U.C. Berkeley Division of Biostatistics Working Paper Series
University of California, Berkeley

Year 2004 *Paper* 166

Multiple Testing Procedures for Controlling Tail Probability Error Rates

Sandrine Dudoit* Mark J. van der Laan†
Merrill D. Birkner‡

*Division of Biostatistics, School of Public Health, University of California, Berkeley
†Division of Biostatistics, School of Public Health, University of California, Berkeley
‡Division of Biostatistics, School of Public Health, University of California, Berkeley

Multiple Testing Procedures for Controlling Tail Probability Error Rates

Abstract

The present article discusses and compares multiple testing procedures (MTP) for controlling Type I error rates defined as tail probabilities for the number (gFWER) and proportion (TPPFP) of false positives among the rejected hypotheses. Specifically, we consider the gFWER- and TPPFP-controlling MTPs proposed recently by Lehmann & Romano (2004) and in a series of four articles by Dudoit et al. (2004), van der Laan et al. (2004b,a), and Pollard & van der Laan (2004). The former Lehmann & Romano (2004) procedures are marginal, in the sense that they are based solely on the marginal distributions of the test statistics, i.e., on cut-off rules for the corresponding unadjusted p-values. In contrast, the procedures discussed in our previous articles take into account the joint distribution of the test statistics and apply to general data generating distributions, i.e., dependence structures among test statistics. The gFWER-controlling common-cut-off and common-quantile procedures of Dudoit et al. (2004) and Pollard & van der Laan (2004) are based on the distributions of maxima of test statistics and minima of unadjusted p-values, respectively. For a suitably chosen initial FWER-controlling procedure, the gFWER- and TPPFP-controlling augmentation multiple testing procedures (AMTP) of van der Laan et al. (2004a) can also take into account the joint distribution of the test statistics. Given a gFWER-controlling procedure, we also propose AMTPs for controlling tail probability error rates, Pr(g(V_n,R_n) > q), for arbitrary functions g(V_n,R_n) of the numbers of false positives V_n and rejected hypotheses R_n. The different gFWER- and TPPFP-controlling procedures are compared in a simulation study, where the tests concern the components of the mean vector of a multivariate Gaussian data generating distribution. Among notable findings are the substantial power gains achieved by joint procedures compared to marginal procedures.

Contents

1 Introduction **3**
- 1.1 Motivation . . . 3
- 1.2 Outline . . . 4

2 Multiple hypothesis testing framework **4**
- 2.1 Basic notions . . . 4
- 2.2 Test statistics null distribution . . . 10
- 2.3 Rejection regions . . . 11

3 General results on augmentation multiple testing procedures **13**
- 3.1 Augmentation multiple testing procedures for controlling tail probability error rates, $Pr(g(V_n, R_n) > q)$ . . . 14
- 3.2 Adjusted p-values for general augmentation multiple testing procedures . . . 16
- 3.3 Examples: gFWER, TPPFP, and gTPPFP control . . . 18
  - 3.3.1 gFWER-controlling augmentation multiple testing procedure . . . 18
  - 3.3.2 TPPFP-controlling augmentation multiple testing procedure . . . 19
  - 3.3.3 gTPPFP-controlling augmentation multiple testing procedure . . . 19

4 gFWER-controlling multiple testing procedures **20**
- 4.1 Marginal procedures of Lehmann & Romano (2004) . . . 20
- 4.2 Common-cut-off and common-quantile procedures of Dudoit et al. (2004) and Pollard & van der Laan (2004) . . . 21
- 4.3 Augmentation procedures of van der Laan et al. (2004a) . . . 23
- 4.4 Comparison of gFWER-controlling procedures . . . 24

5 TPPFP-controlling multiple testing procedures **26**
- 5.1 Marginal procedures of Lehmann & Romano (2004) . . . 26
- 5.2 Augmentation procedures of van der Laan et al. (2004a) . . . 27

6 Simulation study **29**
- 6.1 Simulation study design . . . 29
  - 6.1.1 Simulation model . . . 29
  - 6.1.2 Multiple testing procedures . . . 32

6.1.3 Type I error rate and power comparisons 33
6.2 Simulation study results 37
6.2.1 Results for gFWER-controlling procedures 37
6.2.2 Results for TPPFP-controlling procedures 39

7 Discussion **41**

1 Introduction

1.1 Motivation

The present article discusses and compares multiple testing procedures (MTP) for controlling Type I error rates defined as tail probabilities for the *number* V_n and *proportion* V_n/R_n of false positives among the rejected hypotheses. Specifically, the *generalized family-wise error rate* (gFWER) is a relaxed version of the family-wise error rate (FWER), that allows $k \geq 0$ Type I errors, that is, $gFWER(k)$ is defined as the chance $gFWER(k) = Pr(V_n > k)$ of committing more than k Type I errors ($k = 0$ for the usual FWER). In contrast, the *tail probability for the proportion of false positives* (TPPFP) among the rejected hypotheses allows a proportion $q \in (0, 1)$ of Type I errors, i.e., $TPPFP(q) = Pr(V_n/R_n > q)$. Error rates based on the proportion of false positives are especially appealing for large-scale testing problems, compared to error rates based on the absolute number of false positives, as they remain stable with an increasing number of tested hypotheses. Since the early article of Benjamini and Hochberg (1995), a number of methods have been developed for controlling the *false discovery rate* (FDR), i.e., the expected proportion $FDR = E[V_n/R_n]$ of Type I errors. However, MTPs proposed thus far for controlling a parameter (either FDR or TPPFP) of the distribution of the proportion of false positives typically rely on various assumptions concerning the joint distribution of the test statistics, such as, independence, specific dependence structure (e.g., positive regression dependence, ergodic dependence), or normality (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001; Genovese and Wasserman, 2003, 2004).

Here, we consider the gFWER- and TPPFP-controlling MTPs proposed recently by Lehmann and Romano (2004) and in a series of four articles by Dudoit et al. (2004), van der Laan et al. (2004b,a), and Pollard and van der Laan (2004). The former Lehmann and Romano (2004) procedures are *marginal*, in the sense that they are based solely on the marginal distributions of the test statistics, i.e., on cut-off rules for individual test statistics or their corresponding unadjusted p-values (e.g., classical Bonferroni procedure). In contrast, the procedures discussed in our previous articles take into account the *joint* distribution of the test statistics and apply to general data generating distributions, i.e., dependence structures among test statistics. The gFWER-controlling common-cut-off and common-quantile procedures of Dudoit et al. (2004) and Pollard and van der Laan (2004) are based on the

distributions of maxima of test statistics and minima of unadjusted p-values, respectively. van der Laan et al. (2004a) show that any single-step or stepwise procedure (asymptotically) controlling the FWER can be straightforwardly *augmented*, by adding suitably chosen null hypotheses to the set of rejected hypotheses, to provide (asymptotic) control of the gFWER or TPPFP. By choosing an appropriate initial FWER-controlling procedure (e.g., single-step or step-down maxT or minP), such gFWER- and TPPFP-controlling procedures can take into account the joint distribution of the test statistics and are therefore expected to be less conservative than marginal procedures. Furthermore, as demonstrated in Dudoit and van der Laan (2004), the augmentation approach to multiple testing is very general and can be extended to a broad class of Type I error rates, defined as tail probabilities, $Pr(g(V_n, R_n) > q)$, for arbitrary functions $g(V_n, R_n)$ of the numbers of false positives V_n and rejected hypotheses R_n. The important practical implication of these results is that one can build on the large pool of available FWER-controlling procedures (e.g., maxT and minP procedures) to immediately and conveniently control a wide variety of error rates, such as the gFWER and TPPFP. While the augmentation multiple testing procedures of van der Laan et al. (2004a) can be applied to *any* FWER-controlling procedure, thereby exploiting the joint distribution of the test statistics, this article also considers for comparison purposes conservative versions of the AMTPs based on the marginal Bonferroni single-step and Holm step-down MTPs.

1.2 Outline

The article is organized as follows. Section 2 introduces our general framework for multiple hypothesis testing. Section 3 provides general results on augmentations of gFWER-controlling procedures for control of tail probability error rates, $Pr(g(V_n, R_n) > q)$, for an arbitrary function $g(V_n, R_n)$ of the numbers of false positives V_n and rejected hypotheses R_n. Sections 4 and 5 describe, respectively, recently proposed gFWER- and TPPFP-controlling multiple testing procedures. Section 6 reports the results of a simulation study comparing various gFWER- and TPPFP-controlling procedures. Finally, Section 7 summarizes our findings and outlines ongoing work.

2 Multiple hypothesis testing framework

2.1 Basic notions

Hypothesis testing is concerned with using observed data to test hypotheses, i.e., make decisions, regarding properties of the unknown data generating distribution. Below, we discuss in turn the main ingredients of a multiple testing problem, namely: data, null and alternative hypotheses, test statistics, multiple testing procedure (MTP), Type I and Type II errors, adjusted p-values, test statistics null distribution, choices of rejection regions. Further detail on each of these components can be found in Dudoit and van der Laan (2004) and Dudoit et al. (2004); specific proposals of MTPs are given in Sections 3 – 5.

Data. Let $X_1, \ldots, X_n$ be a *random sample* of n independent and identically distributed (i.i.d.) random variables, $X \sim P \in \mathcal{M}$, where the *data generating distribution* P is known to be an element of a particular *statistical model* $\mathcal{M}$ (i.e., a set of possibly non-parametric distributions).

Null and alternative hypotheses. In order to cover a broad class of testing problems, define M null hypotheses in terms of a collection of *submodels*, $\mathcal{M}(m) \subseteq \mathcal{M}$, $m = 1, \ldots, M$, for the data generating distribution P. The M *null hypotheses* are defined as $H_0(m) \equiv \mathrm{I}(P \in \mathcal{M}(m))$ and the corresponding *alternative hypotheses* as $H_1(m) \equiv \mathrm{I}(P \notin \mathcal{M}(m))$.

In many testing problems, the submodels concern *parameters*, i.e., functions of the data generating distribution P, $\Psi(P) = \psi = (\psi(m) : m = 1, \ldots, M)$, such as means, differences in means, correlation coefficients, and regression parameters in linear models, generalized linear models, survival models, time-series models, dose-response models, etc. One distinguishes between two types of testing problems: *one-sided tests*, where $H_0(m) = \mathrm{I}(\psi(m) \leq \psi_0(m))$, and *two-sided tests*, where $H_0(m) = \mathrm{I}(\psi(m) = \psi_0(m))$. The user-supplied hypothesized *null values*, $\psi_0(m)$, are frequently zero.

Let $\mathcal{H}_0 = \mathcal{H}_0(P) \equiv \{m : H_0(m) = 1\} = \{m : P \in \mathcal{M}(m)\}$ be the set of $h_0 \equiv |\mathcal{H}_0|$ true null hypotheses, where we note that $\mathcal{H}_0$ depends on the data generating distribution P. Let $\mathcal{H}_1 = \mathcal{H}_1(P) \equiv \mathcal{H}_0^c(P) = \{m : H_1(m) = 1\} = \{m : P \notin \mathcal{M}(m)\}$ be the set of $h_1 \equiv |\mathcal{H}_1| = M - h_0$ false null hypotheses, i.e., true positives. The goal of a multiple testing procedure is to accurately estimate the set $\mathcal{H}_0$, and thus its complement $\mathcal{H}_1$, while controlling proba-

bilistically the number of false positives.

Test statistics. A testing procedure is a *data-driven* rule for deciding whether or not to *reject* each of the M null hypotheses $H_0(m)$, i.e., declare that $H_0(m)$ is false (zero) and hence $P \notin \mathcal{M}(m)$. The decisions to reject or not the null hypotheses are based on an M–vector of *test statistics*, $T_n = (T_n(m) : m = 1, \ldots, M)$, that are functions $T_n(m) = T(X_1, \ldots, X_n)(m)$ of the data, $X_1, \ldots, X_n$. Denote the typically unknown (finite sample) *joint distribution* of the test statistics T_n by $Q_n = Q_n(P)$.

Single-parameter null hypotheses are commonly tested using *t-statistics*, i.e., standardized differences,

$$T_n(m) \equiv \frac{\text{Estimator} - \text{Null value}}{\text{Standard error}} = \sqrt{n}\frac{\psi_n(m) - \psi_0(m)}{\sigma_n(m)}. \tag{1}$$

In general, the M–vector $\psi_n = (\psi_n(m) : m = 1, \ldots, M)$ denotes an asymptotically linear *estimator* of the parameter M–vector $\psi = (\psi(m) : m = 1, \ldots, M)$ and $(\sigma_n(m)/\sqrt{n} : m = 1, \ldots, M)$ denote consistent estimators of the *standard errors* of the components of ψ_n. For tests of means, one recovers the usual one-sample and two-sample t-statistics, where the $\psi_n(m)$ and $\sigma_n(m)$ are based on empirical means and variances, respectively. In some settings, it may be appropriate to use (unstandardized) *difference statistics*, $T_n(m) \equiv \sqrt{n}(\psi_n(m) - \psi_0(m))$ (Pollard and van der Laan, 2004). Test statistics for other types of null hypotheses include F-statistics, χ^2-statistics, and likelihood ratio statistics.

Multiple testing procedure. A *multiple testing procedure* (MTP) provides *rejection regions*, $\mathcal{C}_n(m)$, i.e., sets of values for each test statistic $T_n(m)$ that lead to the decision to reject the null hypothesis $H_0(m)$. In other words, a MTP produces a random (i.e., data-dependent) subset $\mathcal{R}_n$ of rejected hypotheses that estimates $\mathcal{H}_1$, the set of true positives,

$$\mathcal{R}_n = \mathcal{R}(T_n, Q_{0n}, \alpha) \equiv \{m : H_0(m) \text{ is rejected}\} = \{m : T_n(m) \in \mathcal{C}_n(m)\}, \tag{2}$$

where $\mathcal{C}_n(m) = \mathcal{C}(T_n, Q_{0n}, \alpha)(m)$, $m = 1, \ldots, M$, denote possibly random rejection regions. The long notation $\mathcal{R}(T_n, Q_{0n}, \alpha)$ and $\mathcal{C}(T_n, Q_{0n}, \alpha)(m)$ emphasizes that the MTP depends on: (i) the *data*, $X_1, \ldots, X_n$, through the M–vector of *test statistics*, $T_n = (T_n(m) : m = 1, \ldots, M)$; (ii) a (estimated) test statistics *null distribution*, Q_{0n}, for deriving rejection regions for each

$T_n(m)$ and the resulting adjusted p-values (Section 2.2); and (iii) the *nominal level* α of the MTP, i.e., the desired upper bound for a suitably defined Type I error rate.

Unless specified otherwise, it is assumed that large values of the test statistic $T_n(m)$ provide evidence against the corresponding null hypothesis $H_0(m)$, that is, we consider rejection regions of the form $\mathcal{C}_n(m) = (c_n(m), \infty)$, where $c_n(m)$ are to-be-determined *critical values*, or *cut-offs*, computed under the null distribution Q_{0n} for the test statistics T_n (Section 2.3).

Type I and Type II errors. In any testing situation, two types of errors can be committed: a *false positive*, or *Type I error*, is committed by rejecting a true null hypothesis, and a *false negative*, or *Type II error*, is committed when the test procedure fails to reject a false null hypothesis. The situation can be summarized by Table 1, below, where the number of Type I errors is $V_n \equiv |\mathcal{R}_n \cap \mathcal{H}_0| = \sum_{m \in \mathcal{H}_0} \mathrm{I}(T_n(m) \in \mathcal{C}_n(m))$ and the number of Type II errors is $U_n \equiv |\mathcal{R}_n^c \cap \mathcal{H}_1| = \sum_{m \in \mathcal{H}_1} \mathrm{I}(T_n(m) \notin \mathcal{C}_n(m))$. Note that both U_n and V_n depend on the unknown data generating distribution P through the unknown set of true null hypotheses $\mathcal{H}_0 = \mathcal{H}_0(P)$. The numbers $h_0 = |\mathcal{H}_0|$ and $h_1 = |\mathcal{H}_1| = M - h_0$ of true and false null hypotheses are *unknown parameters*, the number of rejected hypotheses $R_n \equiv |\mathcal{R}_n| = \sum_{m=1}^{M} \mathrm{I}(T_n(m) \in \mathcal{C}_n(m))$ is an *observable random variable*, and the entries in the body of the table, U_n, $h_1 - U_n$, V_n, and $h_0 - V_n$, are *unobservable random variables* (depending on P, through $\mathcal{H}_0(P)$).

Ideally, one would like to simultaneously minimize both the chances of committing a Type I error and a Type II error. Unfortunately, this is not feasible and one seeks a *trade-off* between the two types of errors. A standard approach is to specify an acceptable level α for the Type I error rate and derive testing procedures, i.e., rejection regions, that aim to minimize the Type II error rate, i.e., maximize *power*, within the class of tests with Type I error rate at most α.

Type I error rates. When testing multiple hypotheses, there are many possible definitions for the Type I error rate (and power) of a test procedure. Accordingly, we adopt the general framework proposed in Dudoit and van der Laan (2004), Dudoit et al. (2004), and van der Laan et al. (2004b,a), and define Type I error rates as *parameters*, $\theta_n = \theta(F_{V_n,R_n})$, of the joint distribution F_{V_n,R_n} of the numbers of Type I errors V_n and rejected hypotheses R_n. Here, we focus primarily on procedures that control the following two broad

classes of Type I error rates: tail probabilities (i.e., survivor function) for the *number* V_n of false positives and for the *proportion* V_n/R_n of false positives among the rejected hypotheses.

The *generalized family-wise error rate* (gFWER), for a user-supplied integer k, $k = 0, \ldots, (h_0 - 1)$, is the probability of at least $(k+1)$ Type I errors. That is,

$$gFWER(k) \equiv Pr(V_n > k) = 1 - F_{V_n}(k), \tag{3}$$

where F_{V_n} is the discrete cumulative distribution function (c.d.f.) on $\{0, \ldots, M\}$ for the number of Type I errors, V_n. When $k = 0$, the gFWER is the usual *family-wise error rate* (FWER), or probability of at least one Type I error,

$$FWER \equiv Pr(V_n > 0) = 1 - F_{V_n}(0). \tag{4}$$

The FWER is controlled, in particular, by the classical Bonferroni procedure.

The *tail probability for the proportion of false positives* (TPPFP) among the rejected hypotheses, for a user-supplied constant $q \in (0, 1)$, is defined as

$$TPPFP(q) \equiv Pr(V_n/R_n > q) = 1 - F_{V_n/R_n}(q), \tag{5}$$

where F_{V_n/R_n} is the c.d.f. for the proportion V_n/R_n of false positives among the rejected hypotheses, with the convention that $V_n/R_n \equiv 0$ if $R_n = 0$. In the remainder of this article, we use the shorter phrase *proportion of false positives* (PFP) to refer to the proportion V_n/R_n of false positives *among the* R_n *rejected hypotheses* and not among the total number M of null hypotheses. Controlling the later proportion would amount to controlling the gFWER.

Note that while the gFWER is a parameter of only the *marginal* distribution F_{V_n} of the number of Type I errors V_n (tail probability, or survivor function, for V_n), the TPPFP is a parameter of the *joint* distribution of (V_n, R_n) (tail probability, or survivor function, for V_n/R_n). Error rates based on the *proportion* of false positives (e.g., TPPFP and false discovery rate, FDR, $E[V_n/R_n]$) are especially appealing for large-scale testing problems such as those encountered in genomics, compared to error rates based on the *number* of false positives (e.g., gFWER), as they do not increase exponentially with the number of tested hypotheses.

The aforementioned error rates are part of the broad class of Type I error rates considered in Dudoit and van der Laan (2004) and defined as tail probabilities $Pr(g(V_n, R_n) > q)$ and expected values $E[g(V_n, R_n)]$ for an arbitrary function $g(V_n, R_n)$ of the numbers of false positives V_n and rejected

hypotheses R_n. The gFWER and TPPFP correspond to the special cases $g(V_n, R_n) = V_n$ and $g(V_n, R_n) = V_n/R_n$, respectively.

Power. As with Type I error rates, the concept of power can be generalized in various ways when moving from single to multiple hypothesis testing. Three common definitions of power are: (i) the probability of rejecting at least one false null hypothesis, $Pr(|\mathcal{R}_n \cap \mathcal{H}_1| \geq 1) = Pr(h_1 - U_n \geq 1) = Pr(U_n \leq h_1 - 1)$, where we recall that U_n is the number of Type II errors; (ii) the expected proportion of rejected false null hypotheses, $E[|\mathcal{R}_n \cap \mathcal{H}_1|]/|\mathcal{H}_1| = E[h_1 - U_n]/h_1$, or *average power*; and (iii) the probability of rejecting all false null hypotheses, $Pr(|\mathcal{R}_n \cap \mathcal{H}_1| = h_1) = Pr(U_n = 0)$ (Shaffer, 1995). When the family of tests consists of pairwise mean comparisons, these quantities have been called any-pair power, per-pair power, and all-pairs power (Ramsey, 1978). In a spirit analogous to the FDR, one could also define power as $E[(h_1 - U_n)/R_n | R_n > 0] Pr(R_n > 0) = E[(R_n - V_n)/R_n | R_n > 0] Pr(R_n > 0) = Pr(R_n > 0) - FDR$; when $h_1 = M$, this is the any-pair power $Pr(U_n \leq h_1 - 1)$.

Adjusted p-values. The notion of p-value extends directly to multiple testing problems, as follows. Given a MTP $\mathcal{R}_n(\alpha) = \mathcal{R}(T_n, Q_{0n}, \alpha)$, the *adjusted p-value* $\widetilde{P}_{0n}(m) = \widetilde{P}(T_n, Q_{0n})(m)$, for null hypothesis $H_0(m)$, is defined as the smallest Type I error level α at which one would reject $H_0(m)$, that is,

$$\widetilde{P}_{0n}(m) \equiv \inf \{\alpha \in [0,1] : m \in \mathcal{R}_n(\alpha)\}, \qquad m = 1, \ldots, M. \tag{6}$$

Note that *unadjusted* or *marginal p-values*, for the test of a single hypothesis, correspond to the special case $M = 1$. For a continuous null distribution Q_{0n}, the unadjusted p-value for null hypothesis $H_0(m)$ is given by $P_{0n}(m) = P(T_n(m), Q_{0n,m}) = \bar{Q}_{0n,m}(T_n(m))$, where $Q_{0n,m}$ and $\bar{Q}_{0n,m}$ denote, respectively, the marginal c.d.f.'s and survivor functions for Q_{0n}.

As in single hypothesis tests, the smaller the adjusted p-value, the stronger the evidence against the corresponding null hypothesis. If $\mathcal{R}_n(\alpha)$ is right-continuous at α, in the sense that $\lim_{\alpha' \downarrow \alpha} \mathcal{R}_n(\alpha') = \mathcal{R}_n(\alpha)$, then one has two equivalent representations for the MTP, in terms of rejection regions for the test statistics and in terms of adjusted p-values,

$$\mathcal{R}_n(\alpha) = \{m : T_n(m) \in \mathcal{C}_n(m)\} = \{m : \widetilde{P}_{0n}(m) \leq \alpha\}. \tag{7}$$

Let $O_n(m)$ denote indices for the *ordered adjusted p-values*, so that $\widetilde{P}_{0n}(O_n(1)) \leq \ldots \leq \widetilde{P}_{0n}(O_n(M))$. Then, the set of rejected hypotheses $\mathcal{R}_n(\alpha)$ consists of

the indices for the $R_n(\alpha) \equiv |\mathcal{R}_n(\alpha)|$ hypotheses with the smallest adjusted p-values, that is, $\mathcal{R}_n(\alpha) = \{O_n(m) : m = 1, \ldots, R_n(\alpha)\}$. For example, the adjusted p-values for the classical Bonferroni procedure for FWER control are given by $\widetilde{P}_{0n}(m) = \min(MP_{0n}(m), 1)$.

Reporting the results of a MTP in terms of adjusted p-values, as opposed to only rejection or not of the hypotheses, offers several advantages. (i) Adjusted p-values can be defined for *any Type I error rate* (gFWER, TPPFP, FDR, etc.). (ii) They reflect the strength of the evidence against each null hypothesis in terms of the Type I error rate for the *entire MTP*. (iii) They are *flexible summaries* of a MTP, in that results are provided for *all levels* α, i.e., the level α need not be chosen ahead of time. (iv) Finally, as seen below, adjusted p-values provide convenient benchmarks to *compare* different MTPs, whereby smaller adjusted p-values indicate a less conservative procedure.

2.2 Test statistics null distribution

One of the main tasks in specifying a MTP is to derive rejection regions for the test statistics such that the Type I error rate is controlled at a desired level α, i.e., such that $\theta(F_{V_n,R_n}) \leq \alpha$, for *finite sample control*, or $\limsup_n \theta(F_{V_n,R_n}) \leq \alpha$, for *asymptotic control*. It is common practice, especially for FWER control, to set $\alpha = 0.05$. However, one is immediately faced with the problem that the *true distribution* $Q_n = Q_n(P)$ of the test statistics T_n is usually *unknown*, and hence, so are the distributions of the numbers of Type I errors, $V_n = \sum_{m \in \mathcal{H}_0} \mathrm{I}(T_n(m) \in \mathcal{C}_n(m))$, and rejected hypotheses, $R_n = \sum_{m=1}^{M} \mathrm{I}(T_n(m) \in \mathcal{C}_n(m))$. In practice, the test statistics *true distribution* $Q_n(P)$ is replaced by a *null distribution* Q_0 (or estimate thereof, Q_{0n}), in order to derive rejection regions and resulting adjusted p-values.

The choice of null distribution Q_0 is crucial, in order to ensure that (finite sample or asymptotic) control of the Type I error rate under the *assumed* null distribution Q_0 does indeed provide the required control under the *true* distribution $Q_n(P)$. For error rates $\theta(F_{V_n})$, defined as arbitrary parameters of the distribution of the number of Type I errors V_n, we propose as null distribution the asymptotic distribution $Q_0 = Q_0(P)$ of the M–vector Z_n of null value shifted and scaled test statistics (Dudoit and van der Laan, 2004; Dudoit et al., 2004; van der Laan et al., 2004b,a; Pollard and van der Laan,

2004),

$$Z_n(m) \equiv \sqrt{\min\left(1, \frac{\tau_0(m)}{Var[T_n(m)]}\right)}\Big(T_n(m) + \lambda_0(m) - E[T_n(m)]\Big). \qquad (8)$$

For the test of single-parameter null hypotheses using t-statistics, the null values are $\lambda_0(m) = 0$ and $\tau_0(m) = 1$. For testing the equality of K population means using F-statistics, the null values are $\lambda_0(m) = 1$ and $\tau_0(m) = 2/(K-1)$, under the assumption of equal variances in the different populations. Dudoit et al. (2004) and van der Laan et al. (2004b) prove that this null distribution does indeed provide the desired asymptotic control of the Type I error rate $\theta(F_{V_n})$, for general data generating distributions (with arbitrary dependence structures among variables), null hypotheses (defined in terms of submodels for the data generating distribution), and test statistics (e.g., t-statistics, F-statistics). In practice, however, since the data generating distribution P is unknown, then so is the proposed null distribution $Q_0 = Q_0(P)$. Resampling procedures, such as the bootstrap procedures proposed in Dudoit et al. (2004) and van der Laan et al. (2004b), may be used to conveniently obtain consistent estimators Q_{0n} of the null distribution Q_0 and of the corresponding test statistic cut-offs and adjusted p-values.

Note that the following important points distinguish our approach from existing approaches to Type I error rate control. Firstly, we are only concerned with Type I error control under the *true data generating distribution* P. The notions of weak and strong control (and associated subset pivotality, Westfall and Young (1993), p. 42–43) are therefore irrelevant to our approach. Secondly, we propose a *null distribution for the test statistics* ($T_n \sim Q_0$), and not a data generating null distribution ($X \sim P_0 \in \cap_{m=1}^{M}\mathcal{M}(m)$). The latter practice does not necessarily provide proper Type I error control, as the test statistics' *assumed* null distribution $Q_n(P_0)$ and their *true* distribution $Q_n(P)$ may have different dependence structures (in the limit) for the true null hypotheses $\mathcal{H}_0$.

The reader is referred to our earlier articles and a book in preparation for a detailed discussion of the choice of test statistics T_n, null distribution Q_0, and approaches for estimating this null distribution (Dudoit and van der Laan, 2004; Dudoit et al., 2004; van der Laan et al., 2004b,a; Pollard and van der Laan, 2004). Accordingly, we take the test statistics T_n and their null distribution Q_0 (or estimate thereof, Q_{0n}) as given, and denote a multiple testing procedure by $\mathcal{R}_n(\alpha) = \mathcal{R}(T_n, Q_{0n}, \alpha)$, to emphasize only the

dependence on the nominal Type I error level α.

2.3 Rejection regions

Having selected a suitable test statistics null distribution, there remains the main task of specifying rejection regions for each null hypothesis, i.e., cut-offs for each test statistic. Among the different approaches for defining rejection regions, we distinguish between the following.

Common-cut-off vs. common-quantile multiple testing procedures. In *common-cut-off procedures*, the same cut-off c_0 is used for each test statistic (cf. maxT procedures, based on maxima of test statistics, Table 2). In contrast, in *common-quantile procedures*, the cut-offs are the δ_0–quantiles of the marginal null distributions of the test statistics (cf. minP procedures, based on minima of unadjusted p-values, Table 2). The latter procedures tend to be more "balanced" than the former, as the transformation to p-values places the null hypotheses on an equal footing. However, this comes at the expense of increased computational complexity.

Single-step vs. stepwise multiple testing procedures. In *single-step procedures*, each null hypothesis is evaluated using a rejection region that is independent of the results of the tests of other hypotheses. Improvement in power, while preserving Type I error rate control, may be achieved by *stepwise procedures*, in which rejection of a particular null hypothesis depends on the outcome of the tests of other hypotheses. That is, the (single-step) test procedure is applied to a sequence of successively smaller nested random (i.e., data-dependent) subsets of null hypotheses, defined by the ordering of the test statistics (common cut-offs) or unadjusted p-values (common-quantile cut-offs). In *step-down procedures*, the hypotheses corresponding to the *most significant* test statistics (i.e., largest absolute test statistics or smallest unadjusted p-values) are considered successively, with further tests depending on the outcome of earlier ones. As soon as one fails to reject a null hypothesis, no further hypotheses are rejected. In contrast, for *step-up procedures*, the hypotheses corresponding to the *least significant* test statistics are considered successively, again with further tests depending on the outcome of earlier ones. As soon as one hypothesis is rejected, all remaining more significant hypotheses are rejected.

Marginal vs. joint multiple testing procedures. *Marginal multiple testing procedures* are based solely on the marginal distributions of the test statistics, i.e., on cut-off rules for individual test statistics or their corresponding unadjusted p-values (e.g., classical Bonferroni FWER-controlling single-step procedure; Lehmann and Romano (2004) gFWER-controlling Procedures 2 and 3). In contrast, *joint multiple testing procedures* take into account the dependence structure of the test statistics (e.g., gFWER-controlling single-step common-cut-off and common-quantile Procedures 4 and 5, based on maxima of test statistics and minima of unadjusted p-values, respectively). Note that while a procedure may be marginal, proof of Type I error control by this MTP may require certain assumptions on the dependence structure of the test statistics (e.g., Hochberg (1988) FWER-controlling step-up MTP; Lehmann and Romano (2004) restricted TPPFP-controlling step-down Procedure 7, below).

Our previous articles and book in preparation (Dudoit and van der Laan, 2004) discuss three main approaches for deriving rejection regions and corresponding adjusted p-values: single-step common-cut-off and common-quantile procedures for control of general Type I error rates $\theta(F_{V_n})$ (Dudoit et al., 2004; Pollard and van der Laan, 2004); step-down common-cut-off (maxT) and common-quantile (minP) procedures for control of the FWER (van der Laan et al., 2004b); augmentation procedures for control of the gFWER and TPPFP, based on an initial FWER-controlling procedure (van der Laan et al., 2004a).

General results on augmentation procedures controlling tail probability error rates $Pr(g(V_n, R_n) > q)$, for an arbitrary function $g(V_n, R_n)$ of the numbers of false positives V_n and rejected hypotheses R_n, are given next, in Section 3. The following two sections summarize different gFWER- and TPPFP-controlling MTPs in terms of their adjusted p-values. Type I error control is typically established for a given null distribution Q_0 (or estimate thereof, Q_{0n}). As discussed above, control of the corresponding Type I error rate under the true, unknown test statistics distribution $Q_n(P)$ follows only for a suitable choice of this null distribution Q_0. For details, proofs, and other available gFWER- and TPPFP-controlling procedures, the reader is referred to Dudoit and van der Laan (2004), Dudoit et al. (2004), van der Laan et al. (2004b,a), and Lehmann and Romano (2004).

3 General results on augmentation multiple testing procedures

Dudoit and van der Laan (2004) and van der Laan et al. (2004a) propose *augmentation multiple testing procedures* (AMTP), obtained by adding suitably chosen null hypotheses to the set of hypotheses already rejected by an initial MTP. Specifically, given any initial procedure controlling the generalized family-wise error rate (gFWER), augmentation procedures are derived for controlling Type I error rates defined as tail probabilities and expected values for arbitrary functions of the numbers of Type I errors and rejected hypotheses. Adjusted p-values for the AMTP are shown to be simply shifted versions of the adjusted p-values of the original MTP. The important practical implication of these results is that *any* FWER-controlling MTP and its corresponding adjusted p-values, provide, without additional work, multiple testing procedures controlling a broad class of Type I error rates and their adjusted p-values. One can therefore build on the large pool of available FWER-controlling procedures, such as the single-step and step-down maxT and minP procedures discussed in Dudoit et al. (2004) and van der Laan et al. (2004b).

The present section summarizes some of the general results of Dudoit and van der Laan (2004) on AMTPs for controlling target Type I error rates θ^+ of the form

$$\theta^+(F_{V_n,R_n}; q) = Pr(g(V_n, R_n) > q), \tag{9}$$

that is, tail probabilities for an arbitrary non-negative function $g(V_n, R_n)$ of the numbers of Type I errors V_n and rejected hypotheses R_n. Error rates covered by this representation include the gFWER, where $g(v, r) = v$, the TPPFP, where $g(v, r) = v/r$, and the generalized TPPFP (gTPPFP), where $g(v, r) = v/r\,\mathrm{I}(k_0/r \leq q)$, for some integer $k_0 \geq 0$. Given a user-supplied non-negative integer r_0, one may also consider error rates based on the function $g(v, r) = v/r\,\mathrm{I}(r > r_0)$. Controlling tail probabilities $Pr(V_n/R_n\,\mathrm{I}(R_n > r_0) > q)$ amounts to considering multiple testing procedures for which the proportion of false positives (PFP), V_n/R_n, does not exceed q when more than r_0 hypotheses are rejected (i.e., when $R_n > r_0$).

Section 3.1 proposes augmentation Procedure 1, for (asymptotic) control of tail probability error rates $\theta^+(F_{V_n,R_n}; q) = Pr(g(V_n, R_n) > q)$, based on an initial gFWER-controlling procedure. In Section 3.2, the adjusted p-values for a general AMTP are shown to be simply shifted versions of the adjusted

p-values of the initial MTP. Three examples of AMTPs are given in Section 3.3, for control of the gFWER, TPPFP, and gTPPFP, respectively. The reader is referred to Dudoit and van der Laan (2004) and van der Laan et al. (2004a) for an in-depth treatment of the augmentation approach to multiple testing and, in particular, for: proofs of finite sample and exact asymptotic Type I error control by Procedure 1, (conservative) procedures controlling the expected value of the function $g(V_n, R_n)$, and detailed studies of gFWER and TPPFP-controlling AMTPs.

3.1 Augmentation multiple testing procedures for controlling tail probability error rates, $Pr(g(V_n, R_n) > q)$

Consider a multiple testing procedure $\mathcal{R}_n(k_0; \alpha)$ (or in short $\mathcal{R}_n$) that controls $gFWER(k_0)$ at level α, i.e., such that $Pr(V_n(k_0; \alpha) > k_0) \leq \alpha$. Assume that the function g and the MTP $\mathcal{R}_n(k_0; \alpha)$ satisfy the following three properties:

(i) $v \rightarrow g(v, r)$ is non-decreasing;

(ii) $k \rightarrow g(k + v, k + r)$ is non-decreasing;

(iii)

$$Pr(g(k_0, R_n(k_0; \alpha)) \leq q) = 1. \tag{10}$$

The first monotonicity assumption is used to prove Type I error control by an AMTP such as Procedure 1, below. The second assumption guarantees that the cardinality of the augmentation set in Equation (15), for Procedure 1, increases with the allowed proportion q of false positives. The third assumption in Equation (10) ensures that the initial gFWER-controlling procedure also controls the target error rate θ^+ at level α, that is, $Pr(g(V_n, R_n) > q) \leq \alpha$. Thus, one can always obtain a θ^+–controlling AMTP, even in the worst case scenario of an empty augmentation set.

For instance, for control of the proportion of false positives V_n/R_n, one can define

$$g(v, r) \equiv v/r \, \mathrm{I}(k_0/r \leq q). \tag{11}$$

We shall refer to the Type I error rate

$$gTPPFP(k_0, q) \equiv Pr(V_n/R_n \, \mathrm{I}(k_0/R_n \leq q) > q), \tag{12}$$

defined by the function g in Equation (11), as *generalized tail probability for the proportion of false positives* (gTPPFP).

Assumption (iii) in Equation (10) allows us to derive an augmentation procedure that controls $gTPPFP(k_0, q)$, even when the PFP exceeds q for the initial $gFWER(k_0)$–controlling procedure, i.e., even when $k_0/R_n > q$. As detailed in Procedure 1, below, one simply keeps rejecting null hypotheses until $g(k_0 + m, R_n + m)$ exceeds q, i.e., the following two conditions are both met: $(k_0 + m)/(R_n + m) > q$ and $k_0/(R_n + m) \leq q$, where m denotes the number of additional rejections. Note that for their TPPFP-controlling AMTP of Theorem 3.3, Genovese and Wasserman (2004) do not enforce control of the TPPFP by the initial $gFWER(k_0)$–controlling procedure.

Procedure 1 provides a specific construction for an augmentation procedure $\mathcal{R}_n^+(k_0, q; \alpha)$, controlling the tail probability error rate $\theta^+(F_{V_n,R_n}; q) = Pr(g(V_n, R_n) > q)$, based on an initial $gFWER(k_0)$–controlling procedure $\mathcal{R}_n(k_0; \alpha)$.

Procedure 1 [Augmentation procedure for controlling the tail probability error rate $Pr(g(V_n, R_n) > q)$ based on a gFWER-controlling procedure]
Consider a multiple testing procedure $\mathcal{R}_n(k_0; \alpha)$ that provides finite sample control of $gFWER(k_0)$ at level α_n and asymptotic control of $gFWER(k_0)$ at level α.

1. *First, order the M null hypotheses according to their gFWER adjusted p-values, $\widetilde{P}_{0n}(m)$, from smallest to largest, that is, define indices $O_n(m)$, so that $\widetilde{P}_{0n}(O_n(1)) \leq \ldots \leq \widetilde{P}_{0n}(O_n(M))$. The gFWER-controlling procedure rejects the $R_n(k_0; \alpha) \equiv |\mathcal{R}_n(k_0; \alpha)|$ null hypotheses*

$$\mathcal{R}_n(k_0; \alpha) \equiv \{m : \widetilde{P}_{0n}(m) \leq \alpha\} = \{O_n(m) : m = 1, \ldots, R_n(k_0; \alpha)\}. \tag{13}$$

2. *For a given q, define an augmentation procedure $\mathcal{R}_n^+(k_0, q; \alpha)$, controlling the tail probability error rate $\theta^+(F_{V_n,R_n}; q) = Pr(g(V_n, R_n) > q)$, by*

$$\mathcal{R}_n^+(k_0, q; \alpha) \equiv \mathcal{R}_n(k_0; \alpha) \cup \mathcal{A}_n(k_0, q; \alpha), \tag{14}$$

where $\mathcal{A}_n(k_0, q; \alpha)$ is an augmentation set of cardinality

$$A_n(k_0, q; \alpha) \equiv \max\{m \in \{0, \ldots, M - R_n(k_0; \alpha)\} : g(k_0 + m, R_n(k_0; \alpha) + m) \leq q\}, \tag{15}$$

defined by

$$\mathcal{A}_n(k_0, q; \alpha) \equiv \{O_n(m) : m = R_n(k_0; \alpha)+1, \ldots, R_n(k_0; \alpha)+A_n(k_0, q; \alpha)\}. \tag{16}$$

That is, the set $\mathcal{A}_n(k_0, q; \alpha)$ corresponds to the $A_n(k_0, q; \alpha)$ most significant null hypotheses that were not rejected *by the gFWER-controlling procedure $\mathcal{R}_n(k_0; \alpha)$.*

3.2 Adjusted p-values for general augmentation multiple testing procedures

This section presents general results concerning the adjusted p-values of an augmentation multiple testing procedure. We stress the level of generality of these results: they apply to any procedure $\mathcal{R}_n$ controlling an *arbitrary initial Type I error rate* $\theta(F_{V_n,R_n})$ (i.e., not only the gFWER) and to augmentation procedures $\mathcal{R}_n^+$ controlling an *arbitrary target Type I error rate* $\theta^+(F_{V_n^+,R_n^+})$ (i.e., not only the gFWER or TPPFP).

Let $O_n(m)$ denote indices for the ordered adjusted p-values, $\widetilde{P}_{0n}(O_n(m))$, of the initial θ–controlling procedure $\mathcal{R}_n(\alpha)$, such that $\widetilde{P}_{0n}(O_n(1)) \leq \ldots \leq \widetilde{P}_{0n}(O_n(M))$. As above, focus on augmentation procedures that reject null hypotheses in order of their *increasing* θ–specific adjusted p-values, i.e., starting with the null hypothesis $H_0(O_n(1))$ with the smallest θ–specific adjusted p-value. If the level α is set at the mth ordered θ–specific adjusted p-value, i.e., $\alpha = \widetilde{P}_{0n}(O_n(m))$, the initial θ–controlling procedure rejects the following $R_n(\widetilde{P}_{0n}(O_n(m))) = m$ null hypotheses

$$\mathcal{R}_n(\widetilde{P}_{0n}(O_n(m))) = \{h : \widetilde{P}_{0n}(h) \leq \widetilde{P}_{0n}(O_n(m))\} = \{O_n(h) : h = 1, \ldots, m\},$$

and the augmentation θ^+–controlling procedure rejects the following $R_n^+(\widetilde{P}_{0n}(O_n(m))) = m + A_n(\widetilde{P}_{0n}(O_n(m)))$ null hypotheses

$$\begin{aligned} \mathcal{R}_n^+(\widetilde{P}_{0n}(O_n(m))) &= \mathcal{R}_n(\widetilde{P}_{0n}(O_n(m))) \cup \mathcal{A}_n(\widetilde{P}_{0n}(O_n(m))) \\ &= \{O_n(h) : h = 1, \ldots, m + A_n(\widetilde{P}_{0n}(O_n(m)))\}. \end{aligned}$$

Hence, as stated in Proposition 1, below, the mth ordered θ–specific adjusted p-value $\widetilde{P}_{0n}(O_n(m))$ is the $(m + A_n(\widetilde{P}_{0n}(O_n(m))))$th ordered θ^+–specific adjusted p-value.

Proposition 1 [Adjusted p-values for a general augmentation multiple testing procedure] *Consider a (single-step or stepwise) multiple testing procedure $\mathcal{R}_n(\alpha)$ that provides (asymptotic) control of a Type I error rate $\theta(F_{V_n,R_n})$ at level α. Let $O_n(m)$ denote indices for the ordered adjusted p-values, $\widetilde{P}_{0n}(O_n(m))$, of this initial θ–controlling MTP, so that $\widetilde{P}_{0n}(O_n(1)) \leq \ldots \leq \widetilde{P}_{0n}(O_n(M))$. Then, the adjusted p-values, $\widetilde{P}^+_{0n}(O_n(m))$, for the θ^+–controlling augmentation procedure $\mathcal{R}^+_n(\alpha) = \mathcal{R}_n(\alpha) \cup \mathcal{A}_n(\alpha)$ satisfy*

$$\widetilde{P}_{0n}(O_n(m)) = \widetilde{P}^+_{0n}(O_n(S(m))), \tag{17}$$

where $S : \{1, \ldots, M\} \to \{1, \ldots, M\}$ is an integer shift function *defined by*

$$S(m) \equiv R^+_n(\widetilde{P}_{0n}(O_n(m))) = m + A_n(\widetilde{P}_{0n}(O_n(m))) \tag{18}$$

and $A_n(\widetilde{P}_{0n}(O_n(m)))$ denotes the cardinality of the augmentation set for a level $\alpha = \widetilde{P}_{0n}(O_n(m))$ test. Thus, the adjusted p-values $\widetilde{P}^+_{0n}(O_n(m))$ for the θ^+–controlling AMTP are simply a shifted *version of the adjusted p-values $\widetilde{P}_{0n}(O_n(m))$ for the initial θ–controlling MTP, according to the shift function S.*

Note that getting a closed form expression for the θ^+–specific p-values $\widetilde{P}^+_{0n}(O_n(m))$ may or may not be straightforward, depending on the complexity of the function $A_n(\alpha)$ for the cardinality of the augmentation set, i.e., on how easily one can invert the shift function $S : m \to m + A_n(\widetilde{P}_{0n}(O_n(m)))$. In general, one cannot simply shift the θ–specific adjusted p-values by the size $A_n(\widetilde{P}_{0n}(O_n(m)))$ of the augmentation set, as this size is also a function of m. Furthermore, the shift function S is not necessarily one-to-one or onto, as illustrated in Section 3.3, below.

Instead, we rely on the general definition of an adjusted p-value, whereby

$$\begin{aligned} \widetilde{P}^+_{0n}(O_n(m)) &\equiv \inf\{\alpha \in [0,1] : O_n(m) \in \mathcal{R}^+_n(\alpha)\} \\ &= \inf\{\alpha \in [0,1] : R^+_n(\alpha) \geq m\} \\ &= \inf\{\alpha \in [0,1] : R_n(\alpha) + A_n(\alpha) \geq m\}. \end{aligned} \tag{19}$$

3.3 Examples: gFWER, TPPFP, and gTPPFP control

3.3.1 gFWER-controlling augmentation multiple testing procedure

Consider a gFWER-controlling AMTP, based on an initial FWER-controlling MTP. The shift function is

$$S(m) = \min\{m + k, M\}, \tag{20}$$

and the adjusted p-values are given by

$$\widetilde{P}^{+}_{0n}(O_n(m)) = \begin{cases} 0, & \text{if } m \leq k \\ \widetilde{P}_{0n}(O_n(m-k)), & \text{if } m > k \end{cases}. \tag{21}$$

Intuitively, the larger the allowed number k of false positives, the larger the augmentation set and the smaller the adjusted p-values for the AMTP. The shift function $S(m)$ is piecewise linear and is neither one-to-one nor onto. The number of rejections and adjusted p-values are simply shifted by the allowed number k of false positives, provided this does not lead to more than M rejected hypotheses. gFWER-controlling AMTPs are discussed further in Section 4.3, below.

3.3.2 TPPFP-controlling augmentation multiple testing procedure

Consider a TPPFP-controlling AMTP, based on an initial FWER-controlling MTP. The shift function is

$$S(m) = \min\left\{ \left\lfloor \frac{m}{1-q} \right\rfloor, M \right\}, \tag{22}$$

where the *floor* $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x, i.e., $\lfloor x \rfloor \leq x < \lfloor x \rfloor + 1$. The adjusted p-values are given by

$$\widetilde{P}^{+}_{0n}(O_n(m)) = \widetilde{P}_{0n}(O_n(\lceil (1-q)m \rceil)), \tag{23}$$

where the *ceiling* $\lceil x \rceil$ denotes the least integer greater than or equal to x, i.e., $\lceil x \rceil - 1 < x \leq \lceil x \rceil$. Intuitively, the larger the allowed proportion q of false positives, the larger the augmentation set and the smaller the adjusted p-values for the AMTP. As in the gFWER case, the shift function is neither

one-to-one nor onto. However, as evidenced in the above equations, TPPFP-controlling AMTPs are more complex than gFWER-controlling AMTPs. As a result of the floor and ceiling functions, the p-value shifts are step functions. In addition, while the adjusted p-values for the gFWER-controlling procedure are shifted by a constant k, the shift mq for the adjusted p-values of the TPPFP-controlling AMTP increases with m, as the hypotheses become less significant. TPPFP-controlling AMTPs are discussed further in Section 5.2, below.

3.3.3 gTPPFP-controlling augmentation multiple testing procedure

Consider a gTPPFP-controlling AMTP, based on an initial gFWER-controlling MTP. The shift function is

$$S(m) = \begin{cases} \min\left\{M, \left\lceil \frac{k_0}{q} - 1 \right\rceil\right\}, & \text{if } m < k_0 + (1-q)\left\lceil \frac{k_0}{q} - 1 \right\rceil \\ \min\left\{M, \left\lfloor \frac{m-k_0}{1-q} \right\rfloor\right\}, & \text{if } m \geq k_0 + (1-q)\left\lceil \frac{k_0}{q} - 1 \right\rceil \end{cases}, \tag{24}$$

and the adjusted p-values are given by

$$\widetilde{P}^{+}_{0n}(O_n(m)) = \begin{cases} 0, & \text{if } m < \lceil \frac{k_0}{q} \rceil \\ \widetilde{P}_{0n}(O_n(\lceil (1-q)m + k_0 \rceil)), & \text{if } m \geq \lceil \frac{k_0}{q} \rceil \end{cases}. \tag{25}$$

The adjusted p-values for the gTPPFP-controlling AMTP are hybrids of the adjusted p-values for the gFWER- and TPPFP-controlling AMTPs in Equations (21) and (23), respectively. In particular, as a result of starting from an initial gFWER-controlling MTP, that allows k_0 false positives, the first $\lceil \frac{k_0}{q} \rceil - 1$ adjusted p-values are set to zero, i.e., one automatically rejects $\lceil \frac{k_0}{q} \rceil - 1$ null hypotheses (compared to k for a $gFWER(k)$–controlling AMTP). The remaining adjusted p-values are similar in form to those of the TPPFP-controlling AMTP in Equation (23), but with a shift of $qm - k_0$ instead of qm. Note that for $k_0 = 0$, i.e., for an initial FWER-controlling procedure, one recovers the TPPFP shift function and adjusted p-values given in Equations (22) and (23), above.

4 gFWER-controlling multiple testing procedures

We consider three broad classes of gFWER-controlling procedures: the marginal single-step and step-down procedures of Lehmann and Romano (2004), the joint single-step common-cut-off $T(k+1)$ and common-quantile $P(k+1)$ procedures of Dudoit et al. (2004) and Pollard and van der Laan (2004), and the general augmentation procedures of van der Laan et al. (2004a). Unlike the Lehmann and Romano (2004) *marginal* MTPs, the latter two types of MTPs take into account the *joint* distribution of the test statistics. Adjusted p-values for gFWER-controlling procedures are listed in Table 3.

4.1 Marginal procedures of Lehmann & Romano (2004)

In their Theorems 2.1 and 2.2, Lehmann and Romano (2004) propose both single-step and step-down marginal gFWER-controlling procedures. Procedures 2 and 3, below, summarize these MTPs and provide their corresponding adjusted p-values.

Procedure 2 [Lehmann and Romano (2004) Bonferroni-like gFWER-controlling single-step procedure] *For controlling the $gFWER(k)$ at level α, the* Lehmann and Romano (2004) single-step procedure *rejects any hypothesis $H_0(m)$ with unadjusted p-value $P_{0n}(m)$ less than or equal to the common single-step cut-off $a_m(\alpha) \equiv (k+1)\alpha/M$. That is, the set of rejected hypotheses is*

$$\mathcal{R}_n(\alpha) \equiv \left\{ m : P_{0n}(m) \leq \frac{(k+1)}{M} \alpha \right\}. \tag{26}$$

The corresponding adjusted p-values are thus given by

$$\widetilde{P}_{0n}(m) = \min \left(\frac{M}{(k+1)} P_{0n}(m), 1 \right), \qquad m = 1, \ldots, M. \tag{27}$$

Note that in the special case of FWER control ($k = 0$), Procedure 2 reduces to the Bonferroni single-step procedure, i.e., the p-value cut-offs are $a_m(\alpha) \equiv \alpha/M$.

Procedure 3 [Lehmann and Romano (2004) Holm-like gFWER-controlling step-down procedure] *Let $P_{0n}(m)$ denote the unadjusted p-value for null hypothesis $H_0(m)$ and let $O_n(m)$ denote indices for the ordered unadjusted p-values, so that $P_{0n}(O_n(1)) \leq \ldots \leq P_{0n}(O_n(M))$. For controlling $gFWER(k)$ at level α, the unadjusted p-value cut-offs for the* Lehmann and Romano (2004) step-down procedure *are as follows*

$$a_m(\alpha) \equiv \begin{cases} \frac{(k+1)}{M}\,\alpha, & \text{if } m \leq (k+1) \\ \frac{(k+1)}{(M+k+1-m)}\,\alpha, & \text{if } m > (k+1) \end{cases}, \qquad m = 1, \ldots, M, \tag{28}$$

and the set of rejected hypotheses is

$$\mathcal{R}_n(\alpha) \equiv \{O_n(m) : P_{0n}(O_n(h)) \leq a_h(\alpha),\ \forall\ h \leq m\}\,. \tag{29}$$

The corresponding adjusted p-values are thus given by

$$\widetilde{P}_{0n}(O_n(m)) = \begin{cases} \min\left(\frac{M}{(k+1)} P_{0n}(O_n(m)),\, 1\right), & \text{if } m \leq (k+1) \\ \max_{h=1,\ldots,m-k}\ \left\{\min\left(\frac{(M-h+1)}{(k+1)} P_{0n}(O_n(h+k)),\, 1\right)\right\}, & \text{if } m > (k+1) \end{cases}. \tag{30}$$

Note that in the special case of FWER control ($k = 0$), Procedure 3 reduces to the Holm step-down procedure, i.e., the p-value cut-offs are $a_m(\alpha) = \alpha/(M - m + 1)$.

4.2 Common-cut-off and common-quantile procedures of Dudoit et al. (2004) and Pollard & van der Laan (2004)

Dudoit et al. (2004) and Pollard and van der Laan (2004) propose single-step common-cut-off and common-quantile procedures for controlling arbitrary parameters $\theta(F_{V_n})$ of the distribution of the number of Type I errors. The main idea is to substitute control of the parameter $\theta(F_{V_n})$, for the *unknown, true distribution* F_{V_n} of the number of Type I errors, by control of the corresponding parameter $\theta(F_{R_0})$, for the *known, null distribution* F_{R_0} of the number of rejected hypotheses. That is, consider single-step procedures of the form $\mathcal{R}_n \equiv \{m : T_n(m) > c_n(m)\}$, where the cut-offs $c_n(m)$ are chosen so that $\theta(F_{R_0}) \leq \alpha$, for $R_0 \equiv \sum_{m=1}^{M} \mathrm{I}(Z(m) > c_n(m))$ and $Z \sim Q_0$. Among

the class of MTPs that satisfy $\theta(F_{R_0}) \leq \alpha$, Dudoit et al. (2004) propose two procedures, based on common cut-offs and common-quantile cut-offs, respectively (Procedures 2 and 1 of this earlier article). The procedures are summarized below and the reader is referred to the articles for proofs and details on the derivation of cut-offs and adjusted p-values.

Procedure 4 [Dudoit et al. (2004) and Pollard and van der Laan (2004) gFWER-controlling single-step common-cut-off $T(k+1)$ procedure] *The set of rejected hypotheses for the general* θ–controlling single-step common-cut-off procedure *is of the form* $\mathcal{R}_n(\alpha) \equiv \{m : T_n(m) > c_0\}$, *where the common cut-off* c_0 *is the* smallest *(i.e., least conservative) value for which* $\theta(F_{R_0}) \leq \alpha$.

For $gFWER(k)$ *control (special case* $\theta(F_{V_n}) = 1 - F_{V_n}(k)$*), the procedure is based on the* $(k+1)$st ordered test statistic. *The adjusted p-values for the* single-step $T(k+1)$ procedure *are given by*

$$\widetilde{p}_{0n}(m) = Pr_{Q_0}\left(Z^\circ(k+1) \geq t_n(m)\right), \qquad m = 1, \ldots, M, \tag{31}$$

where $Z^\circ(m)$ *denotes the* m*th ordered component of* $Z = (Z(m) : m = 1, \ldots, M) \sim Q_0$, *so that* $Z^\circ(1) \geq \ldots \geq Z^\circ(M)$.

For FWER control ($k = 0$*), one recovers the* single-step maxT procedure, *based on the* maximum test statistic, $Z^\circ(1)$.

Procedure 5 [Dudoit et al. (2004) and Pollard and van der Laan (2004) gFWER-controlling single-step common-quantile $P(k+1)$ procedure] *The set of rejected hypotheses for the general* θ–controlling single-step common-quantile procedure *is of the form* $\mathcal{R}_n(\alpha) \equiv \{m : T_n(m) > c_0(m)\}$, *where* $c_0(m) = Q_{0,m}^{-1}(\delta_0)$ *is the* δ_0*–quantile of the marginal null distribution* $Q_{0,m}$ *of the* m*th test statistic, i.e., the smallest value* c *such that* $Q_{0,m}(c) = Pr_{Q_0}(Z(m) \leq c) \geq \delta_0$ *for* $Z \sim Q_0$. *Here,* δ_0 *is chosen as the* smallest *(i.e., least conservative) value for which* $\theta(F_{R_0}) \leq \alpha$.

For $gFWER(k)$ *control, the procedure is based on the* $(k+1)$st ordered unadjusted p-value. *Specifically, let* $\bar{Q}_{0,m} \equiv 1 - Q_{0,m}$ *denote the survivor functions for the marginal null distributions* $Q_{0,m}$ *and define unadjusted p-values* $P_0(m) \equiv \bar{Q}_{0,m}(Z(m))$ *and* $P_{0n}(m) \equiv \bar{Q}_{0,m}(T_n(m))$, *for* $Z \sim Q_0$ *and* $T_n \sim Q_n$, *respectively. Then, the adjusted p-values for the* single-step $P(k+1)$ procedure *are given by*

$$\widetilde{p}_{0n}(m) = Pr_{Q_0}\left(P_0^\circ(k+1) \leq p_{0n}(m)\right), \qquad m = 1, \ldots, M, \tag{32}$$

where $P_0^\circ(m)$ denotes the mth ordered component of the M–vector of unadjusted p-values $P_0 = (P_0(m) : m = 1, \ldots, M)$, so that $P_0^\circ(1) \leq \ldots \leq P_0^\circ(M)$.

For FWER control ($k = 0$), one recovers the single-step minP procedure*, based on the* minimum unadjusted *p*-value*, $P_0^\circ(1)$.*

4.3 Augmentation procedures of van der Laan et al. (2004a)

Procedure 6 [van der Laan et al. (2004a) gFWER-controlling augmentation procedure] *Denote the adjusted p-values for an initial FWER-controlling procedure by $\widetilde{P}_{0n}(m)$. Order the M null hypotheses according to these p-values, from smallest to largest, that is, define indices $O_n(m)$, so that $\widetilde{P}_{0n}(O_n(1)) \leq \ldots \leq \widetilde{P}_{0n}(O_n(M))$. Then, for a nominal level α test, the initial FWER-controlling procedure rejects the $R_n(\alpha)$ null hypotheses $\mathcal{R}_n(\alpha) \equiv \{m : \widetilde{P}_{0n}(m) \leq \alpha\}$. For control of $gFWER(k)$ at level α, reject the $R_n(\alpha)$ hypotheses specified by this MTP, as well as the next $A_n(\alpha)$ most significant null hypotheses,*

$$A_n(\alpha) = \min\{k, M - R_n(\alpha)\}. \tag{33}$$

The adjusted p-values $\widetilde{P}_{0n}^+(O_n(m))$ for the new gFWER-controlling AMTP are simply k–shifted versions of the adjusted p-values of the initial FWER-controlling MTP, with the first k adjusted p-values set to zero. That is,

$$\widetilde{P}_{0n}^+(O_n(m)) = \begin{cases} 0, & \text{if } m \leq k \\ \widetilde{P}_{0n}(O_n(m-k)), & \text{if } m > k \end{cases}. \tag{34}$$

The AMTP thus guarantees at least k rejected hypotheses.

For example, the adjusted p-values for a (conservative) version of the augmentation procedure, based on the marginal Bonferroni single-step procedure, are given by

$$\widetilde{P}_{0n}^+(O_n(m)) = \begin{cases} 0, & \text{if } m \leq k \\ \min\left(M\, P_{0n}(O_n(m-k)), 1\right), & \text{if } m > k \end{cases}, \tag{35}$$

where $P_{0n}(m)$ denotes the unadjusted p-value for null hypothesis $H_0(m)$. Likewise, the adjusted p-values for a (conservative) version of the augmentation procedure, based on the marginal Holm step-down procedure, are given

by

$$\widetilde{P}_{0n}^{+}(O_n(m)) = \begin{cases} 0, & \text{if } m \leq k \\ \max_{h=1,\ldots,m-k} \left\{ \min\left((M-h+1)\, P_{0n}(O_n(h)), 1 \right) \right\}, & \text{if } m > k \end{cases}. \tag{36}$$

One can also consider less conservative gFWER-controlling augmentation procedures based on the single-step and step-down common-cut-off (maxT) and common-quantile (minP) procedures (Table 2). Unlike the *marginal* Bonferroni and Holm AMTPs, such procedures take into account the *joint* distribution of the test statistics.

4.4 Comparison of gFWER-controlling procedures

Proposition 2 *The adjusted p-values $\widetilde{P}_{0n}^{P(k+1)}(m)$ for the single-step common-quantile $P(k+1)$ Procedure 5 are uniformly smaller than the adjusted p-values $\widetilde{P}_{0n}^{LR}(m)$ for the Lehmann and Romano (2004) Bonferroni-like single-step Procedure 2. Hence, the former procedure is more powerful than the latter. As a corollary for FWER control ($k = 0$), the single-step minP procedure is more powerful than the Bonferroni single-step procedure.*

Proof of Proposition 2. Applying Markov's Inequality to the adjusted p-values $\widetilde{p}_{0n}^{P(k+1)}(m)$ yields the following conservative upper bounds, which are precisely the adjusted p-values $\widetilde{p}_{0n}^{LR}(m)$ for the Lehmann and Romano (2004) single-step Procedure 2. Specifically,

$$\begin{aligned}
\widetilde{p}_{0n}^{P(k+1)}(m) &= Pr_{Q_0}\left(P_0^{\circ}(k+1) \leq p_{0n}(m)\right) \\
&= Pr_{Q_0}\left(\sum_{h=1}^{M} \mathrm{I}(P_0(h) \leq p_{0n}(m)) \geq (k+1)\right) \\
&\leq \min\left(\frac{1}{(k+1)} E_{Q_0}\left[\sum_{h=1}^{M} \mathrm{I}(P_0(h) \leq p_{0n}(m))\right],\ 1\right) \\
&= \min\left(\frac{1}{(k+1)} \sum_{h=1}^{M} Pr_{Q_0}(P_0(h) \leq p_{0n}(m)),\ 1\right) \\
&\leq \min\left(\frac{M}{(k+1)} p_{0n}(m),\ 1\right) = \widetilde{p}_{0n}^{LR}(m), \qquad m = 1, \ldots, M.
\end{aligned}$$

That is, $\widetilde{P}_{0n}^{P(k+1)}(m) \leq \widetilde{P}_{0n}^{LR}(m)$, $m = 1, \ldots, M$. □

Hence, the single-step common-quantile $P(k+1)$ procedure is more powerful than the Bonferroni-like single-step procedure of Lehmann and Romano (2004), in the sense that it always leads to a larger number of rejected hypotheses. This example illustrates that *joint* procedures, based on the joint distribution of the test statistics, can lead to gains in power over *marginal* MTPs. Direct comparisons with common-cut-off $T(k+1)$ Procedure 4 are not as straightforward, although, in practice, we also expect gains in power from exploiting the joint distribution of the test statistics.

Analytical comparisons between other pairs of gFWER-controlling procedures are not fully conclusive, in the sense that the adjusted p-values of one procedure cannot be shown to be uniformly smaller than those of the other procedure (Dudoit and van der Laan, 2004). We expect, however, to increase power by taking into account the joint distribution of the test statistics, as in common-cut-off and common-quantile Procedures 4 and 5 and augmentations of joint FWER-controlling MTPs. In addition, the automatic rejection of the k null hypotheses corresponding to the k most significant test statistics may confer an advantage to a gFWER-controlling AMTP, even when the procedure is based on a marginal FWER-controlling MTP. The results of a simulation study comparing different gFWER-controlling MTPs are reported in Section 6.2.1.

5 TPPFP-controlling multiple testing procedures

We consider two broad classes of TPPFP-controlling procedures: the marginal restricted and general step-down procedures of Lehmann and Romano (2004) and the general augmentation procedures of van der Laan et al. (2004a). Unlike the Lehmann and Romano (2004) *marginal* MTPs, the latter augmentation procedures can take into account the *joint* distribution of the test statistics. Adjusted p-values for TPPFP-controlling procedures are listed in Table 4.

5.1 Marginal procedures of Lehmann & Romano (2004)

Lehmann and Romano (2004) propose the following two marginal step-down procedures for controlling the TPPFP. The first procedure is shown to control the TPPFP under either one of two assumptions on the dependence

structure of the unadjusted p-values (Theorems 3.1 and 3.2 in Lehmann and Romano (2004)), while the second and more conservative procedure controls the TPPFP under arbitrary dependence structures (Theorem 3.3).

Let $P_{0n}(m)$ denote the unadjusted p-value for null hypothesis $H_0(m)$ and let $O_n(m)$ denote indices for the ordered unadjusted p-values, so that $P_{0n}(O_n(1)) \leq \ldots \leq P_{0n}(O_n(M))$.

Procedure 7 [Lehmann and Romano (2004) restricted TPPFP-controlling step-down procedure] *For controlling $TPPFP(q)$ at level α, the unadjusted p-value cut-offs for the* Lehmann and Romano (2004) restricted step-down procedure *are as follows*

$$a_m(\alpha) \equiv \frac{(\lfloor qm \rfloor + 1)}{(M + \lfloor qm \rfloor + 1 - m)} \alpha, \qquad m = 1, \ldots, M, \tag{37}$$

where we recall that the floor $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x. The set of rejected hypotheses is

$$\mathcal{R}_n(\alpha) \equiv \{O_n(m) : P_{0n}(O_n(h)) \leq a_h(\alpha), \ \forall \ h \leq m\}. \tag{38}$$

The corresponding adjusted p-values are thus given by

$$\widetilde{P}_{0n}(O_n(m)) = \max_{h=1,\ldots,m} \left\{ \min \left(\frac{(M + \lfloor qh \rfloor + 1 - h)}{(\lfloor qh \rfloor + 1)} P_{0n}(O_n(h)), \, 1 \right) \right\}. \tag{39}$$

Procedure 8 [Lehmann and Romano (2004) general TPPFP-controlling step-down procedure] *For controlling $TPPFP(q)$ at level α, the unadjusted p-value cut-offs for the* Lehmann and Romano (2004) general step-down procedure *are as follows*

$$a_m(\alpha) \equiv \frac{1}{C(\lfloor qM \rfloor + 1)} \frac{(\lfloor qm \rfloor + 1)}{(M + \lfloor qm \rfloor + 1 - m)} \alpha, \qquad m = 1, \ldots, M, \tag{40}$$

where $C(M) \equiv \sum_{m=1}^{M} 1/m$. The set of rejected hypotheses is

$$\mathcal{R}_n(\alpha) \equiv \{O_n(m) : P_{0n}(O_n(h)) \leq a_h(\alpha), \ \forall \ h \leq m\}. \tag{41}$$

The corresponding adjusted p-values are thus given by

$$\widetilde{P}_{0n}(O_n(m)) = \max_{h=1,\ldots,m} \left\{ \min \left(C(\lfloor qM \rfloor + 1) \frac{(M + \lfloor qh \rfloor + 1 - h)}{(\lfloor qh \rfloor + 1)} P_{0n}(O_n(h)), \, 1 \right) \right\}. \tag{42}$$

The conservative cut-offs in Procedure 8 are simply obtained by dividing the p-value cut-offs of Procedure 7 by $C(\lfloor qM \rfloor + 1) \approx \log(qM)$ for large M. It is interesting to note the parallels between the above two TPPFP-controlling step-down procedures and the FDR-controlling step-up procedures of Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001). The adjustment to achieve Type I error control for general dependence structures tends to be more conservative for the FDR-controlling procedures than it is for the TPPFP-controlling procedures, i.e., usually $C(M) \geq C(\lfloor qM \rfloor + 1)$. Note that in the trivial $q = 0$ case, both Procedures 7 and 8 yield, as is intuitive, the Holm step-down procedure for FWER control.

5.2 Augmentation procedures of van der Laan et al. (2004a)

Procedure 9 [van der Laan et al. (2004a) TPPFP-controlling augmentation procedure] *Denote the adjusted p-values for an initial FWER-controlling procedure by $\widetilde{P}_{0n}(m)$. Order the M null hypotheses according to these p-values, from smallest to largest, that is, define indices $O_n(m)$, so that $\widetilde{P}_{0n}(O_n(1)) \leq \ldots \leq \widetilde{P}_{0n}(O_n(M))$. Then, for a nominal level α test, the initial FWER-controlling procedure rejects the $R_n(\alpha)$ null hypotheses $\mathcal{R}_n(\alpha) \equiv \{m : \widetilde{P}_{0n}(m) \leq \alpha\}$. For control of $TPPFP(q)$ at level α, reject the $R_n(\alpha)$ hypotheses specified by this MTP, as well as the next $A_n(\alpha)$ most significant null hypotheses,*

$$\begin{aligned} A_n(\alpha) &= \max\left\{ m \in \{0, \ldots, M - R_n(\alpha)\} : \frac{m}{m + R_n(\alpha)} \leq q \right\} \\ &= \min\left\{ \left\lfloor \frac{qR_n(\alpha)}{1-q} \right\rfloor, M - R_n(\alpha) \right\}. \end{aligned} \tag{43}$$

That is, keep rejecting null hypotheses until the ratio of additional rejections to the total number of rejections reaches the allowed proportion q of false positives. The adjusted p-values $\widetilde{P}^{+}_{0n}(O_n(m))$ for the new TPPFP-controlling AMTP are simply mq–shifted versions of the adjusted p-values of the initial FWER-controlling MTP. That is,

$$\widetilde{P}^{+}_{0n}(O_n(m)) = \widetilde{P}_{0n}(O_n(\lceil (1-q)m \rceil)), \qquad m = 1, \ldots, M. \tag{44}$$

For example, the adjusted p-values for a (conservative) version of the

augmentation procedure, based on the marginal Bonferroni single-step procedure, are given by

$$\widetilde{P}_{0n}^{+}(O_n(m)) = \min\left(M\, P_{0n}(O_n(\lceil (1-q)m \rceil)), 1\right), \tag{45}$$

where $P_{0n}(m)$ denotes the unadjusted p-value for null hypothesis $H_0(m)$. Likewise, the adjusted p-values for a (conservative) version of the augmentation procedure, based on the marginal Holm step-down procedure, are given by

$$\widetilde{P}_{0n}^{+}(O_n(m)) = \max_{h=1,\ldots,\lceil (1-q)m \rceil} \left\{\min\left((M-h+1)\, P_{0n}(O_n(h)), 1\right)\right\}. \tag{46}$$

One can also consider less conservative TPPFP-controlling augmentation procedures, that take into account the joint distribution of the test statistics, and are based, for example, on the single-step and step-down common-cut-off (maxT) and common-quantile (minP) procedures (Table 2).

As for gFWER control, analytical comparisons between the van der Laan et al. (2004a) TPPFP-controlling AMTPs and the Lehmann and Romano (2004) marginal MTPs are not fully conclusive, in the sense that the adjusted p-values of one procedure cannot be shown to be uniformly smaller than those of the other procedure (Dudoit and van der Laan, 2004). We expect, however, to achieve gains in power from taking into account the joint distribution of the test statistics. The results of a simulation study comparing different TPPFP-controlling MTPs are reported in Section 6.2.2.

6 Simulation study

6.1 Simulation study design

The simulation study focuses primarily on *marginal multiple testing procedures*, that is, procedures based solely on the marginal distributions of the test statistics $T_n = (T_n(m) : m = 1, \ldots, M)$. The adjusted p-values for such procedures are simple functions of the unadjusted p-values $P_{0n}(m)$ corresponding to each null hypothesis $H_0(m)$. *Joint multiple testing procedures*, obtained by augmenting the set of rejected hypotheses for the FWER-controlling single-step maxT procedure, are also investigated. The various gFWER- and TPPFP-controlling MTPs compared in the simulation study

are listed in Tables 5 and 6, respectively; adjusted p-values are provided in Tables 2 – 4.

6.1.1 Simulation model

Data. Consider random samples $X_1, \ldots, X_n$, from an M–dimensional Gaussian data generating distribution P. That is, the data are n i.i.d. random variables, $X_i \sim N(\psi, \sigma)$, $i = 1, \ldots, n$, where $\psi = (\psi(m) : m = 1, \ldots, M) = \Psi(P) = E[X]$ and $\sigma = (\sigma(m, m') : m, m' = 1, \ldots, M) = \Sigma(P) = Cov[X]$ denote, respectively, the M–dimensional mean vector and $M \times M$ covariance matrix of $X \sim P$. We may adopt the shorter notation $\sigma^2(m) \equiv \sigma(m, m)$ for variances.

Null and alternative hypotheses. The null hypotheses of interest concern the M components of the mean vector ψ. Specifically, we are interested in two-sided tests of the M null hypotheses $H_0(m) = \mathrm{I}\big(\psi(m) = \psi_0(m)\big)$ vs. the alternative hypotheses $H_1(m) = \mathrm{I}\big(\psi(m) \neq \psi_0(m)\big)$, $m = 1, \ldots, M$. For simplicity, the null values are set to zero, i.e., $\psi_0(m) \equiv 0$.

Test statistics. In the known variance case, one can test the null hypotheses using simple *one-sample z-statistics*,

$$T_n(m) \equiv \sqrt{n}\frac{\psi_n(m) - \psi_0(m)}{\sigma(m)}, \tag{47}$$

where $\psi_n(m) = \sum_i X_i(m)/n$ denote the empirical means for the M components of X. For our simple model, the test statistics $T_n(m)$ can be rewritten as

$$T_n(m) = \sqrt{n}\frac{\psi_n(m) - \psi(m)}{\sigma(m)} + \sqrt{n}\frac{\psi(m) - \psi_0(m)}{\sigma(m)} = Z_n(m) + d_n(m),$$

in terms of random variables

$$Z_n = \left(Z_n(m) \equiv \sqrt{n}\frac{\psi_n(m) - \psi(m)}{\sigma(m)} : m = 1, \ldots, M \right) \sim N(0, \sigma^*) \tag{48}$$

and *shift* parameters

$$d_n = \left(d_n(m) \equiv \sqrt{n}\frac{\psi(m) - \psi_0(m)}{\sigma(m)} : m = 1, \ldots, M \right), \tag{49}$$

where $\sigma^* = \Sigma^*(P) = Cor[X]$ is the correlation matrix of $X \sim P$. Thus, the test statistics T_n have an M–variate Gaussian distribution with mean vector the shift vector d_n and covariance matrix σ^*: $T_n \sim N(d_n, \sigma^*)$.

Under the null hypothesis $H_0(m)$, the shift $d_n(m) = 0$. Note that a given value of the shift $d_n(m)$ corresponds to different combinations of sample size n, mean $\psi(m)$, and variance $\sigma^2(m)$. The larger the absolute value of the difference $(\psi(m) - \psi_0(m))$, the larger the absolute value of the shift $d_n(m)$, and hence the larger the power. Also, the smaller the variance $\sigma^2(m)$ and the larger the sample size n, the larger the shift.

For these simple choices of data generating model and one-sample z-statistics, the test statistics null distribution Q_0 is simply the M–variate Gaussian distribution $N(0, \sigma^*)$, with mean vector zero and covariance matrix σ^* (recall that Q_0 is defined as the asymptotic distribution of the vector of null value shifted and scaled test statistics in Equation (8)).

Unadjusted p-values. For the above Gaussian model, two-sided unadjusted p-values $P_{0n}(m)$ can be obtained straightforwardly from the standard normal c.d.f. Φ, as

$$P_{0n}(m) \equiv 2(1 - \Phi(T_n(m))). \tag{50}$$

Simulation parameters. For our simple choices of data generating model and one-sample z-statistics, one can simulate the test statistics T_n directly (and more efficiently from a computational point of view) from the M–variate Gaussian distribution $T_n \sim N(d_n, \sigma^*)$, where the parameter of interest is now the shift vector d_n, with mth component equal to zero under the corresponding null hypothesis.

The following model parameters are varied in the simulation study.

- *Number of null hypotheses, M*: The following values are considered for the total number of null hypotheses, $M = 24$, 400.

- *Proportion of true null hypotheses, h_0/M*: Complete null hypothesis ($h_0/M = 1$), 50% of true null hypotheses ($h_0/M = 0.50$), and 75% of true null hypotheses ($h_0/M = 0.75$).

- *Shift parameter vector, d_n*: For the true null hypotheses, i.e., for $m \in \mathcal{H}_0$, $d_n(m) = 0$. For the false null hypotheses, i.e., for $m \in \mathcal{H}_1$, the following (common) shift values d_1 are considered: $d_n(m : m \in \mathcal{H}_1) = \texttt{rep}(\texttt{d}_1, \texttt{times} = \texttt{h}_1)$, $d_1 = 2$, 3, 4, 5. Shift parameter vectors

with two equally represented values are also considered: $d_n(m : m \in \mathcal{H}_1) = \texttt{c}(\texttt{rep}(0.50, \texttt{times} = \texttt{h}_1/2), \texttt{rep}(2, \texttt{times} = \texttt{h}_1/2))$ and $d_n(m : m \in \mathcal{H}_1) = \texttt{c}(\texttt{rep}(1, \texttt{times} = \texttt{h}_1/2), \texttt{rep}(3, \texttt{times} = \texttt{h}_1/2))$.

- *Correlation matrix,* σ^*: The following correlation structures are considered.
  - *Uncorrelation*, where $\sigma^* = I_M$, the $M \times M$ identity matrix.
  - *Local correlation*, where the only non-zero elements of σ^* are the diagonal and first off-diagonal elements: $\sigma^*(m, m) = 1$, for $m = 1, \ldots, M$; $\sigma^*(m, m-1) = \sigma^*(m-1, m) = 0.25, 0.50$, for $m = 2, \ldots, M$; and $\sigma^*(m, m') = 0$, for $m, m' = 1, \ldots, M$ and $m' \neq m-1, m, m+1$.
  - *Full correlation*, where all off-diagonal elements of σ^* are set to a common value: $\sigma^*(m, m) = 1$, for $m = 1, \ldots, M$; and $\sigma^*(m, m') = 0.50, 0.85$, for $m \neq m' = 1, \ldots, M$.
  - *Microarray correlation*, where σ^* corresponds to a random $M \times M$ submatrix of the genes $\times$ genes correlation matrix for the Golub et al. (1999) ALL/AML dataset. The following three pre-processing steps were applied to the $7,129 \times 38$ genes $\times$ patients matrix of expression measures corresponding to the training set of 38 patients (object `golubTrain` in the Bioconductor R package golubEsets; `www.bioconductor.org`): (i) *thresholding*, floor of 100 and ceiling of 16,000; (ii) *filtering*, exclusion of genes with $\max / \min \leq 5$ or $(\max - \min) \leq 500$, where max and min refer, respectively, to the maximum and minimum intensities for a particular gene across the 38 mRNA samples; (iii) *base-2 logarithmic transformation*. These pre-processing steps resulted in a $3,051 \times 38$ genes $\times$ patients matrix of expression measures, from which one can compute a $3,051 \times 3,051$ gene correlation matrix and extract a random $M \times M$ submatrix σ^*.

6.1.2 Multiple testing procedures

Control of the generalized family-wise error rate. We compare the Type I error and power properties of the four *marginal* gFWER-controlling multiple testing procedures listed in Table 5: the Lehmann and Romano

(2004) Bonferroni-like and Holm-like MTPs and augmentations of the Bonferroni and Holm MTPs. We also consider a *joint* AMTP based on the FWER-controlling single-step maxT procedure. The following values are considered (when applicable, given the value of M) for the allowed number k of false positives: $k = 1$, 5, 10, 20, 25, 50, 100.

Control of the tail probability for the proportion of false positives. We compare the Type I error and power properties of the four *marginal* TPPFP-controlling multiple testing procedures listed in Table 6: the Lehmann and Romano (2004) restricted and general MTPs and augmentations of the Bonferroni and Holm MTPs. We also consider a *joint* AMTP based on the FWER-controlling single-step maxT procedure. The following values are considered for the allowed proportion q of false positives: $q = 0.01$, 0.05, 0.10, 0.20, 0.50, 0.75.

Note that in our simple example, where each test statistic $T_n(m)$ has a standard normal marginal null distribution $Q_{0,m}$, common-cut-off (maxT) and common-quantile (minP) procedures are equivalent. We therefore only consider augmentations of the maxT procedure.

Adjusted p-values. For *marginal* MTPs, adjusted p-values $\widetilde{P}_{0n}(m)$ are simple functions of the unadjusted p-values $P_{0n}(m)$ in Equation (50) and can be computed as in Tables 2 – 4.

In contrast, the adjusted p-values for (augmentations of) the maxT (and minP) procedure are based on the *joint* distribution of the test statistics. Specifically, the adjusted p-values for the single-step maxT procedure are obtained by setting $k = 0$ in Equation (31), that is,

$$\widetilde{p}_{0n}(m) = Pr_{Q_0}\left(\max_{m\in\{1,\ldots,M\}} Z(m) \geq t_n(m)\right), \tag{51}$$

where $Z \sim Q_0$ and for our simple simulation model the joint null distribution Q_0 is the M–variate Gaussian distribution $N(0, \sigma^*)$, with mean vector zero and covariance matrix σ^*.

The maxT tail probabilities in Equation (51) may be estimated by simulation as follows. Generate $B = 5,000$ random vectors, $Z^1, \ldots, Z^B$, from the M–variate Gaussian distribution $Q_0 = N(0, \sigma^*)$. The distribution of $\max_m Z(m)$ can then be estimated by the empirical distribution of the B maxima, $\max_m Z^b(m)$, $b = 1, \ldots, B$.

6.1.3 Type I error rate and power comparisons

Estimating Type I error rates and power. For each model (i.e., for each choice of the parameters M, h_0/M, d_n, and σ^*), generate $B = 1,000$ M–vectors of test statistics, $T_n^b \sim N(d_n, \sigma^*)$, $b = 1, \ldots, B$. For each such simulated dataset, compute unadjusted p-values $P_{0n}^b(m) = 2(1 - \Phi(T_n^b(m)))$ and adjusted p-values $\widetilde{P}_{0n}^b(m)$ for each of the MTPs in Tables 5 and 6. For a given nominal Type I error level $\alpha \in \{0, 0.01, 0.02, \ldots, 0.50\}$, i.e., values of α in `seq(from = 0, to = 0.50, by=0.01)`, compute the numbers of rejected hypotheses $R_n^b(\alpha)$, Type I errors $V_n^b(\alpha)$, and Type II errors $U_n^b(\alpha)$,

$$R_n^b(\alpha) \equiv \sum_{m=1}^{M} \mathrm{I}(\widetilde{P}_{0n}^b(m) \leq \alpha), \tag{52}$$

$$V_n^b(\alpha) \equiv \sum_{m \in \mathcal{H}_0} \mathrm{I}(\widetilde{P}_{0n}^b(m) \leq \alpha), \tag{53}$$

$$U_n^b(\alpha) \equiv \sum_{m \in \mathcal{H}_1} \mathrm{I}(\widetilde{P}_{0n}^b(m) > \alpha). \tag{54}$$

The *actual Type I error rates* $\theta(F_{V_n,R_n})$, for control of the gFWER and TPPFP, can be estimated as follows and then compared to the *nominal Type I error rate* α,

$$gFWER(k;\alpha) = \frac{1}{B}\sum_{b=1}^{B} \mathrm{I}(V_n^b(\alpha) > k), \tag{55}$$

$$TPPFP(q;\alpha) = \frac{1}{B}\sum_{b=1}^{B} \mathrm{I}(V_n^b(\alpha)/R_n^b(\alpha) > q), \tag{56}$$

with the convention that $V_n^b(\alpha)/R_n^b(\alpha) \equiv 0$ if $R_n^b(\alpha) = 0$. The *average power* of a given MTP can be estimated by

$$AvgPwr(\alpha) = 1 - \frac{1}{h_1}\frac{1}{B}\sum_{b=1}^{B} U_n^b(\alpha). \tag{57}$$

The simulation error for the actual Type I error rates and average power is of the order $1/\sqrt{B} = 1/\sqrt{1000} \approx 0.032$.

Graphical summaries. We consider the following two main types of graphical summaries of the simulation results.

- **Type I error control comparison.** For a given data generating model, plot for each MTP the *difference between the nominal and actual Type I error rates* vs. the *nominal Type I error rate*, i.e., plot

$$\begin{aligned}(\alpha - gFWER(k;\alpha)) \quad &\text{vs.} \quad \alpha \\ (\alpha - TPPFP(q;\alpha)) \quad &\text{vs.} \quad \alpha,\end{aligned}$$

  for $\alpha \in \{0, 0.01, 0.02, \ldots, 0.50\}$. Such plots may be used to compare different MTPs in terms of their Type I error control properties. Positive (negative) differences correspond to (anti-)conservative MTPs; the higher the curve, the more conservative the procedure.

- **Power comparison.** For a given data generating model, *receiver operator characteristic* (ROC) curves, comparing different MTPs in terms of power, can be obtained by plotting for each MTP *power* vs. *actual Type I error rate*, i.e., $AvgPwr(\alpha)$ vs. $gFWER(k;\alpha)$ or $TPPFP(q;\alpha)$, for a range of nominal Type I error rates $\alpha \in \{0, 0.01, 0.02, \ldots, 0.50\}$. The higher the curve, the more powerful the procedure. However, due to possibly large variations in power between simulation models, we consider instead the following modified display which facilitates comparisons across models. For a given model, plot for each MTP the *difference in power* relative to the `Aug. maxT` procedure vs. the *actual Type I error rate*, i.e., plot

$$\begin{aligned}(AvgPwr^{Aug.maxT}(\alpha) - AvgPwr(\alpha)) \quad &\text{vs.} \quad gFWER(k;\alpha) \\ (AvgPwr^{Aug.maxT}(\alpha) - AvgPwr(\alpha)) \quad &\text{vs.} \quad TPPFP(q;\alpha),\end{aligned}$$

  for $\alpha \in \{0, 0.01, 0.02, \ldots, 0.50\}$. The lower the curve, the more powerful the procedure.

For a subset of the simulation models described in Section 6.1.1, Figures 1 – 6, 14 display power and Type I error rates for the five gFWER-controlling MTPs of Table 5. Likewise, Figures 7 – 13 display power and Type I error rates for the five TPPFP-controlling MTPs of Table 6. The three van der Laan et al. (2004a) augmentation procedures and the two Lehmann and Romano (2004) marginal procedures are displayed using red and blue lines, respectively. The following line types are used to distinguish between different classes of MTPs: solid line type for single-step procedures; dashed line type for step-down procedures; dotted line type for joint procedures (i.e.,

augmentations of the joint single-step maxT MTP); dotdashed line type for the Lehmann and Romano (2004) restricted TPPFP-controlling Procedure 7. Friedman (1984)'s "Super Smoother", implemented in the R function `supsmu`, was applied to produce smoothed plots of power and Type I error rate differences (default argument values were used for `supsmu`, `stats` package, R Version 1.9.1).

Comparisons of interest. The following four types of comparisons are of interest. For each of these comparisons, we report in Sections 6.2.1 and 6.2.2, below, combinations of parameter values (i.e., choices of model parameters, M, h_0/M, d_n, and σ^*, and Type I error rate parameters, k or q) corresponding to the greatest differences between MTPs in terms of power and Type I error rate control. The results are summarized in Table 7.

1. **Step-down vs. single-step procedures (`SD` vs. `SS`).** Any step-down procedure should yield a larger set of rejected hypotheses, i.e., should be more powerful, than its single-step counterpart. Directly comparable procedures in our study are: `Aug. Holm` vs. `Aug. Bonf.`, `LR SD` vs. `LR SS` (only for gFWER control). `SD` vs. `SS` comparisons involve contrasting dashed vs. solid lines of the same color, in Figures 1 – 6, 14 for gFWER control and Figures 7 – 13 for TPPFP control.

2. **Joint vs. marginal procedures (`Joint` vs. `Marginal`).** Procedures that take into account the joint distribution of the test statistics should be more powerful than their marginal counterparts. Directly comparable procedures in our study are: `Aug. maxT` vs. `Aug. Bonf.` (recall that for our choices of data generating model and test statistics, maxT and minP procedures are equivalent, and, from Proposition 2, the single-step minP procedure is more powerful than the Bonferroni single-step procedure). `Joint` vs. `Marginal` comparisons involve contrasting red dotted vs. red solid lines, in Figures 1 – 6, 14 for gFWER control and Figures 7 – 13 for TPPFP control.

3. **Augmentations of marginal MTPs vs. Lehmann & Romano marginal MTPs (`AMTP` vs. `LR`).** For gFWER control, it is of interest to compare the marginal Bonferroni-like single-step procedure of Lehmann and Romano (2004), `LR SS`, to the corresponding Bonferroni augmentation procedure of van der Laan et al. (2004a), `Aug. Bonf.`. Likewise, it is of interest to compare the Holm-like step-down procedure

of Lehmann and Romano (2004), `LR SD`, to the corresponding Holm augmentation procedure of van der Laan et al. (2004a), `Aug. Holm`. For TPPFP control, it is of interest to compare the marginal step-down restricted and general procedures of Lehmann and Romano (2004), `LR Restricted` and `LR General`, to the Holm augmentation procedure of van der Laan et al. (2004a), `Aug. Holm`. Note, however, that any comparison to `LR Restricted` should be interpreted with caution, as Type I error control under this MTP requires a number of assumptions on the dependence structure of the test statistics. `AMTP` vs. `LR` comparisons involve contrasting red vs. blue lines of the same type (i.e., solid or dashed), in Figures 1 – 6, 14 for gFWER control and Figures 7 – 13 for TPPFP control.

4. **Lehmann & Romano restricted vs. general TPPFP-controlling MTPs (`LR Restricted` vs. `LR General`).** For TPPFP control, it is of interest to compare the Lehmann and Romano (2004) restricted and general MTPs: How much more conservative is general Procedure 8? For which models, if any, is restricted Procedure 7 anti-conservative? `LR Restricted` vs. `LR General` comparisons involve contrasting blue dotdashed and blue dashed lines in Figures 7 – 13.

Finally, we also examine whether similar patterns are obtained for the two Type I error rates, gFWER and TPPFP.

6.2 Simulation study results

Figures 1 – 14 provide plots of Type I error rate and power for a range of data generating models. None of the procedures display anti-conservative behavior, i.e., the *actual* Type I error rate ($gFWER(k;\alpha)$ or $TPPFP(q;\alpha)$) is always lower than the *nominal* Type I error rate α. However, as detailed below, some MTPs are much more conservative than others. As expected, joint MTPs are more powerful than marginal MTPs, i.e., for virtually all simulation models and both error rates, the most powerful procedure is the augmentation of the single-step maxT procedure (`Aug. maxT`). Also as expected, step-down procedures are more powerful than their single-step analogs, although differences in power vary across models. The relative merits of other procedures depend on the data generating model.

In general, and as expected, the (absolute) power of each MTP increases with increasing values of the shift parameter d_n and increasing values of the Type I error rate parameters k and q (as either k or q increases, Type I error control is relaxed and more false positives are allowed). Some increases in power are also observed for decreasing values of the total number M of null hypotheses and increasing values of the proportion h_0/M of true null hypotheses. The correlation structure σ^* does not appear to affect much absolute power.

Table 7 summarizes the simulation study results in terms of how variations in model parameters (M, h_0/M, d_n, and σ^*) and Type I error rate parameters (k or q) affect the relative merits (i.e., power differences) of the MTPs for the four types of comparisons listed above. Although references to relevant figures are provided for each parameter value (i.e., each row), the conclusions reported in the table are based in some cases on figures that are not shown in this article.

6.2.1 Results for gFWER-controlling procedures

1. **Step-down vs. single-step procedures (`SD` vs. `SS`).** Step-down procedures are more powerful than their single-step analogs. As illustrated in Figures 1 and 2, the largest gains in power for step-down procedures are obtained for a small number M of null hypotheses and a small proportion h_0/M of true null hypotheses. Such a behavior is intuitive if one considers, for example, the adjusted p-values for the Lehmann and Romano (2004) Bonferroni-like single-step Procedure 2 and Holm-like step-down Procedure 3, in Equations (27) and (30), respectively (in the special case $k = 0$, one recovers the FWER-controlling Bonferroni and Holm MTPs, respectively). The greatest differences between the single-step and step-down p-values occur indeed for small values of the total number M of hypotheses and large values of the "current" hypothesis m. In a step-down procedure, the mth most significant null hypothesis $H_0(O_n(m))$ will only be considered for testing provided the $(m-1)$ more significant null hypotheses were already rejected. As the chance of reaching $H_0(O_n(m))$ should increase with the proportion of true positives, h_1/M, one expects gains in power from step-down procedures for small values of h_0/M.

2. **Joint vs. marginal procedures (`Joint` vs. `Marginal`).** For virtu-

ally all data generating models, the most powerful procedure is the augmentation of the joint single-step maxT procedure (`Aug. maxT`). This can be seen from the positive power differences ($AvgPwr^{Aug.maxT}(\alpha) - AvgPwr(\alpha)$) in Figures 1 – 3, 5 – 6, and 14. In addition, the joint `Aug. maxT` procedure is the least conservative of the five MTPs, in the sense that it exhibits the smallest differences ($\alpha - gFWER(k;\alpha)$) between nominal and actual Type I error rates (Figure 4 and other figures not shown). Figure 4 also shows that the Type I error rate difference ($\alpha - gFWER(k;\alpha)$) decreases with increasing values of the common shift parameter d_1. From Figure 5, the largest gains in power for the joint MTP are obtained, as expected, for highly-correlated variables. One can also observe some gains in power for higher values of the number M of null hypotheses (Figures not shown).

3. **Augmentations of marginal MTPs vs. Lehmann & Romano marginal MTPs (`AMTP` vs. `LR`).** The `AMTP` approach tends to be more powerful than the `LR` approach for a broad range of models. However, the magnitude of the power differences varies with the data generating model. The largest gains in power for augmentation procedures tend to occur for a small number M of null hypotheses (Figure 1) and a large proportion h_0/M of true null hypotheses (Figure 2). As illustrated in Figure 6, gains in power for AMTPs are obtained for a large allowed number k of false positives. One possible reason for this behavior is the automatic rejection of the k most significant null hypotheses by the AMTP. For a large number of hypotheses ($M = 400$) and for a large nominal level α, the automatic rejection of $k \ll M$ hypotheses cannot always compensate for the lack of power of the initial marginal (Bonferroni or Holm) FWER-controlling MTP. However, AMTPs generally retain their power advantages for the small nominal levels one would use in practice (e.g., $\alpha < 0.10$) and for large alternative shift parameter values (see Figure 14, which plots power differences vs. both *actual* and *nominal* gFWER, for shift parameter vectors with extreme values of 10). We also note that power differences between `AMTP` and `LR` MTPs tend to decrease with increasing values of the actual Type I error rate $gFWER(k;\alpha)$.

From Figure 3, differences in power between MTPs tend to decrease with increasing values of the common alternative shift parameter d_1. Also, from

Figure 4, the larger the shift parameter d_1, the more conservative `SS` vs. `SD`, `Marginal` vs. `Joint`, and `LR` vs. `AMTP`.

6.2.2 Results for TPPFP-controlling procedures

The simulation study revealed similar trends for gFWER- and TPPFP-controlling MTPs.

1. **Step-down vs. single-step procedures (`SD` vs. `SS`).** Step-down procedures are generally more powerful than their single-step analogs. As for gFWER control, Figures 7 and 8 indicate power gains for step-down procedures for a small number M of null hypotheses and a small proportion h_0/M of true null hypotheses. Figures 12 – 13 also suggest increased power differences for large values of the allowed proportion q of false positives.

2. **Joint vs. marginal procedures (`Joint` vs. `Marginal`).** Again, for virtually all data generating models, the most powerful procedure is the augmentation of the joint single-step maxT procedure (`Aug. maxT`). This can be seen from the positive power differences ($AvgPwr^{Aug.maxT}(\alpha) - AvgPwr(\alpha)$) in Figures 7 – 9 and 11 – 13. In addition, the joint `Aug. maxT` procedure is the least conservative of the five MTPs, in the sense that it exhibits the smallest differences ($\alpha - TPPFP(q;\alpha)$) between nominal and actual Type I error rates (Figure 10 and other figures not shown). Figure 10 also shows that the Type I error rate difference ($\alpha - TPPFP(q;\alpha)$) tends to decrease with increasing values of the common shift parameter d_1, although not as strikingly as for gFWER-controlling MTPs. As illustrated in Figure 11, the largest gains in power for the joint MTP are obtained, as expected, for highly-correlated variables. One can also observe some gains in power for higher values of the number M of null hypotheses (Figure 7) and lower values of the proportion h_0/M of true null hypotheses (Figures not shown).

3. **Augmentations of marginal MTPs vs. Lehmann & Romano marginal MTPs (`AMTP` vs. `LR`).** As for gFWER control, the `AMTP` approach tends to be more powerful than the `LR` approach for a broad range of models. However, the magnitude of the power differences varies with the data generating model. Augmentation procedures tend

to be more powerful for a small number M of null hypotheses (Figure 7) and a large proportion h_0/M of true null hypotheses (Figures not shown). Figures 12 – 13 also suggest increased power differences for large values of the allowed proportion q of false positives. For a large number of hypotheses ($M = 400$) and for a large nominal level α, AMTPs cannot always compensate for the lack of power of the initial marginal (Bonferroni or Holm) FWER-controlling MTP. However, AMTPs generally retain their power advantages for the small nominal levels one would use in practice (e.g., $\alpha < 0.10$) and for large alternative shift parameter values. We also note that power differences between `AMTP` and `LR` MTPs tend to increase with increasing values of the actual Type I error rate $TPPFP(q;\alpha)$ (the reverse behavior as for gFWER-controlling MTPs).

4. **Lehmann & Romano restricted vs. general TPPFP-controlling MTPs (`LR Restricted` vs. `LR General`).** As expected, the Lehmann and Romano (2004) general TPPFP-controlling Procedure 8 (`LR General`) can be much more conservative than the corresponding restricted Procedure 7 (`LR Restricted`), due to the $C(\lfloor qM \rfloor + 1) \approx \log(qM)$ penalty on the adjusted p-values. This behavior is reminiscent of the differences between the FDR-controlling step-up procedures of Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001). Larger values of the allowed proportion q of false positives tend to be associated with increased differences in power between `LR Restricted` and `LR General` (Figures not show). There are also some indications that larger values of the proportion h_0/M of true null hypotheses correspond to increased differences between the two MTPs. Note that Lehmann and Romano (2004) only proved TPPFP control for Procedure 7 under certain assumptions concerning the joint distribution of the test statistics. In our simulation study, the `LR Restricted` MTP always controlled the TPPFP.

From Figure 9, differences in power between MTPs tend to decrease with increasing values of the common alternative shift parameter d_1.

7 Discussion

This article considers multiple testing procedures (MTP) for controlling tail probability error rates $Pr(g(V_n, R_n) > q)$, for some user-supplied non-negative constant q and function $g(V_n, R_n)$ of the numbers of Type I errors V_n and rejected hypotheses R_n. Error rates covered by this representation include the generalized family-wise error rate (gFWER), where $g(v, r) = v$, and tail probabilities for the proportion of false positives (TPPFP) among the rejected hypotheses, where $g(v, r) = v/r$.

Given an arbitrary function g, Section 3 proposes a general class of augmentation multiple testing procedures (AMTP), for controlling tail probability error rates $Pr(g(V_n, R_n) > q)$, by adding suitably chosen null hypotheses to the set of hypotheses already rejected by an initial gFWER-controlling procedure (Procedure 1). The adjusted p-values of the new AMTP are shown to be simply shifted versions of the adjusted p-values of the initial MTP (Theorem 1). The important practical implication of these results is that *any* gFWER-controlling MTP and its corresponding adjusted p-values immediately provide multiple testing procedures controlling a broad class of Type I error rates and their adjusted p-values. One can therefore build on the large pool of available FWER-controlling procedures, such as the joint single-step and step-down maxT and minP procedures discussed in Dudoit et al. (2004) and van der Laan et al. (2004b). The reader is referred to Dudoit and van der Laan (2004) and van der Laan et al. (2004a) for an in-depth treatment of the augmentation approach to multiple testing.

Sections 4 and 5 focus, respectively, on gFWER- and TPPFP-controlling procedures. Specifically, Section 4 describes the marginal single-step and step-down procedures of Lehmann and Romano (2004) (Procedures 2 and 3), the joint single-step common-cut-off $T(k+1)$ and common-quantile $P(k+1)$ procedures of Dudoit et al. (2004) and Pollard and van der Laan (2004) (Procedures 4 and 5), and the general augmentation procedures of van der Laan et al. (2004a) (Procedure 6). Section 5 summarizes the marginal restricted and general step-down procedures of Lehmann and Romano (2004) (Procedures 7 and 8) and the general augmentation procedures of van der Laan et al. (2004a) (Procedure 9).

The gFWER- and TPPFP-controlling MTPs listed in Tables 5 and 6 are compared in the simulation study described in Section 6. Although a simple Gaussian data generating model is used to simulate test statistics, a broad range of testing scenarios (including extreme ones) are covered by varying

the following model parameters: the total number of null hypotheses M, the proportion of true null hypotheses h_0/M, the shift parameter vector d_n, and the correlation matrix σ^*. The simulation results should therefore provide a fairly complete assessment of the performance of the different MTPs.

One of the main findings of the study is the substantial gains in power of joint MTPs compared to marginal MTPs, i.e., for virtually all simulation models and both error rates, the most powerful procedure is the augmentation of the joint single-step maxT procedure (`Aug. maxT`). As expected, step-down procedures are more powerful than their single-step analogs, although differences in power vary across models. While the augmentation approach can be applied to *any* initial FWER-controlling procedure, we consider, for a direct and favorable comparison to the marginal Lehmann and Romano (2004) procedures, conservative AMTPs based on the marginal Bonferroni single-step and Holm step-down MTPs. Augmentations of marginal FWER-controlling MTPs (`Aug. Bonf.` and `Aug. Holm`) tend to be more powerful than the corresponding marginal Lehmann and Romano (2004) procedures for a broad range of models. The magnitude of the power differences depends, however, on the data generating model, with the largest gains in power for marginal AMTPs occurring for a small number M of null hypotheses. For large values of M, the augmentation approach tends to suffer from the lack of power of the initial marginal FWER-controlling MTP, especially for large nominal level α. However, AMTPs generally retain their power advantages for the small nominal levels one would use in practice (e.g., $\alpha < 0.10$) and for large alternative shift parameter values (Figure 14). Large shift values indeed correspond to a larger number of rejections by the initial FWER-controlling MTP and hence to a "real" augmentation. The gains in power of AMTPs are most noticeable when a few alternative hypotheses have large shift values and a majority have moderate shift values, a scenario one would expect, for example, in DNA microarray gene expression studies. Most data generating models considered in the simulation study of Section 6 have small or moderate shift values, for which one expects few rejected hypotheses. Additional simulations would consider larger values for the number M of null hypotheses and a broader range of alternative shift parameters d_n (including more extreme shift values, e.g., $d_n(m) > 5$).

For a large number of null hypotheses, it may be advantageous to consider hybrid approaches, i.e., augmenting an initial $gFWER(k_0)$–controlling MTP as in Section 3. The initial $gFWER(k_0)$–controlling MTP could be either marginal (e.g., Lehmann and Romano (2004) Bonferroni-like single-step

Procedure 2 and Holm-like step-down Procedure 3) or joint (e.g., single-step common-cut-off $T(k+1)$ Procedure 4 and common-quantile $P(k+1)$ Procedure 5). It would be of interest to also examine a new gFWER-controlling step-down common-cut-off procedure of Romano (personal communication).

In addition to exploring the aforementioned hybrid approaches, ongoing efforts include considering other functions g for defining the tail probability error rate $Pr(g(V_n, R_n) > q)$. While not emphasized in this article, a key ingredient of our framework for multiple hypothesis testing is the test statistics null distribution Q_0 used to derive rejection regions and adjusted p-values. The general construction of a null distribution introduced in Section 2.2, along with the augmentation approach of Section 3, provide multiple testing procedures controlling a variety of Type I error rates, for general data generating distributions (with arbitrary dependence structures among variables), null hypotheses (defined in terms of submodels for the data generating distribution), and test statistics (e.g., t-statistics, F-statistics). We are currently applying these general and flexible MTPs to address the following problems in genomics: the identification of differentially expressed genes in DNA microarray experiments; tests of association between gene expression measures and Gene Ontology (GO) annotation (`www.geneontology.org`); the identification of transcription factor binding sites in ChIP-Chip experiments, where chromatin immunoprecipitation (ChIP) of transcription factor bound DNA is followed by microarray hybridization (Chip) of the IP-enriched DNA (Keleş et al., 2004); and the genetic mapping of complex traits using single nucleotide polymorphisms (SNP).

Software

The simulation study of Section 6 was done using the R language and environment for statistical computing (R Version 1.9.1, `www.r-project.org`).

The resampling-based multiple testing procedures discussed in this and related articles (Dudoit et al., 2004; van der Laan et al., 2004b,a; Pollard and van der Laan, 2004) are implemented in the open source R package multtest, released as part of the Bioconductor Project (Pollard et al. (2005); multtest package Version 1.5.0, Bioconductor Release 1.5, `www.bioconductor.org`).

References

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.

S. Dudoit and M. J. van der Laan. *Multiple Testing Procedures and Applications to Genomics*. Springer, 2004. (In preparation).

S. Dudoit, M. J. van der Laan, and K. S. Pollard. Multiple testing. Part I. Single-step procedures for control of general Type I error rates. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 13, 2004. URL `www.bepress.com/sagmb/vol3/iss1/art13`.

J. H. Friedman. A variable span smoother. Technical Report 5, Laboratory for Computational Statistics, Stanford University, 1984.

C. R. Genovese and L. Wasserman. A stochastic process approach to false discovery rates. Technical Report 762, Department of Statistics, Carnegie Mellon University, January 2003. URL `www.stat.cmu.edu/cmu-stats`.

C. R. Genovese and L. Wasserman. Exceedance control of the false discovery proportion. Technical report, Department of Statistics, Carnegie Mellon University, July 2004. URL `www.stat.cmu.edu/cmu-stats`.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M.L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

Y. Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75:800–802, 1988.

S. Keleş, M. J. van der Laan, S. Dudoit, and S. E. Cawley. Multiple testing methods for ChIP-Chip high density oligonucleotide array data. Technical

Report 147, Division of Biostatistics, University of California, Berkeley, 2004. URL www.bepress.com/ucbbiostat/paper147.

E. L. Lehmann and J. P. Romano. Generalizations of the family-wise error rate. Technical report, Department of Statistics, Stanford University, June 2004. To appear, *Annals of Statistics*.

K. S. Pollard, S. Dudoit, and M. J. van der Laan. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, chapter Multiple Testing Procedures and Applications to Genomics. Springer-Verlag, New York, 2005. URL www.bepress.com/ucbbiostat/paper164. (Submitted).

K. S. Pollard and M. J. van der Laan. Choice of a null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference*, 125(1–2):85–100, 2004.

P. H. Ramsey. Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association*, 73:479–485, 1978.

J. P. Shaffer. Multiple hypothesis testing. *Annu. Rev. Psychol.*, 46:561–584, 1995.

M. J. van der Laan, S. Dudoit, and K. S. Pollard. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 15, 2004a. URL www.bepress.com/sagmb/vol3/iss1/art15.

M. J. van der Laan, S. Dudoit, and K. S. Pollard. Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 14, 2004b. URL www.bepress.com/sagmb/vol3/iss1/art14.

P. H. Westfall and S. S. Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons, 1993.

Table 1: *Type I and Type II errors in multiple hypothesis testing.*

		Null hypotheses not rejected	Null hypotheses rejected	
Null hypotheses	true	$\lvert\mathcal{R}_n^c \cap \mathcal{H}_0\rvert$	$V_n = \lvert\mathcal{R}_n \cap \mathcal{H}_0\rvert$ (Type I errors)	$h_0 = \lvert\mathcal{H}_0\rvert$
	false	$U_n = \lvert\mathcal{R}_n^c \cap \mathcal{H}_1\rvert$ (Type II errors)	$\lvert\mathcal{R}_n \cap \mathcal{H}_1\rvert$	$h_1 = \lvert\mathcal{H}_1\rvert$
		$M - R_n$	$R_n = \lvert\mathcal{R}_n\rvert$	M

Table 2: *FWER-controlling procedures.* Summary of FWER-controlling single-step and step-down procedures and corresponding adjusted p-values. Unadjusted and adjusted p-values are denoted by $p_{0n}(m)$ and $\widetilde{p}_{0n}(m)$, respectively. The indices $o_n(m)$ for the ordered unadjusted p-values are such that $p_{0n}(o_n(1)) \leq \ldots \leq p_{0n}(o_n(M))$. For procedures based on common cut-offs (maxT), the indices $o_n(m)$ for the ordered test statistics are such that $t_n(o_n(1)) \geq \ldots \geq t_n(o_n(M))$. Here, $Z^\circ(m)$ denotes the mth ordered component of the M–vector $Z = (Z(m) : m = 1, \ldots, M) \sim Q_0$, so that $Z^\circ(1) \geq \ldots \geq Z^\circ(M)$. The random variable $P_0^\circ(m)$ denotes the mth ordered component of the M–vector of unadjusted p-values $P_0 = (P_0(m) = \bar{Q}_{0,m}(Z(m)) : m = 1, \ldots, M)$, so that $P_0^\circ(1) \leq \ldots \leq P_0^\circ(M)$. The FWER adjusted p-values can be used as inputs to the gFWER- and TPPFP-controlling augmentation MTPs of Tables 3 and 4.

Procedure	**Marginal/Joint**	**Single-step/Step-down**
Bonferroni $\widetilde{p}_{0n}(m) = \min\left(M\, p_{0n}(m), 1\right)$.	Marginal	Single-step
Single-step minP $\widetilde{p}_{0n}(m) = Pr_{Q_0}\left(P_0^\circ(1) \leq p_{0n}(m)\right)$.	Joint	Single-step
Single-step maxT $\widetilde{p}_{0n}(m) = Pr_{Q_0}\left(Z^\circ(1) \geq t_n(m)\right)$.	Joint	Single-step
Holm $\widetilde{p}_{0n}(o_n(m)) = \max_{h=1,\ldots,m}$	Marginal $\{\min\left((M-h+1)\, p_{0n}(o_n(h)), 1\right)\}$.	Step-down
Step-down minP $\widetilde{p}_{0n}(o_n(m)) = \max_{h=1,\ldots,m}$	Joint $\left\{Pr_{Q_0}\left(\min_{l\in\{o_n(h),\ldots,o_n(M)\}} P_0(l) \leq p_{0n}(o_n(h))\right)\right\}$.	Step-down
Step-down maxT $\widetilde{p}_{0n}(o_n(m)) = \max_{h=1,\ldots,m}$	Joint $\left\{Pr_{Q_0}\left(\max_{l\in\{o_n(h),\ldots,o_n(M)\}} Z(l) \geq t_n(o_n(h))\right)\right\}$.	Step-down

Table 3: *gFWER-controlling procedures.* Summary of gFWER-controlling single-step and step-down procedures and corresponding adjusted p-values. Unadjusted and adjusted p-values are denoted by $p_{0n}(m)$ and $\widetilde{p}_{0n}(m)$, respectively. In the case of augmentation MTPs, the adjusted p-values for the initial FWER-controlling procedure are denoted by $\widetilde{p}_{0n}(m)$ (examples in Table 2) and those for the resulting $gFWER(k)$–controlling AMTP by $\widetilde{p}^{+}_{0n}(m)$. The indices $o_n(m)$ for the ordered unadjusted p-values are such that $p_{0n}(o_n(1)) \leq \ldots \leq p_{0n}(o_n(M))$. For procedures based on common cut-offs (maxT), the indices $o_n(m)$ for the ordered test statistics are such that $t_n(o_n(1)) \geq \ldots \geq t_n(o_n(M))$.

Procedure	**Marginal/Joint**	**Single-step/Step-down**
Marginal procedures of Lehmann and Romano (2004)		
Procedure 2 Bonferroni-like $\widetilde{p}_{0n}(m) = \min\left(\frac{M}{(k+1)}\, p_{0n}(m), 1\right)$.	Marginal	Single-step
Procedure 3 Holm-like $\widetilde{p}_{0n}(o_n(m)) = \begin{cases} \min\left(\frac{M}{(k+1)} p_{0n}(o_n(m)),\, 1\right), & \text{if } m \leq (k+1) \\ \max_{h=1,\ldots,m-k}\ \left\{\min\left(\frac{(M-h+1)}{(k+1)} p_{0n}(o_n(h+k)),\, 1\right)\right\}, & \text{if } m > (k+1) \end{cases}$.	Marginal	Step-down
Common-cut-off and common-quantile procedures of Dudoit et al. (2004) and Pollard and van der Laan (2004)		
Procedure 4 Common-cut-off $T(k+1)$ $\widetilde{p}_{0n}(m) = Pr_{Q_0}\left(Z^{\circ}(k+1) \geq t_n(m)\right)$.	Joint	Single-step
Procedure 5 Common-quantile $P(k+1)$ $\widetilde{p}_{0n}(m) = Pr_{Q_0}\left(P_0^{\circ}(k+1) \leq p_{0n}(m)\right)$.	Joint	Single-step
Augmentation procedures of van der Laan et al. (2004a)		
Procedure 6 General augmentation $\widetilde{p}^{+}_{0n}(o_n(m)) = \begin{cases} 0, & \text{if } m \leq k \\ \widetilde{p}_{0n}(o_n(m-k)), & \text{if } m > k \end{cases}$.	Either	Either

Table 4: *TPPFP-controlling procedures.* Summary of TPPFP-controlling single-step and step-down procedures and corresponding adjusted p-values. Unadjusted and adjusted p-values are denoted by $p_{0n}(m)$ and $\widetilde{p}_{0n}(m)$, respectively. In the case of augmentation MTPs, the adjusted p-values for the initial FWER-controlling procedure are denoted by $\widetilde{p}_{0n}(m)$ (examples in Table 2) and those for the resulting $TPPFP(q)$–controlling AMTP by $\widetilde{p}^{+}_{0n}(m)$. The indices $o_n(m)$ for the ordered unadjusted p-values are such that $p_{0n}(o_n(1)) \leq \ldots \leq p_{0n}(o_n(M))$. For procedures based on common cutoffs (maxT), the indices $o_n(m)$ for the ordered test statistics are such that $t_n(o_n(1)) \geq \ldots \geq t_n(o_n(M))$.

Procedure	**Marginal/Joint**	**Single-step/Step-down**
Marginal procedures of Lehmann and Romano (2004)		
Procedure 7 Restricted*	Marginal	Step-down
$\widetilde{p}_{0n}(o_n(m)) = \max_{h=1,\ldots,m} \left\{ \min \left(\frac{(M+\lfloor qh \rfloor + 1 - h)}{(\lfloor qh \rfloor + 1)} p_{0n}(o_n(h)),\, 1 \right) \right\}$.		
Procedure 8 General	Marginal	Step-down
$\widetilde{p}_{0n}(o_n(m)) = \max_{h=1,\ldots,m} \left\{ \min \left(C(\lfloor qM \rfloor + 1) \frac{(M+\lfloor qh \rfloor + 1 - h)}{(\lfloor qh \rfloor + 1)} p_{0n}(o_n(h)),\, 1 \right) \right\}$, where $C(M) \equiv \sum_{m=1}^{M} 1/m$.		
Augmentation procedures of van der Laan et al. (2004a)		
Procedure 9 General augmentation	Either	Either
$\widetilde{p}^{+}_{0n}(o_n(m)) = \widetilde{p}_{0n}(o_n(\lceil (1-q)m \rceil))$.		

* **N.B.** Proof of Type I error control for Procedure 7 requires certain assumptions on the dependence structure of the test statistics.

Table 5: *gFWER-controlling procedures compared in simulation study.* The following five MTPs are compared in the simulation study to control $gFWER(k) = Pr(V_n > k)$, for an allowed number of false positives $k \in \{1, 5, 10, 20, 25, 50, 100\}$ (when applicable, given the value of M). Note that for our simple choices of data generating model and one-sample z-statistics, common-cut-off (maxT) and common-quantile (minP) procedures are equivalent. Adjusted p-values are given in Tables 2 and 3 and in the equations listed below.

Abbreviation	**Procedure**
`LR SS`	Lehmann & Romano Bonferroni-like single-step procedure
	Procedure 2
	Equation (27)
`LR SD`	Lehmann & Romano Holm-like step-down procedure
	Procedure 3
	Equation (30)
`Aug. Bonf.`	Augmentation of Bonferroni single-step procedure
	Procedure 6
	Equation (35)
`Aug. Holm`	Augmentation of Holm step-down procedure
	Procedure 6
	Equation (36)
`Aug. maxT`	Augmentation of single-step maxT procedure
	Procedure 6
	Equation (51)

Table 6: *TPPFP-controlling procedures compared in simulation study.* The following five MTPs are compared in the simulation study to control $TPPFP(q) = Pr(V_n/R_n > q)$, for an allowed proportion of false positives $q \in \{0.01, 0.05, 0.10, 0.20, 0.50, 0.75\}$. Note that for our simple choices of data generating model and one-sample z-statistics, common-cut-off (maxT) and common-quantile (minP) procedures are equivalent. Adjusted p-values are given in Tables 2 and 4 and in the equations listed below.

Abbreviation	**Procedure**
`LR Restricted`	Lehmann & Romano restricted step-down procedure
	Procedure 7
	Equation (39)
`LR General`	Lehmann & Romano general step-down procedure
	Procedure 8
	Equation (42)
`Aug. Bonf.`	Augmentation of Bonferroni single-step procedure
	Procedure 9
	Equation (45)
`Aug. Holm`	Augmentation of Holm step-down procedure
	Procedure 9
	Equation (46)
`Aug. maxT`	Augmentation of single-step maxT procedure
	Procedure 9
	Equation (51)

Table 7: *Summary of power comparisons for gFWER- and TPPFP-controlling procedures.* An $\Uparrow$ ($\Downarrow$) in a cell of the table indicates that the power of the first MTP in the corresponding comparison (column) tends to increase, relative to that of the second MTP, with increasing (decreasing) values of the corresponding parameter (row). An empty cell denotes an inconclusive comparison. `LR`: for gFWER control, `LR` refers to the Lehmann and Romano (2004) single-step and step-down Procedures 2 and 3 (i.e., `LR SS` and `LR SD`, in Table 5); for TPPFP control, `LR` refers to the Lehmann and Romano (2004) restricted and general step-down Procedures 7 and 8 (i.e., `LR Restricted` and `LR General`, in Table 6). `AMTP`: depending on the comparison, `AMTP` refers to the van der Laan et al. (2004a) gFWER-controlling augmentation Procedure 6 or TPPFP-controlling augmentation Procedure 9, based on either the Bonferroni single-step, the Holm step-down, or the single-step maxT MTPs (i.e., `Aug. Bonf.`, `Aug. Holm`, or `Aug. maxT`, in Tables 5 and 6). `SS` and `SD` refer, respectively, to single-step and step-down MTPs.

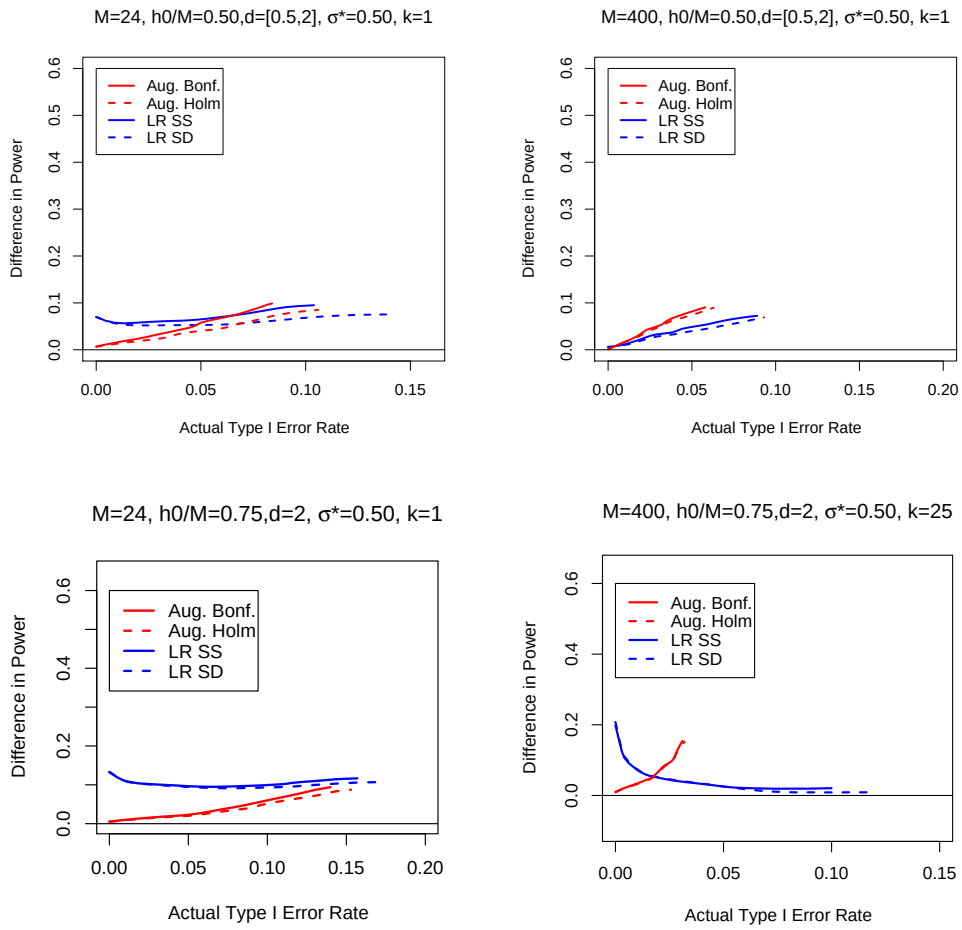
Parameter	`AMTP` **vs.** `LR`	`SD` **vs.** `SS`	`Joint` **vs.** `Marginal`	**Figure**
M	$\Downarrow$	$\Downarrow$	$\Uparrow^\star$	Fig. 1
$\frac{h_0}{M}$	$\Uparrow$	$\Downarrow$		Fig. 2
d_n	$\Downarrow$		$\Downarrow$	Fig. 3
σ^*			$\Uparrow$	Fig. 5
k	$\Uparrow$			Fig. 6

* Figures not shown.

Parameter	`AMTP` **vs.** `LR`	`SD` **vs.** `SS`	`Joint` **vs.** `Marginal`	`LR Restricted` **vs.** `LR General`	**Figure**
M	$\Downarrow$	$\Downarrow$	$\Uparrow$		Fig. 7
$\frac{h_0}{M}$	$\Uparrow^\star$	$\Downarrow$	$\Downarrow^\star$	$\Uparrow^\star$	Fig. 8
d_n	$\Downarrow$		$\Downarrow$	$\Downarrow^\star$	Fig. 9
σ^*			$\Uparrow$		Fig. 11
q	$\Uparrow^\dagger$	$\Uparrow^\dagger$		$\Uparrow^\star$	Figs. 12 – 13

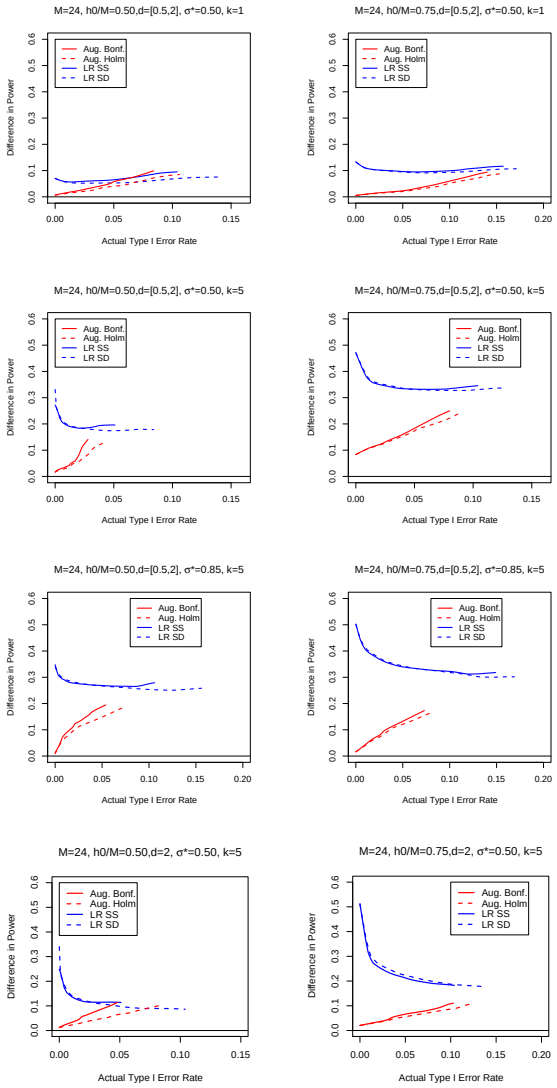
* Figures not shown.

† For large values of q.



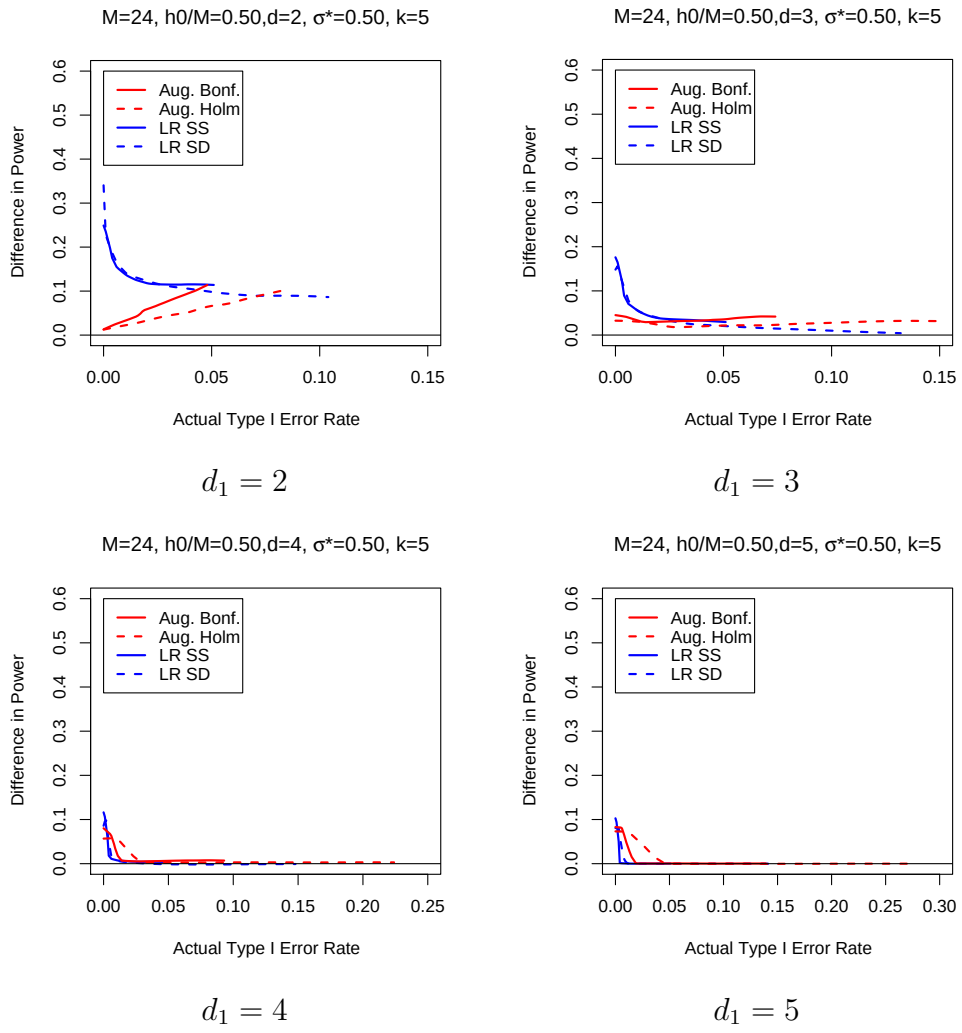
$M = 24$ $M = 400$

Figure 1: *$gFWER(k)$–controlling MTPs, power comparison as M varies.* Plot of differences in power with respect to `Aug. maxT` procedure vs. actual Type I error rate. Model with $M = 24, 400$ null hypotheses; proportion $h_0/M = 0.50, 0.75$ of true null hypotheses; alternative shift parameters $d_n(m : m \in \mathcal{H}_1) = \texttt{c}(\texttt{rep}(0.50, \texttt{times} = \texttt{h}_1/2), \texttt{rep}(2, \texttt{times} = \texttt{h}_1/2))$ and $d_n(m : m \in \mathcal{H}_1) = \texttt{rep}(2, \texttt{times} = \texttt{h}_1)$; full correlation matrix σ^*, with $\sigma^*(m, m') = 0.50$, for $m \neq m' = 1, \ldots, M$; allowed number of false positives $k = 1, 25$. Note that the two lower panels are not directly comparable, in the sense that they correspond to different values of the Type I error parameter k. However, the proportion k/M of allowed Type I errors is similar for both panels.



$h_0/M = 0.50$ $h_0/M = 0.75$

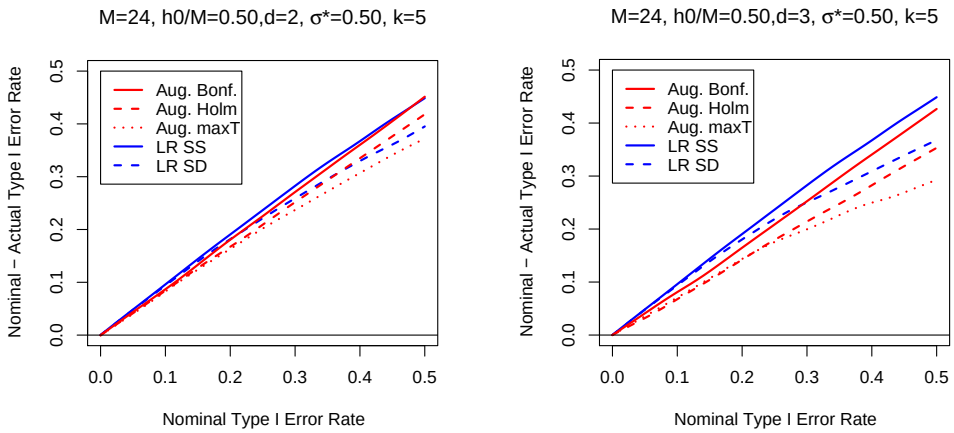
Figure 2: *$gFWER(k)$–controlling MTPs, power comparison as h_0/M varies.* Plot of differences in power with respect to `Aug. maxT` procedure vs. actual Type I error rate. Model with $M = 24$ null hypotheses; proportion $h_0/M = 0.50, 0.75$ of true null hypotheses; alternative shift parameters $d_n(m : m \in \mathcal{H}_1) = \mathtt{c}(\mathtt{rep}(0.50, \mathtt{times} = \mathtt{h_1}/2), \mathtt{rep}(2, \mathtt{times} = \mathtt{h_1}/2))$ and $d_n(m : m \in \mathcal{H}_1) = \mathtt{rep}(2, \mathtt{times} = \mathtt{h_1})$; full correlation matrix σ^*, with $\sigma^*(m, m') = 0.50, 0.85$, for $m \neq m' = 1, \ldots, M$; allowed number of false positives $k = 1, 5$.



$d_1 = 2$ $d_1 = 3$

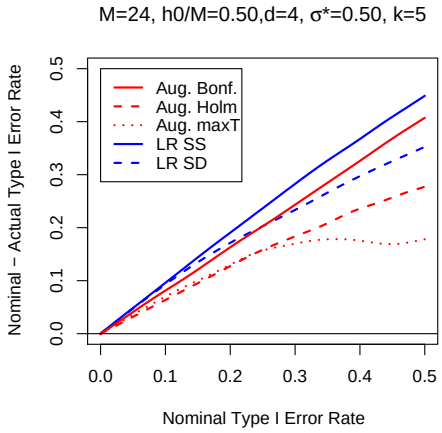
$d_1 = 4$ $d_1 = 5$

Figure 3: *$gFWER(k)$–controlling MTPs, power comparison as d_n varies.* Plot of differences in power with respect to `Aug. maxT` procedure vs. actual Type I error rate. Model with $M = 24$ null hypotheses; proportion $h_0/M = 0.50$ of true null hypotheses; alternative shift parameters $d_n(m : m \in \mathcal{H}_1) = \texttt{rep}(\texttt{d}_1, \texttt{times} = \texttt{h}_1)$, $d_1 = 2, 3, 4, 5$; full correlation matrix σ^*, with $\sigma^*(m, m') = 0.50$, for $m \neq m' = 1, \ldots, M$; allowed number of false positives $k = 5$.

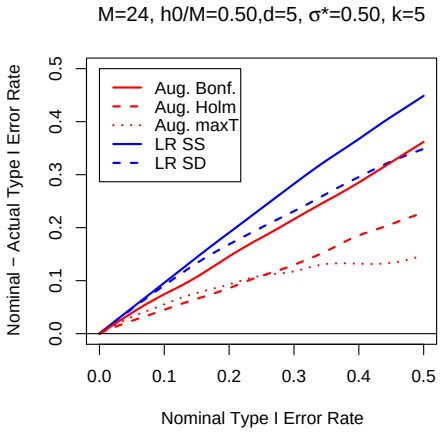


$d_1 = 2$

$d_1 = 3$

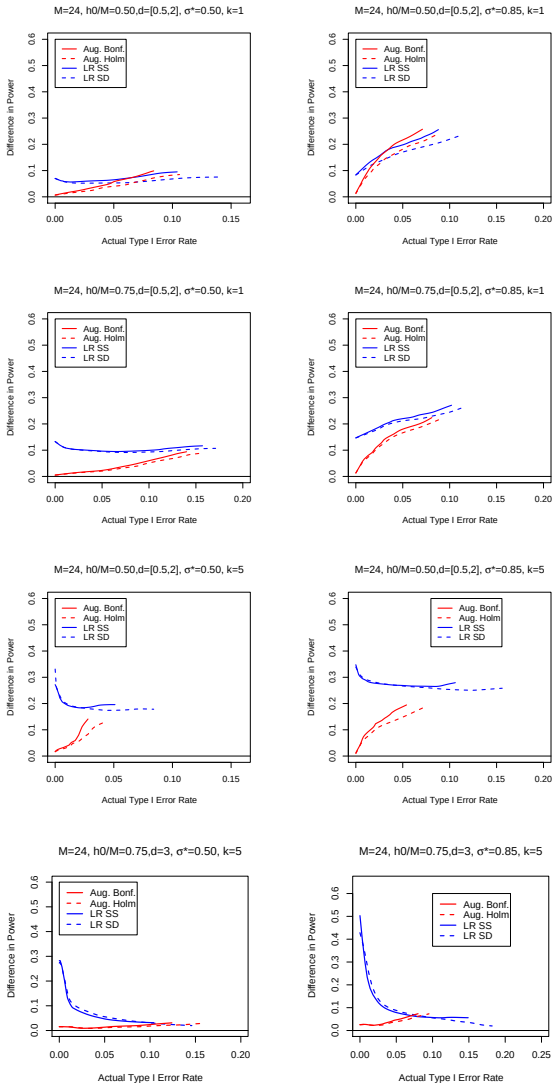


$d_1 = 4$



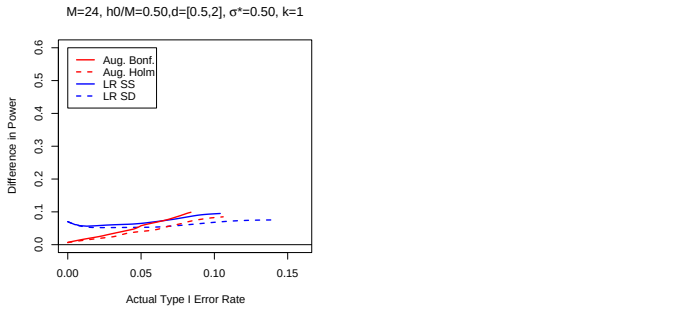
$d_1 = 5$

Figure 4: *$gFWER(k)$–controlling MTPs, Type I error control comparison as d_n varies.* Plot of differences between nominal and actual Type I error rates vs. actual Type I error rate. Model with $M = 24$ null hypotheses; proportion $h_0/M = 0.50$ of true null hypotheses; alternative shift parameters $d_n(m : m \in \mathcal{H}_1) = \texttt{rep}(\texttt{d}_1, \texttt{times} = \texttt{h}_1)$, $d_1 = 2, 3, 4, 5$; full correlation matrix σ^*, with $\sigma^*(m, m') = 0.50$, for $m \neq m' = 1, \ldots, M$; allowed number of false positives $k = 5$.

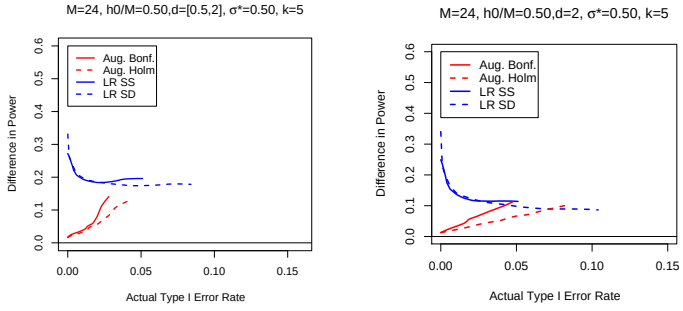


$\sigma^*(m, m') = 0.50$ $\sigma^*(m, m') = 0.85$

Figure 5: *$gFWER(k)$–controlling MTPs, power comparison as σ^* varies.* Plot of differences in power with respect to `Aug. maxT` procedure vs. actual Type I error rate. Model with $M = 24$ null hypotheses; proportion $h_0/M = 0.50, 0.75$ of true null hypotheses; alternative shift parameters $d_n(m : m \in \mathcal{H}_1) = \texttt{c}(\texttt{rep}(0.50, \texttt{times} = \texttt{h}_1/2), \texttt{rep}(2, \texttt{times} = \texttt{h}_1/2))$ and $d_n(m : m \in \mathcal{H}_1) = \texttt{rep}(3, \texttt{times} = \texttt{h}_1)$; full correlation matrix σ^*, with $\sigma^*(m, m') = 0.50, 0.85$, for $m \neq m' = 1, \ldots, M$; allowed number of false positives $k = 1, 5$.

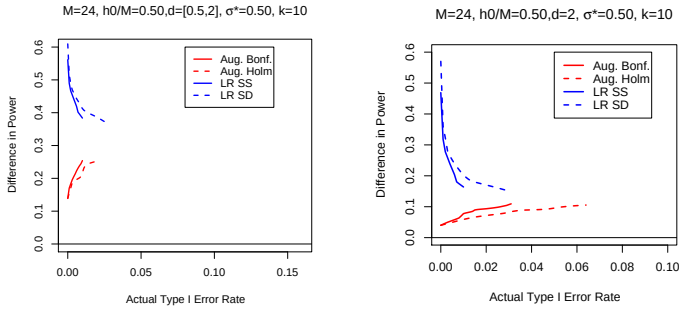


$k = 1$



$k = 5$

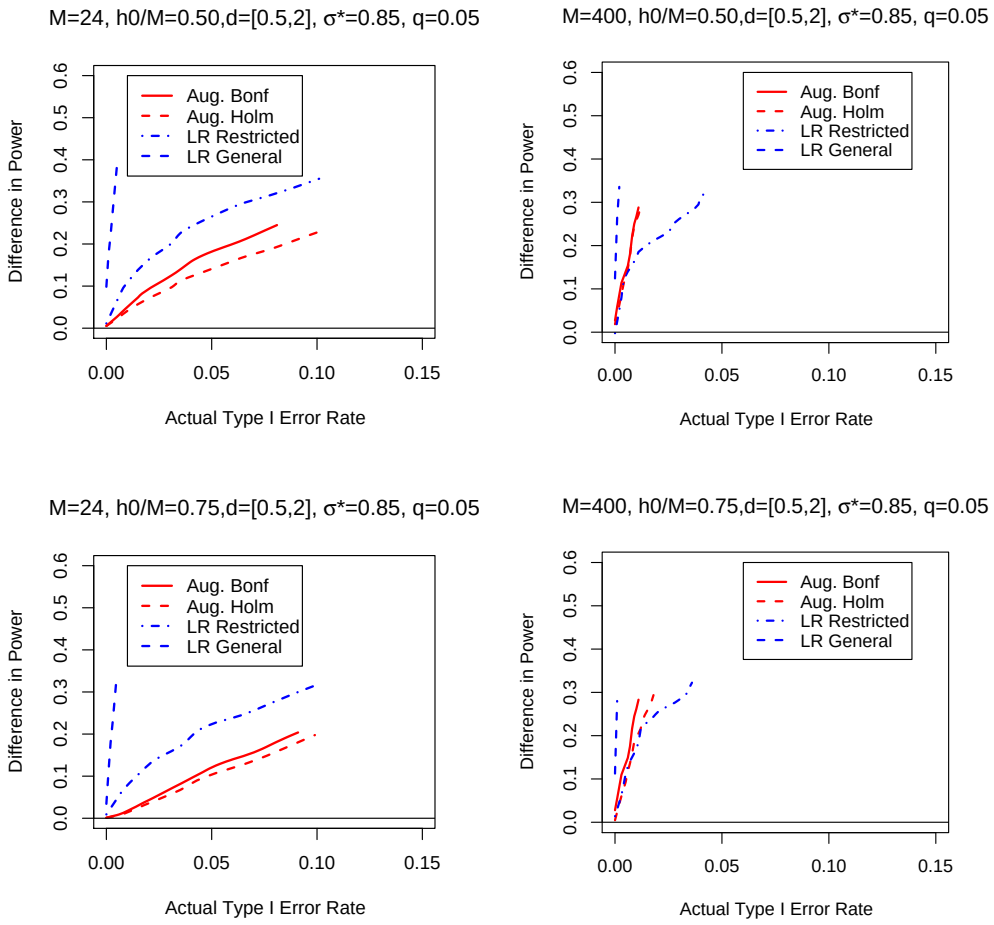
$k = 5$



$k = 10$

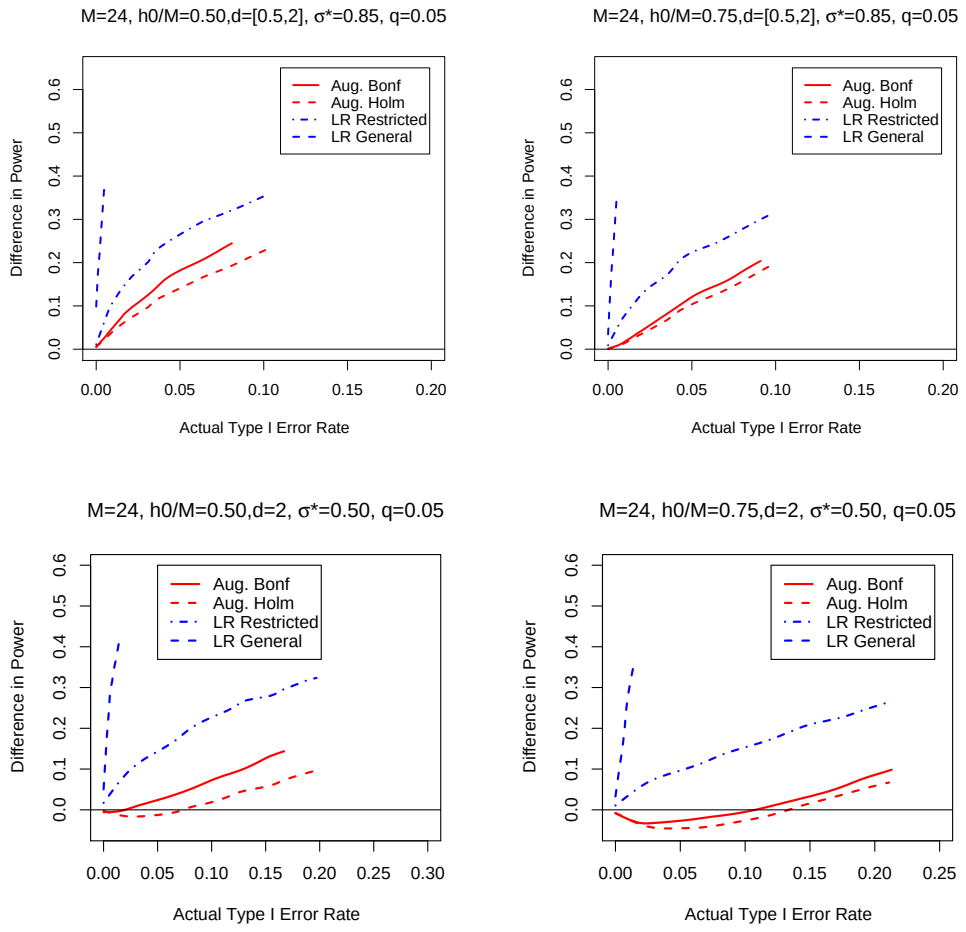
$k = 10$

Figure 6: *$gFWER(k)$–controlling MTPs, power comparison as k varies.* Plot of differences in power with respect to `Aug.maxT` procedure vs. actual Type I error rate. Model with $M = 24$ null hypotheses; proportion $h_0/M = 0.50$ of true null hypotheses; alternative shift parameters $d_n(m : m \in \mathcal{H}_1) = \texttt{c}(\texttt{rep}(0.50, \texttt{times} = \texttt{h}_1/2), \texttt{rep}(2, \texttt{times} = \texttt{h}_1/2))$ and $d_n(m : m \in \mathcal{H}_1) = \texttt{rep}(2, \texttt{times} = \texttt{h}_1)$; full correlation matrix σ^*, with $\sigma^*(m, m') = 0.50$, for $m \neq m' = 1, \ldots, M$; allowed number of false positives $k = 1$, 5, 10.



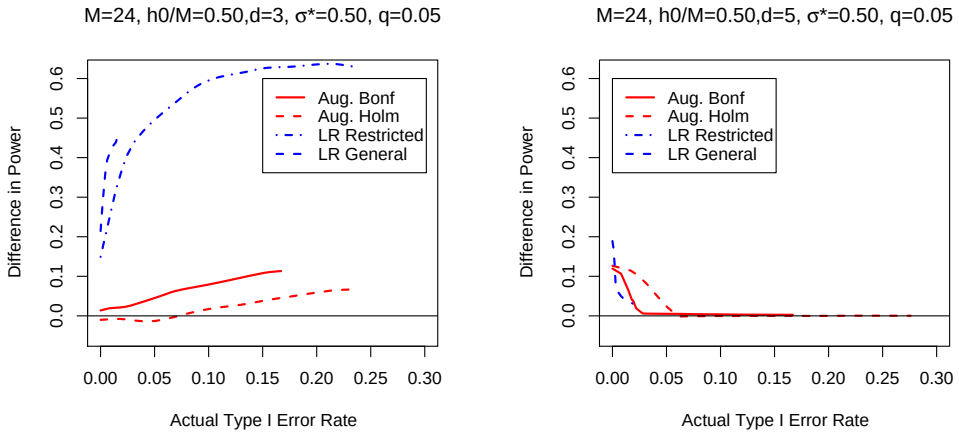
$M = 24$ $M = 400$

Figure 7: *$TPPFP(q)$–controlling MTPs, power comparison as M varies.* Plot of differences in power with respect to `Aug. maxT` procedure vs. actual Type I error rate. Model with $M = 24, 400$ null hypotheses; proportion $h_0/M = 0.50, 0.75$ of true null hypotheses; alternative shift parameters $d_n(m : m \in \mathcal{H}_1) = \texttt{c}(\texttt{rep}(0.50, \texttt{times} = \texttt{h}_1/2), \texttt{rep}(2, \texttt{times} = \texttt{h}_1/2))$; full correlation matrix σ^*, with $\sigma^*(m, m') = 0.85$, for $m \neq m' = 1, \ldots, M$; allowed proportion of false positives $q = 0.05$.



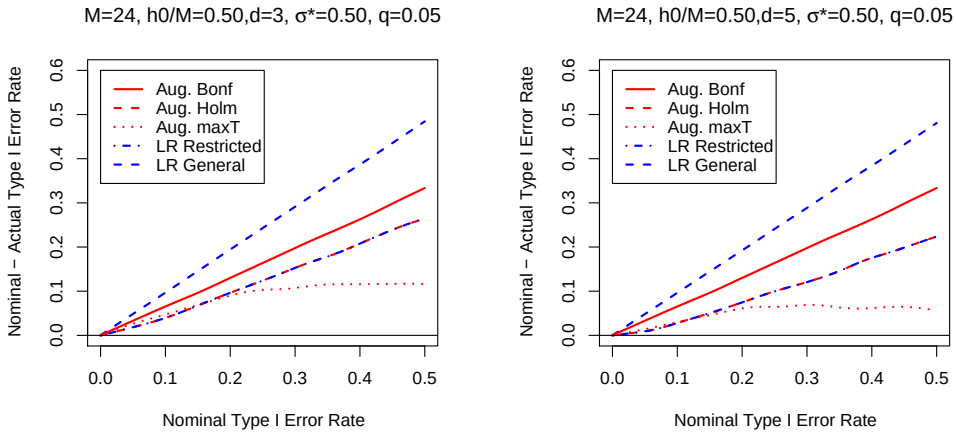
$h_0/M = 0.50$ $h_0/M = 0.75$

Figure 8: *$TPPFP(q)$–controlling MTPs, power comparison as h_0/M varies.* Plot of differences in power with respect to `Aug. maxT` procedure vs. actual Type I error rate. Model with $M = 24$ null hypotheses; proportion $h_0/M = 0.50, 0.75$ of true null hypotheses; alternative shift parameters $d_n(m : m \in \mathcal{H}_1) = \texttt{c}(\texttt{rep}(0.50, \texttt{times} = \texttt{h}_1/2), \texttt{rep}(2, \texttt{times} = \texttt{h}_1/2))$ and $d_n(m : m \in \mathcal{H}_1) = \texttt{rep}(2, \texttt{times} = \texttt{h}_1)$; full correlation matrix σ^*, with $\sigma^*(m, m') = 0.50, 0.85$, for $m \neq m' = 1, \ldots, M$; allowed proportion of false positives $q = 0.05$.



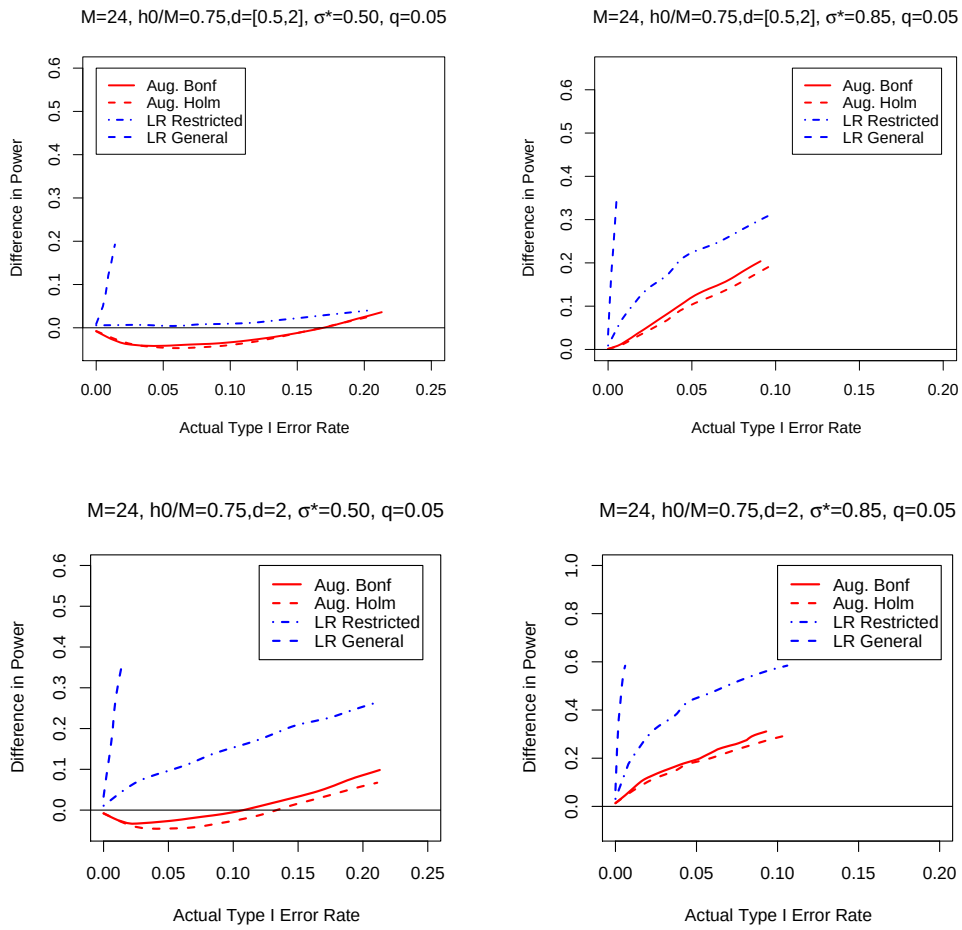
$d_1 = 3$ $d_1 = 5$

Figure 9: *$TPPFP(q)$–controlling MTPs, power comparison as d_n varies.* Plot of differences in power with respect to `Aug.` `maxT` procedure vs. actual Type I error rate. Model with $M = 24$ null hypotheses; proportion $h_0/M = 0.50$ of true null hypotheses; alternative shift parameters $d_n(m : m \in \mathcal{H}_1) =$ $\texttt{rep}(\texttt{d}_1, \texttt{times} = \texttt{h}_1)$, $d_1 = 3, 5$; full correlation matrix σ^*, with $\sigma^*(m, m') = 0.50$, for $m \neq m' = 1, \ldots, M$; allowed proportion of false positives $q = 0.05$.



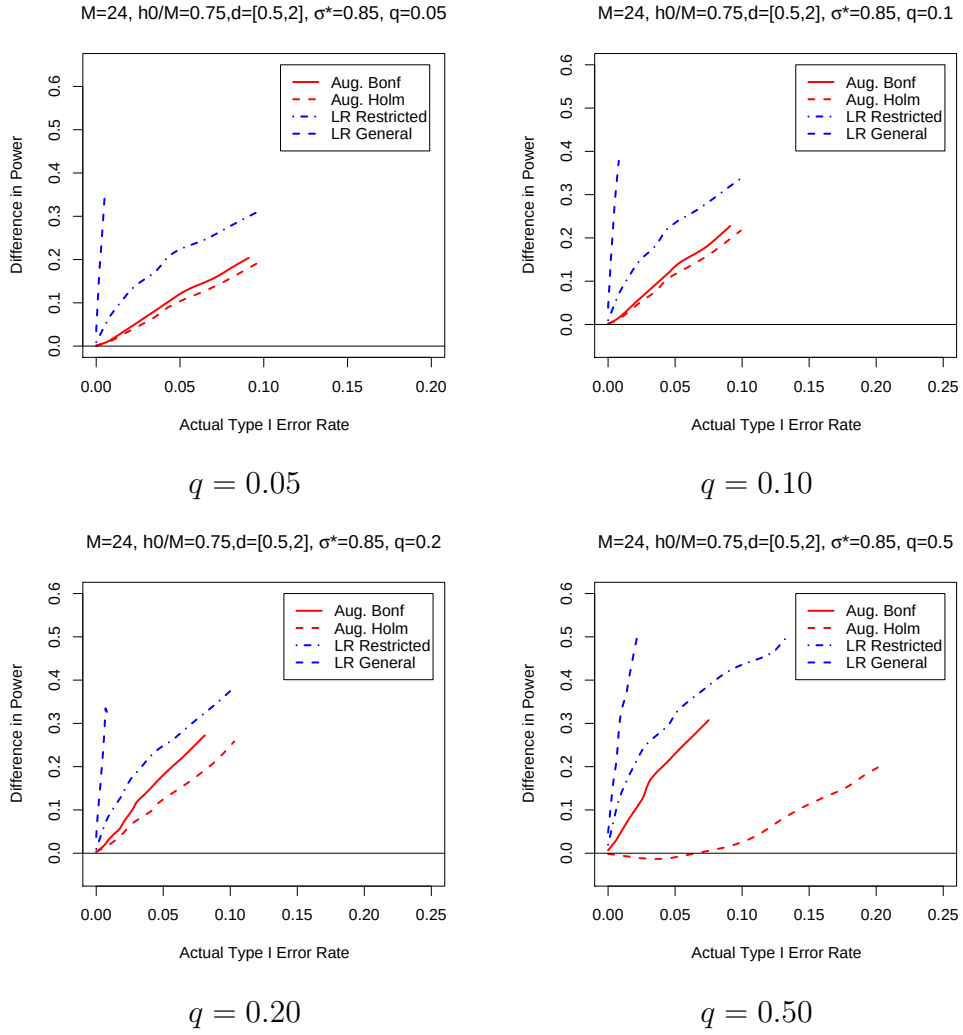
$d_1 = 3$ $d_1 = 5$

Figure 10: *$TPPFP(q)$–controlling MTPs, Type I error control comparison as d_n varies.* Plot of differences between nominal and actual Type I error rates vs. actual Type I error rate. Model with $M = 24$ null hypotheses; proportion $h_0/M = 0.50$ of true null hypotheses; alternative shift parameters $d_n(m : m \in \mathcal{H}_1) = \texttt{rep}(\texttt{d}_1, \texttt{times} = \texttt{h}_1)$, $d_1 = 3$, 5; full correlation matrix σ^*, with $\sigma^*(m, m') = 0.50$, for $m \neq m' = 1, \ldots, M$; allowed proportion of false positives $q = 0.05$.



$\sigma^*(m, m') = 0.50$ $\sigma^*(m, m') = 0.85$

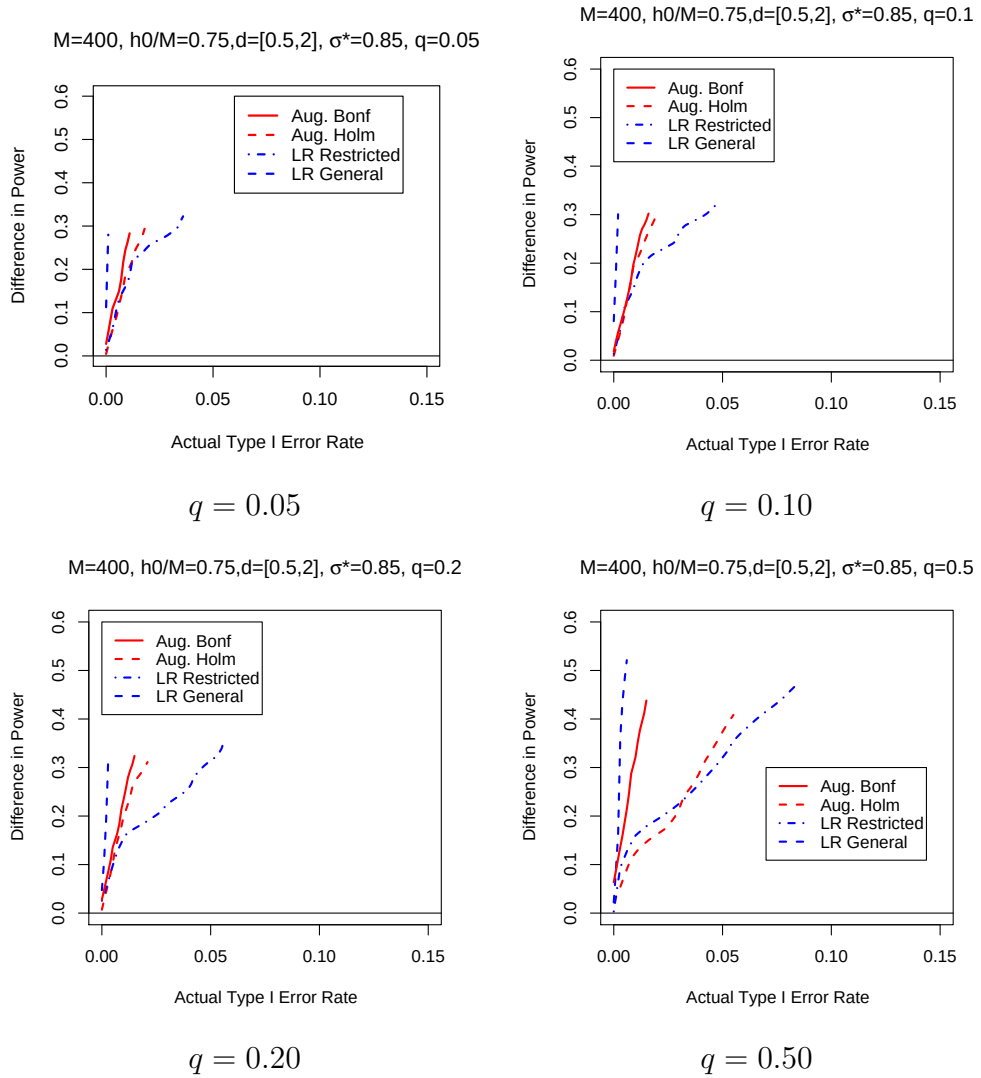
Figure 11: *$TPPFP(q)$–controlling MTPs, power comparison as σ^* varies.* Plot of differences in power with respect to `Aug.  maxT` procedure vs. actual Type I error rate. Model with $M = 24$ null hypotheses; proportion $h_0/M = 0.75$ of true null hypotheses; alternative shift parameters $d_n(m : m \in \mathcal{H}_1) = \texttt{c}(\texttt{rep}(0.50, \texttt{times} = \texttt{h}_1/2), \texttt{rep}(2, \texttt{times} = \texttt{h}_1/2))$ and $d_n(m : m \in \mathcal{H}_1) = \texttt{rep}(2, \texttt{times} = \texttt{h}_1)$; full correlation matrix σ^*, with $\sigma^*(m, m') = 0.50, 0.85$, for $m \neq m' = 1, \ldots, M$; allowed proportion of false positives $q = 0.05$.



$q = 0.05$ $q = 0.10$

$q = 0.20$ $q = 0.50$

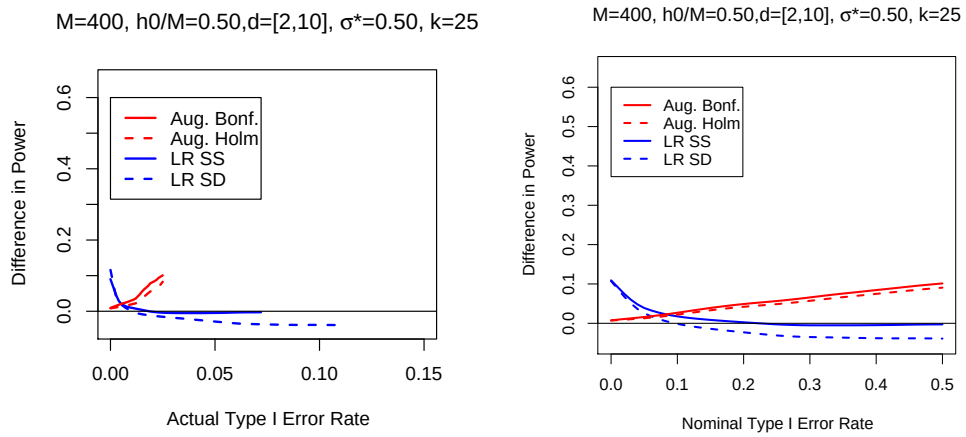
Figure 12: *$TPPFP(q)$–controlling MTPs, power comparison as q varies.* Plot of differences in power with respect to `Aug. maxT` procedure vs. actual Type I error rate. Model with $M = 24$ null hypotheses; proportion $h_0/M = 0.75$ of true null hypotheses; alternative shift parameters $d_n(m : m \in \mathcal{H}_1) = \texttt{c}(\texttt{rep}(0.50, \texttt{times} = \texttt{h}_1/2), \texttt{rep}(2, \texttt{times} = \texttt{h}_1/2))$; full correlation matrix σ^*, with $\sigma^*(m, m') = 0.85$, for $m \neq m' = 1, \ldots, M$; allowed proportion of false positives $q = 0.05$, 0.10, 0.20, 0.50.



$q = 0.05$ $q = 0.10$

$q = 0.20$ $q = 0.50$

Figure 13: *$TPPFP(q)$–controlling MTPs, power comparison as q varies.* Plot of differences in power with respect to `Aug. maxT` procedure vs. actual Type I error rate. Model with $M = 400$ null hypotheses; proportion $h_0/M = 0.75$ of true null hypotheses; alternative shift parameters $d_n(m : m \in \mathcal{H}_1) = \texttt{c}(\texttt{rep}(0.50, \texttt{times} = \texttt{h}_1/2), \texttt{rep}(2, \texttt{times} = \texttt{h}_1/2))$; full correlation matrix σ^*, with $\sigma^*(m, m') = 0.85$, for $m \neq m' = 1, \ldots, M$; allowed proportion of false positives $q = 0.05, 0.10, 0.20, 0.50$.



(a) Actual, $gFWER(k;\alpha)$ (b) Nominal, α

Figure 14: *$gFWER(k)$–controlling MTPs, power comparison in terms of actual and nominal Type I error rates.* Panel (a): Plot of differences in power with respect to `Aug. maxT` procedure vs. *actual* Type I error rate, $gFWER(k;\alpha)$. Panel (b): Plot of differences in power with respect to `Aug. maxT` procedure vs. *nominal* Type I error rate, α. Model with $M = 400$ null hypotheses; proportion $h_0/M = 0.50$ of true null hypotheses; alternative shift parameters $d_n(m : m \in \mathcal{H}_1) = \texttt{c}(\texttt{rep}(10, \texttt{times} = 75), \texttt{rep}(2, \texttt{times} = 125))$; full correlation matrix σ^*, with $\sigma^*(m, m') = 0.50$, for $m \neq m' = 1, \ldots, M$; allowed number of false positives $k = 25$.