Familywise Error Rate Controlling Procedures for Discrete Data

Yalin Zhu
Biostatistics and Research Decision Sciences,
Merck Research Laboratories, Upper Gwynedd, PA, U.S.A.

Wenge Guo
Department of Mathematical Sciences,
New Jersey Institute of Technology, Newark, NJ, U.S.A.

November 26, 2018

Abstract

In applications such as clinical safety analysis, the data of the experiments usually consists of frequency counts. In the analysis of such data, researchers often face the problem of multiple testing based on discrete test statistics, aimed at controlling family-wise error rate (FWER). Most existing FWER controlling procedures are developed for continuous data, which are often conservative when analyzing discrete data. By using minimal attainable p-values, several FWER controlling procedures have been specifically developed for discrete data in the literature. In this paper, by utilizing known marginal distributions of true null p-values, three more powerful stepwise procedures are developed, which are modified versions of the conventional Bonferroni, Holm and Hochberg procedures, respectively. It is shown that the first two procedures strongly control the FWER under arbitrary dependence and are more powerful than the existing Tarone-type procedures, while the last one only ensures control of the FWER in special settings. Through extensive simulation studies, we provide numerical evidence of superior performance of the proposed procedures in terms of the FWER control and minimal power. A real clinical safety data is used to demonstrate applications of our proposed procedures. An R package “MHTdiscrete” and a web application are developed for implementing the proposed procedures.

Keywords: CDF of p-values, clinical safety study, multiple testing, stepwise procedure

1 Introduction

In the applications of clinical trials, multiple hypotheses testing is a very useful statistical tool to analyze efficacy or safety data. Simultaneously testing multiple hypotheses is often required in such applications. For single hypothesis testing, a typical error measure which needs to be controlled is type I error rate, the probability of rejecting the hypothesis while the hypothesis is true. There are several possible measures for overall type I error rate while testing multiple hypotheses. A standard error rate for clinical trials is familywise error rate (FWER), which is the probability of making at least one false rejection.

In the existing literature, most FWER controlling procedures are developed for continuous data and some are widely used in practice such as Bonferroni procedure, Holm procedure [11], Hochberg procedure [10], etc. However, these procedures might be conservative when they are used to analyze discrete data. In the literature, several FWER controlling procedures have been specifically developed for discrete data. Tarone [20] proposed a modified Bonferroni procedure for discrete data, which reduces the number of tested hypotheses by eliminating those hypotheses with relatively large minimal attainable p-values. The Tarone procedure is more powerful than the conventional Bonferroni procedure, but it lacks α-consistency, that is, a hypothesis which is accepted at a given α level may be rejected at a lower α level. To overcome this issue, Hommel and Krummenauer [12] and Roth [17] developed two modified versions of the Tarone procedure, which not only control the FWER, but also satisfy the desired property of α-consistency. By using Tarone's idea, Hommel and Krummenauer [12] also developed a step-down procedure for discrete data, which improves the conventional Holm procedure. By using the similar idea, Roth [17] developed a two-stage step-up procedure for discrete data, which improves the conventional Hochberg procedure by eliminating non-significant tests in the first stage. Westfall and Wolfinger [21] introduced a resampling based approach by simulating the null distribution of minimal p-value, which uses all attainable values for each p-value. Gutman and Hochberg [8] developed new stepwise procedures by using the idea of Tarone and the algorithm of Westfall and Wolfinger, but these procedures are computationally intensive and only ensure asymptotic control of the FWER. For references of recent developments in multiple testing for discrete test statistics, see Heyse [9], Chen et al. [3] and Döhler [5].

For applications of multiple testing procedures in clinical safety studies, see Mehrotra and Heyse [14], Gould [7], Jiang and Xia [13], Dimitrienko et al. [4], and Goeman and Solari [6].

It is noted that these existing procedures for discrete data are mainly developed based on minimal attainable p-values. In practice, if the minimal attainable p-values are known, the corresponding true null distributions of the p-values are often also known. By fully utilizing the true null distributions rather than the minimal attainable p-values, we develop three simple and powerful stepwise procedures for discrete data. Specifically, we develop new single-step, step-down, and step-up procedures for discrete data, which are modified versions of the conventional Bonferroni, Holm, and Hochberg procedures, respectively. Theoretically, we show that the first two procedures strongly control the FWER under arbitrary dependence, whereas the last one only ensures control of the FWER in special settings. We also show that the proposed procedures have several desired properties: (i) the proposed single-step procedure is more powerful than the existing Tarone and modified Tarone procedures, whereas the proposed step-down procedure is more powerful than the existing Tarone-Holm procedure; (ii) the proposed procedures satisfy the properties of α-consistency and p-value monotonicity, which are desired for a multiple testing procedure; (iii) simple formulas for adjusted p-values are given for these proposed procedures. Through extensive simulation studies, we provide numerical evidence of superior performance of the proposed procedures in terms of the FWER control and minimal power. Even for the proposed step-up procedure, although we cannot provide theoretical guarantee of its FWER control for general cases, we find out numerical validation of its FWER control under various simulation settings. A real data set of clinical safety study is also used to demonstrate applications of our proposed procedures.

The rest of the paper is organized as follows. With notations, assumptions and several existing procedures for discrete data given in Section 2, we present our proposed stepwise procedures and discuss their statistical properties in Section 3. The numerical findings from simulation studies are given in Section 4 and a real application of clinical safety study is presented in Section 5. Some concluding remarks are made in Section 6 and all proofs are deferred to the appendix section.

2 Preliminary

Consider the problem of simultaneously testing m hypotheses $H_1, \dots, H_m$, among which there are m_0 true and m_1 false null hypotheses. Suppose 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 values for P_i such that $P_i \in \mathbb{P}_i$. Let F_i denote the cumulative distribution function (CDF) of P_i when H_i is true, that is $F_i(u) = Pr(P_i \le u | H_i \text{ is true})$ for $u \in [0, 1]$. Let $P_{(1)} \le \cdots \le P_{(m)}$ denote the ordered p-values and $H_{(1)}, \dots, H_{(m)}$ denote the corresponding hypotheses, with $F_{(i)}$ denoting the corresponding CDF of $P_{(i)}$ when $H_{(i)}$ is true and $\mathbb{P}_{(i)}$ the corresponding set of all attainable values of $P_{(i)}$. We make the following assumption regarding F_i:

Assumption 2.1. *The marginal distribution functions F_i of all true null p-values P_i are known and satisfy that for any $u \in [0, 1]$, $F_i(u) = u$, if $u \in \mathbb{P}_i$; otherwise, $Pr(P_i \le u) < u$.*

The assumption implies that the marginal distributions of discrete true null p-values are exactly $U(0, 1)$ distributed when u takes attainable p-values of P_i, and stochastically larger than $U(0, 1)$ when u takes unattainable values. For the joint distributions of the p-values, throughout the paper we only consider two types of dependence structure, arbitrary dependence, which allows any joint distribution of the p-values, and positive regression dependence on subset (PRDS) (Benjamini and Yekutieli [1]; Sarkar [18]), which is often satisfied in many multiple testing situations. The PRDS assumption is defined as follow.

Assumption 2.2. *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, $E\{\phi(P_1, \dots, P_m) | P_i \le p\}$ is non-decreasing in p for each true null hypothesis H_i.*

For any multiple testing procedure (MTP), let V denote the number of falsely rejected hypotheses. Then, the FWER of this procedure, defined by $FWER = \Pr(V \ge 1)$ is said to be controlled at level α, strongly unless stated otherwise, if it is bounded above by α for any configuration of hypotheses. That is, for any combination of true and false null hypotheses, the FWER of this procedure is less than or equal to α. In the literature, there are several popular FWER controlling procedures available for any test statistics, such as Bonferroni, Sidak, Holm, Hochberg, and Hommel procedures (Dimitrienko et al. [4]). Specifically, for discrete test statistics, Tarone (1990) introduced a modified Bonferroni procedure below by using the smallest attainable p-values to eliminate the non-significant tests, which has larger critical constant than the conventional Bonferroni procedure.

Procedure 2.1 (Tarone)**.** *Suppose that p_i^* are the smallest attainable p-values for H_i. Let $M(\alpha, k) = \sum_{i=1}^{m} I\{p_i^* \leq \frac{\alpha}{k}\}$ and $K(\alpha) = \min\{1 \leq k \leq m : M(\alpha, k) \leq k\}$. Then, reject H_i if $P_i \leq \frac{\alpha}{K(\alpha)}$.*

As pointed out by Hommel and Krummenauer [12], the Tarone procedure does not satisfy the desired property of α-consistency defined in Section 3.4. In order to overcome this issue, Hommel and Krummenauer developed a modified Tarone procedure as follows, which is shown satisfying the property of α-consistency.

Procedure 2.2 (Modified Tarone)**.** *Suppose that p_i^* are the smallest attainable p-values for H_i. For any $\gamma \in (0, \alpha]$, let $M(\gamma, k) = \sum_{i=1}^{m} I\{p_i^* \leq \frac{\gamma}{k}\}$ and $K(\gamma) = \min\{1 \leq k \leq m : M(\gamma, k) \leq k\}$. Then reject H_i if there exists an $\gamma \in (0, \alpha]$, such that $P_i \leq \frac{\gamma}{K(\gamma)}$.*

By incorporating the idea of Tarone [20] into the conventional Holm procedure, Hommel and Krummenauer [12] also developed a modified Holm procedure as follows for discrete test statistics.

Procedure 2.3 (Tarone-Holm)**.**

1. *Set $I = \{1, \ldots, m\}$.*
2. *For $k = 1, \ldots, |I|$, let $M_I(\gamma, k) = \sum_{i \in I} I\{p_i^* \leq \frac{\gamma}{k}\}$ and $K_I(\gamma) = \min\{k = 1, \ldots, |I| : M_I(\gamma, k) \leq k\}$.*
3. *For $i \in I$, reject H_i if and only if $P_i \leq \frac{\gamma}{K_I(\gamma)}$ for some $0 < \gamma \leq \alpha$. Let J be the index set of the rejected hypotheses.*
4. *If J is empty, stop testing; otherwise, set $I = I - J$ and then return to step 2.*

In addition, by using the similar idea of Tarone [20], Roth [17] developed a modified Hochberg procedure for discrete test statistics based on the conventional Hochberg procedure.

3 Proposed Stepwise Procedures for Discrete Data

Many existing FWER controlling procedures for discrete data are developed based on the idea of Tarone [20], which only utilizes partial information of true null p-values, so these procedures might be conservative. In this section, we develop more powerful stepwise procedures by fully exploiting known marginal distributions of true null p-values.

3.1 A new single-step procedure

By using the CDFs of true null p-values, we develop a new modified Bonferroni procedure for discrete data as follows.

Procedure 3.1 (Modified Bonferroni)**.** *Let* $s^* = \max\{p \in \bigcup_{i=1}^{m} \mathbb{P}_i : \sum_{i=1}^{m} F_i(p) \leq \alpha\}$ *and set* $s^* = \frac{\alpha}{m}$ *if the maximum does not exist. For any hypothesis* H_i*, reject* H_i *if its corresponding* p*-value* $P_i \leq s^*$*.*

It should be noted that the proposed modified Bonferroni procedure for discrete data is a natural extension of the usual Bonferroni method. When all true null p-values are $U[0,1]$, its critical value $s^* = \max\{p \in [0,1] : mp \leq \alpha\} = \frac{\alpha}{m}$, which is the same as that of the usual Bonferroni procedure. Thus, the modified Bonferroni reduces to the conventional Bonferroni procedure under such setting. For the proposed Procedure 3.1, the following result holds.

Theorem 3.1. *Procedure 3.1 (Modified Bonferroni) strongly controls the FWER at level* α *under Assumption 2.1.*

Compared to the existing Tarone procedure (Procedure 2.1) and modified Tarone procedure (Procedure 2.2) for discrete data, we have

Proposition 3.1. *Procedure 3.1 (Modified Bonferroni) is universally more powerful than Procedures 2.1 (Tarone) and 2.2 (Modified Tarone), that is, for any* H_i*, if it is rejected by Procedure 2.1 or 2.2, it is also rejected by Procedure 3.1.*

It is useful to calculate its *adjusted p-values* for a multiple testing procedure, since one can make decisions of rejection and acceptance as in single hypothesis testing by simply comparing the adjusted p-values with the given significance level. By Westfall and Young [22], the adjusted p-value for a hypothesis in multiple testing is the smallest significance level at which one would reject the hypothesis using the given multiple testing procedure. Thus, the adjusted p-values $\tilde{P}_{i,MBonf}$ of Procedure 3.1 for H_i can be derived as follow:

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

It is easy to see that the adjusted p-values of Procedure 3.1 are smaller than or equal to those of the conventional Bonferroni procedure, since for any given p-value P_i and $j = 1, \ldots, m$, $F_j(P_i) \leq P_i$, then $\sum\limits_{j=1}^{m} F_j(P_i) \leq mP_i$. Thus, Procedure 3.1 is uniformly more powerful than the conventional Bonferroni.

3.2 A new step-down procedure

By using the similar idea as in Section 3.1, we develop a new modified Holm procedure for discrete data as follows.

Procedure 3.2 (Modified Holm). *For $i = 1, \ldots, m$, let $\alpha_i = \max\{p \in \bigcup\limits_{j=i}^{m} \mathbb{P}_{(j)} : \sum\limits_{j=i}^{m} F_{(j)}(p) \leq \alpha\}$ if the maximum exists; otherwise, set $\alpha_i = \max\left\{\alpha_{i-1},\ \dfrac{\alpha}{m-i+1}\right\}$ with $\alpha_0 = 0$. Then reject $H_{(1)}, \ldots, H_{(i^*)}$ and retain $H_{(i^*+1)}, \ldots, H_{(m)}$, where $i^* = \max\{i : P_{(1)} \leq \alpha_1, \ldots, P_{(i)} \leq \alpha_i\}$, if the maximum exists; otherwise, accepts all the null hypotheses.*

It should be noted that when all true null p-values are $U[0,1]$, the critical values

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

Thus, the proposed modified Holm procedure reduces to the conventional Holm procedure under such case.

Theorem 3.2. *Procedure 3.2 (Modified Holm) strongly controls the FWER at level α under Assumption 2.1.*

Compared to the existing Tarone-Holm procedure for discrete data (Procedure 2.3), we can show that Procedure 3.2 is universally more powerful than Procedure 2.3. That is, for any H_i, if it is rejected by Procedure 2.3, it is also rejected by Procedure 3.2.

Proposition 3.2. *Procedure 3.2 (Modified Holm) is universally more powerful than Procedure 2.3 (Tarone-Holm).*

Similar to Procedure 3.1, the adjusted p-values $\tilde{P}_{(i),MHolm}$ of Procedure 3.2 for corresponding hypotheses $H_{(i)}$ can be directly calculated as follows.

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

3.3 A new step-up procedure

Similar to Procedures 3.1 and 3.2, by fully exploiting the marginal distributions of true null p-values, we can also develop a new modified Hochberg procedure for discrete data as follows, which uses the same critical constants as Procedure 3.2.

Procedure 3.3 (Modified Hochberg)**.** *For $i = 1, \ldots, m$, let $\alpha_i = \max\{p \in \bigcup_{j=i}^{m} \mathbb{P}_{(j)} : \sum_{j=i}^{m} F_{(j)}(p) \leq \alpha\}$ if the maximum exists; otherwise set $\alpha_i = \max\left\{\alpha_{i-1},\ \frac{\alpha}{m-i+1}\right\}$ with $\alpha_0 = 0$. Then reject $H_{(1)}, \ldots, H_{(i^*)}$ and retain $H_{(i^*+1)}, \ldots, H_{(m)}$, where $i^* = \max\{i : P_{(i)} \leq \alpha_i\}$, if the maximum exists; otherwise accepts all the null hypotheses.*

It should be noted that when all true null p-values are $U[0, 1]$, the above procedure reduces to the conventional Hochberg procedure.

Proposition 3.3. *Suppose that the true null p-values are identically distributed, then*

(i) Procedure 3.3 (Modified Hochberg) strongly controls the FWER at level α under Assumptions 2.1 and 2.2.

(ii) Procedure 3.3 (Modified Hochberg) rejects the same hypotheses as the conventional Hochberg procedure.

When the true null p-values are not identically distributed, let us consider a special case of testing two null hypotheses $H_i, i = 1, 2$ for which corresponding p-values P_i under H_i only take two attainable values in $[0, 1]$. Denote the support of P_i under H_i as

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

Without loss of generality, assume $p_1 < p_2$ and at least one of two hypotheses is true.

Proposition 3.4. *Under the special case of testing two hypotheses described as above, Procedure 3.3 (Modified Hochberg) strongly controls the FWER under Assumption 2.1.*

Similar to Procedures 3.1 and 3.2, the adjusted p-values of Procedure 3.3 for corresponding hypotheses $H_{(i)}$ can be directly calculated as follows.

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

3.4 Statistical property

In multiple testing, α-consistency is a desired statistical property for a multiple testing procedure in terms of the significance level α, which is defined as follow:

Definition 3.1 (Dimitrienko et al. [4])**.** *A multiple testing procedure is called to be α-*consistent *if any hypothesis that is rejected at a given α level by the procedure is always rejected at a higher α level by the same procedure.*

This property of α-consistency implies that for a given $\alpha' > \alpha$, the set of rejections determined at α' level will not become smaller than that at α level. This is a desirable property in practice. For single hypothesis testing, it is trivial that this property is always satisfied by any conventional test. However, for multiple hypotheses testing, not all multiple testing procedures satisfy this property. For example, the Tarone procedure (Procedure 2.1) does not satisfy this property. For our proposed Procedures 3.1, 3.2 and 3.3, it is easy to see that they all satisfy this property.

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

Definition 3.2 (Dimitrienko et al. [4])**.** *A multiple testing procedure is called to be* p-value monotone *if one or more p-values are made smaller, then at least the same or even more hypotheses would be rejected by this procedure.*

The property of p-value monotonicity is always satisfied by conventional p-value based stepwise procedures. It is easy to see that it is also satisfied by all of our proposed procedures. This property helps to avoid logical inconsistency of decisions of rejection and acceptance; as such it is an essential requirement for a multiple testing procedure. Summarizing the above discussion, we have

Proposition 3.5. *Procedures 3.1, 3.2 and 3.3 satisfy the properties of α-consistency and p-value monotonicity.*

4 Simulation studies

In this section, we perform extensive simulation studies to investigate the performances of the proposed procedures in terms of the FWER control and minimal power, the probability of correctly rejecting at least one false null hypotheses. The simulations are conducted based on two typical discrete tests settings: Fisher's Exact Test (FET) and Binomial Exact Test (BET). Suppose that there are two groups of patients, study group (1) and control group (2).

1. FET: There are m independent binomial responses X_{ij} observed for each of N individuals in each group, 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 m one-sided hypotheses $H_i : p_{i1} = p_{i2}$ vs. $H_i' : p_{i1} < p_{i2}$, where p_{ij} is the success probability for the i-th response in group j, and $i = 1, \ldots, m$, $j = 1, 2$. We conduct the experiment using one-sided FET under α level, then the test statistics $T_i \sim Hypergeometric(X_{i1}, N, X_{i1} + X_{i2}, 2N)$.

   In the simulation, we set the number of hypotheses $m = \{5, 10, 15\}$, the true null proportion $\pi_0 = \{0.2, 0.4, 0.6, 0.8\}$, and the sample size for the binomial response per group used is $N = \{25, 50, 75, 100, 125, 150\}$. For true null hypotheses, set the success probability parameter of each 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. Set $\alpha = 0.05$. The observed responses for each individual in the two groups are chosen randomly from Binomial distribution.

2. BET: 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 m one-sided hypotheses $H_i : \lambda_{i1} = \lambda_{i2}$ vs. $H_i' : \lambda_{i1} < \lambda_{i2}$, 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 BET under α level, then the test statistics for study group follow binomial distribution, i.e., $T_i \sim Bin(X_{i1} + X_{i2}, p_i)$, where $p_i = \dfrac{\lambda_{i1}}{\lambda_{i1} + \lambda_{i2}}$.

   In the simulation, we 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 $\lambda_{i1} = \lambda_{i2} = 2$, and for false null hypotheses, set the mean parameter for study group as $\lambda_{i1} = 2$ and for control group as $\lambda_{i2} = 10$. Set $\alpha = \{0.05, 0.1\}$, respectively. The observed responses for each individual in the two groups are also chosen randomly from Binomial distribution.

By using the FET or BET one can calculate the available p-values P_i and all attainable p-values in the set $\mathbb{P}_i$. Then, for each multiple testing procedure, its simulated FWER, minimal power, and the number of rejections can be calculated by taking average of $B = 2000$ iterations.

4.1 Numerical comparisons for single-step procedures

We now present simulation studies comparing the proposed Procedure 3.1 with existing single-step procedures, including the Bonferroni procedure, Sidak procedure and Modified Tarone procedure. Figures 1 and 2 show the simulated FWER levels and minimal powers of all four procedures under the FET setting. The detailed results can be found in Tables S1 and S2 in the supplementary materials. From these simulation results one can observe:

(i) The proposed Procedure 3.1 always controls the FWER at the pre-specified level α and has higher FWER level and greater power than the existing three procedures for different sample size N.

(ii) Compared to the existing procedures, the FWER level of Procedure 3.1 is less conservative and the power advantage is larger for smaller size N, since the data is more discrete under such setting. For example, when testing $m = 10$ hypotheses with

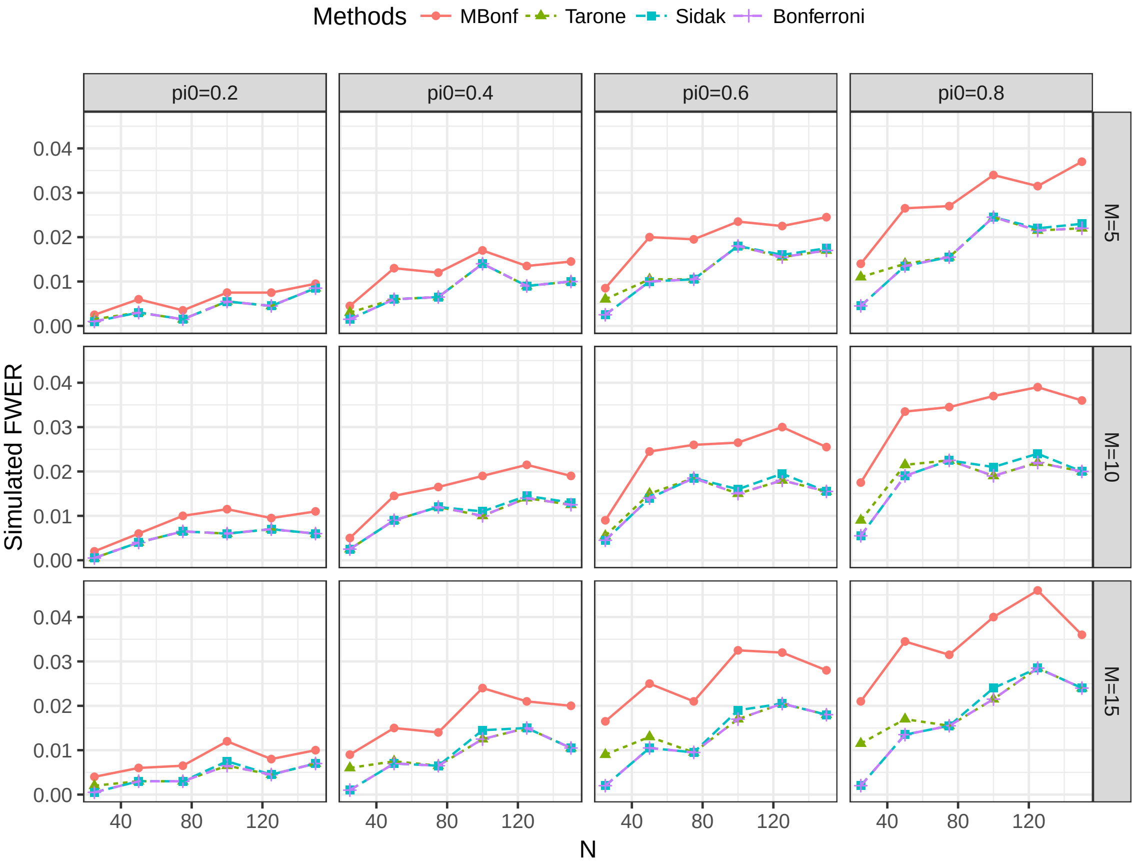

Figure 1: Simulated FWER comparisons for different single-step procedures based on FET, including Procedure 3.1 (MBonf), Procedure 2.1 (Tarone), and the conventional Sidak and Bonferroni procedures.

$\pi_0 = 0.2$, the FWER level of Procedure 3.1 (0.0020) is 300% higher than that of the Tarone procedure (0.0005) when the simulated data is generated from binomial distribution with $N = 5$, however, when $N = 125$, the FWER improvement is only 35.7% (0.0095 versus 0.0070).

(iii) As the proportion of true nulls becomes larger, compared to the existing procedures, the FWER level of Procedure 3.1 is closer to nominal significance level 0.05, but its power performance becomes smaller.

We also conduct simulations by using the BET statistics. The corresponding simulation results are shown in Tables S3 and S4 in the supplementary materials. One can observe from

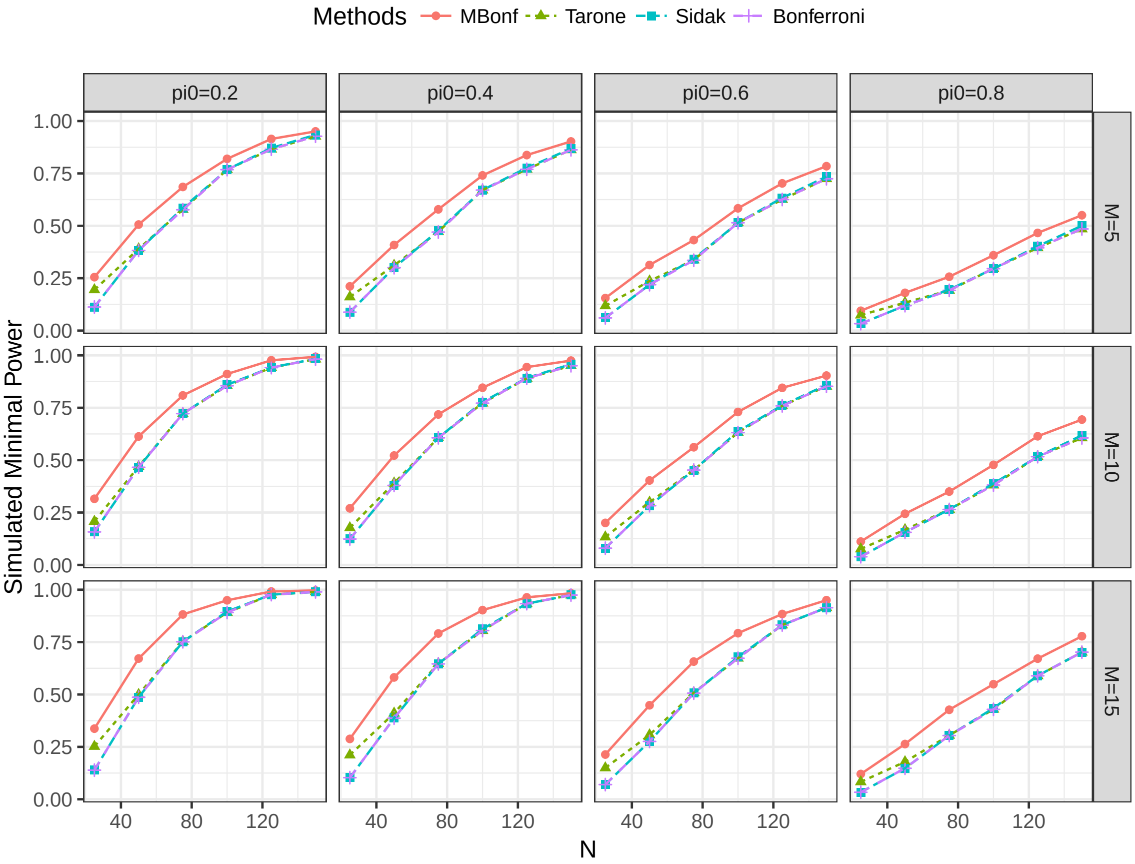

Figure 2: Simulated minimal power comparisons for different single-step procedures based on FET, including Procedure 3.1 (MBonf), Procedure 2.1 (Tarone), and the conventional Sidak and Bonferroni procedures.

Tables S3 and S4 that under the BET setting, the proposed Procedure 3.1 can also control the FWER at level 0.05 and is more powerful than the existing three procedures. For other findings, they are similar to the simulation results obtained under the FET setting.

We also perform simulations to evaluate the effect of dependence among the test statistics on the performance of the proposed Procedure 3.1. The simulation is conducted under block dependence structure for the BET. Details for generating the dependent simulation data can be found in Section S2 of the supplementary materials. In the simulations, we set the number of hypotheses $m = \{5, 10\}$, the true null proportion $\pi_0 = \{0.4, 0.6, 0.8\}$, and the correlation $\rho = \{0, 0.1, \ldots, 0.9\}$. The simulation results under such setting are shown in Figures 3 and 4 and those for other different settings can be found in Tables S9 and S10 of the supplementary materials. These simulation results show that the simulated FWERs of all the four procedures are lower than the pre-specified level 0.05, and the minimal powers for these procedures are decreasing as the correlation ρ becomes larger. More importantly, the proposed Procedure 3.1 is always more powerful than the existing three procedures no matter how the correlation ρ changes.

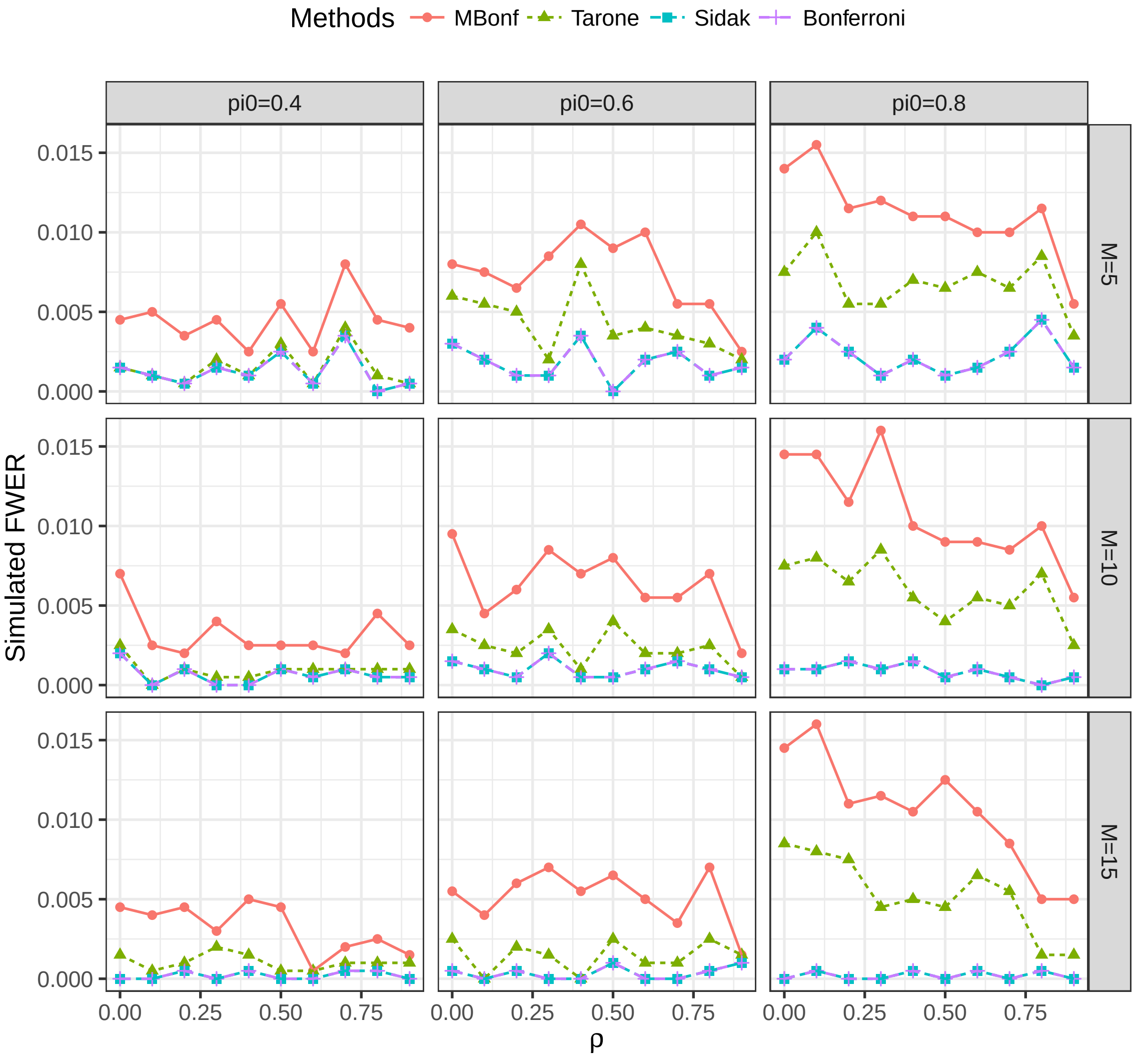

Figure 3: Simulated FWER comparisons for different single-step procedures based on the blocking dependent BET, including Procedure 3.1 (MBonf), Procedure 2.1 (Tarone), and the conventional Sidak and Bonferroni procedures.

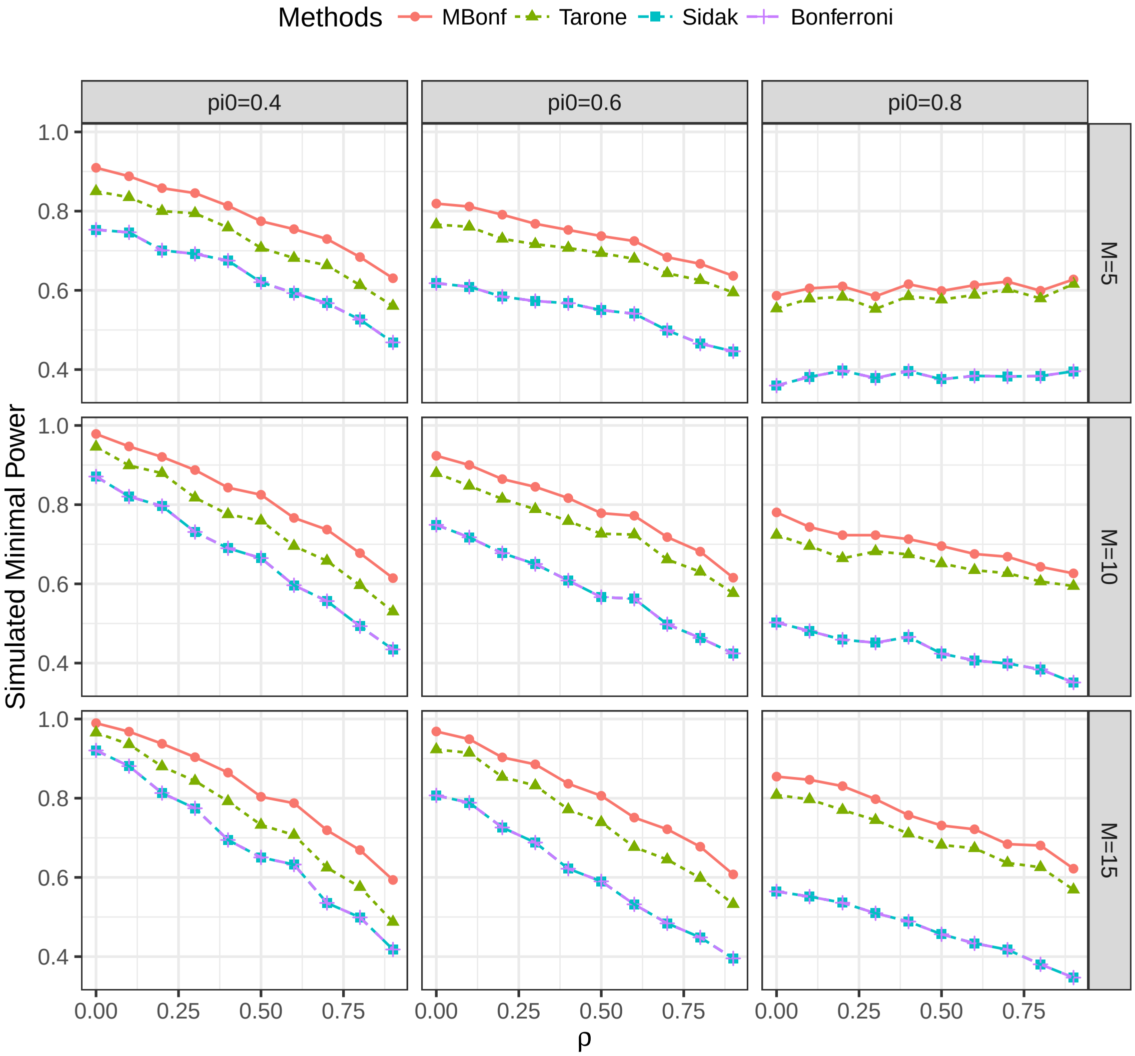

Figure 4: Simulated minimal power comparisons for different single-step procedures based on the blocking dependent BET, including Procedure 3.1 (MBonf), Procedure 2.1 (Tarone), and the conventional Sidak and Bonferroni procedures.

4.2 Numerical comparisons for step-down procedures

Similar to numerical comparisons for the single-step procedures, we also conduct simulation studies to evaluate the proposed step-down Procedure 3.2 in terms of the FWER control and minimal power compared with two existing step-down procedures: the conventional Holm procedure and the Tarone-Holm procedure in Hommel and Krummenauer [12]. We only use FET in the simulations for step-down procedures, since using binomial exact test produces similar patterns. Figures S1 and S2 in the supplementary materials show the simulation results of these step-down procedures under the FET setting. The detailed results can also be found in Tables S5 and S6 in the supplementary materials. These simulation results show that the proposed Procedure 3.2 always controls the FWER at the pre-specified level 0.05 and is more powerful than the existing step-down procedures. Other numerical findings are similar to those of the single-step procedures. In addition, as seen in Tables S1 and S2, the proposed step-down Procedure 3.2 is more powerful than the proposed single-step Procedure 3.1.

We also conduct simulations for step-down procedures under the block dependence structure for BET. In the simulations, we set the number of hypotheses $m = \{5, 10\}$, the true null proportion $\pi_0 = \{0.4, 0.6, 0.8\}$, and the correlation $\rho = \{0, 0.1, \dots, 0.9\}$. The simulation results under such setting are displayed in Figures S5 and S6. Additional simulation results for other different settings can be found in Tables S11 and S12 in the supplementary materials. From these simulation results, one can observe that the simulated FWERs of all these three procedures are lower than the pre-specified level 0.05; the simulated power for each procedure is decreasing in terms of the correlation ρ, and the proposed Procedure 3.2 is always more powerful than the two existing step-down procedures no matter how the correlation ρ changes.

4.3 Numerical comparisons for step-up procedures

In this subsection, we conduct simulation studies to evaluate the proposed step-up Procedure 3.3 in terms of the FWER control and minimal power compared with two existing step-up procedures: the conventional Hochberg procedure and the Roth procedure in Roth (1999). Similar as in Section 4.2, we only use FET in the simulations for step-up procedures, since using binomial exact test produces similar patterns. Figures S3 and S4 in the supplementary materials show the simulation results of these step-up procedures under the FET setting. The detailed results can be found in Tables S7 and S8 in the supplementary materials. These simulation results show that the proposed Procedure 3.3 always controls the FWER at the pre-specified level 0.05 and has greater power than the existing procedures. Other numerical findings are similar to those of the single-step procedures. In addition, as seen in Tables S7 and S8, the proposed step-up Procedure 3.3 is always more powerful than the proposed single-step and step-down procedures 3.1 and 3.2 for different sample size N.

We also conduct simulations for step-up procedures under the block dependence structure for BET. In the simulations, we set the number of hypotheses $m = \{5, 10\}$, the true null proportion $\pi_0 = \{0.4, 0.6, 0.8\}$, and the correlation $\rho = \{0, 0.1, \ldots, 0.9\}$. The simulation results are displayed in Figures S7 and S8. Additional simulation results for other different settings can be found in Tables S13 and S14 in the supplementary materials. From these simulation results, one can observe that the simulated FWERs of all the three step-up procedures are lower than the pre-specified level 0.05; the power for each procedure is decreasing in terms of the correlation ρ, and the proposed Procedure 3.3 is always more powerful than the two existing procedures no matter how the correlation ρ changes.

It should be noted that as the sample size N is larger than some cutoff value, the power improvement for these proposed stepwise procedures could be negligible compared with other existing procedures. Through the simulations studies under the FET setting, we find out that the cutoff value is mainly determined by the sample size N and the number of tested hypotheses m; specifically, when the ratio N/m is larger than 50, the power improvement of the proposed procedures can be relatively minimal.

5 A clinical safety example

The proposed stepwise procedures can be applied in clinical safety studies, since clinical safety data is usually based on the count of patients of having the adverse events exposures. The following example is from Table 1 of Mehrotra and Heyse [15], which reports the AE types of Body System 10 (skin) for two groups of toddlers. The goal of this clinical safety study is to detect significant AEs (so-called “flagging”). For illustration purpose, we reorder the data based on the rank order of the p-values corresponding to the AE types, which are shown in the first three columns of Tables 1-3.

Our analysis includes nine AE types of Body System 10, which are those have a relatively large number of AEs. Thus, they are possibly detected to be significant at level $\alpha = 0.05$ by using Fisher's Exact Test (FET), conditional on the fixed marginal totals. In the data, X_{ij}

Table 1: A comparison of adjusted p-values for the Bonferroni Procedure ($\tilde{P}_{i,Bonf}$), Sidak Procedure ($\tilde{P}_{i,Sidak}$), Procedure 2.2 ($\tilde{P}_{i,T^*}$) and Procedure 3.1 ($\tilde{P}_{i,MBonf}$) when testing the hypotheses for nine AE types of Body System 10 in the clinical safety data example from Mehrotra and Heyse [15], where the numbers of patients for two groups are $N_1 = 148$ and $N_2 = 132$.

i	X_{i1}	X_{i2}	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

is the observed number of the j-th group toddlers experiencing the i-th AE for $i = 1, \ldots, 9$ and $j = 1, 2$ ("1" denotes the control group receiving MMR and "2" the study group receiving the candidate vaccine MMRV), and N_j is the total number of toddlers in group j with $N_1 = 148$ and $N_2 = 132$. In Tables 1-3, the first column shows the indices of the AE types after reordering the data, and the second and third columns show the numbers of toddlers experiencing the corresponding AE in the control and study groups. By using the two-sided FET T, the available p-value P_i and minimal attainable p-value p_i^* for i-th AE type can be calculated as follows:

$$P_i = \Pr\{T \geq X_{i2}\} = \sum_{k=X_{i2}}^{X_{\cdot i}} \frac{\binom{N_2}{k}\binom{N_1}{X_{i\cdot}-k}}{\binom{N_1+N_2}{k}} \tag{3}$$

and

$$p_i^* = \frac{\binom{N_2}{X_{i\cdot}}}{\binom{N_1+N_2}{X_{i\cdot}}}, \tag{4}$$

where $X_{i\cdot} = X_{i1} + X_{i2}$.

Based on the calculated p-values, we apply our proposed stepwise procedures to the clinical safety data analysis. We can make decisions of rejection and acceptance by calculating and comparing the adjusted p-values for these procedures with the given significance

Table 2: A comparison of adjusted p-values for the Holm Procedure ($\tilde{P}_{(i),Holm}$), Procedure 2.3 ($\tilde{P}_{(i),TH}$) and Procedure 3.2 ($\tilde{P}_{(i),MHolm}$) when testing the hypotheses for AE types of Body System 10 in the clinical safety data example from Mehrotra and Heyse [15], where the numbers of patients for two groups are $N_1 = 148$ and $N_2 = 132$.

(i)	X_{i1}	X_{i2}	$P_{(i)}$	$\tilde{P}_{(i),Holm}$	$\tilde{P}_{(i),TH}$	$\tilde{P}_{(i),MHolm}$
(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

level. Table 1 shows that for the first three AE types, the corresponding adjusted p-values of Procedure 3.1 are smaller than those of the Bonferroni, Sidak, and Tarone procedures, which implies the corresponding hypotheses $H_{(1)}, H_{(2)}$ and $H_{(3)}$ are more likely to be rejected by Procedure 3.1 than these existing single-step procedures, that is, those AEs are more easily flagged by using Procedure 3.1.

Table 2 shows that for the AEs corresponding to hypotheses $H_{(1)}, H_{(2)}$ and $H_{(3)}$, the adjusted p-values of Procedure 3.2 are smaller than those of the Holm and Tarone-Holm procedures. It implies that our proposed Procedure 3.2 has more chances to reject these three hypotheses than these two procedures, which in turn implies Procedure 3.2 could be more powerful than the existing step-down methods.

Table 3 shows that for the aforementioned AEs, the adjusted p-values of Procedure 3.3 are smaller than those of Hochberg and Roth procedures. It implies that Procedure 3.3 has more chances to reject $H_{(1)}, \ldots, H_{(3)}$ than these two procedures, which in turn implies our proposed Procedure 3.3 could be more powerful than the existing step-up methods.

Table 3: A comparison of adjusted p-values for the Hochberg Procedure ($\tilde{P}_{(i),Hochberg}$), Roth Procedure ($\tilde{P}_{(i),Roth}$) and Procedure 3.3 ($\tilde{P}_{(i),MHoch}$) when testing the hypotheses for AE types of Body System 10 in the clinical safety data example from Mehrotra and Heyse [15], where the numbers of patients for two groups are $N_1 = 148$ and $N_2 = 132$.

(i)	X_{i1}	X_{i2}	$P_{(i)}$	$\tilde{P}_{(i),Hochberg}$	$\tilde{P}_{(i),Roth}$	$\tilde{P}_{(i),MHoch}$
(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

6 Conclusions

In this paper, we have developed three new FWER controlling procedures for discrete data by fully utilizing marginal distributions of true null p-values rather than minimal attainable p-values, which is often used in the developments of existing procedures for discrete data. We have shown that the proposed modified Bonferroni and Holm procedures strongly control the FWER under arbitrary dependence and are more powerful than the existing Tarone-type procedures, whereas the proposed modified Hochberg procedure ensures control of the FWER in special scenarios. Through extensive simulation studies, we have provided numerical evidence of superior performance of the proposed procedures in terms of the FWER control and minimal power, even for the modified Hochberg. We have also developed an R package “MHTdiscrete” and a web application for implementing the proposed procedures.

A limitation for the proposed methods is when the sample size proportional to the number of tested hypotheses is considerable large, their power improvements are relatively minimal compared to the existing methods. It should be also noted that these proposed procedures are developed under a weak assumption of the marginal distributions (CDF) of true null p-values being known. When the marginal distributions are unknown, the proposed procedures are not directly applicable; some estimate of the marginal distributions is needed. In practice, some resampling methods such as permutation or bootstrap can be potentially considered to estimate the distributions. We leave it for future work. Another possible future work is to explore optimality of the suggested modified Bonferroni and Holm procedures under arbitrary dependence, in the sense of that one cannot increase even one of the critical constants while keeping the remaining fixed without losing control of the FWER.

7 Appendix

7.1 Proof of Theorem 3.1.

Proof. Let V denote the number of falsely rejected hypotheses and I_0 the index set of true null hypotheses, then

$$\begin{aligned} FWER = Pr\{V \geq 1\} &= Pr\left\{\bigcup_{i\in I_0}\{P_i \leq s^*\}\right\} \\ &\leq \sum_{i\in I_0} Pr\{P_i \leq s^*\} = \sum_{i\in I_0} F_i(s^*) \\ &\leq \sum_{i=1}^{m} F_i(s^*) \leq \alpha. \end{aligned} \tag{5}$$

The first inequality follows from Bonferroni inequality. The last inequality follows from the following arguments on the definition of s^*: (i) if the maximum exists, the inequality automatically holds; (ii) if the maximum does not exist, $s^* = \dfrac{\alpha}{m}$ and thus by Assumption 2.1,

$$\sum_{i=1}^{m} F_i(s^*) \leq \sum_{i=1}^{m} s^* = m \cdot \frac{\alpha}{m} = \alpha.$$

The proof is complete. □

7.2 Proof of Proposition 3.1.

Proof. Firstly, we prove that Procedure 3.1 is universally more powerful than Procedures 2.1.

For Procedure 2.1, let $R_{K(\alpha)} = \{i : p_i^* \leq \frac{\alpha}{K(\alpha)}\}$, then $|R_{K(\alpha)}| = M(\alpha, K(\alpha)) \leq K(\alpha)$. Thus,

$$\sum_{i=1}^{m} F_i(\frac{\alpha}{K(\alpha)}) = \sum_{i \in R_{K(\alpha)}} F_i(\frac{\alpha}{K(\alpha)}) \leq |R_{K(\alpha)}| \cdot \frac{\alpha}{K(\alpha)} \leq \alpha. \tag{6}$$

Let

$$t^* = \min \left\{ p \in \bigcup_{i=1}^{m} \mathbb{P}_i : p > s^* \right\}.$$

Then, by the definition of s^* in Procedure 3.1, we have

$$\sum_{i=1}^{m} F_i(t^*) > \alpha. \tag{7}$$

Combining (6) and (7), we have $\frac{\alpha}{K(\alpha)} < t^*$. Then there are two cases regarding the critical values s^* and $\frac{\alpha}{K(\alpha)}$ of Procedures 3.1 and 2.1:

(i) if $\frac{\alpha}{K(\alpha)} \leq s^*$, it is trivial that the set of rejections by Procedures 2.1 is no larger than that of Procedure 3.1;

(ii) if $s^* < \frac{\alpha}{K(\alpha)} < t^*$, by the definition of t^*, it follows that

$$\{H_i : P_i \leq s^*\} = \{H_i : P_i < t^*\} = \{H_i : P_i \leq \frac{\alpha}{K(\alpha)}\},$$

which implies that the rejection sets for these two methods are the same.

Summarizing the above two cases, Procedure 3.1 always rejects any hypotheses rejected by Procedures 2.1. That is, Procedure 3.1 is universally more powerful than Procedures 2.1.

Secondly, we prove that Procedure 3.1 is universally more powerful than Procedure 2.2.

We show that for any $\gamma \in (0, \alpha]$, $\sum_{i=1}^{m} F_i(\frac{\gamma}{K(\gamma)}) \leq \alpha$. Let $R_{K(\gamma)} = \{i : p_i^* \leq \frac{\gamma}{K(\gamma)}\}$, then $|R_{K(\gamma)}| = M(\gamma, K(\gamma)) \leq K(\gamma)$. Thus,

$$\begin{aligned} &\sum_{i=1}^{m} F_i(\frac{\gamma}{K(\gamma)}) = \sum_{i \in R_{K(\gamma)}} F_i(\frac{\gamma}{K(\gamma)}) \\ &\leq |R_{K(\gamma)}| \cdot \frac{\gamma}{K(\gamma)} = M(\gamma, K(\gamma)) \cdot \frac{\gamma}{K(\gamma)} \\ &\leq \gamma \leq \alpha. \end{aligned} \tag{8}$$

For the rest of proof, it is similar to the arguments used in the first part and the conclusion follows. □

7.3 Proof of Theorem 3.2.

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

Let k be the smallest (random) index of the minimal true null p-values $\hat{P}_{(1)}$, i.e., $P_{(k)} = \min_{i \in I_0} P_i$ and

$$I_0 \subseteq \{(k), \ldots, (m)\}. \tag{9}$$

Define $\alpha_{I_0} = \max\{p \in \bigcup_{i=1}^{m} \mathbb{P}_i : \sum_{i \in I_0} F_i(p) \leq \alpha\}$ if the maximum exists; otherwise set $\alpha_{I_0} = \frac{\alpha}{m_0}$. It is easy to see that α_{I_0} is constant. In the following, we show by using induction that for $i = 1, \ldots, k$,

$$\alpha_i \leq \max\left\{\alpha_{I_0}, \frac{\alpha}{m_0}\right\}. \tag{10}$$

When $i = 1$, by the definition of α_1, if the maximum exists, then by (9), we have $\alpha_1 \leq \alpha_{I_0}$; if the maximum does not exist, then $\alpha_1 = \frac{\alpha}{m} \leq \frac{\alpha}{m_0}$. Therefore, (10) holds for the case of $i = 1$.

Assume that the inequality (10) holds for $i = l < k$. In the following, we show that (10) also holds for $i = l + 1$. Note that by (9) and $l < k$, we have $m_0 \leq m - k + 1 \leq m - l$ and then $\dfrac{\alpha}{m-l} \leq \dfrac{\alpha}{m_0}$. Thus, if the maximum does not hold, by the induction assumption,

$$\alpha_{l+1} = \max\left\{\alpha_l, \frac{\alpha}{m-l}\right\} \leq \max\left\{\alpha_{I_0}, \frac{\alpha}{m_0}\right\}.$$

If the maximum exists, by using the similar argument as in the case of $i = 1$, the inequality (10) also holds. Therefore, (10) holds for the case of $i = l+1$. By induction, the inequality (10) holds for $i = 1, \ldots, k$. Thus,

$$\begin{aligned} FWER &= Pr\{V \geq 1\} \leq Pr\{\hat{P}_{(1)} \leq \alpha_k\} \\ &\leq \sum_{i \in I_0} Pr\left\{P_i \leq \max\left\{\alpha_{I_0}, \frac{\alpha}{m_0}\right\}\right\} \\ &= \max\left\{\sum_{i \in I_0} F_i(\alpha_{I_0}), \sum_{i \in I_0} F_i\left(\frac{\alpha}{m_0}\right)\right\} \\ &\leq \alpha. \end{aligned} \tag{11}$$

Here, the second inequality follows from (10) and the Bonferroni inequality, and the third inequality follows from the same argument as in Theorem 3.1 along with the definition of α_{I_0} and Assumption 2.1. The proof is complete. □

7.4 Proof of Proposition 3.2.

Proof. For $i = 1, \ldots, m$, denote I_i as the index set of the ordered p-values starting from $P_{(i)}$, i.e., $I_i = \{(i), \ldots, (m)\}$. Let $R_{K_{I_i}(\gamma)} = \{j \in I_i : p_j^* \leq \dfrac{\gamma}{K_{I_i}(\gamma)}\}$, then $|R_{K_{I_i}(\gamma)}| = M_{I_i}(\gamma, K_{I_i}(\gamma)) \leq K_{I_i}(\gamma)$. Thus, by using the similar argument as in the proof of Proposition 3.1, we have

$$\begin{aligned} &\sum_{j=i}^{m} F_{(j)}(\frac{\gamma}{K_{I_i}(\gamma)}) = \sum_{j \in I_i} F_j(\frac{\gamma}{K_{I_i}(\gamma)}) \\ &= \sum_{j \in R_{K_{I_i}(\gamma)}} F_j(\frac{\gamma}{K_{I_i}(\gamma)}) \leq |R_{K_{I_i}(\gamma)}| \cdot \frac{\gamma}{K_{I_i}(\gamma)} \\ &= M_{I_i}(\gamma, K_{I_i}(\gamma)) \cdot \frac{\gamma}{K_{I_i}(\gamma)} \leq \gamma \leq \alpha. \end{aligned} \tag{12}$$

For the rest of proof, it is similar to that of Proposition 3.1 and the conclusion follows. □

7.5 Proof of Proposition 3.3.

Proof. Since the true null p-values are identically distributed, let us assume that the true null p-values P_i have the same support $\mathbb{P}$ and CDF $F(\cdot)$. Thus, for each $i = 1, \ldots, m$,

$$\begin{aligned} \alpha_i &= \max\{p \in \bigcup_{j=i}^{m} \mathbb{P}_{(j)} : \sum_{j=i}^{m} F_{(j)}(p) \leq \alpha\} \\ &= \max\{p \in \mathbb{P} : (m-i+1)F(p) \leq \alpha\} \\ &= \max\left\{p \in \mathbb{P} : p \leq \frac{\alpha}{m-i+1}\right\}. \end{aligned} \tag{13}$$

The last equality follows from Assumption 2.1. Obviously, $\alpha_i \leq \frac{\alpha}{m-i+1}$, that is, α_i is always smaller than or equal to the critical value $\frac{\alpha}{m-i+1}$ of the conventional Hochberg. By the FWER control of the Hochberg procedure under under Assumption 2.2, we have that Procedure 3.3 also controls the FWER under Assumption 2.2.

To prove (ii), let $R = \max\{i : P_{(i)} \leq \frac{\alpha}{m-i+1}\}$ be the number of rejections by the Hochberg procedure, then for each H_i, H_i is rejected by the Hochberg procedure if $P_i \leq P_{(R)}$. Thus, by (13),

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

We also note that

$$P_{(i)} > \frac{\alpha}{m-i+1} \geq \alpha_i \;\text{ for } i = R+1, \ldots, m.$$

Thus, the number of rejections by Procedure 3.3 is $\max\{i : P_{(i)} \leq \alpha_i\} = R$. Therefore, Procedure 3.3 rejects the same hypotheses $H_{(1)}, \ldots, H_{(R)}$ as the conventional Hochberg procedure. □

7.6 Proof of Proposition 3.4.

Proof. By the definition of the critical values of Procedure 3.3, the critical values $\alpha_i, i = 1, 2$ for this procedure under the special case of two null hypotheses are calculated as

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

and if $P_1 \leq P_2$,

$$\alpha_2 = \begin{cases} \alpha, & \alpha < p_2 \\ p_2, & \alpha \geq p_2; \end{cases}$$

otherwise,

$$\alpha_2 = \begin{cases} \alpha, & \alpha < p_1 \\ p_1, & \alpha \geq p_1. \end{cases}$$

In the following, we prove control of the FWER for Procedure 3.3 for different combinations of true and false null hypotheses.

Case 1. H_1 and H_2 are both true.

There are three attainable p-value settings in which at least one hypothesis is rejected.

(i) $P_1 = p_1$ and $P_2 = 1$. Since $P_2 = 1 > \alpha_2$, accept H_2. To reject H_1, one needs to check if $P_1 \leq \alpha_1$, i.e., $p_1 \leq \alpha$. Thus, H_1 is rejected iff $p_1 \leq \alpha$.

(ii) $P_1 = 1$ and $P_2 = p_2$. Similarly, H_1 is accepted since $P_1 > \alpha_2$. To reject H_2, one needs to check if $P_2 \leq \alpha_1$, which is equivalent to $p_1 + p_2 \leq \alpha$. Thus, H_2 is rejected iff $p_1 + p_2 \leq \alpha$.

(iii) $P_1 = p_1$ and $P_2 = p_2$. By the definition of step-up procedure, it is easy to check that H_1 and H_2 are both rejected iff $p_2 \leq \alpha$; only H_1 is rejected iff $p_1 \leq \alpha < p_2$. Thus, by $p_1 < p_2$, we have that at least one hypothesis is rejected iff $p_1 \leq \alpha$.

Therefore, if $p_1 \leq \alpha$ but $p_1 + p_2 > \alpha$,

$$\begin{aligned} FWER &= \Pr\{H_1 \text{ or } H_2 \text{ rejected}\} \\ &= \Pr(P_1 = p_1, P_2 = 1) + \Pr(P_1 = p_1, P_2 = p_2) \\ &= Pr(P_1 = p_1) = p_1 \leq \alpha. \end{aligned} \tag{14}$$

If $p_1 + p_2 \leq \alpha$,

$$\begin{aligned} FWER &= \Pr\{H_1 \text{ or } H_2 \text{ rejected}\} \\ &= \Pr(P_1 = p_1, P_2 = 1) + \Pr(P_1 = 1, P_2 = p_2) + \Pr(P_1 = p_1, P_2 = p_2) \\ &\leq Pr(P_1 = p_1) + \Pr(P_2 = p_2) \\ &= p_1 + p_2 \leq \alpha. \end{aligned} \tag{15}$$

If $p_1 > \alpha$,

$$FWER = 0. \tag{16}$$

Combining (14)-(16), the desired result follows under Case 1.

Case 2. H_1 is true but H_2 is false.

By the p-value monotonicity of Procedure 3.3, its FWER is maximized when $P_2 = 0$ with probability 1. Thus, H_1 is rejected iff $P_1 \leq \alpha_2$, which is equivalent to $p_1 \leq \alpha$. Therefore, if $p_1 \leq \alpha$,

$$\begin{aligned} FWER = \Pr\{H_1 \text{ rejected}\} &= \Pr(P_1 = p_1) \\ &= p_1 \leq \alpha; \end{aligned} \tag{17}$$

otherwise,

$$FWER = 0. \tag{18}$$

Combining (17)-(18), the desired result follows under Case 2.

Case 3. H_1 is false but H_2 is true.

By using the similar arguments as in Case 2, we have

$$FWER = \Pr\{H_2 \text{ rejected}\} \leq \alpha. \tag{19}$$

Summarizing the above discussions under Cases 1-3, we have that Procedure 3.3 strongly controls the FWER, which completes the proof. □

Supplementary Materials

The online supplementary materials contain additional simulation results for independence and dependence settings, and a brief description of statistical computing softwares for the proposed methods.

Acknowledgements

The research of Wenge Guo was supported in part by NSF Grant DMS-1309162.

References

[1] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**, 1165–1188.

[2] Block, H. W., Savits, T. H., and Shaked, M. (1985). A concept of negative dependence using stochastic ordering. *Statistics & Probability Letters*, **3**, 81–86.

[3] Chen, X., Doerge, R. W., and Heyse, J. F. (2018). Multiple testing with discrete data: proportion of true null hypotheses and two adaptive FDR procedures. *Biometrical Journal*, **60**, 761–779.

[4] Dmitrienko, A., Tamhane, A. C., and Bretz, F. (2009). *Multiple Testing Problems in Pharmaceutical Statistics*. CRC Press.

[5] Döhler, S. (2018). A discrete modification of the Benjamini–Yekutieli procedure. *Econometrics and Statistics*, **5**, 137–147.

[6] Goeman, J. J. and Solari, A. (2014). Multiple hypothesis testing in genomics. *Statistics in Medicine*, **33**, 1946–1978.

[7] Gould, A. L. (2015). *Statistical Methods for Evaluating Safety in Medical Product Development*. John Wiley & Sons.

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

[9] Heyse, J. F. (2011). A false discovery rate procedure for categorical data. *Resent Advances in Biostatistics: False Discovery Rates, Survival Analysis, and Related Topics*, **4**, 43–58.

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

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

[12] Hommel, G. and Krummenauer, F. (1998). Improvements and modifications of Tarone's multiple test procedure for discrete data. *Biometrics*, **54**, 673–681.

[13] Jiang, Q. and Xia, H. A. (2014). *Quantitative Evaluation of Safety in Drug Development: Design, Analysis and Reporting*. Chapman and Hall/CRC.

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

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

[16] R Development Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

[18] Sarkar, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Annals of Statistics*, **30**, 239–257.

[19] Sarkar, S. K. (2008). On the simes inequality and its generalization. In *Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen*, pages 231–242. Institute of Mathematical Statistics.

[20] Tarone, R. E. (1990). A modified Bonferroni method for discrete data. *Biometrics*, **46**, 515–522.

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

[22] Westfall, P. H. and Young, S. S. (1993). *Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment*. John Wiley & Sons.