

## CONTROL OF DIRECTIONAL ERRORS IN FIXED SEQUENCE MULTIPLE TESTING

Anjana Grandhi, Wenge Guo and Joseph P. Romano

*Merck & Co., New Jersey Institute of Technology,  
Stanford University*

*Abstract:* In this paper, we consider the problem of simultaneously testing many two-sided hypotheses when rejections of null hypotheses are accompanied by claims of the direction of the alternative. The fundamental goal is to construct methods that control the mixed directional familywise error rate (mdFWER), which is the probability of making any type 1 or type 3 (directional) error. In particular, attention is focused on cases where the hypotheses are ordered as  $H_1, \dots, H_n$ , so that  $H_{i+1}$  is tested only if  $H_1, \dots, H_i$  have all been previously rejected. In this situation, one can control the usual familywise error rate under arbitrary dependence by the basic procedure which tests each hypothesis at level  $\alpha$ , and no other multiplicity adjustment is needed. However, we show that this is far too liberal if one also accounts for directional errors. But, by imposing certain dependence assumptions on the test statistics, one can retain the basic procedure. Through a simulation study and a clinical trial example, we numerically illustrate good performance of the proposed procedures compared to the existing mdFWER controlling procedures. The proposed procedures are also implemented in the R-package FixSeqMTP.

*Key words and phrases:* Directional error, fixed sequence multiple testing, mixed directional familywise error rate, monotone likelihood ratio, positive dependence, type 1 error.

### 1. Introduction

Directional errors or type 3 errors occur in testing situations with two-sided alternatives when rejections are accompanied by additional directional claims. For example, when testing a null hypothesis  $\theta = 0$  against  $\theta \neq 0$ , rejection of the null hypothesis is often augmented with the decision of whether  $\theta > 0$  or  $\theta < 0$ . In the case of testing a single hypothesis, type 3 error is generally controlled at level  $\alpha$  when type 1 error is controlled at level  $\alpha$  (and sometimes type 3 error is controlled at level  $\alpha/2$ ). However, in the case of simultaneously testing multiple hypotheses, it is often not known whether additional directional decisions can be made without losing control of the mixed directional familywise error rate (mdFWER), the probability of at least one type 1 or type 3 error. Some methods

have been developed in the literature by augmenting additional directional decisions to the existing  $p$ -value based stepwise procedures. Shaffer (1980) showed that Holm's procedure (Holm (1979)), augmented with decisions on direction based on the values of test statistics, can strongly control mdFWER under the assumption that the test statistics are independent and under specified conditions on the marginal distributions of the test statistics, but she also showed that counterexamples exist even with two hypotheses. Finner (1994) and Liu (1997) independently proved the same result for the Hochberg procedure (Hochberg (1988)). Finner (1999) generalized the result of Shaffer (1980) to a large class of stepwise or closed multiple test procedures under the same assumptions. Some recent results have been obtained in Guo and Romano (2015).

Several situations occur in practice where hypotheses are ordered in advance, based on relative importance by some prior knowledge (for example in a dose-response study, hypotheses of higher dose vs. a placebo are tested before those of lower dose vs. placebo), or there exists a natural hierarchy in tested hypotheses (for example in a clinical trial, secondary endpoints are tested only when the associated primary endpoints are significant), and so on. In such fixed sequence multiple testing situations, it is also desired to make further directional decisions once significant differences are observed. For example, in dose response studies, once the hypothesis of no difference between a dose and placebo is rejected, it is of interest to decide whether the new treatment dose is more or less effective than the placebo. In such cases, the possibility of making type 3 errors must be taken into account.

For control of the usual familywise error rate (FWER) (which does not account for the possibility of additional type 3 errors), the conventional *fixed sequence* multiple testing procedure that strongly controls the FWER under arbitrary dependence, is known to be a powerful procedure in testing situations with pre-ordered hypotheses (Maurer, Hothorn and Lehman (1995)). For reviews on recent relevant developments of fixed sequence multiple testing procedures for testing strictly pre-ordered hypotheses and gatekeeping strategies for testing partially pre-ordered hypotheses, see Dmitrienko, Tamhane and Bretz (2009), and Dmitrienko and Huque (2013). Indeed, suppose null hypotheses  $H_1, \dots, H_n$  are pre-ordered, so that  $H_{i+1}$  is tested only if  $H_1, \dots, H_i$  have all been rejected. The probability mechanism generating the data is  $P$  and  $H_i$  asserts that  $P \in \omega_i$ , some family of data generating distributions. In such case, it is easy to see that each  $H_i$  can be tested at level  $\alpha$  in order to control the FWER at level  $\alpha$ , so that no adjustment for multiplicity is required. The argument is simple and goes

as follows. Fix any given  $P$  such that at least one  $H_i$  is true (or otherwise the FWER is 0 anyway). If  $H_1$  is true,  $P \in \omega_1$ , then a type 1 error occurs if and only if  $H_1$  is rejected, and so the FWER is just the probability  $H_1$  is rejected, which is assumed controlled at level  $\alpha$  when testing  $H_1$ . If  $H_1$  is false, let  $f$  be the smallest index corresponding to a true null hypothesis,  $H_f$  is true but  $H_1, \dots, H_{f-1}$  are all false. In this case, a type 1 error occurs if and only if  $H_f$  is rejected, which is assumed to be controlled at level  $\alpha$ .

In situations where ordering is not specified, this result suggests it may be worthwhile to think about hypotheses in order of importance so that potentially false hypotheses are more easily detected. Indeed, as is well-known, when the number  $n$  of tested hypotheses is large, control of the FWER is often so stringent that often no rejections can be detected, largely due to the multiplicity of tests and the need to find significance at very low levels (as required, for example, in the Bonferroni method with  $n$  large). On the other hand, under a specified ordering, each test is carried out at the same conventional level.

To our knowledge, no one has explored the possibility of making additional directional decisions for such fixed sequence procedures. We introduce a fixed sequence procedure augmented with additional directional decisions and discuss its mdFWER control under independence and some dependence. For such directional procedures, the simple fixed sequence structure of the tested hypotheses makes the notoriously challenging problem of controlling the mdFWER under dependence a little easier to handle than stepwise procedures.

Throughout this work, we consider the problem of testing  $n$  two-sided hypotheses  $H_1, \dots, H_n$  specified as follows:

$$H_i : \theta_i = 0 \quad \text{vs.} \quad H'_i : \theta_i \neq 0, \quad i = 1, \dots, n. \quad (1.1)$$

We assume the hypotheses are ordered in advance, either using some prior knowledge about the importance of the hypotheses or by some other specified criteria, so that  $H_1$  is tested first and  $H_i$  is only tested if  $H_1, \dots, H_{i-1}$  are all rejected. We also assume that, for each  $i$ , a test statistic  $T_i$  and  $p$ -value  $P_i$  are available to test  $H_i$  (as a single test). For a rejected hypothesis  $H_i$ , we decide on the sign of the parameter  $\theta_i$  by the sign of the corresponding test statistic  $T_i$ , i.e., we conclude  $\theta_i > 0$  if  $T_i > 0$  and vice versa. The errors that might occur while testing these hypotheses are type 1 and type 3 errors. A *type 1 error* occurs when a true  $H_i$  is falsely rejected. A *type 3 error* occurs when a false  $H_i$  is correctly rejected but the claimed sign of the parameter  $\theta_i$  is wrong. Then, the mdFWER is the probability of making at least a type 1 or type 3 error, and it is desired that this

error rate is no bigger than  $\alpha$  for all possible data generating distributions in the model.

We make a few standard assumptions about the test statistics. Let  $T_i \sim F_{\theta_i}(\cdot)$  for some continuous cumulative distribution function  $F_{\theta_i}(\cdot)$  having parameter  $\theta_i$ . In general, most of our results also apply through the same arguments when the family of distributions of  $T_i$  depends on  $i$ , though for simplicity of notation, the notation is suppressed. We assume that  $F_0$  is symmetric about 0 and  $F_{\theta_i}$  is stochastically increasing in  $\theta_i$ . Various dependence assumptions between the test statistics will be used throughout the paper. (Some of the results can generalize outside this parametric framework. Of course, for many problems, approximations are used to construct marginal tests and the approximate distributions of the  $T_i$  are often normal, in which case our exact finite sample results hold approximately as well.) Let  $c_1 = F_0^{-1}(\alpha/2)$  and  $c_2 = F_0^{-1}(1 - \alpha/2)$ , so that a marginal level  $\alpha$  test of  $H_i$  rejects if  $T_i < c_1$  or  $T_i > c_2$ . For testing  $H_i$  vs.  $H'_i$ , rejections are based on large values of  $|T_i|$  and the corresponding two-sided  $p$ -value is defined by

$$P_i = 2 \min(F_0(T_i), 1 - F_0(T_i)), \quad i = 1, \dots, n. \quad (1.2)$$

We assume that the  $p$ -value  $P_i$  is distributed as  $U(0,1)$  when  $\theta_i = 0$ .

The rest of the paper is organized as follows. In Section 2, we consider the problem of mdFWER control under no dependence assumptions on the test statistics. Unlike control of the usual FWER where each test can be constructed at level  $\alpha$ , it is seen that  $H_i$  can only be tested at a much smaller level  $\alpha/2^{i-1}$ . This rapid decrease in the critical values used motivates the study of the problem under various dependence assumptions. In Section 3 we introduce a directional fixed sequence procedure and prove that it controls the mdFWER under independence. In Sections 4 and 5 we further discuss its mdFWER control under positive dependence. In Section 6 we numerically evaluate the performances of the proposed procedure through a simulation study. In Section 7 we illustrate an application of the proposed procedures through a clinical trial example. Section 8 makes some concluding remarks. Proofs are relegated to the online supplementary materials.

## 2. The mdFWER Control Under Arbitrary Dependence

A general fixed sequence procedure based on marginal  $p$ -values must specify the critical level  $\alpha_i$  that is used for testing  $H_i$ , in order for the resulting procedure to control the mdFWER at level  $\alpha$ . When controlling the FWER without regard

to type 3 errors, each  $\alpha_i$  can be as large as  $\alpha$ . Our Theorem 1 shows that by using the critical constant  $\alpha_i = \alpha/2^{i-1}$ , the mdFWER is controlled at level  $\alpha$ , and that these critical constants are unimprovable. Formally, the optimal procedure is defined as follows.

**Procedure 1 (Directional fixed sequence procedure under arbitrary dependence)**

- Step 1: If  $P_1 \leq \alpha$  then reject  $H_1$  and continue to test  $H_2$  after making directional decision on  $\theta_1$ : conclude  $\theta_1 > 0$  if  $T_1 > 0$  or  $\theta_1 < 0$  if  $T_1 < 0$ . Otherwise, accept all the hypotheses and stop.
- Step  $i$ : If  $P_i \leq \alpha/2^{i-1}$  then reject  $H_i$  and continue to test  $H_{i+1}$  after making directional decision on  $\theta_i$ : conclude  $\theta_i > 0$  if  $T_i > 0$  or  $\theta_i < 0$  if  $T_i < 0$ . Otherwise, accept the remaining hypotheses  $H_i, \dots, H_n$ .

In the following, we discuss the mdFWER control of Procedure 1 under arbitrary dependence of the  $p$ -values. When testing a single hypothesis, the mdFWER of Procedure 1 reduces to the type 1 or type 3 error rate depending on whether  $\theta = 0$  or  $\theta \neq 0$ , and Procedure 1 reduces to the usual  $p$ -value based method along with the directional decision for the two-sided test.

**Lemma 1.** *Consider testing the single hypothesis  $H : \theta = 0$  against  $H' : \theta \neq 0$  at level  $\alpha$ , using the usual  $p$ -value based method along with a directional decision. If  $H$  is a false null hypothesis, then the type 3 error rate is bounded above by  $\alpha/2$ .*

Generally, when simultaneously testing  $n$  hypotheses, by using Lemma 1 and mathematical induction, we have the following result.

**Theorem 1.** *For Procedure 1 the following conclusions hold.*

- (i) *The procedure strongly controls the mdFWER at level  $\alpha$  under arbitrary dependence of the  $p$ -values.*
- (ii) *One cannot increase even one of the critical constants  $\alpha_i = \alpha/2^{i-1}$ ,  $i = 1, \dots, n$ , while keeping the remaining fixed without losing control of the mdFWER.*

In fact, the proof shows that no strong parametric assumptions are required. However, the rapid decrease in critical values  $\alpha/2^{i-1}$  makes rejection of additional hypotheses difficult. Thus, it is of interest to explore how dependence assumptions can be used to increase these critical constants while maintaining control

of the mdFWER. The assumptions and methods are described in the remaining sections.

Instead of Procedure 1 with rapidly decreasing critical constants, consider the conventional fixed sequence procedure with the same critical constant  $\alpha$  augmented with additional directional decisions, defined in Section 3 as Procedure 2. By using the Bonferroni inequality and Lemma 1, we can prove that the mdFWER of this procedure is bounded above by  $\{(n+1)/2\}\alpha$ .

**Proposition 1.** *The conventional fixed sequence procedure with the same critical constant  $2\alpha/(n+1)$  augmented with additional directional decisions strongly controls the mdFWER at level  $\alpha$  under arbitrary dependence of  $p$ -values.*

It is unclear if the critical constant of the fixed sequence procedure defined in Proposition 1 can be further improved without losing the control of the mdFWER.

### 3. The mdFWER Control Under Independence

We make further assumptions on the distribution of the test statistics.

**Assumption 1** (Independence). *The test statistics,  $T_1, \dots, T_n$ , are mutually independent.*

Of course, it follows that the  $p$ -values  $P_1, \dots, P_n$  are mutually independent as well.

As will be seen, it is necessary to make further assumptions on the family of distributions for each marginal test statistic.

**Definition 1** (Monotone Likelihood Ratio (MLR)). *A family of probability density functions  $f_\delta(\cdot)$  is said to have a monotone likelihood ratio property if, for any two values of the parameter  $\delta$ ,  $\delta_2 > \delta_1$  and any two points  $x_2 > x_1$ ,*

$$\frac{f_{\delta_2}(x_2)}{f_{\delta_1}(x_2)} \geq \frac{f_{\delta_2}(x_1)}{f_{\delta_1}(x_1)}, \quad (3.1)$$

or equivalently,

$$\frac{f_{\delta_1}(x_1)}{f_{\delta_1}(x_2)} \geq \frac{f_{\delta_2}(x_1)}{f_{\delta_2}(x_2)}. \quad (3.2)$$

Definition 1 means that, for fixed  $x_1 < x_2$ , the ratio  $f_\delta(x_1)/f_\delta(x_2)$  is non-increasing in  $\delta$ . Direct implications of Definition 1 in terms of the cdf  $F_\delta(\cdot)$  are

$$\frac{F_{\delta_1}(x_2)}{F_{\delta_1}(x_1)} \leq \frac{F_{\delta_2}(x_2)}{F_{\delta_2}(x_1)}, \quad (3.3)$$

$$\frac{1 - F_{\delta_1}(x_2)}{1 - F_{\delta_1}(x_1)} \leq \frac{1 - F_{\delta_2}(x_2)}{1 - F_{\delta_2}(x_1)}. \tag{3.4}$$

**Assumption 2** (MLR Assumption). *The family of marginal distributions of the  $T_i$  has monotone likelihood ratio.*

Based on the conventional fixed sequence multiple testing procedure, we define a directional fixed sequence procedure as the conventional fixed sequence procedure augmented with directional decisions. Thus, any hypothesis is tested at level  $\alpha$ , and under the specified conditions, no reduction in critical values is necessary in order to achieve mdFWER control.

**Procedure 2 (Directional fixed sequence procedure)**

- Step 1: If  $P_1 \leq \alpha$ , then reject  $H_1$  and continue to test  $H_2$  after making a directional decision on  $\theta_1$ : conclude  $\theta_1 > 0$  if  $T_1 > 0$  or  $\theta_1 < 0$  if  $T_1 < 0$ . Otherwise, accept all the hypotheses and stop.
- Step  $i$ : If  $P_i \leq \alpha$ , then reject  $H_i$  and continue to test  $H_{i+1}$  after making a directional decision on  $\theta_i$ : conclude  $\theta_i > 0$  if  $T_i > 0$  or  $\theta_i < 0$  if  $T_i < 0$ . Otherwise, accept the remaining hypotheses,  $H_i, \dots, H_n$ .

For Procedure 2, in the case of  $n = 2$ , we derive a simple expression for the mdFWER in Lemma 2 below and prove its mdFWER control in Lemma 3 by using such simple expression.

**Lemma 2.** *Consider testing two hypotheses  $H_1 : \theta_1 = 0$  and  $H_2 : \theta_2 = 0$ , against both sided alternatives, using Procedure 2 at level  $\alpha$ . Let  $c_1 = F_0^{-1}(\alpha/2)$  and  $c_2 = F_0^{-1}(1 - \alpha/2)$ . When  $\theta_2 = 0$ ,*

$$mdFWER = \begin{cases} \alpha + F_{\theta_1}(c_1) - F_{\theta_1}(c_2) + F_{(\theta_1,0)}(c_2, c_2) - F_{(\theta_1,0)}(c_2, c_1) & \text{if } \theta_1 > 0 \\ 1 + F_{\theta_1}(c_1) - F_{\theta_1}(c_2) + F_{(\theta_1,0)}(c_1, c_1) - F_{(\theta_1,0)}(c_1, c_2) & \text{if } \theta_1 < 0. \end{cases} \tag{3.5}$$

Here  $F_{\theta_1, \theta_2}(\cdot, \cdot)$  refers to the joint c.d.f. of  $(T_1, T_2)$ . Then, under Assumption 1, (3.5) can be simplified as

$$mdFWER = \begin{cases} \alpha + F_{\theta_1}(c_1) - \alpha F_{\theta_1}(c_2) & \text{if } \theta_1 > 0 \\ 1 + \alpha F_{\theta_1}(c_1) - F_{\theta_1}(c_2) & \text{if } \theta_1 < 0. \end{cases} \tag{3.6}$$

**Lemma 3.** *Under Assumptions 1 and 2, Procedure 2 strongly controls the mdFWER when  $n = 2$ .*

Generally, for testing any  $n$  hypotheses, by using mathematical induction and Lemma 3, we also prove the mdFWER control of Procedure 2 under the same assumptions as in the case of  $n = 2$ .

**Theorem 2.** *Under Assumptions 1 and 2, Procedure 2 strongly controls the mdFWER at level  $\alpha$ .*

Many families of distributions have the MLR property: normal, uniform, logistic, Laplace, Student's t, generalized extreme value, exponential families of distributions, etc. It is also important to know whether or not our results fail without the MLR assumption. A natural family of distributions to consider without the MLR property is the Cauchy family; indeed, Shaffer (1980) used this family to obtain a counterexample for the directional Holm procedure while testing  $p$ -value ordered hypotheses. We now show that Procedure 2 fails to control the mdFWER for this family of distributions with corresponding cdf  $F_\theta(x) = 0.5 + 1/\pi \arctan(x - \theta)$ , even under independence.

Lemma 2 can be used to verify the calculation for the case of  $n = 2$  with  $\theta_1 > 0$  and  $\theta_2 = 0$ ; specifically, see (3.6). Indeed, we just need to show

$$F_{\theta_1}(-c) = F_0(-c - \theta_1) > \alpha F_{\theta_1}(c) = \alpha F_0(c - \theta_1), \quad (3.7)$$

where  $c$  is the  $1 - \alpha/2$  quantile of the standard Cauchy distribution, given by  $\tan[\pi(1 - \alpha)/2]$ . Taking  $\alpha = 0.05$ , so  $c = 12.7062$ , this inequality (3.7) holds, for example, by  $\theta_1 = 100$ . The left side is approximately  $F(-112.7) \approx 0.002824$  while the right side is

$$0.05 \times F(-87.3) = 0.05 \times 0.0036 = 0.00018.$$

#### 4. Extension to Positive Dependence

Clearly, the assumption of independence is of limited utility in multiple testing, as many tests are usually carried out on the same data set. Thus, it is important to generalize the results of the previous section to cover some more general cases. As is typical in the multiple testing literature (Benjamini and Yekutieli (2001); Sarkar (2002); Sarkar and Guo (2010), etc), assumptions of positive regression dependence will be used.

We introduce several notations. Among the prior-ordered hypotheses  $H_1, \dots, H_n$ , let  $i_0$  denote the index of the first true null hypothesis,  $n_1$  denote the number of all false nulls, and  $T_{i_1}, \dots, T_{i_{n_1}}$  denote the corresponding false null test statistics. Specifically, if all  $H_i$ 's are false, let  $i_0 = n + 1$ .

**Assumption 3** (Weak PRD). *The false null test statistics along with parameters,  $\theta_{i_1} T_{i_1}, \dots, \theta_{i_{n_1}} T_{i_{n_1}}$ , are positively regression dependent in the sense of*

$$E \left\{ \phi(\theta_{i_1} T_{i_1}, \dots, \theta_{i_{n_1}} T_{i_{n_1}}) \mid \theta_{i_k} T_{i_k} \geq u \right\} \uparrow u, \quad (4.1)$$

for each  $\theta_{i_k} T_{i_k}$  and any (coordinatewise) non-decreasing function  $\phi$ .

**Assumption 4** (Weak Independence). *The first true null statistic,  $T_{i_0}$ , is independent of all false null statistics  $T_{i_k}, k = 1, \dots, n_1$  with  $i_k < i_0$ .*

**Theorem 3.** *Under Assumptions 2, 3 (weak PRD), and 4, Procedure 2 strongly controls the mdFWER at level  $\alpha$ .*

**Corollary 1.** *When all tested hypotheses are false, Procedure 2 strongly controls the mdFWER at level  $\alpha$  under Assumptions 2 and 3.*

**Remark 1.** When all of the tested hypotheses are false, Assumption 4 is automatically satisfied. Generally, consider the case of any combination of true and false null hypotheses where Assumption 4 is not imposed. Without loss of generality, suppose the first  $n - 1$  hypotheses are false and the last one is true. Under Assumptions 2-3, if the true null statistic  $T_n$  (or  $-T_n$ ) and the false null statistics  $T_1, \dots, T_{n-1}$  are positively regression dependent (a slightly weak version of Assumption 5 in Section 5), the mdFWER of Procedure 2, when testing  $H_1, \dots, H_n$  is, for any  $n$ , bounded above by

$$\begin{aligned} & \Pr(\text{make at least one type 3 error when testing } H_1, \dots, H_{n-1} \text{ or } T_n \notin (c_1, c_2)) \\ & \leq \lim_{\theta_n \rightarrow 0^+} \Pr(\text{make at least one type 3 error when testing } H_1, \dots, H_n) \\ & \quad + \lim_{\theta_n \rightarrow 0^+} \Pr(T_n \geq c_2) \\ & \leq \alpha + \frac{\alpha}{2} = 3\frac{\alpha}{2}. \end{aligned}$$

The first inequality follows from the fact that when  $\theta_n \rightarrow 0^+$ ,  $H_n$  can be interpreted as a false null hypothesis with  $\theta_n > 0$ , and thus one type 3 error is made if  $H_n$  is rejected and  $T_n \leq c_1$ . The second inequality follows from Corollary 1 and Lemma 1.

Based on this inequality, a modified version of Procedure 2, the directional fixed sequence procedure with the critical constant  $2\alpha/3$ , strongly controls the mdFWER at level  $\alpha$  under Assumptions 2-3 and the additional assumption of positive regression dependence.

**Remark 2.** If we do not make any assumption regarding dependence between the true null statistic  $T_n$  and the false null statistics  $T_1, \dots, T_{n-1}$ , then, by Theorem 3, the mdFWER of Procedure 2 when testing  $H_1, \dots, H_n$  is bounded above by

$$\begin{aligned} & \Pr(\text{make at least one type 3 error when testing } H_1, \dots, H_{n-1}) \\ & \quad + \Pr(\text{make type 1 error when testing } H_n) \end{aligned}$$

$$\leq \alpha + \alpha = 2\alpha.$$

Therefore, an alternative modified version of Procedure 2, the directional fixed sequence procedure with the critical constant  $\alpha/2$ , strongly controls the mdFWER at level  $\alpha$  only under Assumptions 2-3.

## 5. Further Extensions to Positive Dependence

We develop alternative results to show that Procedure 2 can control mdFWER even under certain dependence between the false null and true null statistics. We consider a slightly stronger version of the conventional positive regression dependence on subset of true null statistics (PRDS) (Benjamini and Yekutieli (2001)).

**Assumption 5** (PRD). *The false null test statistics,  $T_1, \dots, T_{i_0-1}$  and the first true null statistic  $T_{i_0}$ , are positive regression dependent in the sense of*

$$E\{\phi(T_1, \dots, T_{i_0-1}) \mid T_{i_0} \geq u, T_1, \dots, T_j\} \uparrow u, \quad (5.1)$$

for any given  $j = 1, \dots, i_0 - 1$ , any given values of  $T_1, \dots, T_j$  and any (coordinatewise) non-decreasing function  $\phi$ .

We first consider the case of testing two hypotheses, and show control of the mdFWER of Procedure 2 when the test statistics are positively regression dependent in the sense of Assumption 5.

**Proposition 2.** *Under Assumptions 2 and 5, the mdFWER of Procedure 2 is strongly controlled at level  $\alpha$  when  $n = 2$ .*

Specifically, in the case of the bivariate normal distribution, Assumption 2 is satisfied and test statistics  $T_1$  and  $T_2$  are always positively or negatively regression dependent. As in the proof of Proposition 2, to show the mdFWER control of Procedure 2, we only need to consider the case of  $\theta_1 \neq 0$  and  $\theta_2 = 0$ . Thus, if  $T_1$  and  $T_2$  are negatively regression dependent, we can choose  $-T_2$  as the statistic for testing  $H_2$  and Assumption 5 is still satisfied. By Proposition 2, we have the following.

**Corollary 2.** *Under the bivariate normal case, the mdFWER of Procedure 2 is strongly controlled at level  $\alpha$  when  $n = 2$ .*

Consider the case of three hypotheses. The general case will ultimately be considered, but it is instructive to discuss this case separately due to the added MLR condition, is described as follows.

Let  $f(x|T_1)$  and  $g(x|T_1)$  denote the probability density functions of  $T_2$  and  $T_3$  conditional on  $T_1$ , respectively.

**Assumption 6** (Bivariate Monotone Likelihood Ratio (BMLR)). *For any given value of  $T_1$ ,  $f(x|T_1)$  and  $g(x|T_1)$  have the monotone likelihood ratio (MLR) property in  $x$  if, for any  $x_2 > x_1$ ,*

$$\frac{f(x_2|T_1)}{g(x_2|T_1)} \geq \frac{f(x_1|T_1)}{g(x_1|T_1)}. \quad (5.2)$$

**Proposition 3.** *Under Assumptions 2, 3, 5, and 6, the mdFWER of Procedure 2 is strongly controlled at level  $\alpha$  when  $n = 3$ .*

**Remark 3.** In the case of three hypotheses, suppose that the test statistics  $T_i, i = 1, 2, 3$  are trivariate normally distributed with the mean  $\theta_i$ . Without loss of generality, assume  $\theta_i > 0, i = 1, 2$  and  $\theta_3 = 0$ , that is,  $H_1$  and  $H_2$  are false and  $H_3$  is true. Let  $\Sigma = (\sigma_{ij}), i, j = 1, \dots, 3$ , denote the variance-covariance matrix of  $T_i$ 's. It is easy to see that Assumption 2 is always satisfied. Also, when  $\sigma_{ij} \geq 0$  for  $i \neq j$ , Assumption 3 and Assumption 5 are satisfied. Finally, when  $\sigma_{22} = \sigma_{33}$  and  $\sigma_{12} = \sigma_{13}$ , Assumption 6 is satisfied.

Finally, We consider the general case of  $n$  hypotheses. Here we must consider the multivariate monotone likelihood ratio property, described as follows. For any given  $j = 1, \dots, i_0 - 1$ , let  $f(x|T_1, \dots, T_{j-1})$  and  $g(x|T_1, \dots, T_{j-1})$  denote the probability density functions of  $T_j$  and  $T_{i_0}$  conditional on  $T_1, \dots, T_{j-1}$ , respectively.

**Assumption 7** (Multivariate Monotone Likelihood Ratio (MMLR)). *For any given values of  $T_1, \dots, T_{j-1}$ ,  $f(x|T_1, \dots, T_{j-1})$  and  $g(x|T_1, \dots, T_{j-1})$  satisfy, for any  $x_2 > x_1$ ,*

$$\frac{f(x_2|T_1, \dots, T_{j-1})}{g(x_2|T_1, \dots, T_{j-1})} \geq \frac{f(x_1|T_1, \dots, T_{j-1})}{g(x_1|T_1, \dots, T_{j-1})}. \quad (5.3)$$

**Theorem 4.** *Under Assumptions 2, 3, 5, and 7, the mdFWER of Procedure 2 is strongly controlled at level  $\alpha$ .*

## 6. A Simulation Study

We conducted a simulation study to illustrate the performance of Procedures 1 and 2 in terms of mdFWER control and average power, and compared them with the directional Bonferroni procedure, directional Holm procedure and directional Hochberg procedure. We studied two simulation settings for evaluating the

effects of proportion of false nulls and dependence on the performance of these procedures, respectively. We generated  $n$ -dimensional normal random vectors  $(T_1, \dots, T_n)$  where the components were  $N(\theta_i, 1)$  with common pairwise correlation  $\rho$ . We considered simultaneously testing  $n$  two-sided hypotheses using  $T_i$  along with making directional decisions on  $\theta_i$  based on the sign of  $T_i$ :

$$H_i : \theta_i = 0 \quad \text{vs.} \quad H_i' : \theta_i \neq 0, \quad i = 1, \dots, n. \quad (6.1)$$

In simulations, we set the first  $n_1$  of the  $n$  hypotheses  $H_i$  to be false null and the rest to be true null. The true null test statistics were generated from  $N(0, 1)$  and the false null test statistics were generated from  $N(\theta_i, 1)$  with  $\theta_i \neq 0$ . The simulation results are obtained under the significance level  $\alpha = 0.05$  and based on 10,000 replicates. The “power” of a procedure at a replication is the proportion of non-null  $\theta_i$  to be rejected along with correct directional decisions on  $\theta_i$  to be made among all non-null  $\theta_i$  out of  $n$  hypotheses. The “average power” is the average of the power for the 10,000 replications. The mdFWER is estimated as the proportion of replications where at least one true null hypothesis is falsely rejected or at least one false null hypothesis is correctly rejected but a wrong directional decision is made regarding the corresponding  $\theta_i$ .

### 6.1. Simulation setting 1

In this setting, we set the number of tested hypotheses  $n = 20$ , the common correlations as  $\rho = 0$  or  $\rho = 0.5$ , and the proportion of false null hypotheses  $\pi_1$  to be between 0.05 and 1.0. For the values of  $\theta_i$ , we set  $\theta_i = 3$  for false null and  $\theta_i = 0$  for true null.

Figure 1 shows the plots of mdFWER and average power of all five directional procedures plotted against  $\pi_1$ , the fraction of false null hypotheses. All five procedures control mdFWER at level 0.05 and Procedure 1 has the lowest mdFWER. When the test statistics are independent, the mdFWER of Procedure 2 is also lower than those of the existing procedures, whereas when the test statistics are positively correlated, the mdFWER of Procedure 2 is generally higher than that of the directional Bonferroni procedure but lower than those of the directional Holm and directional Hochberg procedures, except for very high fractions of false nulls.

When the fraction of false nulls is low or moderate ( $\pi_1 \leq 0.4$ ), as is usually expected in practical applications, Procedure 2 has the highest power followed by Procedure 1, both when the test statistics are independent or positively correlated. However, when the fraction of false nulls is high, even Procedure 2 loses its

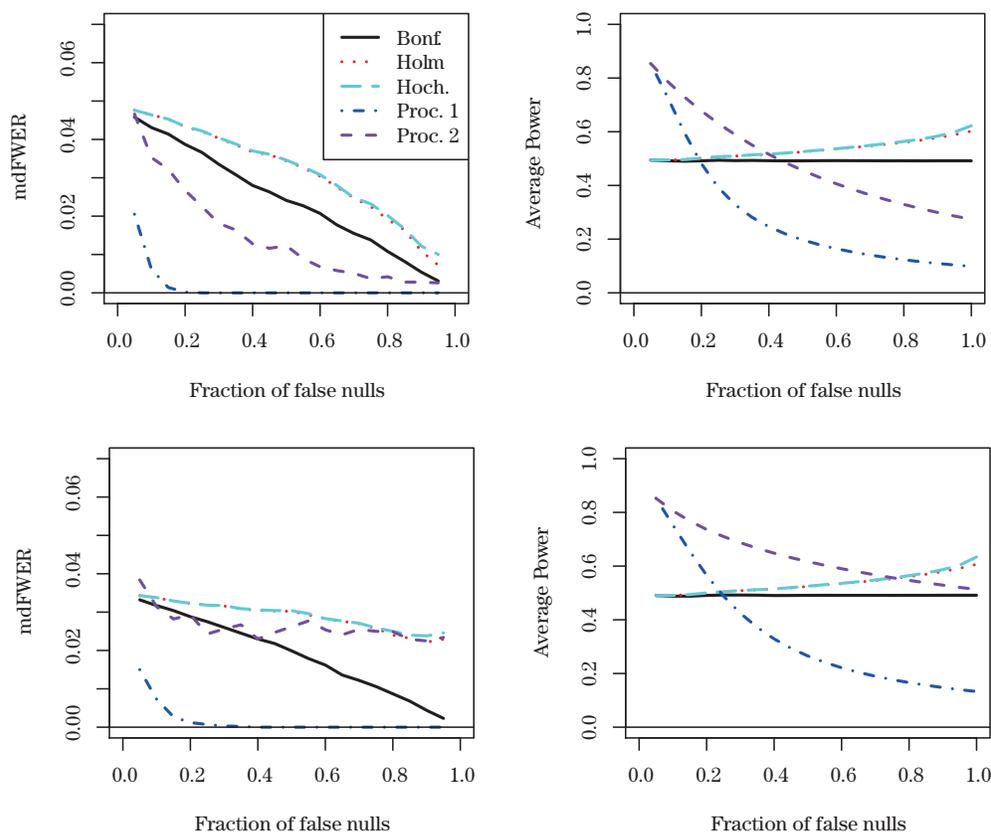


Figure 1. Estimated mdFWER and average powers of Procedures 1 and 2 along with existing directional Bonferroni procedure, directional Holm procedure, and directional Hochberg procedure for  $n = 20$  hypotheses with the fraction of false nulls  $\pi_1$  from 0.05 to 1.0 and common correlation  $\rho = 0$  (upper panel) or  $\rho = 0.5$  (bottom panel).

edge over the existing procedures. Figure 1 shows that the proposed procedures and the existing procedures have different power performances with increasing proportion of false nulls. The average powers of Procedures 1 and 2 are decreasing in terms of the proportion of false nulls, whereas the average powers of the existing Procedures are slightly increasing in the proportion of false nulls.

**6.2. Simulation setting 2**

In this setting, took the number of tested hypotheses to be 20, the number of false null hypotheses  $n_1 = 5$ , and the common correlation  $\rho$  to be between 0 and 1. For the values of non-null  $\theta_i$ , we set  $\theta_i = \theta_0 r^{i-1}, i = 1, \dots, n_1$ , with the values  $(\theta_0, r) = (5, 0.8)$  or  $(\theta_0, r) = (8, 0.5)$ , and for the values of null  $\theta_i$ , we set

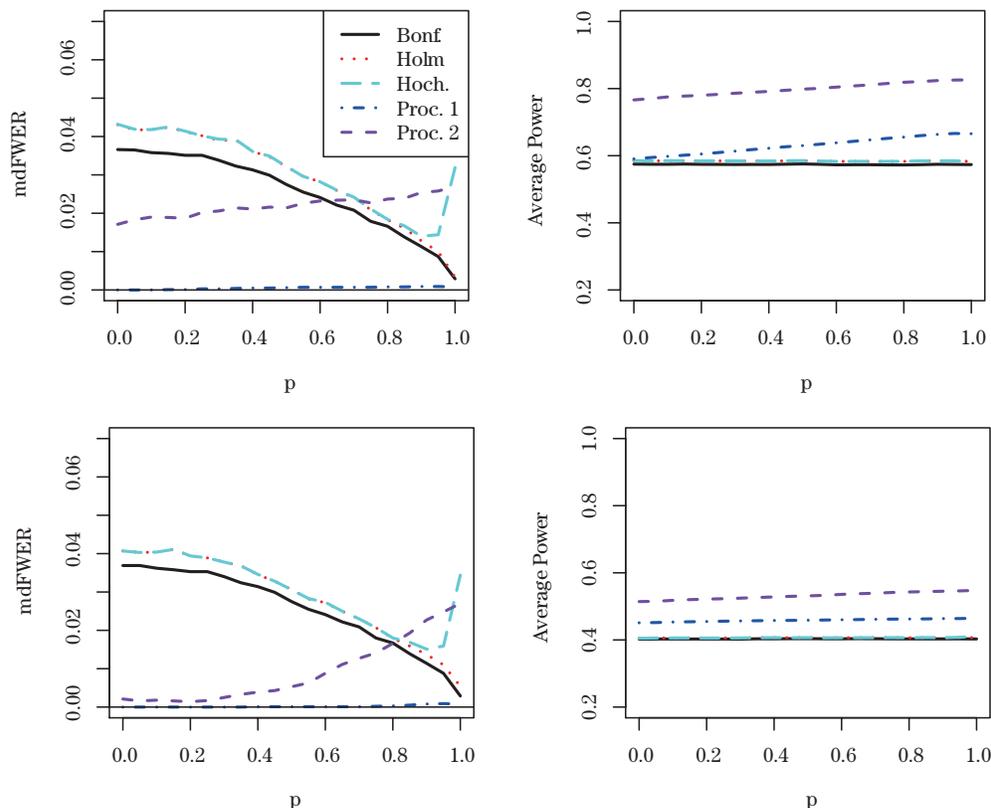


Figure 2. Estimated mdFWER and average powers of Procedures 1 and 2 along with existing directional Bonferroni procedure, directional Holm procedure, and directional Hochberg procedure for  $n = 20$  hypotheses with common correlation  $\rho$  between 0 and 1 and  $n_1 = 5$  non-null  $\theta_i = \theta_0 r^{i-1}$  with  $(\theta_0, r) = (5, 0.8)$  (upper panel) or  $(\theta_0, r) = (8, 0.5)$  (bottom panel).

$\theta_i = 0$ .

Figure 2 shows the plots of mdFWER and average power of all five directional procedures plotted against  $\rho$ , the common correlation. As seen from Figure 2, all the five procedures control the mdFWER at level  $\alpha$  and Procedure 2 has the highest average power followed by Procedure 1 for different values of  $\rho$ . Our procedures have different performances with respect to common correlation compared to the existing procedures. The mdFWER and average powers of Procedures 1 and 2, and their power improvements over the existing three procedures are all increasing in terms of correlation, whereas the mdFWERs of the existing three procedures are basically decreasing in terms of correlation, except for the directional Hochberg procedure, its mdFWER increasing when  $\rho$  is very

large.

Our simulation studies were conducted with false nulls ordered ahead of the true nulls, which may give some advantage in power of Procedures 1 and 2 over the existing procedures.

## 7. Clinical Trial Example

The directional fixed sequence procedure comes in handy in dose-response studies or studies with multiple endpoints where hypotheses are ordered in advance. To illustrate our procedure we use the hypertension trial example considered in (Dmitrienko et al. (2005), p. 118). This clinical trial was conducted to test the efficacy and safety of four doses of an investigational drug versus placebo. The four doses, from lowest to highest, were respectively labeled as D1, D2, D3 and D4 and the placebo was labeled P. The primary efficacy endpoint was the reduction in diastolic blood pressure (measured in mm Hg). Dose D4 was believed to be the most efficacious one (in terms of its effect on diastolic blood pressure), followed by doses D3 and D2 while dose D1 was expected to be marginally efficacious.

The original analysis had eight two sided hypotheses, four were dose-placebo contrasts and four dose-dose contrasts. For our analysis, we use these comparisons to test the hypotheses in the order mentioned to conclude on the direction of efficacy. We applied Procedures 1 and 2 and, for comparison, we also include the results of the Bonferroni single-step procedure appended with directional decisions. Table 1 shows the results of our analysis done at level  $\alpha = 0.05$ .

As seen in Table 1, both the Bonferroni single-step procedure (appended with directional decisions) as well as Procedure 2 reject the most number of hypotheses, namely three. However, we believe that Procedure 2 makes more sense because one can conclude with some statistical validity that D4, D3, and D2 outperform the placebo. Indeed, the sequential nature of our method inherently allows for some kind of internal consistency among the rejections. On the other hand, the results from Bonferroni are less interpretable: one can conclude D4 is superior to the placebo, but not D3 (or D2), though it does allow the conclusion that  $D4 > D1$  and  $D3 > D1$ . Certainly the conclusion that  $D3 > D1$  seems less interesting if one cannot establish that D3 or D1 is any better than the placebo.

While Procedure 1 assumes nothing about the dependence structure of the  $p$ -values, it is obviously more conservative than Procedure 2. However, in the context where each  $p$ -value corresponds to a different dose of the same drug, it is

Table 1. Results of directional fixed sequence procedures in the hypertension trial example (R: Rejected and NR: Not rejected) with  $\alpha = 0.05$ .

Test Contrast	Test statistic	Raw $p$ -value	Procedure 1 Decision (Direction)	Procedure 2 Decision (Direction)	Bonferroni Decision (Direction)
D4-P	3.4434	0.0008	R (More Effective)	R (More Effective)	R (More Effective)
D3-P	2.5085	0.0135	R (More Effective)	R (More Effective)	NR (More Effective)
D2-P	2.3642	0.0197	NR	R (More Effective)	NR
D1-P	-0.3543	0.7237	-	NR	NR
D4-D1	3.7651	0.0003	-	-	R (More Effective)
D4-D2	1.0900	0.2779	-	-	NR
D3-D1	2.8340	0.0054	-	-	R (More Effective)
D3-D2	0.1930	0.8473	-	-	NR
Number Rejected			2	3	3

reasonable to assume positive dependence of the outcomes. In such case, results based on Procedure 2 are valid and indicate that even Dose 2 is significantly beneficial as compared to the placebo.

## 8. Conclusions

In this paper, we consider the problem of simultaneously testing multiple prior-ordered hypotheses accompanied by directional decisions. The conventional fixed sequence procedure augmented with additional directional decisions is shown to control the mdFWER under independence and some dependence, whereas, it is also shown to be far too liberal to control the mdFWER if no dependence assumptions are imposed on the test statistics. Through a simulation study, we numerically showed the good performances of the proposed procedures in terms of the mdFWER control and average power as compared to the existing directional Bonferroni, Holm, and Hochberg procedures. The proposed procedures are also implemented in the R-package FixSeqMTP.

In the existing literature, to our knowledge, only the directional Bonferroni procedure is theoretically shown to strongly control the mdFWER under de-

pendence (Hochberg and Tamhane (1987); Shaffer (1995)). It is still an open problem whether the directional Holm and Hochberg procedures control the mdFWER under certain dependence. Our directional fixed sequence procedure can be a powerful alternative for the problem of directional errors control under dependence. We hope our paper will shed some light on attacking the challenging problem of controlling the mdFWER under dependence for these  $p$ -value ordered stepwise procedures.

## Supplementary Materials

Proofs of theorems and propositions are in the online supplementary materials.

## Acknowledgment

The research of Wenge Guo was supported in part by NSF Grant DMS-1309162 and the research of Joseph Romano was supported in part by NSF Grant DMS-0707085. We sincerely thank three referees for giving helpful and insightful comments and Yalin Zhu for implementing the proposed procedures in the R package FixSeqMTP.

## References

- Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–1188.
- Dmitrienko, A., Molenberghs, G., Chuang-Stein, C. and Offen, W. (2005). *Analysis of Clinical Trials Using SAS: A Practical Guide*. SAS Press, Cary, NC.
- Dmitrienko, A., Tamhane, A. and Bretz, F. (2009). *Multiple Testing Problems in Pharmaceutical Statistics*. Chapman and Hall/CRC Press, New York.
- Dmitrienko, A., D’Agostino, R. and Huque, M. (2013). Key multiplicity issues in clinical drug development. *Stat. Med.* **32**, 1079–1111.
- Finner, H. (1994). *Testing Multiple Hypotheses: General theory, Specific problems, and Relationships to Other Multiple Decision Procedures*. Habilitationsschrift, Fachbereich IV Mathematik, Univ. Trier.
- Finner H. (1999). Stepwise multiple test procedures and control of directional errors. *Ann. Statist.* **27**, 274–289.
- Guo, W. and Romano, J. (2015). On stepwise control of directional errors under independence and some dependence. *J. Statist. Plann. Inference* **163**, 21–33.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–802.
- Hochberg, Y. and Tamhane, A. (1987). *Multiple Comparison Procedures*. John Wiley, New York.

- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70.
- Liu, W. (1997). Control of directional errors with step-up multiple tests. *Statist. Probab. Lett.* **31**, 239–242.
- Maurer, W., Hothorn, L. and Lehmacher, W. (1995). Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. In *Biometrie in der Chemisch-pharmazeutischen Industrie*, 6 (Edited by J. Vollmar), 3–18. Fischer Verlag, Stuttgart.
- Sarkar S. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.* **30**, 239–257.
- Sarkar, S. and Guo, W. (2010). Procedures controlling generalized false discovery rate using bivariate distributions of the null  $p$ -values. *Statist. Sinica* **20**, 1227–1238.
- Shaffer, J. P. (1980). Control of directional errors with stagewise multiple test procedures. *Ann. Statist.* **8**, 1342–1347.
- Shaffer, J. P. (1995). Multiple hypothesis testing. *Annu. Rev. Psychol.* **46**, 561–584.

Biostatistics and Research Decision Sciences, Merck & Co., Rahway, NJ 07065, USA.

E-mail: anjana.grandhi@merck.com

Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ 07102-1982, USA.

E-mail: wenge.guo@njit.edu

Departments of Statistics and Economics, Stanford University, Stanford, CA 94305-4065, USA.

E-mail: romano@stanford.edu

(Received May 2017; accepted November 2017)