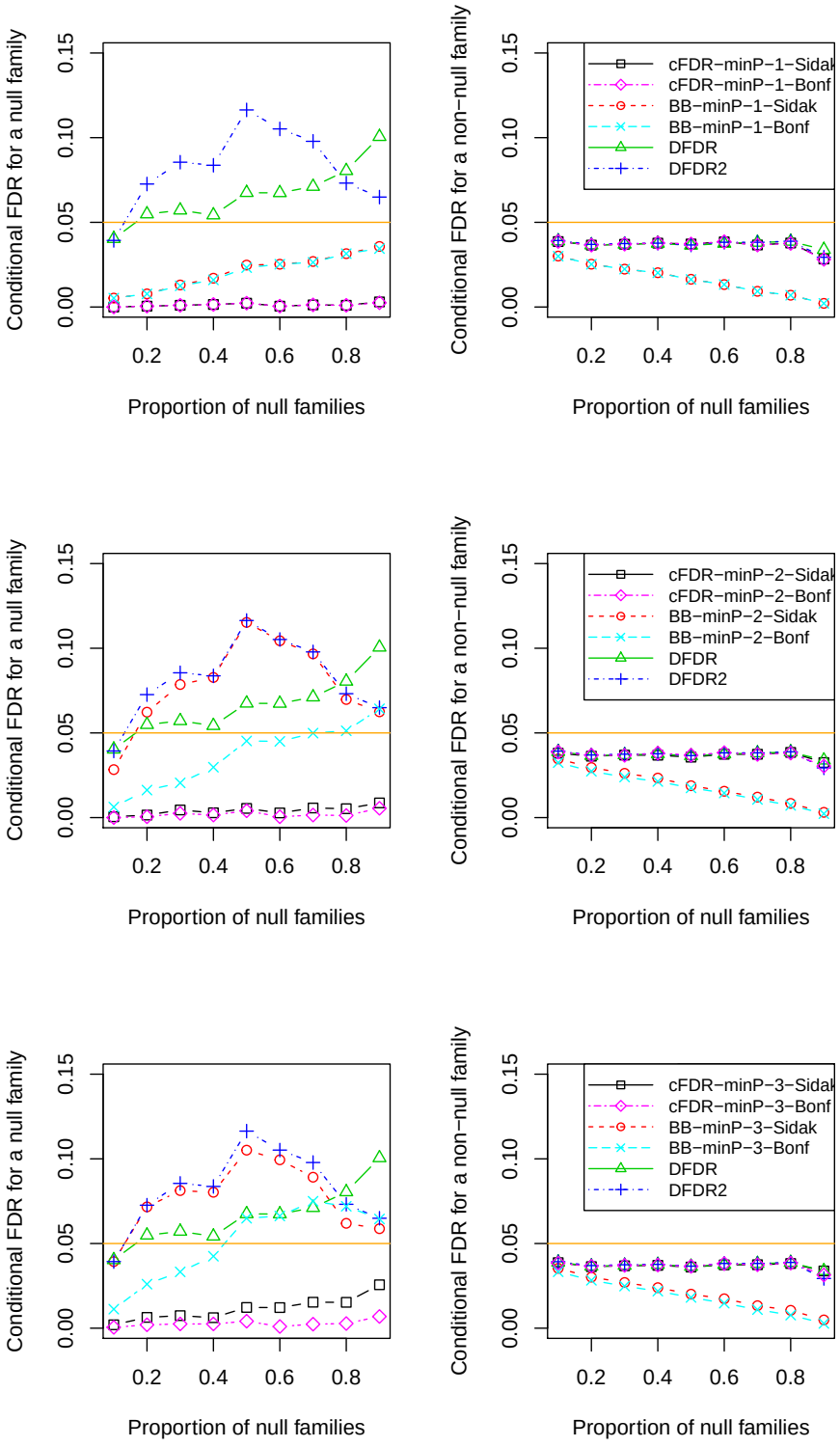

Figure 3.15 Comparisons for using generalized Bonferroni and generalized Sidak selection rules with $k = 1, 2, 3$ under independence, the plots show the conditional FDR for null or non-null family versus the proportion of null families, $m = 10$, $n = 20$, $n_0 = 15$, $\alpha = \alpha_1 = 0.05$.

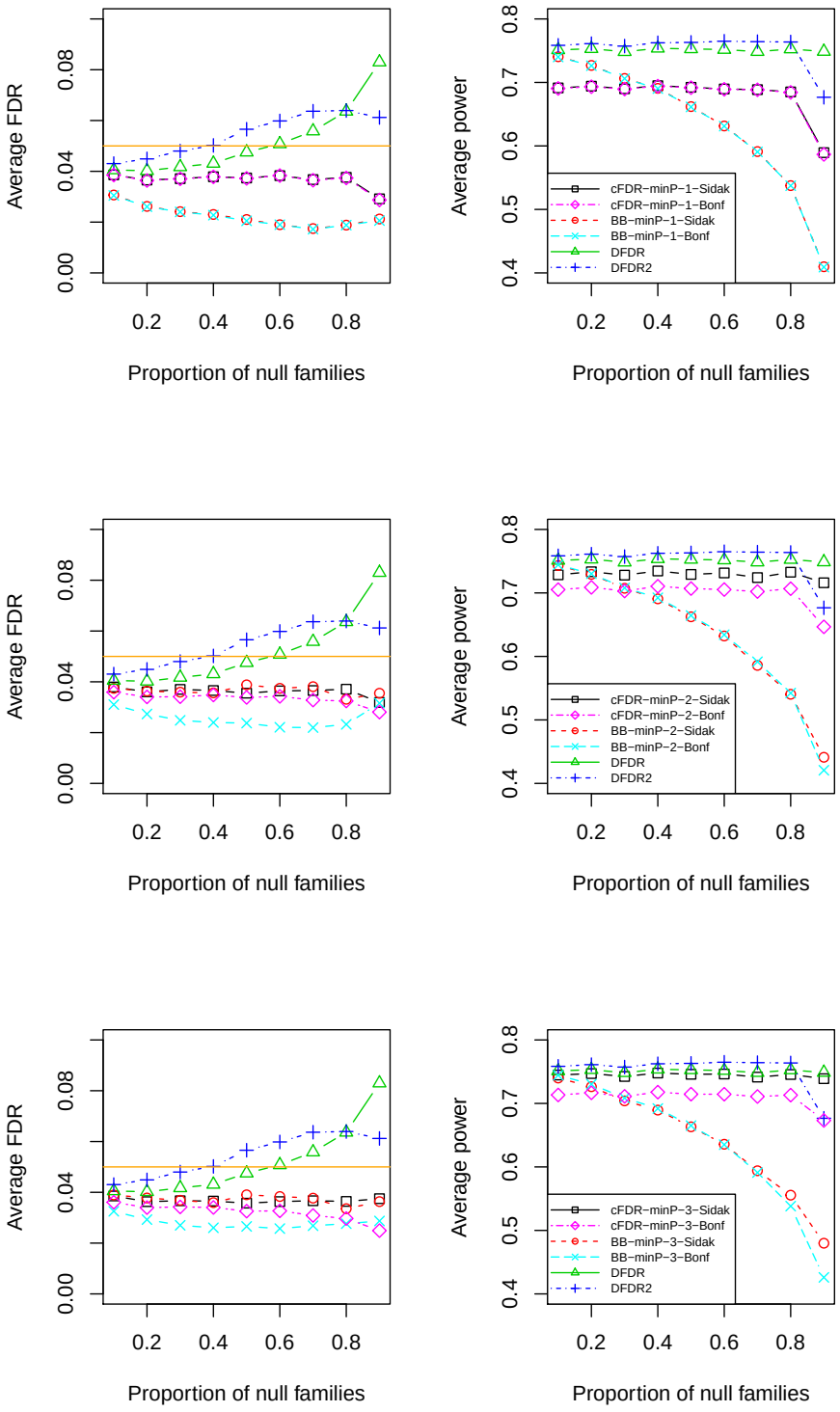

Figure 3.16 Comparisons for using generalized Bonferroni and generalized Sidak selection rules with $k = 1, 2, 3$ under independence, the plots show the average FDR and average power versus the proportion of null families, $m = 10$, $n = 20$, $n_0 = 15$, $\alpha = \alpha_1 = 0.05$.

than 0.4. When k becomes larger, the procedures using generalized Sidak selection rule are more powerful than the procedures using generalized Bonferroni rule.

3.6.2 Simulations for the Dependence Settings

We also perform the simulations under two types of dependence settings: (I) the p-values are dependent within each body system and independent of the p-values in other body systems; (II) the p-values are independent within each body system and dependent of the p-values in other body systems. We consider equal correlated dependence setting with $\rho = 0, 0.1, \ldots, 0.9$ in the following simulation studies.

From Figures 3.17 and 3.18, we can observe when $m_0 = 4$ out of total 10 families, the proposed procedure can control average FDR and maintain high powers. When and correlation ρ is less than 0.4 for $m_0 = 8$ out of total 10 families, the proposed cFDR controlling procedure can control the average FDR and more powerful than BB procedure by using generalized Sidak selection rule with fold $k = 3$. But when correlation becomes larger, the proposed procedure using generalized Sidak selection rule cannot control the average FDR, while modified BB procedure using the same selection rule can control average FDR. Therefore, we recommend to use the cFDR controlling procedure when proportion of true null families is small (about 40%) and use the modified BB procedure when proportion of true null families is large (about 80%).

3.7 Real Data Analysis: Clinical Safety Studies

In this section, we apply the proposed cFDR-minP-k-FWER controlling procedure (Procedure 3.6) in the clinical safety studies to flag the significant AE types. The data analysis is conducted in the following three steps.

Step 1 (Select body systems) Selecting body system BS_i if $f(p_{i1}, \ldots, p_{in_i}) \leq t_i$, where t_i can be the same value for any i or a sequence of thresholds.

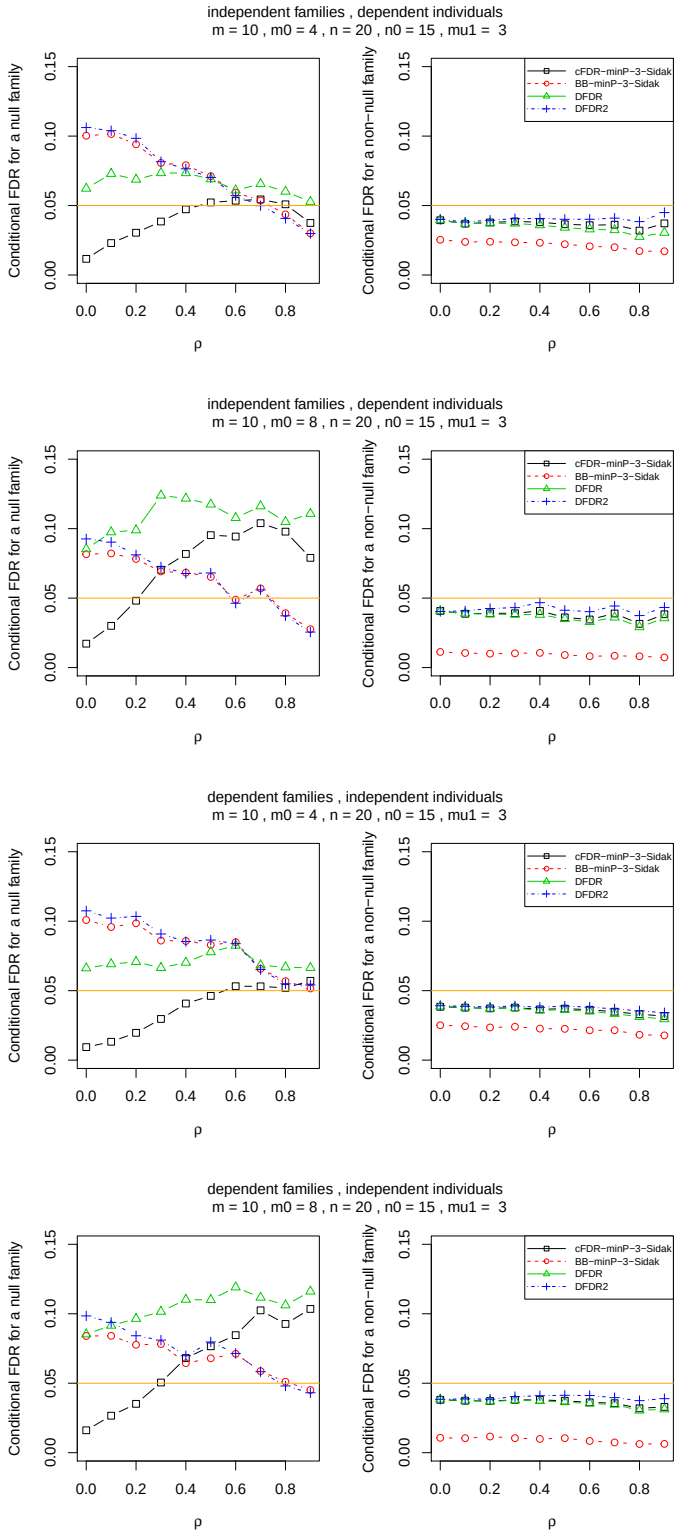

Figure 3.17 Comparisons of conditional FDR's with respect to ρ for different dependent structures and different numbers of null families ($m_0 = 4, 8$) by using different multiple families testing procedures. $\alpha = \alpha_1 = 0.05$.

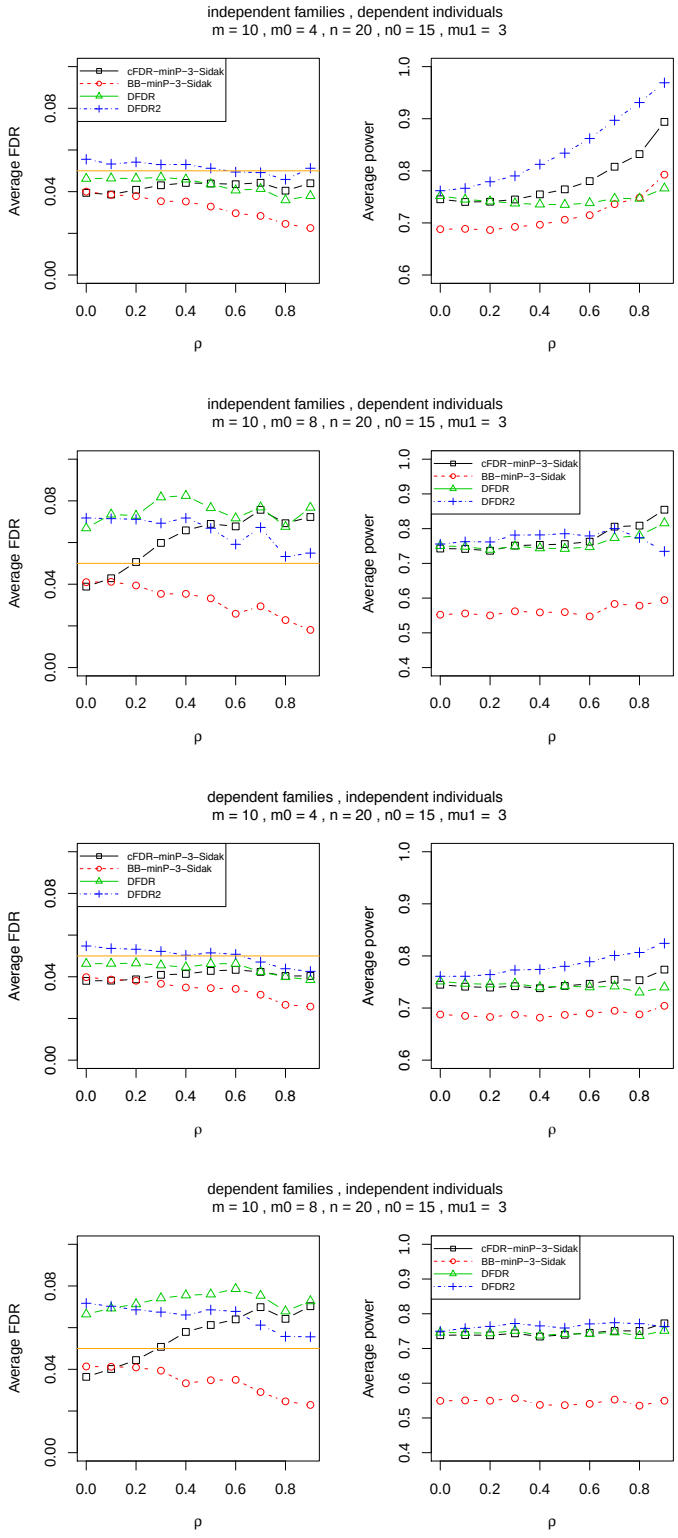

Figure 3.18 Comparisons of average FDR's and powers with respect to ρ for different dependent structures and different numbers of null families ($m_0 = 4, 8$) by using different multiple families testing procedures. $\alpha = \alpha_1 = 0.05$.

We want to control some error rate when selecting the body systems. In different phases of clinical trials, the adverse experience (AEs) usually are across 5-50 body systems, and the trials allow to more than one false discoveries on body systems selection when there are more classified body systems. Thus, we suggest if the number of body systems is no more than 30, consider generalized FWER (k-FWER) controlling procedure to select body systems. Otherwise we can consider FDR controlling procedures. Note for generalized FWER selection rule, k can be decided based on clinicians' experience or protocol information.

Step 2 (Conditional inference for flagging AEs) Within each selected body system $\mathcal{F}_i : i \in S$, calculate the conditional p-value for each hypothesis P'_{ij}, and calculate corresponding BH adjusted conditional p-value ${P'_{ij}}^{\text{BH-adj}}$.

Step 3 (Make decision) For $i \in \mathcal{S}$ and $j = 1, \ldots, n_i$. If ${P'_{ij}}^{\text{BH-adj}} \leq \alpha$, then reject H_{ij}, that is, flag the j-th AE type in the i-th body system.

We consider the following two cases in reality to select body systems of interest, then flag the significant AE types within the selected body systems.

Case 1: Fixed selection rule and different combining methods are considered. For instance, let $t_i = 0.1$ for minP combining method. Such a fixed selection rule cannot guarantee type 1 error rate control across family level.

Case 2: A sequence of data-adaptive thresholds t_i to select body systems. For instance, for i-th family $\mathcal{F}_i$, $i = 1, \ldots, m$, we set the global null p-value as $\tilde{p}_i = 1 - (1 - \min_{1 \leq j \leq n_i} \{p_{ij}\})^{n_i}$ for minP combining method. If the p-values in the family are uniformly distributed in $(0, 1)$, then $\tilde{p}_i \sim U(0, 1)$ for $i = 1, \ldots, m$. Consider generalized Bonferroni procedure on the global null p-values. The family is selected if $\tilde{p}_i \leq \frac{k\alpha_1}{m}$, where $k \in \{1, \ldots, m\}$. If $k = 1$, the procedure reduces to Bonferroni procedure, which controls FWER on family level. We can also apply generalized Sidak procedure on the global null p-values. For fixed $k \in \{1, \ldots, m\}$, the family is selected if $\tilde{p}_i \leq \tilde{t} = \tilde{t}_{k,m}(\alpha)$, where $\tilde{t}$ satisfies $\sum_{l=k}^{m} \tilde{t}^l (1 - \tilde{t})^{m-l} = \alpha_1$.

Besides considering different selection rules for proposed procedure (fixed selection threshold $t_i = 0.1$, generalized Sidak with $k = 1$ and $k = 3$, minP and Fisher's combining methods), we also compare them with (I) simply applying BH procedure on pooled p-values (Naive BH); (II) ad-hoc version of original double FDR procedure (DFDR, $\alpha_1 = 0.05$, $\alpha = 0.1$); (III) modified double FDR procedure (DFDR2); (IV) average FDR controlling procedure (Original BB) and (V) group BH procedure (GBH).

3.7.1 Example 3.1

This example is from Mehrotra and Heyse (2001). The trial involved a quadrivalent vaccine containing measles, mumps, rubella, and varicella (MMRV). Participants were 296 healthy toddlers aged 12-18 months who were randomly assigned to two groups (148 for treatment group, 132 for control group). The treatment group received MMRV on day 0 and controls received MMR on day 0 followed by V on day 42. All participants received PedvaxHIB on day 0. Safety follow-up used standard AE reporting and the primary question was to assess local and systemic reactions for the varicella component. The comparison of AEs was between the treatment group during days 0-42 with the control group during days 42-84. There are 40 AE types across eight body systems. The Fisher's exact test two-sided p-values are calculated based on the counts.

If the selection threshold is fixed $t = 0.05$, the 5th body system are selected but there is no AE in this body system detected. If $t = 0.1$, the 2nd, 5th and 7th body system are selected and there is one AE in the 5th family is flagged. Even consider an extreme case, set $t = 1$, then all families are selected to make the inference, that is, there is no selection effect on families and no conditional inference. Still only the AE in the 7th family are detected.

In Example 3.1, there are 40 AE types across eight body systems. The AE types in the result details are denoted by BS+AE, for example, 503 means Body system No. 5, the third AE. The no multiplicity adjustment approach flags only four AE types (204, 503, 704, 706) with $p \leq 0.05$. The numbers in the parenthesis are numbers of

Table 3.1 Example of Clinical Safety Study from Mehrotra and Heyse (2001), where "BS" is Abbreviate of "Body System" and "No." is the Type of AEs in Each Body System

BS	Family	No.	AE name	X_1	X_1/N_1	X_2	X_2/N_2	p-value
1	1	1	Asthenia/fatigue	57	0.385	40	0.303	0.167
1	1	2	Fever	34	0.230	26	0.197	0.561
1	1	3	Infection.fungal	2	0.014	0	0.000	0.500
1	1	4	Infection.viral	3	0.020	1	0.008	0.625
1	1	5	Malaise	27	0.182	20	0.152	0.525
3	2	1	Anorexia	7	0.047	2	0.015	0.179
3	2	2	Candidiasis.oral	2	0.014	0	0.000	0.500
3	2	3	Constipation	2	0.014	0	0.000	0.500
3	2	4	Diarrhea	24	0.162	10	0.076	0.029
3	2	5	Gastroenteritis	3	0.020	1	0.008	0.625
3	2	6	Nausea	2	0.014	7	0.053	0.089
3	2	7	Vomiting	19	0.128	19	0.144	0.730
5	3	1	Lymphadenopathy	3	0.020	2	0.015	1.000
6	4	1	Dehydration	0	0.000	2	0.015	0.221
8	5	1	Crying	2	0.014	0	0.000	0.500
8	5	2	Insomnia	2	0.014	2	0.015	1.000
8	5	3	Irritability	75	0.507	43	0.326	0.002
9	6	1	Bronchitis	4	0.027	1	0.008	0.375
9	6	2	Congestion.nasal	4	0.027	1	0.008	0.375
9	6	3	Congestion.resp	1	0.007	2	0.015	0.603
9	6	4	Cough	13	0.088	8	0.061	0.497
9	6	5	Infection.resp	28	0.189	20	0.152	0.431
9	6	6	Laryn	2	0.014	1	0.008	1.000
9	6	7	Pharyngitis	13	0.088	8	0.061	0.497
9	6	8	Rhinorrhea	15	0.101	14	0.106	1.000
9	6	9	Sinusitis	3	0.020	1	0.008	0.625
9	6	10	Tonsillitis	2	0.014	1	0.008	1.000
9	6	11	Wheezing	3	0.020	1	0.008	0.625
10	7	1	Bite/sting	4	0.027	0	0.000	0.125
10	7	2	Eczema	2	0.014	0	0.000	0.500
10	7	3	Pruritus	2	0.014	1	0.008	1.000
10	7	4	Rash	13	0.088	3	0.023	0.021
10	7	5	Rash.diaper	6	0.041	2	0.015	0.288
10	7	6	Rash.measles	8	0.054	1	0.008	0.039
10	7	7	Rash.varicella-like	4	0.027	2	0.015	0.687
10	7	8	Urticaria	0	0.000	2	0.015	0.221
10	7	9	Viral.exanthema	1	0.007	2	0.015	0.603
11	8	1	Conjunctivitis	0	0.000	2	0.015	0.221
11	8	2	Otitis.media	18	0.122	14	0.106	0.711
11	8	3	Otorrhea	2	0.014	1	0.008	1.000

Table 3.2 Flagging AE Types for Example 3.1 under $\alpha_1 = 0.05$ for Selecting BSoI and $\alpha = 0.1$ for Detecting AEs

Approach	BSoI	Flagging AE Types
Naive BH	NA	503 (1)
DFDR	5 (1)	503 (1)
DFDR2	5 (1)	503 (1)
GBH	NA	503 (1)
Original BB	5 (1)	503 (1)
cFDR-minP-0.1	2,5,7 (3)	503 (1)
cFDR-minP-Sidak	0	0
cFDR-minP-3-Sidak	5 (1)	0
cFDR-Fisher-Sidak	0	0
cFDR-Fisher-3-Sidak	5, 7 (2)	0

selected body systems and numbers of flagging AE Types. Note that Naive BH and GBH procedure do not provide selection function, so by using these two procedures the BSoI selections are not applicable (NA).

cFDR-minP-Sidak and cFDR-Fisher-Sidak procedures do not select any body systems and flag no AE, but cFDR-minP-3-Sidak procedure selects the 5th body system and cFDR-Fisher-3-Sidak select the 5th and 7th body systems. There is still no AE flagged. DFDR, DFDR2 and Original BB procedures select the 5th body system and flag one AE in the body system.

3.7.2 Example 3.2

This example is from Example 4.1 in Mehrotra and Adewale (2012), there are 42 AE types across six body systems. Figure 1 in that paper shows typical summaries of tier 2 AE counts from a (hypothetical) clinical trial. The p-values and corresponding 95% confidence intervals for a difference between two independent binomial proportions using the Miettinen and Nurminen method. The no multiplicity adjustment approach flags nine AE types (101-106, 305, 404 and 507)

Table 3.3 Flagging AE Types for Example 3.2 under $\alpha_1 = 0.05$ for Selecting BSoI and $\alpha = 0.1$ for Detecting AEs

Approach	BSoI	Flagging AE Types
Naive BH	NA	106 (1)
DFDR	1,3,4,5 (4)	101-106, 305, 404 (8)
DFDR2	1 (1)	101-106 (6)
GBH	NA	101-106, 305, 404, 507 (9)
Original BB	1 (1)	106 (1)
cFDR-minP-0.1	1,3,4,5 (4)	101-106 (6)
cFDR-minP-Sidak	1 (1)	101-106 (6)
cFDR-minP-3-Sidak	1,3,4 (3)	101-106 (6)
cFDR-Fisher-Sidak	1 (1)	101-106 (6)
cFDR-Fisher-3-Sidak	1 (1)	101-106 (6)

From Table 3.3, we can see the the naive BH procedure only flags one AE types. Although double FDR and modified double FDR methods flag eight and six AE types, but it cannot guarantee FDR control. GBH method can only ensure FDR control under asymptotic case, but not finite number of AE types. Average FDR controlling procedure only select one body system and flag one AE type, which is too conservative.

The cFDR with fixed threshold $t = 0.1$ cannot provide any error control on body system level, although it selected four body systems. Our proposed cFDR using generalized Sidak selection rule and minP combining method can select one body system with $k = 1$ and three body systems with $k = 3$. When using Fisher combining method, the procedures with $k = 1$ and $k = 3$ select one body system, which guarantee FWER control and 3-FWER control at $\alpha_1 = 0.05$ across the body systems.

3.7.3 Example 3.3

This example is from Example 4.2 in Mehrotra and Adewale (2012), there are 49 AE types across nine body systems. Figure 2 in that paper shows tier 2 AE counts and related summaries for a double-blind, randomized clinical trial that was designed, in

Table 3.4 Flagging AE Types for Example 3.3 under $\alpha_1 = 0.05$ for Selecting BSoI and $\alpha = 0.1$ for Detecting AEs

Procedures	BSoI	Flagging AE Types
Naive BH	NA	703 (1)
DFDR	3, 7 (2)	301, 704 (2)
DFDR2	7 (1)	703, 704 (2)
GBH	NA	301, 401, 703, 704 (4)
Original BB	7 (1)	703 (1)
cFDR-minP-0.1	1, 2, 3, 4, 6, 7 (6)	301, 703, 704(3)
cFDR-minP-Sidak	7 (1)	703, 704 (2)
cFDR-minP-2-Sidak	3, 7 (2)	703, 704 (2)
cFDR-minP-3-Sidak	3, 4, 7 (3)	703, 704 (2)
cFDR-Fisher-Sidak	7 (1)	703, 704 (2)
cFDR-Fisher-2-Sidak	7 (1)	703, 704 (2)
cFDR-Fisher-3-Sidak	3, 4, 6, 7 (4)	703, 704 (2)

part, to compare the safety and efficacy of two medical treatments. The no multiplicity adjustment approach flags nine AE types (101, 209, 301, 305, 401, 602, 703, and 704).

Table 3.4 shows that the naive BH and original BB procedures flag one AE type (703), and double FDR and modified double FDR the other methods flag five AE types (203, 206, 211, 604 and 702).

3.7.4 Example 3.4

This example is from Example 4.3 in Mehrotra and Adewale (2012), there are 64 AE types across eight body systems. Figure 3 in that paper shows tier 2 AE counts and related summaries for another clinical trial, using the same format as in the previous three examples.The no multiplicity adjustment approach flags eight AE types (203, 206, 211, 213, 305, 603, 604, and 702).

From Table 3.5, we can see the naive BH and average FDR methods flag four AE types (203, 211, 604 and 702), and the other methods flag five AE types (203, 206, 211, 604 and 702).

Table 3.5 Flagging AE Types for Example 3.4 under $\alpha_1 = 0.05$ for Selecting BSoI and $\alpha = 0.1$ for Detecting AEs

Approach	BSoI	Flagging AE Types
Naive BH	NA	203, 211, 604, 702 (4)
DFDR	2, 3, 6, 7 (4)	203, 206, 211, 604, 702 (5)
DFDR2	2, 6, 7 (3)	203, 206, 211, 604, 702 (5)
GBH	NA	203, 206, 211, 604, 702 (5)
Origianl BB	2, 6, 7 (3)	203, 211, 604, 702 (4)
cFDR-minP-0.1	1, 2, 3, 6, 7 (5)	203, 206, 211, 604, 702 (5)
cFDR-minP-Sidak	2, 6, 7 (3)	203, 206, 211, 604, 702 (5)
cFDR-minP-3-Sidak	2, 6, 7 (3)	203, 206, 211, 604, 702 (5)
cFDR-Fisher-Sidak	2, 6, 7 (3)	203, 206, 211, 604, 702 (5)
cFDR-Fisher-3-Sidak	2, 6, 7 (3)	203, 206, 211, 604, 702 (5)

3.8 Concluding Remarks

Most existing approaches for two-stage multiple families procedures such as double FDR, modified double FDR, BB and GBH procedures do not consider the type 1 error control in both family level and individual level. Selection bias is always existing in those procedures. In the past, it is also challenge to separate selection effect and multiplicity effect if the test statistics and select statistics are dependent. In this chapter, by using conditional inference, we can make valid selective inferences. In clinical safety studies, the existing double FDR and modified double FDR procedure fail to control FDR based on our simulation studies. However, similar as these two procedures, by using minimum p-values, the proposed procedure can guarantee overall FDR control. The procedure also guarantee k-FWER control for Body System level and conditional FDR for Adverse Event level.

We summarize the comparisons for different approaches used for multiple families multiple testing procedures (MTPs) in the Table 3.6. In practice, based on the clinical safety experience, discoveries across body systems are considerable important and should be given more attentions for future research. The proposed procedure can

Table 3.6 Error Rates Control for Different MTPs with Multiple Families Structure

Approach	Family level	Within selected family	Overall
Original DFDR	×	×	×
DFDR2	FDR	×	×
GBH	×	×	global-FDR
p-filter	FDR	×	global-FDR
Original BB	FDR	×	average-FDR
cFDR-minP-t (fixed)	×	cFDR	average-FDR
cFDR-minP-k-Sidak	k-FWER	cFDR	average-FDR

provide suitable type 1 error controls across family level, within selected families and overall on all families. Based on the simulation studies, the proposed procedures using specific selection rules can outperform other existing procedures. In clinical safety studies, the proposed procedures can select some body systems of interest and efficiently flag the AEs in these body systems.

In this chapter, the recommended procedure using conditional p-value based on minP combination and generalized Sidak selection rule, which requires p-value within body system must be independent. But for dependent p-values within body system, minP or Fisher's combining method cannot be used. We can consider Brown's combining method, which is an extension of Fisher combination, but for dependent p-values combination. For any dependent global p-values across body system, generalized Bonferroni or generalized Holm procedure can be considered for selecting the body systems; for positive dependent global p-values, the generalized step-up k-FWER procedures in Sarkar (2006) can also be considered. Moreover, we can also consider applying adaptive procedures (Storey et al., 2004; Sarkar, 2008) on the conditional p-values to get more powerful procedures. Other problems related to how selection rule affects the procedures are also interest to solve, such as estimating the proportion of non-null families (containing signals) among the selected families, and the proportion of true rejections (detecting signals) among the selected non-null families.

3.9 Software

The multiple families error rate control methods described in this chapter have been implemented as a part of the MHTmult R package [Zhu and Guo, 2017], which is available online at https://cran.r-project.org/web/packages/MHTmult.

CHAPTER 4

MULTIVARIATE LOGISTIC-TYPE MODELS BASED ON AN INVERSE SAMPLING SCHEME

4.1 Introduction

In the past half a century there have been many contributions to generalized linear models (GLMs). The logistic model with the logit link function developed by Cox (1958) is a very popular generalized linear model. This model is applicable when the response variable is binary, such as taking two qualitative values (e.g. male/female, low/high, dead/survived). It is the simplest classification model. A natural extension here is the multinomial logistic model, where the dependent variable is more than two categories. The multinomial logistic regression model has been a fundamental model for developing research in deep learning or softmax regression. The moments and properties of the negative binomial (NB) distribution are given in Johnson et al. (1992). In the generalized linear model, when the response variable follows NB distribution, the variable measures the number of failures until k successes have been observed. Johnson et al. (1997) describes and analyzes a generalization of the NB distribution called the negative multinomial distribution (NMn), which can be used to develop GLM models. Bringing this distribution in GLM, Bonett (1985) proposes NMn GLM models with linear link and logit link. The GLM considered by Evans and Bonett (1989) defines a log-linear model for the multilevel contingency tables with negative multinomial frequency counts and also gives the maximum likelihood estimators.

Dhar (1995) introduces the concept of a generalized inverse sampling scheme which can be used to study several special events at a time. He derives the Extended Negative Multinomial distribution (ENMn), the distribution of the frequency counts under a generalized inverse sampling scheme. Zelterman (1997) proposes an estimate of the shape parameter based on the mean and quartiles of Pearson's χ^2 statistic. They also show that the maximum likelihood estimator (MLE) of the shape parameter

of the negative multinomial distribution cannot be obtained by directly maximizing the log-likelihood function. Using the EM algorithm, Adamidis (1999) derives the MLE of the NB distribution's shape parameter. Dhar and Lahiri (2014) proposes a log-linear GLM under the ENMn distribution used to study the incidence of cancer. The parameters of this new model are estimated by the quasi-likelihood method and the corresponding score function gives a close form estimate of the regression parameters.

Subsequently, the chapter is organized starting with Section 4.2 that introduces basic notations, concepts and desired statistical properties for the inverse sampling scheme and multivariate GLM models. In Section 4.3, a new multivariate logistic-type model is proposed based on the inverse sampling scheme and desired statistical properties of this model are discussed. Maximum likelihood estimation of the regression parameters and further inferences, such as confidence intervals, are derived in Section 4.4. Section 4.5 provides model diagnostics and application of this new model. Section 4.6 summarizes findings and discusses potential future work.

4.2 Preliminaries

Basic notations and definitions are introduced to present the multivariate logistic-type model. Many types of multivariate discrete models are seen in clinical trial and biomedical research. In particular, categorical data can arise in the experiments where the distinct outcomes are classified by factors at several levels that consists of the count number of experimental units formed by these categories. Under these settings, consider the multinomial (Mn) distribution with exactly G distinct categories and let the probability of a sample falling in the j-th category in a trial be $p_j (j = 1, ..., G)$ where $\sum_{j=1}^{G} p_j = 1$. Further, for a fixed number of independent identically distributed trials $(n > 1)$, the probability of observing exactly $y_1, \ldots, y_{G-1}$ occurrences of category $1, \ldots, G-1$, respectively, is given by $Mn(n, p_1, \ldots, p_{G-1})$. In this multinomial trials setting, consider the G-th category to be of special interest. With this background, consider the model that counts the number of individuals that fall in each of the 1 to

$G-1$ categories until exactly k individuals of the G-th category have been observed. Then the probability of exactly $y_1, \ldots, y_{G-1}$ individuals of categories $1 \ldots, G-1$, receptively, is given by the negative multinomial distribution $NMn(k, p_1, \ldots, p_{G-1})$. This model is also known as the inverse sampling scheme.

This model was further generalized by Dhar(1995) introducing the extended negative multinomial (ENMn) model, which is a generalized inverse sampling scheme. Several special events can be simultaneously analyzed using ENMn. To see its definition, draw samples until a pre-determined total number of $k \geq 1$ special events of different types that occur are observed out of total distinct types of events $G > k$. So the model is called ENMn$(k, p_1, \ldots p_G)$ model. What is interesting is that the ENMn model also deals with the response vector that counts the various categories. Bringing this feature of the ENMn distribution into the logistic model, one can propose a GLM of the logistic-type with random samples of response vector that follows this generalized inverse sampling scheme. The new GLM developed in this chapter considers log ratio of expected counts of response categories equal to the linear regression similar to the multinomial logistic model. The properties of the ENMn distribution within the multinomial logistic-type model framework makes it applicable to analyze more practical data sets in the health field. The following section formally introduces the ENMn distribution.

4.2.1 Generalized Inverse Sampling Scheme

The ENMn distribution also known as generalized inverse sampling scheme, Dhar (1995), is formally introduced. Consider a multivariate response variable to be the G-categories Mn distribution. In these multinomial categories, without loss of generality, consider the first G_0 groups of events $(E_1, \ldots, E_{G_0})$ as common (non-special) events, and the remaining $G_1 = G - G_0$ groups of events $(E_{G_0+1}, \ldots, E_G)$ as special events. Then in the multinomial trials, keep observing events until k events from the special group $1, \ldots, G_0$ are observed. The count vector of different groups (special or

non-special), denoted by $\boldsymbol{y}$ is distributed as ENMn with parameters $(k, p_1, p_2, \ldots, p_{G-1})$. The mean vector of the ENMn distribution with parameters $p_{1,i}, p_{2,i}, \ldots, p_{G,i}$ and $k_i = \sum_{j=G_0+1}^{G} y_{j,i}$ is given by

$$E\{(y_{1,i}, y_{2,i}, \ldots, y_{G,i})'\} = \frac{k_i}{\sum_{j=G_0+1}^{G} p_{j,i}} (p_{1,i}, p_{2,i}, \ldots, p_{G,i})'. \tag{4.2.1}$$

Lahiri et al. (2008) also give the variance-covariance of $\boldsymbol{y_i}$, which is a blocking diagonal matrix. Generalized inverse sampling scheme distribution is part of the exponential family as can be seen in the following section.

4.2.2 Multivariate Exponential Family

Jorgensen (1983) studies the response variable y follows the distribution from an exponential family. The exponential family of distribution has the following form

$$f(y, \theta, \kappa) = c(y, \kappa) \exp\{a(\kappa) t(y, \theta)\},$$

where y, θ, κ can be vectors. Thus, the above exponential family can be viewed as a multivariate generalization of the univariate case. Here $a(\kappa) > 0$ and θ is an m-dimensional parameter.

The ENMn distribution belongs to the exponential family since its probability distribution function can be expressed as

$$\begin{aligned} & f(y_{1,i}, \ldots, y_{G-1,i}, p_{1i}, \ldots, p_{G-1,i}, k_i) \\ =& \frac{(y_{1,i} + \cdots + y_{G_0,i} + k_i - 1)! k_i}{y_{1,i}! \cdots y_{G,i}!} p_{1,i}^{y_{1,i}} \cdots p_{G,i}^{y_{G,i}} \\ =& \frac{(y_{1,i} + \cdots + y_{G_0,i} + k_i - 1)! k_i}{y_{1,i}! \cdots y_{G,i}!} \exp\left\{ \sum_{j=1}^{G} y_{j,i} \ln(p_{j,i}) \right\}, \end{aligned} \tag{4.2.2}$$

where $y_{G,i} = k_i - \sum_{j=G_0+1}^{G-1} y_{j,i}$ and $p_{G,i} = 1 - \sum_{j=1}^{G-1} p_{j,i}$. Thus, the logistic-type GLM under ENMn is the class of GLM as considered by Jorgensen (1983). The following

section introduces a motivating example for the ENMn distribution and describes the data structure used in a logistic-type GLM under this distribution.

4.2.3 A Motivating Example of ENMn and Data

Therefore, keeping a concrete example in mind when reading this research makes it easier, consider a sequence of independent trials, which contains several kinds of liver disease diagnosis as described in Plomteux (1980). This data set is in the form of multinomial records of one set of liver disease group followed by that of another, along with the covariates. The study includes 57 cases of acute viral hepatitis (Group 1), 44 cases of persistent chronic hepatitis (Group 2), 40 cases of aggressive chronic hepatitis (Group 3), and 77 cases of post-necrotic cirrhosis (Group 4). Further, this data set consists of enzymatic activity measured for the 218 patients giving four liver enzymes as covariates: aspartate aminotransferase (AST), alanine aminotransferase (ALT), glutamate dehydrogenase (GLDH) and ornithine carbamyltransferase (OCT).

The four liver disease groups naturally form the categories of the Mn distribution. Then GLM with response vector following ENMn distribution is simulated as follows. The four groups counts are aggregated until either Group 3 or Group 4 is observed and the covariate for a sample here is taken to be that corresponding to average covariate of the Mn samples involved in it. The independent samples are achieved by randomly reordering the Plomteux (1980) data. The counts of different groups give rise to a sample of ENMn distribution with parameters $(k = 1, p_1, p_2, p_3, p_4)$ and $k = y_3 + y_4 = 1$. Since the available data was randomly reordered, one observes that this reordered data contains $n = 117$ independent observations from $ENMn(k = 1, p_1, p_2, p_3)$. In the following section, the multivariate logistic-type GLM under ENMn distribution is formally introduced.

4.3 A Multivariate Logistic-type Model under the ENMn Distribution

Now suppose there are m covariates and $G = G_0 + G_1$ distinct attributes for a response, where G_1 attributes are of special interest and the remaining are not. Then the multivariate logistic-type model under the ENMn distribution is defined as follows.

Definition 4.1 (Multivariate Logistic-type GLM)**.** *Without loss of generality, let the last G-th group be a reference group. Then,*

$$E(y_{j,i}) = \begin{cases} k_i \dfrac{\exp[\boldsymbol{x_i'}\boldsymbol{\beta^{(j)}}]}{1+\exp[\boldsymbol{x_i'}\boldsymbol{\beta^{(G-1)}}]}, & j = 1, \ldots, G-1, \\ \dfrac{k_i}{1+\exp[\boldsymbol{x_i'}\boldsymbol{\beta^{(G-1)}}]}, & j = G, \end{cases} \tag{4.3.1}$$

where vector $\boldsymbol{x_i} = \left(1, x_{i,1}, x_{i,2}, \cdots, x_{i,m}\right)'$ consists of the m covariates and"1" gives rise to the intercept regression parameter in the vector product, $i = 1, \ldots, n$. Here, $\boldsymbol{\beta^{(j)}} = \left(\beta_0^{(j)}, \beta_1^{(j)}, \cdots, \beta_m^{(j)}\right)$, $j = 1, \ldots, G-1$, are the regression coefficients.

Then the estimated response for the i-th sample and j-th category is given by using the estimators $\boldsymbol{\beta^{(j)}} = \hat{\boldsymbol{\beta}}^{\boldsymbol{(j)}}$, $j = 1, \ldots, G-1$.

$$\hat{y}_{j,i} = \begin{cases} k_i \dfrac{\exp[\boldsymbol{x_i'}\hat{\boldsymbol{\beta}}^{\boldsymbol{(j)}}]}{1+\exp[\boldsymbol{x_i'}\hat{\boldsymbol{\beta}}^{\boldsymbol{(G-1)}}]}, & j = 1, \ldots, G-1, \\ \dfrac{k_i}{1+\exp[\boldsymbol{x_i'}\hat{\boldsymbol{\beta}}^{\boldsymbol{(G-1)}}]}, & j = G. \end{cases} \tag{4.3.2}$$

Remark 4.1. The model in Definition 4.1 and (4.2.1) gives rise to the equations

$$\ln\left[\frac{E(y_{j,i})}{E(y_{G,i})}\right] = \ln\left[\frac{p_{j,i}}{p_{G,i}}\right] = \boldsymbol{x_i'}\boldsymbol{\beta^{(j)}}, \tag{4.3.3}$$

where $j = 1, \ldots, G-1$. Note that these equations are also used to define the traditional multinomial logistic model. Further, note that in the proposed model, the regression part is equal to the log ratio of the expectation of a response category to that of the baseline response category G, which is a log odds ratio.

Remark 4.2. The reference group can be any group in the categories, so it can be either a common or a special event group in the ENMn model. Without loss of generality,

in the following equations we use the last G-th group from special categories as the reference.

Hence, using (4.3.3),

$$p_{j,i} = \begin{cases} \dfrac{\exp[\boldsymbol{x}_i'\boldsymbol{\beta}^{(j)}]}{1 + \sum_{l=1}^{G-1} \exp[\boldsymbol{x}_i'\boldsymbol{\beta}^{(l)}]}, & j = 1, \ldots, G-1, \\ \dfrac{1}{1 + \sum_{l=1}^{G-1} \exp[\boldsymbol{x}_i'\boldsymbol{\beta}^{(l)}]}, & j = G, \end{cases} \tag{4.3.4}$$

which satisfies $\sum_{j=1}^{G} p_{j,i} = 1$, $i = 1, \ldots, n$, similar to the multinomial logistic regression. In this case, the plugin estimator of $p_{j,i}$ is given by

$$\hat{p}_{j,i} = \begin{cases} \dfrac{\exp[\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}^{(j)}]}{1 + \sum_{l=1}^{G-1} \exp[\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}^{(l)}]}, & j = 1, \ldots, G-1, \\ \dfrac{1}{1 + \sum_{l=1}^{G-1} \exp[\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}^{(l)}]}, & j = G, \end{cases} \tag{4.3.5}$$

which is used to describe the model diagnostic procedure in Section 4.3. Note that by (4.3.2) and (4.3.5) $\hat{p}_{j,i} = \hat{y}_{j,i}/k_i$, and a straightforward fact is that when $k_i = 1$, then $\hat{p}_{j,i} = \hat{y}_{j,i}$.

So far, the logistic-type GLM model based on an inverse sampling scheme has been defined. One can now develop the inference and diagnostics for the proposed model.

4.4 Model Inferences and Diagnostics

4.4.1 Maximum Likelihood Estimation

Estimation of the regression parameters of the proposed model using MLE theory is developed in this section. The calculation of the MLE for the regression parameter Fisher's scoring method is equivalent to an iterative weighted least squares procedure is proved by Nelder and Wedderburn (1972). Moreover, calculation of the MLE of the

regression parameter by Fisher's scoring method is equivalent to the generalized Gauss-Newton method for calculation of the least squares estimator is shown by Jorgensen (1983). These approaches are used to do inference for the regression parameter under multivariate logistic regression models by Glonek and McCullagh (1995). Similar ideas to obtain the MLE of regression parameter in the proposed model are used. The likelihood function of the ENMn model is

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\beta^{(1)}}, \ldots, \boldsymbol{\beta^{(G-1)}}, k_1, \ldots, k_n, \boldsymbol{y_1}, \ldots, \boldsymbol{y_n}) \\
&= \prod_{i=1}^{n} f(y_{1,i}, \ldots, y_{G,i}, k_i, p_{1i}, \ldots, p_{Gi}) \\
&= \prod_{i=1}^{n} \frac{(\sum_{j=1}^{G_0} y_{j,i} + k_i - 1)! k_i}{y_{1,i}! \ldots y_{G,i}!} p_{1,i}^{y_{1,i}} \cdot \ldots \cdot p_{G,i}^{y_{G,i}},
\end{aligned} \tag{4.4.1}$$

where $y_{G,i} = k_i - \sum_{j=G_0+1}^{G-1} y_{j,i}$ and $p_{G,i} = 1 - \sum_{j=1}^{G-1} p_{j,i}$. In the special case $k_i \equiv 1$, that is, one stops Mn trials when one observes an event from either one of the special events group. Then, the likelihood function becomes

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\beta^{(1)}}, \ldots, \boldsymbol{\beta^{(G-1)}}, \boldsymbol{y_1}, \ldots, \boldsymbol{y_n}) \\
&= \prod_{i=1}^{n} f(y_{1,i}, \ldots, y_{G,i}, p_{1i}, \ldots, p_{Gi}) \\
&= \prod_{i=1}^{n} \frac{\left(\sum_{j=1}^{G_0} y_{j,i}\right)! p_{1,i}^{y_{1,i}} \cdot \ldots \cdot p_{G-1,i}^{y_{G-1,i}} \cdot \left(1 - \sum_{j=1}^{G-1} p_{j,i}\right)^{1 - \sum_{j=G_0+1}^{G-1} y_{j,i}}}{y_{1,i}! \cdot \ldots \cdot y_{G-1,i}! \cdot \left(1 - \sum_{j=G_0+1}^{G-1} y_{j,i}\right)!}.
\end{aligned} \tag{4.4.2}$$

The score equations can be derived by taking first-order derivative of $\ln \mathcal{L}$ in (4.4.2) with respect to $\boldsymbol{\beta}$, giving

$$\frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta^{(1)}}} \\ \cdots \\ \frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta^{(j)}}} \\ \cdots \\ \frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta^{(G-1)}}} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} \left\{ y_{1,i} - \left(\sum_{l=1}^{G} y_{l,i} \right) p_{1,i} \right\} \boldsymbol{x_i} \\ \cdots \\ \sum_{i=1}^{n} \left\{ y_{j,i} - \left(\sum_{l=1}^{G} y_{l,i} \right) p_{j,i} \right\} \boldsymbol{x_i} \\ \cdots \\ \sum_{i=1}^{n} \left\{ y_{G-1,i} - \left(\sum_{l=1}^{G} y_{l,i} \right) p_{G-1,i} \right\} \boldsymbol{x_i} \end{pmatrix}, \tag{4.4.3}$$

where $p_{j,i}$ is as expressed in (4.3.4).

Setting the derivative in (4.4.3) as equal to $\mathbf{0}_{(G-1)\times 1}$ gives the score equations. MLE $\hat{\boldsymbol{\beta}}$ is now obtained by solving these score equations. MLE of $\boldsymbol{\beta}$ can also be iteratively obtained by the Newton-Raphson's algorithm. This algorithm is described as follows.

Algorithm 4.1.

Step 1: Start with an initial estimate $\hat{\boldsymbol{\beta}}_{(0)} = \left(\hat{\boldsymbol{\beta}}_{\boldsymbol{(0)}}^{\boldsymbol{(1)}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{(0)}}^{\boldsymbol{(2)}}, \cdots, \hat{\boldsymbol{\beta}}_{\boldsymbol{(0)}}^{\boldsymbol{(G-1)}} \right)'$. *For example, one can set initial estimate as the MLE of the multinomial logistic regression parameter.*

Step 2: Take

$$\hat{\boldsymbol{\beta}}_{(i+1)} = \hat{\boldsymbol{\beta}}_{(i)} - \left\{ \left(\frac{\partial^2 \ln \mathcal{L}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta'}} \right)^{-1} \frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta}} \right\}_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{(i)}}, \tag{4.4.4}$$

where $\hat{\boldsymbol{\beta}}_{(i)}$ *is the i-th iterated vector* $\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{(i)}}^{\boldsymbol{(1)}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{(i)}}^{\boldsymbol{(2)}}, \cdots, \hat{\boldsymbol{\beta}}_{\boldsymbol{(i)}}^{\boldsymbol{(G-1)}} \right)'$.

Step 3: Iterate Steps 1 and 2 until the sequence in (4.4.4) convergence.

Remark 4.3. The estimators of the proposed model are the same as the conventional multinomial logistic model estimators of the regression parameters. Since the kernel of the likelihood function (4.4.1) and equations in (4.3.5) are the same as multinomial logistic regression. Additional inference is developed in the following section.

4.4.2 Confidence Intervals and Tests

The Fisher scoring method is used to develop the inference. Using this method, the asymptotic variance-covariance matrix of the $\hat{\boldsymbol{\beta}}$ for logistic-type GLM under ENMn is derived Thus, second-order partial derivatives of the log likelihood with respect to $\boldsymbol{\beta}$ is computed to obtain the Hessian matrix. The components of Hessian matrix is given by the expectations of

$$\frac{\partial^2 \ln \mathcal{L}}{\partial \boldsymbol{\beta^{(j)}}^2} = -\sum_{i=1}^{n} \left\{ \left(\sum_{l=1}^{G} y_{l,i} \right) \left[p_{j,i}(1 - p_{j,i}) \right] \boldsymbol{x_i} \boldsymbol{x_i'} \right\}, \tag{4.4.5}$$

and

$$\frac{\partial^2 \ln \mathcal{L}}{\partial \boldsymbol{\beta^{(j)}} \partial \boldsymbol{\beta^{(l)}}} = \sum_{i=1}^{n} \left\{ \left(\sum_{l=1}^{G} y_{l,i} \right) \left(p_{j,i} p_{l,i} p_{G,i} \right) \boldsymbol{x_i} \boldsymbol{x_i'} \right\}, \tag{4.4.6}$$

where $j = 1, \ldots, G-1, l = 1, \ldots, G-1, j \neq l$. Since $\boldsymbol{x_i}$ is $(m+1)$-dimensional vector, each component as described in (4.4.5) and (4.4.6) of the Hessian is $(m+1) \times (m+1)$ matrix. Thus, the Hessian $\mathbf{H}$ in (4.4.7) is $(m+1)(G-1) \times (m+1)(G-1)$ matrix.

$$\mathbf{H} = E \begin{bmatrix} \dfrac{\partial^2 \ln \mathcal{L}}{\partial \boldsymbol{\beta^{(1)}}^2} & \dfrac{\partial^2 \ln \mathcal{L}}{\partial \boldsymbol{\beta^{(1)}} \partial \boldsymbol{\beta^{(2)}}} & \cdots & \dfrac{\partial^2 \ln \mathcal{L}}{\partial \boldsymbol{\beta^{(1)}} \partial \boldsymbol{\beta^{(G-1)}}} \\ \dfrac{\partial^2 \ln \mathcal{L}}{\partial \boldsymbol{\beta^{(2)}} \partial \boldsymbol{\beta^{(1)}}} & \dfrac{\partial^2 \ln \mathcal{L}}{\partial \boldsymbol{\beta^{(2)}}^2} & \cdots & \dfrac{\partial^2 \ln \mathcal{L}}{\partial \boldsymbol{\beta^{(2)}} \partial \boldsymbol{\beta^{(G-1)}}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 \ln \mathcal{L}}{\partial \boldsymbol{\beta^{(G-1)}} \partial \boldsymbol{\beta^{(1)}}} & \dfrac{\partial^2 \ln \mathcal{L}}{\partial \boldsymbol{\beta^{(G-1)}} \partial \boldsymbol{\beta^{(2)}}} & \cdots & \dfrac{\partial^2 \ln \mathcal{L}}{\partial \boldsymbol{\beta^{(G-1)}}^2} \end{bmatrix}. \tag{4.4.7}$$

Therefore, the estimator of the variance-covariance matrix based on the MLE is

$$\hat{\boldsymbol{\Sigma}} = -\mathbf{H}^{-1}{}_{|\boldsymbol{\beta^{(j)}} = \hat{\boldsymbol{\beta}}^{(j)}}. \tag{4.4.8}$$

The diagonal elements of the matrix $\hat{\boldsymbol{\Sigma}}$ are used to get the estimate of the individual variances of $\hat{\boldsymbol{\beta}}$. The confidence interval and hypothesis test for each regression parameter can be now developed.

Further, the fact that $\hat{\boldsymbol{\beta}}$ is asymptotically multivariate Gaussian is used. Then, the two-sided $100(1-\alpha)\%$ confidence intervals of the regression parameters can be derived from (4.4.8) as $\hat{\boldsymbol{\beta}}^{(j)} \pm z_{1-\alpha/2}\sqrt{diag\{\hat{Var}(\hat{\boldsymbol{\beta}}^{(j)})\}}$, where $j = 1, \ldots, G-1$. Similarly, to test $H_0 : \hat{\boldsymbol{\beta}}^{(j)} = \hat{\boldsymbol{\beta}}_{\mathbf{0}}^{(j)}$ versus $H_1 : \hat{\boldsymbol{\beta}}^{(j)} \neq \hat{\boldsymbol{\beta}}_{\mathbf{0}}^{(j)}$ for each component, $Z = \dfrac{|\hat{\boldsymbol{\beta}}^{(j)} - \hat{\boldsymbol{\beta}}_{\mathbf{0}}^{(j)}|}{\sqrt{diag\{\hat{Var}(\hat{\boldsymbol{\beta}}^{(j)})\}}}$ component-wise is used.

Remark 4.4. The variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ of the proposed model is different from that of the traditional logistic model. So, confidence intervals, tests of the regression parameters and model diagnostics are also different. Moreover, when the proposed model's distribution is correctly specified as the ENMn distribution, the model fitting will benefit from the information of special events and the stopping rule.

The goodness-of-fit of the multivariate GLM is developed in the next section.

4.4.3 Model Diagnostics

Similar to Myers et al. (2012), the deviance of the proposed model is computed as follows.

$$\begin{aligned} D &= 2\ln\mathcal{L}(\text{Satuated Model}) - 2\ln\mathcal{L}(\text{Full Model}) \\ &= 2\ln\prod_{i=1}^{n}\frac{(\sum_{j=1}^{G_0} y_{j,i}+k_i-1)!k_i}{y_{1,i}!\ldots y_{G,i}!}\tilde{p}_{1,i}^{y_{1,i}}\tilde{p}_{2,i}^{y_{2,i}}\cdot\cdots\cdot\tilde{p}_{G,i}^{y_{G,i}} \\ &\quad - 2\ln\prod_{i=1}^{n}\frac{(\sum_{j=1}^{G_0} y_{j,i}+k_i-1)!k_i}{y_{1,i}!\ldots y_{G,i}!}\hat{p}_{1,i}^{y_{1,i}}\hat{p}_{2,i}^{y_{2,i}}\cdot\cdots\cdot\hat{p}_{G,i}^{y_{G,i}} \\ &= 2\ln\prod_{i=1}^{n}\prod_{j=1}^{G}\left(\frac{\tilde{p}_{j,i}}{\hat{p}_{j,i}}\right)^{y_{j,i}}, \end{aligned} \tag{4.4.9}$$

where $\tilde{p}_{j,i} = \dfrac{y_{j,i}}{\sum_{j=1}^{G} y_{j,i}}$ is MLE of $p_{j,i}$, Dhar (1995), and $\hat{p}_{j,i}$ is as in (4.3.5). A new model fitting diagnostic based on deviance residuals is now proposed. The deviance residual for the i-th sample is

$$d_{j,i} = sgn(\tilde{p}_{j,i} - \hat{p}_{j,i})\sqrt{2\ln\left(\frac{\tilde{p}_{j,i}}{\hat{p}_{j,i}}\right)^{y_{j,i}}}, i = 1, \ldots, n; j = 1, \ldots, G. \tag{4.4.10}$$

The normality of the deviance residuals $d_{j,i}$ is used to diagnose the model fit, similar to Myers et al. (2012). Another approach is to do the model fitting diagnostic based on the Pearson residuals, instead that of the deviance, which is also used in Dhar et al. (2014). The Pearson residual is defined as

$$r_{j,i} = \frac{y_{j,i} - \hat{y}_{j,i}}{\sqrt{\hat{Var}(\hat{y}_{j,i})}}, i = 1, \ldots, n; j = 1, \ldots, G, \tag{4.4.11}$$

where $\hat{y}_{j,i}$ is obtained from (4.3.2). These model fitting are used in the application described in Section 4.5 to compare the proposed model with the multinomial logistic regression model. Also, the confidence intervals for the different models are computed.

4.5 An Application for the Proposed Model

The multivariate logistic-type GLM under ENMn is fitted to the data explained in Section 4.2.3. This section demonstrates the virtues of the proposed model in comparison with the multinomial logistic regression model. In the example, there are $G = 4$ groups (Groups 1 to 4), including $G_0 = 2$ (Groups 1 and 2) common event groups and the rest of $G_1 = 2$ (Groups 3 and 4) special event groups. Denote the covariates in the proposed model as x_1, x_2, x_3 and x_4 which are averages of the enzymes: AST, ALT, GLDH and OCT corresponding to an ENMn sample as described in Section 4.2.3. Set the last group (Group 4) as a reference group.

In order to do the model comparison, the Section 4.2.3 data is further reduced as follows. The multinomial logistic regression model is fitted to the data in Plomteux (1980) (n=218). Covariates based on large p-values are eliminated. In Tables 4.1 and 4.2 "j:covariate" represents the description of the regression parameter for the j-th category and corresponding covariate. The analysis result is shown as follows.

Table 4.1 Regression Results Applying Multinomial Logistic GLM

	Estimate	Std. Error	t-value	Pr(>|t|)
1:(intercept)	-10.734182	3.058415	-3.509721	0.000449
2:(intercept)	5.804643	2.598270	2.234041	0.025480
3:(intercept)	-6.276608	1.773263	-3.539580	0.000401
1:log(X1)	-5.171071	1.131643	-4.569525	0.000005
2:log(X1)	-6.369987	1.098285	-5.799940	0.000000
3:log(X1)	-1.439686	0.648000	-2.221736	0.026301
1:log(X2)	9.300834	1.346797	6.905891	0.000000
2:log(X2)	6.506386	1.073560	6.060569	0.000000
3:log(X2)	2.081067	0.618276	3.365919	0.000763
1:log(X3)	-2.025177	1.124202	-1.801434	0.071634
2:log(X3)	-1.880089	1.035320	-1.815949	0.069378
3:log(X3)	1.192963	0.590522	2.020183	0.043364
1:log(X4)	-1.275437	1.007589	-1.265830	0.205574
2:log(X4)	-0.671354	0.795664	-0.843766	0.398800
3:log(X4)	0.010497	0.500557	0.020971	0.983269

From Table 4.1, one can observe the covariate of $\ln(x_4)$ is not significant at all categories, so the predictor x_4 is eliminated. Thus, keeping the three covariates $\ln(x_1), \ln(x_2)$ and $\ln(x_3)$, that is, $m = 3$, the parsimonious model is obtained.

Table 4.2 Fitted Multinomial Logistic GLM on Parsimonious Model

	Estimate	Std. Error	t-value	Pr(>|t|)
1:(intercept)	-11.589083	3.087192	-3.753924	0.000174
2:(intercept)	5.509834	2.577567	2.137610	0.032548
3:(intercept)	-6.186569	1.591427	-3.887435	0.000101
1:log(X1)	-5.544992	1.139901	-4.864450	0.000001
2:log(X1)	-6.533954	1.076117	-6.071785	0.000000
3:log(X1)	-1.425277	0.639885	-2.227395	0.025921
1:log(X2)	8.784099	1.228232	7.151826	0.000000
2:log(X2)	6.217821	1.007456	6.171801	0.000000
3:log(X2)	2.063563	0.603589	3.418822	0.000629
1:log(X3)	-2.728094	0.980222	-2.783138	0.005384
2:log(X3)	-2.377281	0.786379	-3.023075	0.002502
3:log(X3)	1.185168	0.407927	2.905341	0.003669

Table 4.2 shows the three covariates are all significant. The results indicate that $\hat{\boldsymbol{\beta}} = (-11.59, 5.51, -6.19, -5.55, -6.53, -1.43, 8.78, 6.22, 2.06, -2.73, -2.38, 1.19)$. The MLE of regression parameters for the proposed model are the same as that of multinomial logistic model as mentioned in Remark 4.3. The confidence interval results for the multinomial logistic regression model are shown in Table 4.3.

Table 4.3 MLE and 95% Two-sided Confidence Interval of the Regression Parameters for Multinomial Logistic Regression Model

Index of the parameter β	MLE	Lower bound	Upper bound
1	-11.589083	-17.639172	-5.540112
2	5.509834	0.433103	10.529090
3	-6.186569	-9.304570	-3.067024
4	-5.544992	-7.786518	-3.314742
5	-6.533954	-8.643587	-4.424091
6	-1.425277	-2.678925	-0.170815
7	8.784099	6.383295	11.202516
8	6.217821	4.249069	8.203398
9	2.063563	0.880889	3.247565
10	-2.728094	-4.656752	-0.812217
11	-2.377281	-3.923184	-0.838680
12	1.185168	0.383946	1.982532

In order to estimate confidence interval of the parameters for the proposed model, one needs to calculate the variance-covariance matrix based on (4.4.7) and (4.4.8), the results are shown in (4.5.1).

$$\begin{bmatrix}
9.65 & -0.01 & -1.39 & -0.67 & 0.11 & 0.00 & -0.01 & -0.03 & 0.12 & -0.01 & -0.00 & -0.01 \\
-0.01 & 0.48 & -0.45 & 0.05 & -0.01 & 0.01 & -0.01 & -0.00 & -0.02 & 0.01 & -0.00 & -0.01 \\
-1.39 & -0.45 & 0.74 & -0.19 & -0.00 & -0.01 & 0.01 & 0.00 & 0.00 & -0.00 & -0.00 & 0.01 \\
-0.67 & 0.05 & -0.19 & 0.57 & -0.01 & -0.00 & -0.00 & 0.01 & -0.01 & -0.00 & 0.00 & 0.01 \\
0.11 & -0.01 & -0.00 & -0.01 & 7.22 & -2.07 & 0.86 & -0.84 & 0.14 & -0.02 & -0.00 & -0.01 \\
0.00 & 0.01 & -0.01 & -0.00 & -2.07 & 0.96 & -0.58 & 0.24 & 0.00 & 0.01 & -0.00 & -0.01 \\
-0.01 & -0.01 & 0.01 & -0.00 & 0.86 & -0.58 & 0.49 & -0.30 & -0.03 & -0.00 & 0.00 & 0.00 \\
-0.03 & -0.00 & 0.00 & 0.01 & -0.84 & 0.24 & -0.30 & 0.52 & 0.00 & -0.01 & 0.00 & 0.01 \\
0.12 & -0.02 & 0.00 & -0.01 & 0.14 & 0.00 & -0.03 & 0.00 & 2.43 & -0.32 & -0.03 & -0.24 \\
-0.01 & 0.01 & -0.00 & -0.00 & -0.02 & 0.01 & -0.00 & -0.01 & -0.32 & 0.24 & -0.15 & -0.03 \\
-0.00 & -0.00 & -0.00 & 0.00 & -0.00 & -0.00 & 0.00 & 0.00 & -0.03 & -0.15 & 0.16 & -0.01 \\
-0.01 & -0.01 & 0.01 & 0.01 & -0.01 & -0.01 & 0.00 & 0.01 & -0.24 & -0.03 & -0.01 & 0.17
\end{bmatrix} \tag{4.5.1}$$

Now one can derive the 95% two-sided confidence interval for the regression parameters of the proposed model, which is shown in Table 4.4.

Table 4.4 MLE and 95% Two-sided Confidence Interval of the Regression Parameters for the Proposed Logistic-type GLM using ENMn Model

Index of parameters	MLE	Lower bound	Upper bound
1	-11.589083	-17.678111	-5.500056
2	5.509834	4.158039	6.861628
3	-6.186569	-7.868941	-4.504198
4	-5.544992	-7.021851	-4.068132
5	-6.533954	-11.801088	-1.266820
6	-1.425277	-3.348624	0.498070
7	8.784099	7.411104	10.157095
8	6.217821	4.806222	7.629419
9	2.063563	-0.991926	5.119052
10	-2.728094	-3.678583	-1.777606
11	-2.377281	-3.157388	-1.597175
12	1.185168	0.381420	1.988916

To compare the above two models, the MLE and corresponding 95% two-sided confidence interval are plotted side-by-side in Figure 4.1. From Figure 4.1, one can

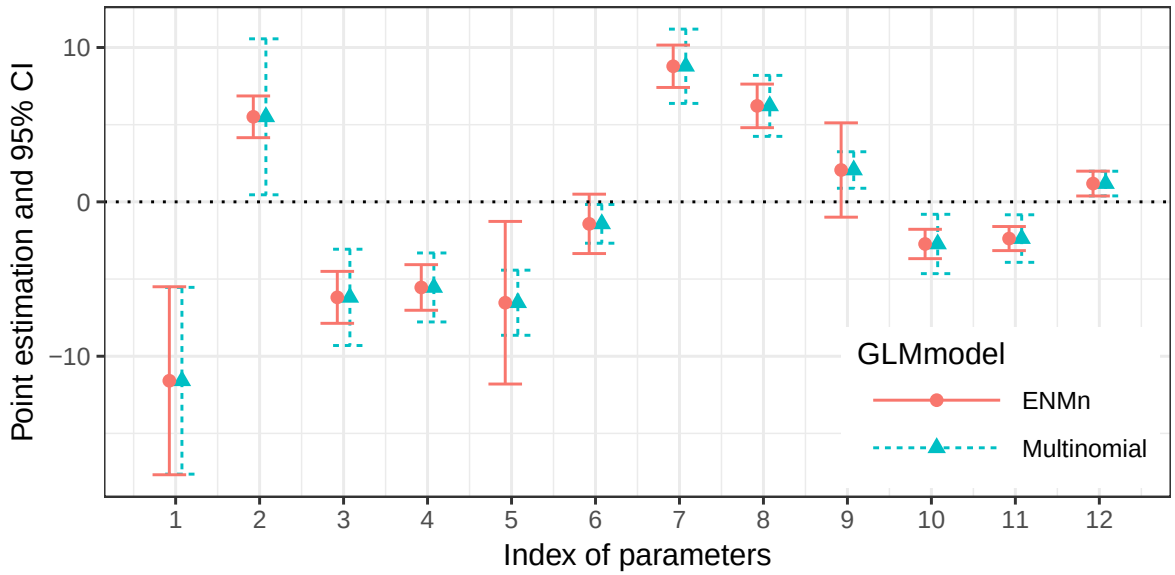

Figure 4.1 MLE and confidence interval comparisons between the proposed model and multinomial logistic regression model.

see most confidence intervals for the proposed model are shorter than the multinomial

logistic model except the β's with indices 5, 6 and 9, whose covariates are $\ln(x_1)$ for Group 2, $\ln(x_1)$ for Group 3 and $\ln(x_2)$ for Group 3, respectively. Please note here that the confidence intervals for both β_6 and β_9 show that the parameters are not significant different from zero at level 0.05. If one therefore takes zero for these two regression parameters, the proposed GLM model under the inverse sampling scheme can produce more accurate estimation than conventional model when the true response distribution is ENMn. To illustrate the deviance calculation as shown in (4.4.9), using the implemented R package, the deviance is 231.1164 for conventional multinomial logistic model and 226.7065 for the proposed model. The proposed model again shows to be a better fit than conventional one since the latter has a smaller deviance.

To make comparisons between models, the normal probability plots of the deviance residual is used. We can compare the models using normal probability or QQ-plot for deviance residuals and Pearson residuals.

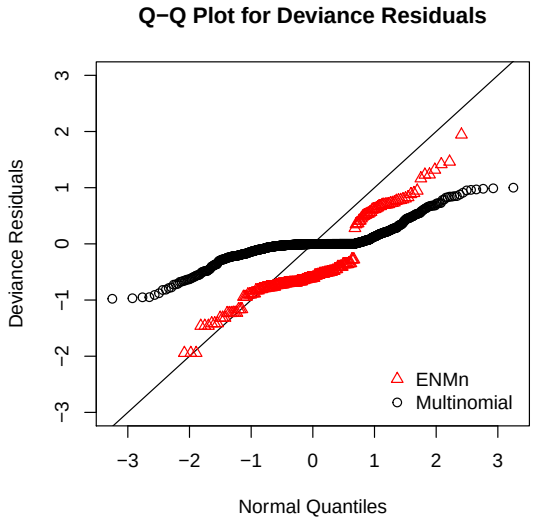

Figure 4.2 Normal probability plot for deviance residuals comparisons between multinomial logistic regression model and the proposed model.

From Figures 4.2 and 4.3, one can see that the deviance and Pearson residuals for the proposed model are spreads closer to the diagonal line than that for the multinomial logistic model, indicating the former shows a better fit. Since Pearson residual is

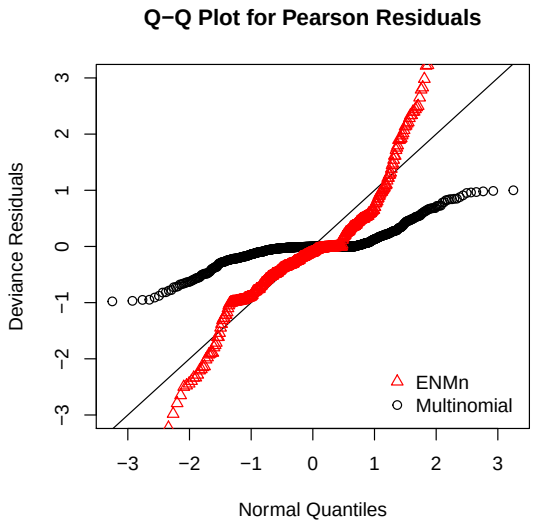

Figure 4.3 Normal probability plot for Pearson residuals comparisons between multinomial logistic regression model and the proposed model.

standardized, ideally it should be well spread from -3 to 3. However, from Figures 4.2 and 4.3, the Pearson residuals are spread out from -1 to 1 for the multinomial logistic regression model, again indicating a better fit for the proposed model.

4.6 Conclusion

This chapter presents an applicable generalized linear model using extended negative multinomial distribution (inverse sampling scheme) with the known multiple categories and log odds ratio of expected counts. The proposed model is suitable for the data that is described by recordings of count of different categories of special or non-special type occurrences and corresponding covariates. By comparison with the multinomial logistic regression, the proposed model has several benefits such as more accurate estimation (providing shorter confidence intervals), providing better goodness-of-fit, model diagnostics, and deviance.

The estimation algorithms currently uses Newton-Raphson's method to calculate the estimator of the proposed model. This method is not the only one, Qaqish and Ivanova (2006) presented an efficient algorithm for parameter estimation. A potential

work would be to develop an efficient algorithm using similar ideas, which will result in an alternative inference and model diagnostic. If the inverse sampling scheme is implemented in clinical settings, then the health industry can benefit from the optimality of the sampling design that is incorporated in the proposed model.

4.7 Software

The proposed GLM model in this chapter has been implemented in mvlogit R package [Zhu and Dhar, 2017], which will be available online at https://cran.r-project.org/web/packages/mvlogit.

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this dissertation, we considered testing multiple hypotheses with discrete data and multiple families structure and developed novel methods which exploit these inherent data structures. We also considered categorical data which contains several special events of interest and developed a logistic-type multivariate model by using inverse sampling scheme. Those multiple testing and generalized linear model methodology can be applied in clinical trial and biomedical health care research.

In Chapter 2, we have developed several FWER controlling procedures for discrete data by exploiting the information of discreteness for test statistics. The proposed procedures control FWER under various dependence structure. Real data analysis in both clinical safety studies and cDNA transcript data reveals that the proposed procedures have more chance to detect signals than conventional procedures in terms of adjusted p-values. The simulation results imply that the proposed stepwise procedures outperform the discrete procedures in terms of minimal power. A future work is to explore optimality of the suggested Procedure 2.1 and 2.3 under arbitrary dependence, which means for some joint distribution of the discrete p-values, one cannot increase even one of the critical constants while keeping the remaining fixed without losing control of the FWER. Another possible future work is to incorporate some data driven weights into our current proposed procedures to develop more powerful FWER controlling procedures for discrete data, such as weighted Hochberg type procedure.

In Chapter 3, by using conditional inference and tool of conditional p-value, we can make valid selective inferences. In clinical safety studies, the existing double FDR and modified double FDR procedure fail to control average and conditional FDR based on our simulation studies. However, by using similar screening with minimum p-values, the proposed procedure can guarantee overall FDR control. In this chapter, we

recommend a two-stage procedure using conditional p-value based on minP combination and generalized Sidak selection rule, which requires p-value within body system must be independent. But for dependent p-values within body system, minP or Fisher's combining method cannot be used. We can consider Brown's combining method, which is an extension of Fisher combination, but for dependent p-values combination. For any dependent global p-values across body system, generalized Bonferroni or generalized Holm procedure can be considered for selecting the body systems; for positive dependent global p-values, the generalized step-up k-FWER procedures in Sarkar (2006) [62] can also be considered. Moreover, we can also consider applying adaptive procedures (Storey et al., 2004; Sarkar, 2008) on the conditional p-values to get more powerful procedures. Other problems related to how selection rule affects the procedures such as estimating the proportion of non-null families (containing signals) among the selected families, the proportion of true rejections (detecting signals) among the selected non-null families, are also interesting.

In Chapter 4, we present a novel generalized linear model using extended negative multinomial (inverse sampling scheme) with the known multiple categories log ratio of expected count. The proposed model is suitable for different categories in several samples, the observations of each sample share common covariates information. By comparing with the conventional multinomial logistic regression model, the proposed model have several benefits such as more accurate estimation (providing shorter confidence intervals), have a better goodness-of-fit in model diagnostics. A potential work is to apply the inverse sampling scheme on ordinal multinomial logistic regression model, which is also widely used in social science research, such as marketing survey. One can compare the new model with the conventional ordinal multinomial logistic model in terms of MLE, confidence interval and model diagnostics.

APPENDIX A

SIMULATION RESULTS IN CHAPTER 2

This appendix contains the simulation results stated but not shown in Chapter 2.

A.1 Independent Simulation Results

The simulation results under independent setting for stepwise procedures comparisons are shown in this section. Table A.1 and Table A.2 are single-step procedures comparisons using Fisher Exact Test, and Table A.3 and Table A.4 are single-step procedures comparisons Binomial Exact Test. Table A.5 and Table A.6 are step-down procedures comparisons using Fisher Exact Test. Table A.7 and Table A.8 are step-up procedures comparisons using Fisher Exact Test.

A.2 Dependent Simulation Data Generation and Results

Step 1: Generate dependent Poisson observed counts for each group

In order to generate m dependent EBT statistics T_i, we use the following algorithm to generate m dependent Poisson random variables within each group, note the Poisson random variables between two groups are independent.

1. Generate m independent Poisson random variable $Y_{1i} \sim Poi((1-\rho)\lambda_{1i}) = Poi(2(1-\rho))$ and one $Y_{01} \sim Poi(\rho\lambda_{1i}) = Poi(2\rho)$.

2. Let $X_{1i} = Y_{1i} + Y_{10}$, then $X_{1i} \sim Poi(2)$ are dependent for $i = 1, \ldots, m$, and the correlation between X_{1i} and X_{j1} is $\dfrac{Cov(X_{1i}, X_{j1})}{\sqrt{Var(X_{1i})}\sqrt{Var(X_{j1})}} = \dfrac{Var(Y_{10})}{\sqrt{2}\sqrt{2}} = \dfrac{2\rho}{2} = \rho$ for $i, j = 1, \ldots, m$ and $i \neq j$.

3. For $i = 1, \ldots, m_0$, generate m_0 independent Poisson random variable $Y_{2i} \sim Poi((1-\rho)\lambda_{2i}) = Poi(2(1-\rho))$ and one $Y_{20} \sim Poi(\rho\lambda_{2i}) = Poi(2\rho)$. For $i = m_0 + 1, \ldots, m$, generate $m - m_0$ independent Poisson random variable $Y_{2i} \sim Poi((1-\rho)\lambda_{2i}) = Poi(10(1-\rho))$ and one $Y'_{20} \sim Poi(\rho\lambda_{2i}) = Poi(10\rho)$.

4. Let $X_{2i} = Y_{2i} + Y_{20}$ for $i = 1, \ldots, m_0$, then $X_{2i} \sim Poi(2)$ are dependent for $i = 1, \ldots, m_0$, and the correlation between X_{2i} and X_{2j} is $\frac{Cov(X_{2i}, X_{2j})}{\sqrt{Var(X_{2i})}\sqrt{Var(X_{2j})}} = \frac{Var(Y_{20})}{\sqrt{2}\sqrt{2}} = \frac{2\rho}{2} = \rho$ for $i, j = 1, \ldots, m_0$ and $i \neq j$; let $X_{2i} = Y_{2i} + Y'_{20}$ for $i = m_0 + 1, \ldots, m$, then $X_{2i} \sim Poi(10)$ are dependent for $i = m_0 + 1, \ldots, m$, and the correlation between X_{2i} and X_{2j} is $\frac{Cov(X_{2i}, X_{2j})}{\sqrt{Var(X_{2i})}\sqrt{Var(X_{2j})}} = \frac{Var(Y'_{20})}{\sqrt{10}\sqrt{10}} = \frac{10\rho}{10} = \rho$ for $i, j = m_0 + 1, \ldots, m$ and $i \neq j$.

Step 2: Obtain the conditional test statistics

Since the generated Poisson random variables between two groups are independent, we can directly conduct EBT for each hypothesis. after generating Poisson observed counts x_{1i} and x_{2i}, let $c_i = x_{1i} + x_{2i}$ be the total observed count for two groups. Then the test statistics T_i is conditional test statistics X_{i1} given $X_{1i} + X_{2i} = c_i$. Then the critical value is observed count x_{1i}.

Step 3: Conditional distribution of the test statistics

Based on the conditional inference in Lehman and Romano (2005). which is the EBT in our paper, the conditional distribution of X_{1i} given $X_{1i} + X_{2i} = c_i$ is binomial $Bin(c_i, p_i)$, where $p_i = \frac{\lambda_{1i}}{\lambda_{1i} + \lambda_{2i}}$.

Step 4: Calculate available p-value P_i and attainable p-values When H_i is true, $\lambda_{1i} = \lambda_{2i}$, then $p_i = 0.5$. That is, $X_{1i}|X_{1i} + X_{2i} = c_i \sim Bin(c_i, 0.5)$ under null H_i. Therefore, the available conditional p-value for H_i can be calculated by

$$\begin{aligned} P_i &= \Pr_{H_i}\{X_{1i} \geq x_{1i} | X_{1i} + X_{2i} = c_i\} \\ &= \sum_{j=x_{1i}}^{c_i} \binom{c_i}{j} 0.5^i (1 - 0.5)^{c_i - j} \\ &= \sum_{j=x_{1i}}^{c_i} \binom{c_i}{j} 0.5^{c_i}. \end{aligned} \tag{A.2.1}$$

The corresponding attainable p-values can be calculated by

$$\Pr_{H_i}\{X_{1i} \geq x | X_{1i} + X_{2i} = c_i\} = \sum_{j=x}^{c_i} \binom{c_i}{j} 0.5^{c_i} \text{ for } x = 0, 1, \ldots, c_i. \tag{A.2.2}$$

Table A.1 Simulated FWER Comparisons for Single-step Procedures with Independent p-values Generated from Fisher's Exact Test Statistics

		$N = 25$	$N = 50$	$N = 75$	$N = 100$	$N = 125$	$N = 150$
$m = 5$	MBonf	0.0025	0.0060	0.0035	0.0075	0.0075	0.0095
	Tarone	0.0015	0.0030	0.0015	0.0055	0.0045	0.0085
	Bonf	0.0010	0.0030	0.0015	0.0055	0.0045	0.0085
$\pi_0 = 0.2$	Sidak	0.0010	0.0030	0.0015	0.0055	0.0045	0.0085
$m = 5$	MBonf	0.0045	0.0130	0.0120	0.0170	0.0135	0.0145
	Tarone	0.0030	0.0060	0.0065	0.0140	0.0090	0.0100
	Bonf	0.0015	0.0060	0.0065	0.0140	0.0090	0.0100
$\pi_0 = 0.4$	Sidak	0.0015	0.0060	0.0065	0.0140	0.0090	0.0100
$m = 5$	MBonf	0.0085	0.0200	0.0195	0.0235	0.0225	0.0245
	Tarone	0.0060	0.0105	0.0105	0.0180	0.0155	0.0170
	Bonf	0.0025	0.0100	0.0105	0.0180	0.0155	0.0170
$\pi_0 = 0.6$	Sidak	0.0025	0.0100	0.0105	0.0180	0.0160	0.0175
$m = 5$	MBonf	0.0140	0.0265	0.0270	0.0340	0.0315	0.0370
	Tarone	0.0110	0.0140	0.0155	0.0245	0.0215	0.0220
	Bonf	0.0045	0.0135	0.0155	0.0245	0.0215	0.0220
$\pi_0 = 0.8$	Sidak	0.0045	0.0135	0.0155	0.0245	0.0220	0.0230
$m = 10$	MBonf	0.0020	0.0060	0.0100	0.0115	0.0095	0.0110
	Tarone	0.0005	0.0040	0.0065	0.0060	0.0070	0.0060
	Bonf	0.0005	0.0040	0.0065	0.0060	0.0070	0.0060
$\pi_0 = 0.2$	Sidak	0.0005	0.0040	0.0065	0.0060	0.0070	0.0060
$m = 10$	MBonf	0.0050	0.0145	0.0165	0.0190	0.0215	0.0190
	Tarone	0.0025	0.0090	0.0120	0.0100	0.0140	0.0125
	Bonf	0.0025	0.0090	0.0120	0.0100	0.0140	0.0125
$\pi_0 = 0.4$	Sidak	0.0025	0.0090	0.0120	0.0110	0.0145	0.0130
$m = 10$	MBonf	0.0090	0.0245	0.0260	0.0265	0.0300	0.0255
	Tarone	0.0055	0.0150	0.0185	0.0150	0.0180	0.0155
	Bonf	0.0045	0.0140	0.0185	0.0150	0.0180	0.0155
$\pi_0 = 0.6$	Sidak	0.0045	0.0140	0.0185	0.0160	0.0195	0.0155
$m = 10$	MBonf	0.0175	0.0335	0.0345	0.0370	0.0390	0.0360
	Tarone	0.0090	0.0215	0.0225	0.0190	0.0220	0.0200
	Bonf	0.0055	0.0190	0.0225	0.0190	0.0220	0.0200
$\pi_0 = 0.8$	Sidak	0.0055	0.0190	0.0225	0.0210	0.0240	0.0200
$m = 15$	MBonf	0.0040	0.0060	0.0065	0.0120	0.0080	0.0100
	Tarone	0.0020	0.0030	0.0030	0.0065	0.0045	0.0070
	Bonf	0.0005	0.0030	0.0030	0.0065	0.0045	0.0070
$\pi_0 = 0.2$	Sidak	0.0005	0.0030	0.0030	0.0075	0.0045	0.0070
$m = 15$	MBonf	0.0090	0.0150	0.0140	0.0240	0.0210	0.0200
	Tarone	0.0060	0.0075	0.0065	0.0125	0.0150	0.0105
	Bonf	0.0010	0.0070	0.0065	0.0125	0.0150	0.0105
$\pi_0 = 0.4$	Sidak	0.0010	0.0070	0.0065	0.0145	0.0150	0.0105
$m = 15$	MBonf	0.0165	0.0250	0.0210	0.0325	0.0320	0.0280
	Tarone	0.0090	0.0130	0.0095	0.0170	0.0205	0.0180
	Bonf	0.0020	0.0105	0.0095	0.0170	0.0205	0.0180
$\pi_0 = 0.6$	Sidak	0.0020	0.0105	0.0095	0.0190	0.0205	0.0180
$m = 15$	MBonf	0.0210	0.0345	0.0315	0.0400	0.0460	0.0360
	Tarone	0.0115	0.0170	0.0155	0.0215	0.0285	0.0240
	Bonf	0.0020	0.0135	0.0155	0.0215	0.0285	0.0240
$\pi_0 = 0.8$	Sidak	0.0020	0.0135	0.0155	0.0240	0.0285	0.0240

Table A.2 Simulated Minimal Power Comparisons for Single-step Procedures with Independent p-values Generated from Fisher's Exact Test Statistics

		$N = 25$	$N = 50$	$N = 75$	$N = 100$	$N = 125$	$N = 150$
$m = 5$	MBonf	0.2550	0.5060	0.6855	0.8195	0.9145	0.9505
	Tarone	0.1945	0.3900	0.5775	0.7680	0.8655	0.9275
$\pi_0 = 0.2$	Bonf	0.1125	0.3825	0.5765	0.7680	0.8655	0.9275
	Sidak	0.1125	0.3825	0.5850	0.7680	0.8710	0.9340
$m = 5$	MBonf	0.2110	0.4085	0.5785	0.7405	0.8375	0.9025
	Tarone	0.1605	0.3110	0.4715	0.6705	0.7695	0.8625
$\pi_0 = 0.4$	Bonf	0.0880	0.3000	0.4700	0.6705	0.7695	0.8625
	Sidak	0.0880	0.3000	0.4770	0.6705	0.7765	0.8680
$m = 5$	MBonf	0.1550	0.3130	0.4320	0.5835	0.7025	0.7845
	Tarone	0.1180	0.2365	0.3370	0.5145	0.6255	0.7245
$\pi_0 = 0.6$	Bonf	0.0605	0.2190	0.3355	0.5145	0.6255	0.7245
	Sidak	0.0605	0.2190	0.3420	0.5145	0.6330	0.7345
$m = 5$	MBonf	0.0945	0.1800	0.2570	0.3595	0.4660	0.5505
	Tarone	0.0740	0.1330	0.1920	0.2955	0.3950	0.4850
$\pi_0 = 0.8$	Bonf	0.0330	0.1190	0.1920	0.2955	0.3950	0.4850
	Sidak	0.0330	0.1190	0.1955	0.2955	0.4025	0.5005
$m = 10$	MBonf	0.3155	0.6130	0.8090	0.9110	0.9765	0.9930
	Tarone	0.2075	0.4695	0.7220	0.8550	0.9415	0.9820
$\pi_0 = 0.2$	Bonf	0.1575	0.4660	0.7220	0.8550	0.9415	0.9820
	Sidak	0.1575	0.4660	0.7220	0.8595	0.9425	0.9830
$m = 10$	MBonf	0.2700	0.5220	0.7180	0.8455	0.9440	0.9750
	Tarone	0.1770	0.3905	0.6065	0.7720	0.8905	0.9505
$\pi_0 = 0.4$	Bonf	0.1235	0.3795	0.6065	0.7720	0.8905	0.9505
	Sidak	0.1235	0.3795	0.6065	0.7775	0.8920	0.9575
$m = 10$	MBonf	0.2005	0.4030	0.5615	0.7300	0.8450	0.9035
	Tarone	0.1330	0.2990	0.4525	0.6315	0.7590	0.8525
$\pi_0 = 0.6$	Bonf	0.0800	0.2825	0.4525	0.6315	0.7590	0.8525
	Sidak	0.0800	0.2825	0.4525	0.6375	0.7615	0.8585
$m = 10$	MBonf	0.1115	0.2440	0.3500	0.4775	0.6140	0.6935
	Tarone	0.0760	0.1680	0.2645	0.3810	0.5165	0.6060
$\pi_0 = 0.8$	Bonf	0.0390	0.1555	0.2645	0.3810	0.5165	0.6060
	Sidak	0.0390	0.1555	0.2645	0.3880	0.5170	0.6185
$m = 15$	MBonf	0.3370	0.6715	0.8820	0.9495	0.9915	0.9965
	Tarone	0.2520	0.4995	0.7530	0.8910	0.9765	0.9895
$\pi_0 = 0.2$	Bonf	0.1390	0.4870	0.7515	0.8910	0.9765	0.9895
	Sidak	0.1390	0.4870	0.7515	0.8960	0.9765	0.9895
$m = 15$	MBonf	0.2880	0.5815	0.7910	0.9025	0.9635	0.9830
	Tarone	0.2110	0.4105	0.6475	0.8050	0.9335	0.9745
$\pi_0 = 0.4$	Bonf	0.1030	0.3870	0.6460	0.8050	0.9335	0.9745
	Sidak	0.1030	0.3870	0.6460	0.8125	0.9335	0.9745
$m = 15$	MBonf	0.2135	0.4485	0.6570	0.7925	0.8840	0.9500
	Tarone	0.1495	0.3070	0.5085	0.6730	0.8315	0.9140
$\pi_0 = 0.6$	Bonf	0.0700	0.2760	0.5065	0.6730	0.8315	0.9140
	Sidak	0.0700	0.2760	0.5065	0.6790	0.8315	0.9140
$m = 15$	MBonf	0.1205	0.2635	0.4270	0.5490	0.6710	0.7780
	Tarone	0.0830	0.1785	0.3050	0.4295	0.5890	0.7020
$\pi_0 = 0.8$	Bonf	0.0335	0.1480	0.3040	0.4290	0.5890	0.7020
	Sidak	0.0335	0.1480	0.3040	0.4345	0.5895	0.7020

Table A.3 Simulated FWER Comparisons for Single-step Procedures with Independent p-values Generated from Binomial Exact Test Statistics

		$\pi_0 = 0.2$	$\pi_0 = 0.4$	$\pi_0 = 0.6$	$\pi_0 = 0.8$
	MBonf	0.0020	0.0060	0.0075	0.0165
$m = 5$	Tarone	0.0010	0.0030	0.0055	0.0105
$\alpha = 0.05$	Bonf	0.0010	0.0020	0.0025	0.0030
	Sidak	0.0010	0.0020	0.0025	0.0030
	MBonf	0.0010	0.0045	0.0130	0.0160
$m = 10$	Tarone	0.0000	0.0010	0.0050	0.0115
$\alpha = 0.05$	Bonf	0.0000	0.0005	0.0025	0.0025
	Sidak	0.0000	0.0005	0.0025	0.0025
	MBonf	0.0010	0.0065	0.0045	0.0150
$m = 15$	Tarone	0.0000	0.0010	0.0020	0.0070
$\alpha = 0.05$	Bonf	0.0000	0.0005	0.0000	0.0000
	Sidak	0.0000	0.0005	0.0000	0.0000
	MBonf	0.0070	0.0125	0.0200	0.0365
$m = 5$	Tarone	0.0020	0.0065	0.0110	0.0285
$\alpha = 0.1$	Bonf	0.0020	0.0055	0.0065	0.0130
	Sidak	0.0020	0.0055	0.0065	0.0130
	MBonf	0.0040	0.0080	0.0275	0.0350
$m = 10$	Tarone	0.0000	0.0030	0.0165	0.0195
$\alpha = 0.1$	Bonf	0.0000	0.0015	0.0055	0.0060
	Sidak	0.0000	0.0015	0.0055	0.0060
	MBonf	0.0060	0.0155	0.0185	0.0315
$m = 15$	Tarone	0.0005	0.0060	0.0045	0.0200
$\alpha = 0.1$	Bonf	0.0000	0.0010	0.0020	0.0025
	Sidak	0.0000	0.0010	0.0020	0.0025

Table A.4 Simulated Minimal Power Comparisons for Single-step Procedures with Independent p-values Generated from Binomial Exact Test Statistics

		$\pi_0 = 0.2$	$\pi_0 = 0.4$	$\pi_0 = 0.6$	$\pi_0 = 0.8$
	MBonf	0.9205	0.8805	0.7845	0.5565
$m = 5$	Tarone	0.8815	0.8240	0.7395	0.5235
$\alpha = 0.05$	Bonf	0.8735	0.8055	0.6610	0.4045
	Sidak	0.8735	0.8055	0.6610	0.4045
	MBonf	0.9850	0.9635	0.9035	0.7390
$m = 10$	Tarone	0.9470	0.9240	0.8630	0.6855
$\alpha = 0.05$	Bonf	0.9315	0.8635	0.7050	0.4775
	Sidak	0.9315	0.8635	0.7050	0.4775
	MBonf	0.9925	0.9810	0.9555	0.8210
$m = 15$	Tarone	0.9825	0.9500	0.9095	0.7845
$\alpha = 0.05$	Bonf	0.9820	0.9475	0.8560	0.6135
	Sidak	0.9820	0.9475	0.8560	0.6135
	MBonf	0.9680	0.9415	0.8615	0.6330
$m = 5$	Tarone	0.9410	0.9140	0.8240	0.5920
$\alpha = 0.1$	Bonf	0.9050	0.8375	0.7040	0.4520
	Sidak	0.9050	0.8375	0.7040	0.4520
	MBonf	0.9965	0.9875	0.9620	0.8315
$m = 10$	Tarone	0.9885	0.9660	0.9170	0.7835
$\alpha = 0.1$	Bonf	0.9870	0.9565	0.8690	0.6600
	Sidak	0.9870	0.9565	0.8690	0.6600
	MBonf	0.9995	0.9970	0.9830	0.9030
$m = 15$	Tarone	0.9960	0.9930	0.9605	0.8400
$\alpha = 0.1$	Bonf	0.9880	0.9615	0.8830	0.6515
	Sidak	0.9895	0.9635	0.8880	0.6590

Table A.5 Simulated FWER Comparisons for Step-down Procedures with Independent p-values Generated from Fisher's Exact Test Statistics

		$N = 25$	$N = 50$	$N = 75$	$N = 100$	$N = 125$	$N = 150$
$m = 5$	MHolm	0.0030	0.0090	0.0065	0.0115	0.0150	0.0150
	TH^*	0.0015	0.0045	0.0030	0.0075	0.0090	0.0140
$\pi_0 = 0.2$	Holm	0.0010	0.0045	0.0030	0.0075	0.0090	0.0140
$m = 5$	MHolm	0.0055	0.0155	0.0135	0.0230	0.0225	0.0225
	TH^*	0.0030	0.0080	0.0080	0.0185	0.0140	0.0180
$\pi_0 = 0.4$	Holm	0.0020	0.0075	0.0080	0.0185	0.0140	0.0180
$m = 5$	MHolm	0.0100	0.0215	0.0215	0.0290	0.0305	0.0320
	TH^*	0.0065	0.0115	0.0115	0.0220	0.0185	0.0205
$\pi_0 = 0.6$	Holm	0.0030	0.0110	0.0115	0.0220	0.0185	0.0205
$m = 5$	MHolm	0.0155	0.0285	0.0285	0.0360	0.0375	0.0440
	TH^*	0.0115	0.0145	0.0160	0.0260	0.0240	0.0270
$\pi_0 = 0.8$	Holm	0.0050	0.0140	0.0160	0.0260	0.0240	0.0270
$m = 10$	MHolm	0.0020	0.0070	0.0125	0.0130	0.0160	0.0185
	TH^*	0.0005	0.0040	0.0070	0.0080	0.0115	0.0125
$\pi_0 = 0.2$	Holm	0.0005	0.0040	0.0070	0.0080	0.0115	0.0125
$m = 10$	MHolm	0.0050	0.0155	0.0200	0.0215	0.0280	0.0265
	TH^*	0.0025	0.0090	0.0125	0.0125	0.0200	0.0175
$\pi_0 = 0.4$	Holm	0.0025	0.0090	0.0125	0.0125	0.0200	0.0175
$m = 10$	MHolm	0.0095	0.0250	0.0285	0.0290	0.0360	0.0350
	TH^*	0.0060	0.0150	0.0185	0.0155	0.0220	0.0215
$\pi_0 = 0.6$	Holm	0.0045	0.0140	0.0185	0.0155	0.0220	0.0215
$m = 10$	MHolm	0.0175	0.0340	0.0360	0.0380	0.0420	0.0405
	TH^*	0.0090	0.0215	0.0235	0.0195	0.0255	0.0230
$\pi_0 = 0.8$	Holm	0.0055	0.0190	0.0225	0.0195	0.0255	0.0230
$m = 15$	MHolm	0.0045	0.0070	0.0070	0.0140	0.0125	0.0120
	TH^*	0.0025	0.0035	0.0030	0.0090	0.0060	0.0085
$\pi_0 = 0.2$	Holm	0.0005	0.0030	0.0030	0.0090	0.0060	0.0085
$m = 15$	MHolm	0.0095	0.0165	0.0145	0.0255	0.0255	0.0285
	TH^*	0.0060	0.0080	0.0075	0.0160	0.0175	0.0165
$\pi_0 = 0.4$	Holm	0.0010	0.0070	0.0075	0.0160	0.0175	0.0165
$m = 15$	MHolm	0.0165	0.0260	0.0215	0.0345	0.0350	0.0345
	TH^*	0.0090	0.0130	0.0105	0.0190	0.0215	0.0195
$\pi_0 = 0.6$	Holm	0.0020	0.0105	0.0100	0.0190	0.0215	0.0195
$m = 15$	MHolm	0.0215	0.0350	0.0315	0.0415	0.0465	0.0390
	TH^*	0.0120	0.0170	0.0165	0.0225	0.0290	0.0260
$\pi_0 = 0.8$	Holm	0.0020	0.0135	0.0165	0.0225	0.0290	0.0260

Table A.6 Simulated Minimal Power Comparisons for Step-down Procedures with Independent p-values Generated from Fisher's Exact Test Statistics

		$N = 25$	$N = 50$	$N = 75$	$N = 100$	$N = 125$	$N = 150$
$m = 5$	MHolm	0.2555	0.5070	0.6855	0.8200	0.9145	0.9505
	TH^*	0.1945	0.3905	0.5780	0.7680	0.8660	0.9280
$\pi_0 = 0.2$	Holm	0.1130	0.3830	0.5770	0.7680	0.8660	0.9280
$m = 5$	MHolm	0.2120	0.4090	0.5790	0.7405	0.8375	0.9030
	TH^*	0.1605	0.3115	0.4725	0.6705	0.7695	0.8630
$\pi_0 = 0.4$	Holm	0.0880	0.3005	0.4710	0.6705	0.7695	0.8630
$m = 5$	MHolm	0.1555	0.3150	0.4330	0.5855	0.7035	0.7855
	TH^*	0.1185	0.2365	0.3375	0.5160	0.6265	0.7260
$\pi_0 = 0.6$	Holm	0.0605	0.2190	0.3360	0.5160	0.6265	0.7260
$m = 5$	MHolm	0.0950	0.1815	0.2585	0.3615	0.4690	0.5530
	TH^*	0.0745	0.1330	0.1920	0.2965	0.3960	0.4860
$\pi_0 = 0.8$	Holm	0.0330	0.1190	0.1920	0.2965	0.3960	0.4860
$m = 10$	MHolm	0.3160	0.6130	0.8095	0.9120	0.9765	0.9930
	TH^*	0.2075	0.4700	0.7220	0.8550	0.9415	0.9820
$\pi_0 = 0.2$	Holm	0.1575	0.4660	0.7220	0.8550	0.9415	0.9820
$m = 10$	MHolm	0.2705	0.5220	0.7185	0.8455	0.9445	0.9750
	TH^*	0.1770	0.3905	0.6065	0.7720	0.8905	0.9505
$\pi_0 = 0.4$	Holm	0.1235	0.3795	0.6065	0.7720	0.8905	0.9505
$m = 10$	MHolm	0.2010	0.4035	0.5615	0.7300	0.8450	0.9035
	TH^*	0.1330	0.2990	0.4525	0.6315	0.7590	0.8525
$\pi_0 = 0.6$	Holm	0.0800	0.2825	0.4525	0.6315	0.7590	0.8525
$m = 10$	MHolm	0.1115	0.2440	0.3500	0.4780	0.6145	0.6935
	TH^*	0.0760	0.1680	0.2645	0.3810	0.5175	0.6065
$\pi_0 = 0.8$	Holm	0.0390	0.1555	0.2645	0.3810	0.5175	0.6065
$m = 15$	MHolm	0.3375	0.6715	0.8820	0.9495	0.9915	0.9965
	TH^*	0.2520	0.4995	0.7530	0.8910	0.9765	0.9895
$\pi_0 = 0.2$	Holm	0.1390	0.4870	0.7515	0.8910	0.9765	0.9895
$m = 15$	MHolm	0.2885	0.5825	0.7915	0.9025	0.9635	0.9830
	TH^*	0.2110	0.4105	0.6475	0.8055	0.9335	0.9745
$\pi_0 = 0.4$	Holm	0.1030	0.3870	0.6460	0.8055	0.9335	0.9745
$m = 15$	MHolm	0.2135	0.4495	0.6575	0.7930	0.8840	0.9500
	TH^*	0.1495	0.3070	0.5085	0.6730	0.8315	0.9140
$\pi_0 = 0.6$	Holm	0.0700	0.2760	0.5065	0.6730	0.8315	0.9140
$m = 15$	MHolm	0.1205	0.2645	0.4280	0.5495	0.6730	0.7780
	TH^*	0.0835	0.1785	0.3055	0.4295	0.5890	0.7030
$\pi_0 = 0.8$	Holm	0.0335	0.1480	0.3045	0.4290	0.5890	0.7030

Table A.7 Simulated FWER Comparisons for Step-up Procedures with Independent p-values Generated from Fisher's Exact Test Statistics

		$N = 25$	$N = 50$	$N = 75$	$N = 100$	$N = 125$	$N = 150$
$m = 5$	MHoch	0.0030	0.0090	0.0070	0.0115	0.0150	0.0155
	Roth	0.0020	0.0045	0.0040	0.0085	0.0115	0.0155
$\pi_0 = 0.2$	Hoch	0.0015	0.0045	0.0040	0.0085	0.0115	0.0155
$m = 5$	MHoch	0.0060	0.0155	0.0140	0.0235	0.0230	0.0245
	Roth	0.0035	0.0080	0.0085	0.0185	0.0160	0.0200
$\pi_0 = 0.4$	Hoch	0.0025	0.0075	0.0085	0.0185	0.0160	0.0200
$m = 5$	MHoch	0.0105	0.0215	0.0215	0.0290	0.0305	0.0325
	Roth	0.0065	0.0115	0.0115	0.0220	0.0195	0.0215
$\pi_0 = 0.6$	Hoch	0.0030	0.0110	0.0115	0.0220	0.0195	0.0215
$m = 5$	MHoch	0.0160	0.0285	0.0285	0.0360	0.0380	0.0445
	Roth	0.0115	0.0145	0.0160	0.0265	0.0245	0.0280
$\pi_0 = 0.8$	Hoch	0.0050	0.0140	0.0160	0.0265	0.0245	0.0280
$m = 10$	MHoch	0.0025	0.0070	0.0125	0.0140	0.0170	0.0200
	Roth	0.0005	0.0040	0.0070	0.0080	0.0120	0.0135
$\pi_0 = 0.2$	Hoch	0.0005	0.0040	0.0070	0.0080	0.0120	0.0135
$m = 10$	MHoch	0.0055	0.0155	0.0200	0.0225	0.0290	0.0275
	Roth	0.0025	0.0090	0.0125	0.0125	0.0200	0.0185
$\pi_0 = 0.4$	Hoch	0.0025	0.0090	0.0125	0.0125	0.0200	0.0185
$m = 10$	MHoch	0.0095	0.0250	0.0285	0.0290	0.0360	0.0350
	Roth	0.0060	0.0150	0.0185	0.0155	0.0220	0.0215
$\pi_0 = 0.6$	Hoch	0.0045	0.0140	0.0185	0.0155	0.0220	0.0215
$m = 10$	MHoch	0.0180	0.0340	0.0360	0.0380	0.0420	0.0405
	Roth	0.0095	0.0210	0.0235	0.0195	0.0255	0.0235
$\pi_0 = 0.8$	Hoch	0.0055	0.0190	0.0225	0.0195	0.0255	0.0235
$m = 15$	MHoch	0.0045	0.0070	0.0070	0.0140	0.0125	0.0130
	Roth	0.0020	0.0035	0.0030	0.0090	0.0060	0.0095
$\pi_0 = 0.2$	Hoch	0.0005	0.0030	0.0030	0.0090	0.0060	0.0095
$m = 15$	MHoch	0.0100	0.0165	0.0145	0.0255	0.0255	0.0290
	Roth	0.0060	0.0080	0.0075	0.0160	0.0175	0.0165
$\pi_0 = 0.4$	Hoch	0.0010	0.0070	0.0075	0.0160	0.0175	0.0165
$m = 15$	MHoch	0.0175	0.0260	0.0215	0.0345	0.0350	0.0345
	Roth	0.0090	0.0130	0.0105	0.0190	0.0220	0.0195
$\pi_0 = 0.6$	Hoch	0.0020	0.0105	0.0100	0.0190	0.0220	0.0195
$m = 15$	MHoch	0.0215	0.0350	0.0315	0.0415	0.0465	0.0390
	Roth	0.0120	0.0170	0.0165	0.0225	0.0290	0.0265
$\pi_0 = 0.8$	Hoch	0.0020	0.0135	0.0165	0.0225	0.0290	0.0265

Table A.8 Simulated Minimal Power Comparisons for Step-up Procedures with Independent p-values Generated from Fisher's Exact Test Statistics

		$N = 25$	$N = 50$	$N = 75$	$N = 100$	$N = 125$	$N = 150$
$m = 5$	MHoch	0.2600	0.5075	0.6885	0.8240	0.9170	0.9525
	Roth	0.1975	0.3915	0.5820	0.7685	0.8695	0.9300
$\pi_0 = 0.2$	Hoch	0.1170	0.3845	0.5810	0.7685	0.8695	0.9300
$m = 5$	MHoch	0.2155	0.4105	0.5810	0.7410	0.8400	0.9055
	Roth	0.1630	0.3115	0.4755	0.6705	0.7715	0.8660
$\pi_0 = 0.4$	Hoch	0.0885	0.3010	0.4740	0.6705	0.7715	0.8660
$m = 5$	MHoch	0.1580	0.3155	0.4340	0.5860	0.7045	0.7875
	Roth	0.1200	0.2365	0.3380	0.5165	0.6280	0.7275
$\pi_0 = 0.6$	Hoch	0.0605	0.2190	0.3365	0.5165	0.6280	0.7275
$m = 5$	MHoch	0.0955	0.1815	0.2585	0.3615	0.4695	0.5535
	Roth	0.0745	0.1330	0.1920	0.2970	0.3965	0.4870
$\pi_0 = 0.8$	Hoch	0.0330	0.1190	0.1920	0.2970	0.3965	0.4870
$m = 10$	MHoch	0.3215	0.6155	0.8110	0.9130	0.9765	0.9930
	Roth	0.2080	0.4685	0.7225	0.8555	0.9420	0.9820
$\pi_0 = 0.2$	Hoch	0.1580	0.4660	0.7225	0.8555	0.9420	0.9820
$m = 10$	MHoch	0.2735	0.5245	0.7200	0.8465	0.9450	0.9755
	Roth	0.1770	0.3840	0.6070	0.7720	0.8920	0.9510
$\pi_0 = 0.4$	Hoch	0.1240	0.3795	0.6065	0.7720	0.8920	0.9510
$m = 10$	MHoch	0.2030	0.4045	0.5615	0.7310	0.8450	0.9045
	Roth	0.1335	0.2910	0.4525	0.6315	0.7600	0.8530
$\pi_0 = 0.6$	Hoch	0.0800	0.2825	0.4525	0.6315	0.7600	0.8530
$m = 10$	MHoch	0.1135	0.2440	0.3500	0.4780	0.6150	0.6935
	Roth	0.0765	0.1625	0.2645	0.3810	0.5175	0.6075
$\pi_0 = 0.8$	Hoch	0.0390	0.1555	0.2645	0.3810	0.5175	0.6075
$m = 15$	MHoch	0.3405	0.6720	0.8830	0.9505	0.9915	0.9965
	Roth	0.2520	0.5010	0.7545	0.8910	0.9765	0.9900
$\pi_0 = 0.2$	Hoch	0.1390	0.4875	0.7535	0.8910	0.9765	0.9900
$m = 15$	MHoch	0.2895	0.5830	0.7925	0.9025	0.9635	0.9830
	Roth	0.2110	0.4115	0.6485	0.8060	0.9335	0.9745
$\pi_0 = 0.4$	Hoch	0.1030	0.3870	0.6470	0.8060	0.9335	0.9745
$m = 15$	MHoch	0.2150	0.4500	0.6595	0.7935	0.8845	0.9505
	Roth	0.1495	0.3080	0.5095	0.6730	0.8320	0.9150
$\pi_0 = 0.6$	Hoch	0.0700	0.2760	0.5075	0.6730	0.8320	0.9150
$m = 15$	MHoch	0.1210	0.2645	0.4285	0.5500	0.6730	0.7780
	Roth	0.0835	0.1785	0.3055	0.4295	0.5895	0.7035
$\pi_0 = 0.8$	Hoch	0.0335	0.1480	0.3045	0.4290	0.5895	0.7035

Table A.9 Simulated FWER Comparisons for Single-step Procedures with Dependent p-values Generated from Binomial Exact Test Statistics

	ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$m = 5$ $\pi_0 = 0.4$	MBonf	0.0045	0.0050	0.0035	0.0045	0.0025	0.0055	0.0025	0.0080	0.0045	0.0040
	Tarone	0.0015	0.0010	0.0005	0.0020	0.0010	0.0030	0.0005	0.0040	0.0010	0.0005
	Bonf	0.0015	0.0010	0.0005	0.0015	0.0010	0.0025	0.0005	0.0035	0.0000	0.0005
	Sidak	0.0015	0.0010	0.0005	0.0015	0.0010	0.0025	0.0005	0.0035	0.0000	0.0005
$m = 5$ $\pi_0 = 0.6$	MBonf	0.0100	0.0055	0.0080	0.0085	0.0080	0.0050	0.0095	0.0070	0.0045	0.0040
	Tarone	0.0060	0.0025	0.0060	0.0050	0.0050	0.0045	0.0045	0.0055	0.0040	0.0025
	Bonf	0.0025	0.0015	0.0020	0.0035	0.0020	0.0010	0.0010	0.0000	0.0020	0.0010
	Sidak	0.0025	0.0015	0.0020	0.0035	0.0020	0.0010	0.0010	0.0000	0.0020	0.0010
$m = 5$ $\pi_0 = 0.8$	MBonf	0.0125	0.0145	0.0155	0.0160	0.0115	0.0135	0.0110	0.0060	0.0105	0.0045
	Tarone	0.0090	0.0105	0.0065	0.0095	0.0080	0.0085	0.0070	0.0045	0.0065	0.0030
	Bonf	0.0035	0.0005	0.0020	0.0015	0.0025	0.0025	0.0025	0.0010	0.0025	0.0000
	Sidak	0.0035	0.0005	0.0020	0.0015	0.0025	0.0025	0.0025	0.0010	0.0025	0.0000
$m = 10$ $\pi_0 = 0.4$	MBonf	0.0035	0.0030	0.0025	0.0035	0.0030	0.0025	0.0025	0.0030	0.0025	0.0035
	Tarone	0.0005	0.0010	0.0010	0.0020	0.0010	0.0010	0.0010	0.0005	0.0010	0.0010
	Bonf	0.0005	0.0005	0.0010	0.0005	0.0010	0.0010	0.0000	0.0005	0.0005	0.0005
	Sidak	0.0005	0.0005	0.0010	0.0005	0.0010	0.0010	0.0000	0.0005	0.0005	0.0005
$m = 10$ $\pi_0 = 0.6$	MBonf	0.0080	0.0065	0.0095	0.0095	0.0065	0.0035	0.0025	0.0055	0.0070	0.0030
	Tarone	0.0020	0.0030	0.0060	0.0050	0.0020	0.0020	0.0010	0.0030	0.0030	0.0025
	Bonf	0.0000	0.0015	0.0020	0.0015	0.0000	0.0010	0.0005	0.0010	0.0010	0.0005
	Sidak	0.0000	0.0015	0.0020	0.0015	0.0000	0.0010	0.0005	0.0010	0.0010	0.0005
$m = 10$ $\pi_0 = 0.8$	MBonf	0.0185	0.0105	0.0115	0.0135	0.0135	0.0150	0.0100	0.0090	0.0050	0.0085
	Tarone	0.0120	0.0080	0.0075	0.0095	0.0075	0.0100	0.0065	0.0030	0.0040	0.0045
	Bonf	0.0005	0.0005	0.0005	0.0010	0.0010	0.0000	0.0020	0.0005	0.0010	0.0010
	Sidak	0.0005	0.0005	0.0005	0.0010	0.0010	0.0000	0.0020	0.0005	0.0010	0.0010

Table A.10 Simulated Minimal Power Comparisons for Single-step Procedures with Dependent p-values Generated from Binomial Exact Test Statistics

	ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$m = 5$	MBonf	0.9095	0.8880	0.8580	0.8455	0.8135	0.7745	0.7545	0.7295	0.6840	0.6305
$\pi_0 = 0.4$	Tarone	0.8500	0.8350	0.8000	0.7945	0.7585	0.7070	0.6815	0.6630	0.6130	0.5605
	Bonf	0.7530	0.7465	0.7010	0.6920	0.6750	0.6205	0.5930	0.5675	0.5260	0.4685
	Sidak	0.7530	0.7465	0.7010	0.6920	0.6750	0.6205	0.5930	0.5675	0.5260	0.4685
$m = 5$	MBonf	0.8150	0.8020	0.7755	0.7740	0.7655	0.7195	0.7255	0.6790	0.6750	0.6260
$\pi_0 = 0.6$	Tarone	0.7635	0.7435	0.7210	0.7075	0.7185	0.6775	0.6815	0.6425	0.6255	0.5865
	Bonf	0.6135	0.5985	0.5740	0.5615	0.5725	0.5410	0.5270	0.5070	0.4715	0.4365
	Sidak	0.6135	0.5985	0.5740	0.5615	0.5725	0.5410	0.5270	0.5070	0.4715	0.4365
$m = 5$	MBonf	0.5965	0.6055	0.5955	0.5925	0.6075	0.6095	0.5960	0.5960	0.6075	0.6120
$\pi_0 = 0.8$	Tarone	0.5635	0.5730	0.5675	0.5600	0.5755	0.5880	0.5730	0.5730	0.5935	0.5985
	Bonf	0.3825	0.3845	0.3805	0.3760	0.3875	0.4000	0.3690	0.3735	0.3820	0.3810
	Sidak	0.3825	0.3845	0.3805	0.3760	0.3875	0.4000	0.3690	0.3735	0.3820	0.3810
$m = 10$	MBonf	0.9760	0.9460	0.9175	0.8925	0.8570	0.8250	0.7885	0.7270	0.6895	0.6090
$\pi_0 = 0.4$	Tarone	0.9470	0.8940	0.8535	0.8295	0.7875	0.7585	0.7120	0.6525	0.6040	0.5260
	Bonf	0.8805	0.8260	0.7625	0.7500	0.6845	0.6695	0.6075	0.5550	0.5045	0.4410
	Sidak	0.8805	0.8260	0.7625	0.7500	0.6845	0.6695	0.6075	0.5550	0.5045	0.4410
$m = 10$	MBonf	0.9250	0.9125	0.8845	0.8425	0.8300	0.7920	0.7470	0.7160	0.6590	0.6180
$\pi_0 = 0.6$	Tarone	0.8820	0.8630	0.8370	0.7705	0.7680	0.7285	0.6925	0.6645	0.6090	0.5745
	Bonf	0.7420	0.7260	0.7030	0.6220	0.6285	0.5710	0.5425	0.4995	0.4440	0.4155
	Sidak	0.7420	0.7260	0.7030	0.6220	0.6285	0.5710	0.5425	0.4995	0.4440	0.4155
$m = 10$	MBonf	0.7675	0.7595	0.7390	0.7330	0.7320	0.6975	0.6865	0.6665	0.6340	0.6160
$\pi_0 = 0.8$	Tarone	0.7145	0.7055	0.6910	0.6860	0.6885	0.6540	0.6445	0.6310	0.5925	0.5875
	Bonf	0.4935	0.4880	0.4655	0.4710	0.4630	0.4235	0.4145	0.3975	0.3725	0.3495
	Sidak	0.4935	0.4880	0.4655	0.4710	0.4630	0.4235	0.4145	0.3975	0.3725	0.3495

Table A.11 Simulated FWER Comparisons for Step-down Procedures with Dependent p-values Generated from Binomial Exact Test Statistics

	ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	MHolm	0.0085	0.0090	0.0070	0.0090	0.0105	0.0120	0.0095	0.0150	0.0115	0.0055
$m = 5$	TH	0.0055	0.0060	0.0040	0.0050	0.0065	0.0100	0.0040	0.0130	0.0095	0.0045
$\pi_0 = 0.4$	Holm	0.0025	0.0040	0.0035	0.0035	0.0035	0.0055	0.0015	0.0075	0.0040	0.0035
	MHolm	0.0170	0.0095	0.0125	0.0140	0.0125	0.0095	0.0160	0.0150	0.0090	0.0065
$m = 5$	TH	0.0110	0.0065	0.0095	0.0095	0.0090	0.0070	0.0115	0.0125	0.0075	0.0055
$\pi_0 = 0.6$	Holm	0.0030	0.0020	0.0035	0.0055	0.0040	0.0030	0.0045	0.0025	0.0040	0.0030
	MHolm	0.0160	0.0205	0.0200	0.0190	0.0170	0.0175	0.0135	0.0075	0.0120	0.0070
$m = 5$	TH	0.0110	0.0145	0.0120	0.0150	0.0120	0.0120	0.0115	0.0065	0.0100	0.0060
$\pi_0 = 0.8$	Holm	0.0035	0.0015	0.0020	0.0015	0.0025	0.0025	0.0025	0.0010	0.0025	0.0000
	MHolm	0.0130	0.0150	0.0115	0.0095	0.0090	0.0100	0.0115	0.0130	0.0095	0.0090
$m = 10$	TH	0.0060	0.0030	0.0075	0.0040	0.0040	0.0075	0.0080	0.0105	0.0070	0.0075
$\pi_0 = 0.4$	Holm	0.0005	0.0005	0.0015	0.0005	0.0015	0.0015	0.0010	0.0030	0.0010	0.0010
	MHolm	0.0160	0.0150	0.0185	0.0165	0.0175	0.0105	0.0125	0.0130	0.0140	0.0055
$m = 10$	TH	0.0055	0.0085	0.0115	0.0100	0.0125	0.0075	0.0065	0.0100	0.0125	0.0045
$\pi_0 = 0.6$	Holm	0.0000	0.0020	0.0020	0.0025	0.0000	0.0015	0.0005	0.0015	0.0025	0.0005
	MHolm	0.0230	0.0160	0.0195	0.0195	0.0200	0.0215	0.0145	0.0120	0.0090	0.0130
$m = 10$	TH	0.0160	0.0130	0.0145	0.0130	0.0145	0.0165	0.0130	0.0100	0.0075	0.0120
$\pi_0 = 0.8$	Holm	0.0005	0.0005	0.0005	0.0010	0.0010	0.0000	0.0020	0.0005	0.0010	0.0010

Table A.12 Simulated Minimal Power Comparisons for Step-down Procedures with Dependent p-values Generated from Binomial Exact Test Statistics

	ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	MHolm	0.9095	0.8880	0.8580	0.8455	0.8135	0.7745	0.7545	0.7300	0.6840	0.6305
$m = 5$	TH	0.8500	0.8350	0.8000	0.7945	0.7585	0.7070	0.6815	0.6630	0.6130	0.5605
$\pi_0 = 0.4$	Holm	0.7535	0.7465	0.7010	0.6920	0.6750	0.6205	0.5930	0.5675	0.5260	0.4690
	MHolm	0.8155	0.8020	0.7755	0.7740	0.7655	0.7195	0.7255	0.6795	0.6750	0.6260
$m = 5$	TH	0.7640	0.7440	0.7210	0.7080	0.7185	0.6775	0.6815	0.6425	0.6255	0.5865
$\pi_0 = 0.6$	Holm	0.6135	0.5985	0.5750	0.5615	0.5725	0.5415	0.5270	0.5070	0.4715	0.4365
	MHolm	0.5980	0.6060	0.5975	0.5930	0.6080	0.6095	0.5965	0.5960	0.6080	0.6125
$m = 5$	TH	0.5640	0.5740	0.5685	0.5600	0.5755	0.5880	0.5735	0.5730	0.5940	0.5990
$\pi_0 = 0.8$	Holm	0.3830	0.3845	0.3805	0.3760	0.3875	0.4000	0.3690	0.3735	0.3820	0.3810
	MHolm	0.9760	0.9460	0.9175	0.8925	0.8570	0.8250	0.7885	0.7270	0.6895	0.6090
$m = 10$	TH	0.9470	0.8940	0.8535	0.8295	0.7875	0.7585	0.7120	0.6525	0.6040	0.5260
$\pi_0 = 0.4$	Holm	0.8805	0.8260	0.7625	0.7500	0.6845	0.6695	0.6075	0.5550	0.5045	0.4410
	MHolm	0.9265	0.9125	0.8845	0.8425	0.8300	0.7920	0.7470	0.7160	0.6590	0.6180
$m = 10$	TH	0.8820	0.8630	0.8380	0.7705	0.7680	0.7285	0.6925	0.6645	0.6090	0.5745
$\pi_0 = 0.6$	Holm	0.7420	0.7260	0.7030	0.6220	0.6285	0.5710	0.5425	0.4995	0.4440	0.4155
	MHolm	0.7680	0.7600	0.7390	0.7330	0.7325	0.6975	0.6870	0.6665	0.6340	0.6160
$m = 10$	TH	0.7165	0.7060	0.6925	0.6865	0.6895	0.6540	0.6450	0.6310	0.5925	0.5875
$\pi_0 = 0.8$	Holm	0.4935	0.4880	0.4655	0.4710	0.4630	0.4235	0.4145	0.3975	0.3725	0.3495

Table A.13 Simulated FWER Comparisons for Step-up Procedures with Dependent p-values Generated from Binomial Exact Test Statistics

	ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	MHoch	0.0085	0.0090	0.0070	0.0090	0.0105	0.0120	0.0110	0.0175	0.0130	0.0080
$m = 5$	Roth	0.0060	0.0070	0.0050	0.0065	0.0075	0.0095	0.0060	0.0140	0.0105	0.0070
$\pi_0 = 0.4$	Hoch	0.0025	0.0040	0.0035	0.0035	0.0035	0.0060	0.0020	0.0090	0.0055	0.0060
	MHoch	0.0170	0.0100	0.0125	0.0160	0.0130	0.0105	0.0165	0.0170	0.0120	0.0120
$m = 5$	Roth	0.0120	0.0075	0.0095	0.0110	0.0085	0.0070	0.0125	0.0140	0.0095	0.0105
$\pi_0 = 0.6$	Hoch	0.0030	0.0020	0.0035	0.0055	0.0040	0.0030	0.0050	0.0030	0.0050	0.0075
	MHoch	0.0165	0.0210	0.0200	0.0200	0.0170	0.0180	0.0135	0.0095	0.0140	0.0115
$m = 5$	Roth	0.0110	0.0145	0.0130	0.0150	0.0115	0.0130	0.0100	0.0075	0.0120	0.0085
$\pi_0 = 0.8$	Hoch	0.0035	0.0015	0.0020	0.0015	0.0025	0.0025	0.0030	0.0010	0.0030	0.0035
	MHoch	0.0145	0.0165	0.0115	0.0100	0.0105	0.0105	0.0135	0.0155	0.0110	0.0125
$m = 10$	Roth	0.0075	0.0045	0.0080	0.0035	0.0045	0.0075	0.0095	0.0120	0.0080	0.0115
$\pi_0 = 0.4$	Hoch	0.0005	0.0005	0.0015	0.0010	0.0015	0.0015	0.0015	0.0035	0.0025	0.0060
	MHoch	0.0165	0.0150	0.0185	0.0165	0.0185	0.0115	0.0150	0.0155	0.0190	0.0085
$m = 10$	Roth	0.0060	0.0070	0.0110	0.0100	0.0110	0.0070	0.0070	0.0105	0.0135	0.0070
$\pi_0 = 0.6$	Hoch	0.0000	0.0020	0.0020	0.0025	0.0005	0.0015	0.0010	0.0020	0.0045	0.0015
	MHoch	0.0235	0.0170	0.0205	0.0200	0.0210	0.0225	0.0165	0.0140	0.0105	0.0170
$m = 10$	Roth	0.0145	0.0120	0.0125	0.0130	0.0155	0.0155	0.0130	0.0075	0.0080	0.0110
$\pi_0 = 0.8$	Hoch	0.0005	0.0005	0.0005	0.0010	0.0010	0.0000	0.0020	0.0005	0.0010	0.0035

Table A.14 Simulated Minimal Power Comparisons for Step-up Procedures with Dependent p-values Generated from Binomial Exact Test Statistics

	ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	MHoch	0.9110	0.8905	0.8595	0.8490	0.8155	0.7780	0.7595	0.7370	0.6985	0.6525
$m = 5$	Roth	0.8565	0.8430	0.8075	0.8020	0.7640	0.7200	0.6945	0.6755	0.6300	0.5955
$\pi_0 = 0.4$	Hoch	0.7640	0.7555	0.7105	0.7025	0.6835	0.6310	0.6050	0.5860	0.5435	0.4945
	MHoch	0.8165	0.8065	0.7785	0.7775	0.7690	0.7255	0.7290	0.6880	0.6835	0.6425
$m = 5$	Roth	0.7670	0.7515	0.7245	0.7115	0.7250	0.6845	0.6840	0.6505	0.6325	0.5990
$\pi_0 = 0.6$	Hoch	0.6210	0.6045	0.5780	0.5665	0.5800	0.5460	0.5345	0.5140	0.4845	0.4530
	MHoch	0.5985	0.6065	0.5975	0.5940	0.6080	0.6095	0.5965	0.5965	0.6085	0.6130
$m = 5$	Roth	0.5490	0.5590	0.5520	0.5490	0.5610	0.5755	0.5560	0.5570	0.5775	0.5800
$\pi_0 = 0.8$	Hoch	0.3830	0.3845	0.3805	0.3760	0.3875	0.4000	0.3690	0.3735	0.3820	0.3815
	MHoch	0.9770	0.9465	0.9205	0.8945	0.8630	0.8300	0.7925	0.7380	0.7055	0.6305
$m = 10$	Roth	0.9495	0.8980	0.8595	0.8325	0.7910	0.7610	0.7190	0.6605	0.6190	0.5625
$\pi_0 = 0.4$	Hoch	0.8815	0.8260	0.7630	0.7520	0.6850	0.6705	0.6095	0.5565	0.5075	0.4485
	MHoch	0.9285	0.9155	0.8855	0.8475	0.8340	0.7985	0.7540	0.7235	0.6735	0.6405
$m = 10$	Roth	0.8870	0.8650	0.8425	0.7760	0.7735	0.7335	0.7000	0.6700	0.6145	0.5875
$\pi_0 = 0.6$	Hoch	0.7425	0.7260	0.7030	0.6225	0.6290	0.5710	0.5430	0.5000	0.4465	0.4205
	MHoch	0.7725	0.7625	0.7400	0.7380	0.7350	0.6995	0.6890	0.6710	0.6400	0.6285
$m = 10$	Roth	0.7200	0.7055	0.6945	0.6895	0.6905	0.6515	0.6460	0.6365	0.5970	0.5965
$\pi_0 = 0.8$	Hoch	0.4935	0.4880	0.4655	0.4710	0.4630	0.4235	0.4145	0.3975	0.3725	0.3495

APPENDIX B

PROOFS IN CHAPTER 3

This appendix contains the proofs of the theorems and lemmas stated but not proved in Chapter 3.

B.1 Proof of Theorem 3.1

Proof. Let I_0 denote the index of true null hypothesis, $n_0 = |I_0|$ denote the number of true nulls, R denote the the number of rejected hypothesis. Denote $I = 1$ if the first true null hypothesis H_1 is rejected.

Without loss of generality, replace first true null p-value by 0 and order the p-values such that $p_1 = 0 \leq p_2 \leq \cdots \leq p_n$. Let $J = \max\{k : p_k \leq \frac{k}{n}\alpha\}$. Since the number of rejections R is a non-increasing function of each p-values, $R = R(p'_1, p'_2, \ldots, p'_n) \leq R(p_1, p_2, \ldots, p_n) \leq J(0, p_2, \ldots, p_n) = J$.

So $I = 1$ implies $p_1 \leq p'_1 \leq \frac{R}{n}\alpha \leq \frac{J}{n}\alpha$. For $1 < r \leq J$, there are three cases as the inflation factor b_r and b_1 vary as follows.

- **Case 1:** $b_r = 1$**.**

  Since $p_1 \leq \frac{J}{n}\alpha$, $p_r \leq \frac{J}{n}\alpha$ for $r \leq J$. Hence $p'_r = p_r \leq \frac{J}{n}\alpha$.

- **Case 2:** $b_1 = 1, b_r = \dfrac{e^{-t/2}}{\prod_{l=1(\neq r)}^{n} p_l} < 1$**.**

  Since $b_1 = 1$ implies $\dfrac{e^{-t/2}}{\prod_{l=2}^{n} p_l} \geq 1$, $p'_r = p_r/b_r = \dfrac{\prod_{l=1}^{n} p_l}{e^{-t/2}} = p_1 \dfrac{\prod_{l=2}^{n} p_l}{e^{-t/2}} \leq p_1 = p'_1 \leq \dfrac{J}{n}\alpha$.

- **Case 3:** $b_1 = \dfrac{e^{-t/2}}{\prod_{l=2}^{n} p_l} < 1, b_r = \dfrac{e^{-t/2}}{\prod_{l=1(\neq r)}^{n} p_l} < 1$**,**

  then $p'_r = p'_1 = \dfrac{\prod_{l=1}^{n} p_l}{e^{-t/2}}$. Since $p'_1 \leq \frac{J}{n}\alpha$, $p'_r \leq \frac{J}{n}\alpha$.

Based on the above analysis, we can conclude $I = 1$ implies $p'_r \leq \frac{J}{n}\alpha$ for $r \leq J$, that is, at least J conditional p-values are no more than $\frac{J}{n}\alpha$. Since R is the maximal number r satisfying $p'_r \leq \frac{r}{n}\alpha$, so $J \leq R$. Therefore, we can conclude $J = R$ since $J \geq R$.

Therefore,

$$\begin{aligned}
cFDR &= E\left(\frac{V}{R \vee 1}\Big| f(\mathbf{P}) > t\right) \\
&= n_0 E\left\{\sum_{r=1}^{n} \frac{I(H_1 \text{is rejected}, R = r | f(\mathbf{P}) > t)}{r}\right\} \\
&= n_0 E_{P_2,\ldots,P_n}\left\{E_{P_1|P_2,\ldots,P_n}\left[\sum_{r=1}^{n} \frac{I(H_1 \text{is rejected}, R = r | f(P_1, \mathbf{P}^{(-\mathbf{1})} = \mathbf{p}^{(-\mathbf{1})}) > t)}{r}\right]\right\} \\
&= n_0 E_{P_2,\ldots,P_n}\left\{\sum_{r=1}^{n} \frac{1}{r}\Pr(I = 1, R = r | f(P_1, \mathbf{P}^{(-\mathbf{1})} = \mathbf{p}^{(-\mathbf{1})}) > t)\right\} \\
&= n_0 E_{P_2,\ldots,P_n}\left\{\frac{1}{J}\Pr(P'_1 \leq \frac{J}{n}\alpha | f(P_1, \mathbf{P}^{(-\mathbf{1})} = \mathbf{p}^{(-\mathbf{1})}) > t)\right\} \\
&= \frac{n_0}{n} E(\alpha) \leq \alpha.
\end{aligned} \tag{B.1.1}$$

Then the proof is complete. □

B.2 Proof of Lemma 3.1

Proof. For any $1 \leq j \leq n-1$, we have $p_j \leq p_{j+1}$, and

$$p'_j = \begin{cases} \dfrac{\prod_{l=1}^{n} p_l}{e^{-t/2}} & \text{if } \prod_{\substack{l=1 \\ (\neq j)}}^{n} p_l > e^{-t/2} \\ p_j & \text{otherwise.} \end{cases}$$

p'_{j+1} can only take the value of $\dfrac{\prod_{l=1}^{n} p_l}{e^{-t/2}}$ or p_{j+1}.

- **Case 1** If $p'_{j+1} = \dfrac{\prod_{l=1}^{n} p_l}{e^{-t/2}}$, which implies $\prod_{\substack{l=1 \\ (\neq j+1)}}^{n} p_l > e^{-t/2}$.

So $\prod_{\substack{l=1\\(\neq j)}}^{n} p_l = p_{j+1} \prod_{\substack{l=1\\(\neq j,j+1)}}^{n} p_l \geq p_j \prod_{\substack{l=1\\(\neq j,j+1)}}^{n} p_l = \prod_{\substack{l=1\\(\neq j+1)}}^{n} p_l > e^{-t/2}$, where the first inequality follows from $p_{j+1} \geq p_j$. Thus $p'_j = \dfrac{\prod_{l=1}^{n} p_l}{e^{-t/2}} = p'_{j+1}$.

- **Case 2** If $p'_{j+1} = p_{j+1}$, which implies $\prod_{\substack{l=1\\(\neq j+1)}}^{n} p_l \leq e^{-t/2}$, p'_j can be p_j or $\dfrac{\prod_{l=1}^{n} p_l}{e^{-t/2}}$.

  If $p'_j = p_j$, then it is trivial that $p'_j \leq p'_{j+1}$ since $p_j \leq p_{j+1}$.

  Otherwise, $p'_j = \dfrac{\prod_{l=1}^{n} p_l}{e^{-t/2}} = p_{j+1} \dfrac{\prod_{\substack{l=1\\(\neq j+1)}}^{n} p_l}{e^{-t/2}} \leq p_{j+1} = p'_{j+1}$, since $\prod_{\substack{l=1\\(\neq j+1)}}^{n} p_l \leq e^{-t/2}$.

  Therefore, for any $1 \leq j \leq n-1$, if $p_j \leq p_{j+1}$, then $p'_j \leq p'_{j+1}$, the proof is complete. □

B.3 Proof of Theorem 3.2

Proof. By using similar arguments as the proof of Theorem 1, we can conclude reject H_1, i.e. $I = 1$, implies $p_1 \leq p'_1 \leq \dfrac{R}{n}\alpha \leq \dfrac{J}{n}\alpha$. For $1 < r \leq J$, we also consider the following three cases.

- **Case 1:** $b_r = 1$**.**

  Since $p_1 \leq p'_1 \leq \dfrac{J}{n}\alpha$, $p_r \leq \dfrac{J}{n}\alpha$ for $r \leq J$. Hence $p'_r = p_r \leq \dfrac{J}{n}\alpha$.

- **Case 2:** $b_r = t < 1$**.**

  $b_r = t$ implies $\min\{p_1, p_2, \ldots, p_{r-1}, p_{r+1}, \ldots, p_n\} = \min\{p_1, p_2\} > t$, which means $p_1 > t$. Since $p_1 \leq \dfrac{J}{n}\alpha \leq \alpha$ and $\alpha \leq t$, $p_1 \leq t$, which leads to a contradiction.

  Based on the above analysis, we can conclude $I = 1$ implies $p'_r \leq \dfrac{J}{n}\alpha$ for some $r \leq J$, that is, at least J conditional p-values less than or equal to $\dfrac{J}{n}\alpha$. Since R is the maximal number r satisfying $p'_r \leq \dfrac{r}{n}\alpha$, then $J \leq R$. So $J = R$ since $J \geq R$.

  Therefore,

$$
\begin{aligned}
cFDR &= E\left(\frac{V}{R\vee 1}\Big|\min\{\mathbf{P}\}\le t\right)\\
&= n_0 E\left\{\sum_{r=1}^{n}\frac{I(H_1\text{is rejected }, R=r|\min\{\mathbf{P}\}\le t)}{r}\right\}\\
&= E_{P_2,\ldots,P_n}\left\{E_{P_1|P_2,\ldots,P_n}\left[\sum_{r=1}^{n}\frac{I(H_1\text{is rejected }, R=r|\min\{P_1,\mathbf{P}^{(-\mathbf{1})}=\mathbf{p}^{(-\mathbf{1})}\}\le t)}{r}\right]\right\}\\
&= n_0 E_{P_2,\ldots,P_n}\left\{\sum_{r=1}^{n}\frac{1}{r}\Pr(I=1,R=r|\min\{P_1,\mathbf{P}^{(-\mathbf{1})}=\mathbf{p}^{(-\mathbf{1})}\}\le t)\right\}\\
&= n_0 E_{P_2,\ldots,P_n}\left\{\frac{1}{J}\Pr(P_1'\le\frac{J}{n}\alpha|\min\{P_1,\mathbf{P}^{(-\mathbf{1})}=\mathbf{p}^{(-\mathbf{1})}\}\le t)\right\}\\
&= n_0 E\left(\frac{\alpha}{n}\right)\le\alpha
\end{aligned}
\tag{B.3.1}
$$

□

BIBLIOGRAPHY

[1] A Agresti and M Kateri. *Categorical Data Analysis*. New York, NY: Springer, Berlin, Heidelberg, 2011.

[2] A Bate, M Lindquist, I R Edwards, S Olsson, R Orre, A Lansner, and R M De F. A bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology*, 54(4):315–321, 1998.

[3] Y Benjamini and M Bogomolov. Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):297–318, 2014.

[4] Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[5] Y Benjamini and Y Hochberg. Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24(3):407–418, 1997.

[6] Y Benjamini, A M Krieger, and D Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.

[7] Y Benjamini and W Liu. A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference*, 82(1):163–170, 1999.

[8] Y Benjamini and D Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, pages 1165–1188, 2001.

[9] S M Berry and D A Berry. Accounting for multiplicities in assessing drug safety: A three-level hierarchical mixture model. *Biometrics*, 60(2):418–426, 2004.

[10] D Böhning. Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics*, 44(1):197–200, 1992.

[11] D G Bonett. A linear negative multinomial model. *Statistics and Probability Letters*, 3(3):127–129, 1985a.

[12] D G Bonett. The negative multinomial logit model. *Communications in Statistics - Theory and Methods*, 14(7):1713–1717, 1985b.

[13] M Chen, L Zhu, P Chiruvolu, and Q Jiang. Evaluation of statistical methods for safety signal detection: a simulation study. *Pharmaceutical Statistics*, 14(1):11–19, 2015.

[14] S K Dhar. Extension of a negative multinomial model. *Communications in Statistics - Theory and Methods*, 24(1):39–57, 1995.

[15] S K Dhar and S Lahiri. Generalized linear model under the extended negative multinomial model and cancer incidence. *Journal of the Indian Statistical Association*, 52(1):125–140, 2014.

[16] I Dialsingh, S R Austin, and N S Altman. Estimating the proportion of true null hypotheses when the statistics are discrete. *Bioinformatics*, page btv104, 2015.

[17] A Dmitrienko, A C Tamhane, and F Bretz. *Multiple Testing Problems in Pharmaceutical Statistics*. New York, NY: CRC Press, 2009.

[18] M A Evans and D G Bonett. Maximum likelihood estimation for the negative multinomial log-linear model. *Communications in Statistics - Theory and Methods*, 18(11):4059–4065, 1989.

[19] S Evans, P C Waller, and S Davis. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and Drug Safety*, 10(6):483–486, 2001.

[20] L Fahrmeir and G Tutz. *Multivariate statistical modelling based on generalized linear models*. New York, NY: Springer Science and Business Media, 2013.

[21] A Farcomeni. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, 2007.

[22] L Finos and L Salmaso. FDR-and FWE-controlling methods using data-driven weights. *Journal of Statistical Planning and Inference*, 137(12):3859–3870, 2007.

[23] Y Gavrilov, Y Benjamini, and S K Sarkar. An adaptive step-down procedure with proven fdr control under independence. *The Annals of Statistics*, pages 619–629, 2009.

[24] Y Ge, S Dudoit, and T P Speed. Resampling-based multiple testing for microarray data analysis. *Test*, 12(1):1–77, 2003.

[25] P B Gilbert. A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):143–158, 2005.

[26] G F V Glonek and P McCullagh. Multivariate logistic models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 533–546, 1995.

[27] R Gueorguieva. A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Statistical Modelling*, 1(3):177–193, 2001.

[28] R Gutman and Y Hochberg. Improved multiple test procedures for discrete distributions: New ideas and analytical review. *Journal of Statistical Planning and Inference*, 137(7):2380–2393, 2007.

[29] J D Habiger. Multiple test functions and adjusted p-values for test statistics with discrete distributions. *Journal of Statistical Planning and Inference*, 167:1–13, 2015.

[30] T Hastie, R Tibshirani, and J Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer, 2001.

[31] G Heimann, M Von Tress, and M Gasparini. Exact and asymptotic inference in clinical trials with small event rates under inverse sampling. *Statistics in Medicine*, 34(19):2708–2724, 2015.

[32] R Heller, N Chatterjee, A Krieger, and J Shi. Post-selection inference following aggregate level hypothesis testing in large scale genomic data. *bioRxiv*, 2017.

[33] R Heller and H Gur. False discovery rate controlling procedures for discrete tests. *arXiv preprint arXiv:1112.4627*, 2011.

[34] J M Henshall and M E Goddard. Multiple-trait mapping of quantitative trait loci after selective genotyping using logistic regression. *Genetics*, 151(2):885–894, 1999.

[35] J F Heyse. A false discovery rate procedure for categorical data. *Resent Advances in Biostatistics: False Discovery Rates, Survival Analysis, and Related Topics*, pages 43–58, 2011.

[36] J F Heyse and D Rom. Adjusting for multiplicity of statistical tests in the analysis of carcinogenicity studies. *Biometrical Journal*, 30(8):883–896, 1988.

[37] Y Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[38] S Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70, 1979.

[39] G Hommel. A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, 75(2):383–386, 1988.

[40] G Hommel and F Krummenauer. Improvements and modifications of tarone's multiple test procedure for discrete data. *Biometrics*, pages 673–681, 1998.

[41] L Huang, J Zalkikar, and R C Tiwari. A likelihood ratio test based method for signal detection with application to FDA's drug safety data. *Journal of the American Statistical Association*, 106(496):1230–1241, 2011.

[42] N Ignatiadis, B Klaus, J Zaugg, and W Huber. Data-driven hypothesis weighting increases detection power in big data analytics. *bioRxiv*, page 034330, 2015.

[43] G James, D Witten, T Hastie, and R Tibshirani. *An Introduction to Statistical Learning*, volume 112. New York, NY: Springer, 2013.

[44] B Jørgensen. Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika*, 70(1):19–28, 1983.

[45] E Kulinskaya and A Lewin. On fuzzy familywise error rate and false discovery rate procedures for discrete distributions. *Biometrika*, 96(1):201–211, 2009.

[46] S Lahiri and S K Dhar. Log-linear modeling under generalized inverse sampling scheme. *Communications in Statistics - Theory and Methods*, 37(8):1237–1244, 2008.

[47] E L Lehmann and J P Romano. *Testing Statistical Hypotheses (3rd Edition)*. New York, NY: Springer, 2005.

[48] E Lesaffre and A Albert. Multiple-group logistic regression diagnostics. *Applied Statistics*, pages 425–440, 1989.

[49] P McCullagh and J A Nelder. *Generalized Linear Models*, volume 37. New York, NY: CRC press, 1989.

[50] C E McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, 92(437):162–170, 1997.

[51] D V Mehrotra and A J Adewale. Flagging clinical adverse experiences: reducing false discoveries without materially compromising power for detecting true signals. *Statistics in Medicine*, 31(18):1918–1930, 2012.

[52] D V Mehrotra and J F Heyse. Use of the false discovery rate for evaluating clinical safety data. *Statistical Methods in Medical Research*, 13(3):227–238, 2004.

[53] J E Mosimann. On the compound negative multinomial distribution and correlations among inversely sampled pollen counts. *Biometrika*, 50(1/2):47–54, 1963.

[54] R H Myers, D C Montgomery, G G Vining, and T J Robinson. *Generalized Linear Models: with Applications in Engineering and the Sciences*, volume 791. Hoboken, NJ: John Wiley and Sons, 2012.

[55] J Panaretos. A characterization of the negative multinomial distribution. In *Statistical Distributions in Scientific Work*, pages 331–339. Springer, New York, NY, USA, 1981.

[56] G Plomteux. Multivariate analysis of an enzymic profile for the differential diagnosis of viral hepatitis. *Clinical Chemistry*, 26(13):1897–1899, 1980.

[57] D M Rom. Strengthening some common multiple test procedures for discrete data. *Statistics in Medicine*, 11(4):511–514, 1992.

[58] J P Romano and A M Shaikh. On stepdown control of the false discovery proportion. In *Optimality*, pages 33–50. Institute of Mathematical Statistics, 2006.

[59] A J Roth. Multiple comparison procedures for discrete test statistics. *Journal of Statistical Planning and Inference*, 82(1):101–117, 1999.

[60] K J Rothman, S Lanes, and S T Sacks. The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiology and Drug Safety*, 13(8):519–523, 2004.

[61] S K Sarkar. Some results on false discovery rate in stepwise multiple testing procedures. *Annals of Statistics*, pages 239–257, 2002.

[62] S K Sarkar. Stepup procedures controlling generalized fwer and generalized fdr. *The Annals of Statistics*, pages 2405–2420, 2007.

[63] J P Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46:561, 1995.

[64] Z Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

[65] R J Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

[66] W P Stephenson and M Hauben. Data mining for signals in spontaneous reporting databases: proceed with caution. *Pharmacoepidemiology and Drug Safety*, 16(4):359–365, 2007.

[67] J D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[68] J D Storey, J E Taylor, and D Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004.

[69] A Szarfman, S G Machado, and R T Oneill. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. *Drug Safety*, 25(6):381–392, 2002.

[70] A C Tamhane and L Liu. On weighted hochberg procedures. *Biometrika*, 95(2):279–294, 2008.

[71] R E Tarone. A modified bonferroni method for discrete data. *Biometrics*, pages 515–522, 1990.

[72] L A Waller and D Zelterman. Log-linear modeling with the negative multinomial distribution. *Biometrics*, pages 971–982, 1997.

[73] P H Westfall and R D Wolfinger. Multiple tests with discrete distributions. *The American Statistician*, 51(1):3–8, 1997.

[74] P H Westfall and S S Young. *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*, volume 279. Hoboken, NJ: John Wiley and Sons, 1993.

[75] J Zhu, J C Eickhoff, and P Yan. Generalized linear latent variable models for repeated measures of spatially correlated multivariate data. *Biometrics*, 61(3):674–683, 2005.