The Annals of Statistics
2011, Vol. 39, No. 1, 556–583
DOI: 10.1214/10-AOS844

POWER-ENHANCED MULTIPLE DECISION FUNCTIONS CONTROLLING FAMILY-WISE ERROR AND FALSE DISCOVERY RATES

BY EDSEL A. PEÑA[1], JOSHUA D. HABIGER AND WENSONG WU

University of South Carolina, Columbia, Oklahoma State University and University of South Carolina, Columbia

Improved procedures, in terms of smaller missed discovery rates (MDR), for performing multiple hypotheses testing with weak and strong control of the family-wise error rate (FWER) or the false discovery rate (FDR) are developed and studied. The improvement over existing procedures such as the Šidák procedure for FWER control and the Benjamini–Hochberg (BH) procedure for FDR control is achieved by exploiting possible differences in the powers of the individual tests. Results signal the need to take into account the powers of the individual tests and to have multiple hypotheses decision functions which are not limited to simply using the individual p-values, as is the case, for example, with the Šidák, Bonferroni, or BH procedures. They also enhance understanding of the role of the powers of individual tests, or more precisely the receiver operating characteristic (ROC) functions of decision processes, in the search for better multiple hypotheses testing procedures. A decision-theoretic framework is utilized, and through auxiliary randomizers the procedures could be used with discrete or mixed-type data or with rank-based nonparametric tests. This is in contrast to existing p-value based procedures whose theoretical validity is contingent on each of these p-value statistics being stochastically equal to or greater than a standard uniform variable under the null hypothesis. Proposed procedures are relevant in the analysis of high-dimensional "large M, small n" data sets arising in the natural, physical, medical, economic and social sciences, whose generation and creation is accelerated by advances in high-throughput technology, notably, but not limited to, microarray technology.

1. Introduction and motivation. The advent of modern technology, epitomized by the microarray, has led to the generation of very high-dimensional data pertaining to characteristics of a large number, M, of attributes, hereon called

Received March 2010; revised July 2010.

[1]Supported by NSF Grant DMS-08-05809, NIH Grant RR17698 and EPA Grant RD-83241902-0 to University of Arizona with subaward number Y481344 to the University of South Carolina.

AMS 2000 subject classifications. Primary 62F03; secondary 62J15.

Key words and phrases. Benjamini–Hochberg procedure, Bonferroni procedure, decision process, false discovery rate (FDR), family wise error rate (FWER), Lagrangian optimization, Neyman–Pearson most powerful test, microarray analysis, reverse martingale, missed discovery rate (MDR), multiple decision function and process, multiple hypotheses testing, optional sampling theorem, power function, randomized p-values, generalized multiple decision p-values, ROC function, Šidák procedure.

genes, associated with usually a small number, n, of units or subjects. Several such data sets are, for example, described in [10], and these are the inputs to so-called parallel inference problems. The most common form of inference is multiple hypotheses testing, wherein for the mth gene there are two competing hypotheses, a null hypothesis H_{m0} and an alternative hypothesis H_{m1}, for which a decision is to be made based on the data. In such multiple decision-making, there is a need to be cognizant and cautious of the *Hyde*-ian nature of multiplicity, while also exploiting the *Jekyll*-ian potentials of multiplicity [39]. Furthermore, this entails a tenuous balance between two competing desires: controlling the rate of rejection of correct null hypotheses, while at the same time maintaining the rate of discovery of correct alternative hypotheses.

As in single-pair hypothesis testing, a type I error occurs when a correct null hypothesis is rejected, while a type II error occurs when a false null hypothesis is not rejected. Several type I errors have been proposed in multiple testing; see [6] and [7]. Our focus is on the weak family wise error rate (FWER), the probability of rejecting at least one null hypothesis when all the nulls are correct; strong FWER, the probability of rejecting at least one correct null hypothesis; and false discovery rate (FDR), the expected proportion of the number of false rejections of nulls relative to the number of rejections [1, 37]. Our type II error rate is the missed discovery rate (MDR), the expected number of false nonrejections of null hypotheses. Other type II errors have been discussed in [5–7, 9, 41]. The usual framework in developing multiple decision functions is to bound the chosen type I error rate, and then minimize or make small the MDR. For example, a procedure controlling weak FWER, under an independence assumption, is that of Šidák [36]; while a conservative one not requiring independence is the Bonferroni procedure [3]. For FDR control, the most common procedure is the BH procedure [1]. Control of type I error measures related to the FDR have also been discussed in [8–10, 12, 15, 40, 41, 45], while [20, 23, 34] focused on estimation of the proportion of correct null hypotheses.

Procedures like the Šidák, Bonferroni and BH, rely on the set of p-values of individual tests. Their validity hinges on each p-value statistic being stochastically equal to or greater than a standard uniform variable under the null hypothesis. This fails, however, with noncontinuous variables or when rank-based nonparametric tests are used. Crucially, p-value based procedures also do not exploit the power characteristics of the individual tests, contrary to Neyman and Pearson's [27] adage that such considerations are germane in constructing optimal tests. Such p-value based procedures are fine in exchangeable settings where power characteristics of the individual tests are identical, but not in situations where genes or subclasses of genes have different structures; see [11, 13, 29].

Some papers dealing with procedures exploiting the power functions are [38, 49]. The use of weighted p-values to improve type II performance have also been explored in [16, 21, 29, 30, 46]. Other approaches for optimal procedures are those in [42, 43] which employ a Neyman–Pearson approach and [45] where oracle

and adaptive compound rules were obtained. Compound rules are characterized by information borrowing from each of the genes, so a decision function for a specific gene utilizes information from other genes. Decision-theoretic and Bayesian approaches were also implemented in [10, 17, 26, 33, 35]. More recently, [11] argues for separate subclass analysis, while [13] proposed use of external covariates, with the procedures having a Bayes and empirical Bayes flavor.

The main goal of this paper is to develop better multiple testing procedures controlling weak FWER, strong FWER and FDR by taking into account the individual powers of the tests. We focus on the most fundamental setting where the null and alternative hypotheses for each gene are both simple. This is also the setting in [29]. This admits, as starting point, the Neyman–Pearson most powerful (MP) test for each pair of hypotheses. Each MP test will have a power, but we will see that it is beneficial to look at each of these powers as function of their MP test's size, their so-called receiver operating characteristic (ROC) function.

The paper proceeds as follows. Section 2 presents the decision-theoretic elements. Section 3 reviews and reexamines MP tests, p-value statistics and ROC functions. Section 4 develops the optimal weak FWER-controlling procedure, with existence and uniqueness established in Section 4.2. Section 4.3 analytically describes the procedure for differentiable ROC functions. Section 4.4 provides a concrete example using normal distributions, while Section 4.5 discusses a size-investing strategy for optimality. Section 5 discusses limitations, extensions and connections: Section 5.1 deals with the restriction to the class of simple procedures; Section 5.2 deals with extensions to the composite hypotheses setting in the presence of the monotone likelihood ratio (MLR) property; and Section 5.3 relates the optimal procedure to weighted p-value based procedures. Section 6 develops an improved procedure which strongly controls the FWER, whereas Section 7 develops an improved procedure which controls FDR. The development of these new procedures is anchored on the weak FWER-controlling optimal procedure. We establish that the sequential Šidák and BH procedures are special cases of these more general procedures. Section 8 provides a modest simulation study demonstrating that the new FDR-controlling procedure improves on the BH procedure. Section 9 contains a summary and some concluding remarks.

To manage the length of the paper and provide more focus on the main ideas and results, technical proofs of lemmas, propositions, theorems and corollaries are all gathered in the supplemental article [28].

2. Mathematical setting. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $\mathcal{M} = \{1, 2, \ldots, M\}$ an index set with M a known positive integer. For each $m \in \mathcal{M}$, let $X_m : (\Omega, \mathcal{F}) \to (\mathcal{X}_m, \mathcal{B}_m)$, $\mathcal{X}_m$ some space with σ-field of subsets $\mathcal{B}_m$. Form the product space $(\mathcal{X}, \mathcal{B})$ with $\mathcal{X} = \times_{m\in\mathcal{M}}\mathcal{X}_m$ and $\mathcal{B} = \sigma(\times_{m\in\mathcal{M}}\mathcal{B}_m)$ so $X = (X_1, X_2, \ldots, X_M) : (\Omega, \mathcal{F}) \to (\mathcal{X}, \mathcal{B})$. The probability measure of X is $Q = \mathbf{P}X^{-1}$, while the (marginal) probability measure of X_m is $Q_m = PX_m^{-1}$. For each $m \in \mathcal{M}$, let Q_{m0} and Q_{m1} be two known probability measures on $(\mathcal{X}_m, \mathcal{B}_m)$. We

assume that $Q \in \mathcal{Q}$, a class of probability measures on $(\mathcal{X}, \mathcal{B})$ with marginal probability measure $Q_m \in \{Q_{m0}, Q_{m1}\}$ for each $m \in \mathcal{M}$. Let $\theta = (\theta_1, \ldots, \theta_M) : \mathcal{Q} \to \Theta \equiv \{0,1\}^M$ with $\theta_m(Q) = I\{Q_m = Q_{m1}\}$, $I\{\cdot\}$ denoting indicator function. Define, for each $Q \in \mathcal{Q}$, the subcollections $\mathcal{M}_0 \equiv \mathcal{M}_0(Q) = \{m \in \mathcal{M} : \theta_m(Q) = 0\}$ and $\mathcal{M}_1 \equiv \mathcal{M}_1(Q) = \{m \in \mathcal{M} : \theta_m(Q) = 1\}$. In this paper, we shall impose an *independence condition* given by:

CONDITION (I). $(X_m, m \in \mathcal{M}_0(Q))$ is an independent collection of random entities, that is, $\forall B_m \in \mathcal{B}_m$, $Q(\times_{m \in \mathcal{M}_0(Q)} B_m) = \prod_{m \in \mathcal{M}_0(Q)} Q_m(B_m)$.

However, the collection $(X_m, m \in \mathcal{M}_1(Q))$ need not be an independent collection, but it is independent of $(X_m, m \in \mathcal{M}_0(Q))$. Two extreme subcollections of $\mathcal{Q}$ are $\mathcal{Q}_0 = \{Q \in \mathcal{Q} : \theta_m(Q) = 0, \forall m \in \mathcal{M}\}$ and $\mathcal{Q}_1 = \{Q \in \mathcal{Q} : \theta_m(Q) = 1, \forall m \in \mathcal{M}\}$. By Condition (I), $\mathcal{Q}_0$ is a singleton set, Q_0 will denote its element; while $\mathcal{Q}_1$ need not be a singleton set. The decision problem is to determine $\mathcal{M}_0(Q)$ and $\mathcal{M}_1(Q)$ based on X, which is equivalent to simultaneously testing the M pairs of hypotheses $H_{m0} : Q_m = Q_{m0}$ versus $H_{m1} : Q_m = Q_{m1}$ for $m \in \mathcal{M}$.

We adopt a decision-theoretic framework similar to [33]. The *action space* is $\mathcal{A} = \{0,1\}^M$ with generic element $a = (a_1, a_2, \ldots, a_M)^t \in \mathcal{A}$ with $a_m = 0(1)$ meaning H_{m0} is accepted (rejected). The *parameter space* is $\mathcal{Q}$, though the effective parameter space is $\Theta = \{0,1\}^M$ with generic element $\theta = (\theta_1, \theta_2, \ldots, \theta_M)^t$. We introduce several *loss functions*, $L : \mathcal{A} \times \mathcal{Q} \to \Re_+$, defined via

$$L_0(a, Q) = I\{a^t(1 - \theta(Q)) \geq 1\}; \tag{2.1}$$

$$L_1(a, Q) = \left[\frac{a^t(1 - \theta(Q))}{a^t 1}\right] I\{a^t 1 > 0\}; \tag{2.2}$$

$$L_2(a, Q) = (1 - a)^t \theta(Q), \tag{2.3}$$

with the convention that $0/0 = 0$ and 1 is an $M \times 1$ vector of 1's. The loss function $L_0(a, Q)$ equals 1 if and only if at least one false discovery is committed. The loss $L_1(a, Q)$ is the *false discovery proportion*, being the ratio between the number of false discoveries and the number of discoveries; whereas the loss $L_2(a, Q)$ is the *number of missed discoveries* being the number of true alternative hypotheses that were not discovered. We focus on this missed discovery number since the relevant question is how many correct alternatives $[\theta(Q)^t 1]$ were missed by using the action a? See also [29] which essentially uses this loss function to induce their power metric. Other types of losses, such as the false negative proportion with $(a, Q) \mapsto [(1 - a)^t \theta(Q)]/[(1 - a)^t 1] I\{(1 - a)^t 1 > 0\}$, have also been considered; see [15, 33].

A *nonrandomized* multiple decision function (MDF) is a $\delta : (\mathcal{X}, \mathcal{B}) \to (\mathcal{A}, \sigma(\mathcal{A}))$, where $\sigma(\mathcal{A})$ is the power set of $\mathcal{A}$. Such an MDF may be represented by $\delta(x) = (\delta_1(x), \delta_2(x), \ldots, \delta_M(x))^t$, where $\delta_m(x) \in \{0,1\}$. In general, each δ_m could be made to depend on the full data x instead of just x_m. We denote by $\mathcal{D}$ the

class of all nonrandomized MDFs. A *randomized* MDF may also be considered. Denote by $\mathcal{P}(\mathcal{A})$ the space of all probability measures over $(\mathcal{A}, \sigma(\mathcal{A}))$. A randomized MDF is a $\delta^* : (\mathcal{X}, \mathcal{B}) \to (\mathcal{P}(\mathcal{A}), \sigma(\mathcal{P}(\mathcal{A})))$. For a realization $X = x$, an action is chosen from $\mathcal{A}$ according to the probability measure $\delta^*(x)$. Denote by $\mathcal{D}^*$ the space of all randomized MDFs. Clearly, $\mathcal{D} \subset \mathcal{D}^*$. By augmenting data X with a randomizer $U \sim U(0, 1)$ which is independent of X, randomized MDFs could be made nonrandomized with respect to the *augmented data* (X, U). Henceforth, $\mathcal{D}$ represents all *non*randomized MDFs $\delta(X, U)$'s based on (X, U).

For brevity of notation, $\mathbf{P}_Q\{f(X, U) \in B\}$ and $E_Q\{f(X, U)\}$ represent probability and expectation with respect to (X, U) with $X \sim Q$, $U \sim U(0, 1)$ and X and U independent. For $\delta \in \mathcal{D}$ and the loss functions defined earlier, we have the *risk functions*

$$R_0(\delta, Q) = E_Q\{L_0(\delta(X, U), Q)\}; \tag{2.4}$$

$$R_1(\delta, Q) = E_Q\{L_1(\delta(X, U), Q)\}; \tag{2.5}$$

$$R_2(\delta, Q) = E_Q\{L_2(\delta(X, U), Q)\}. \tag{2.6}$$

Given a $\delta = (\delta_1, \delta_2, \ldots, \delta_M)^t$, let $\pi_\delta(Q) = (\pi_{\delta_1}(Q), \pi_{\delta_2}(Q), \ldots, \pi_{\delta_M}(Q))^t$ with $\pi_{\delta_m}(Q) = E_Q\{\delta_m(X, U)\}$ be its vector of power functions. Then (2.6) becomes $R_2(\delta, Q) = (1 - \pi_\delta(Q))^t\theta(Q)$. In terms of these risk functions, for $\delta \in \mathcal{D}$, its weak FWER is $\mathrm{FWER}(\delta) = R_0(\delta, Q_0)$. If each δ_m depends only on X_m and U, by Condition (I),

$$\mathrm{FWER}(\delta) = 1 - E\left\{\prod_{m\in\mathcal{M}} [1 - \mathbf{P}_{Q_{m0}}\{\delta_m(X_m, U) = 1|U\}]\right\}, \tag{2.7}$$

where the expectation is with respect to U. When $Q = Q_0$ and with the mth component δ_m^* of the randomized MDF depending only on X_m, an alternative formulation is to have $U = (U_1, U_2, \ldots, U_M)$ a vector of i.i.d. $U(0, 1)$ variables which is independent of the X_m's. The mth component may then be redefined via $\delta_m(X_m, U_m) = I\{U_m \le \delta_m^*(X_m)\}$. Then (2.7) becomes $\mathrm{FWER}(\delta) = 1 - \prod_{m\in\mathcal{M}}[1 - \mathbf{P}_{Q_{m0}}\{\delta_m(X_m, U_m) = 1\}]$.

The risk function $R_1(\delta, Q)$ is the false discovery *rate* (FDR) of δ at Q [1]; while the risk function $R_2(\delta, Q)$ will be called the missed discovery *rate* (MDR) of δ at Q. The adjective "rate" is somewhat misleading since $R_2(\delta, Q)$ takes values in $[0, |\mathcal{M}_1(Q)|]$ instead of $[0, 1]$; however, this does not cause difficulty since, given the true underlying probability measure Q of X, $|\mathcal{M}_1(Q)|$ is constant. This risk is related to the expected number of true positives (ETP), an error measure used in [38, 42], via $\mathrm{ETP}(\delta, Q) = |\mathcal{M}_1(Q)| - R_2(\delta, Q)$.

To find an optimal MDF *weakly* controlling FWER in a subclass $\mathcal{D}_0 \subseteq \mathcal{D}$, a threshold $\alpha \in (0, 1)$ is specified and then we seek a $\delta^* \in \mathcal{D}_0$ with $R_0(\delta^*, Q_0) = \mathrm{FWER}(\delta^*) \le \alpha$, and such that for any $\delta \in \mathcal{D}_0$ satisfying $R_0(\delta, Q_0) = \mathrm{FWER}(\delta) \le \alpha$, we have $\sup_{Q\in\mathcal{Q}} R_2(\delta^*, Q) \le \sup_{Q\in\mathcal{Q}} R_2(\delta, Q)$. This criterion has a minimax

flavor. One may require only that $R_2(\delta^*, Q^*) \le R_2(\delta, Q^*)$ where Q^* is the true, but unknown, probability law of X; but this may be too strong to preclude a solution to the optimization problem. However, see [42] for a situation with a different type I error and where an optimal, albeit an oracle, solution for minimizing $R_2(\delta, Q^*)$ is possible. Observe that for $\delta \in \mathcal{D}$, by using the representation of $R_2(\delta, Q)$ in terms of the powers, $\sup_{Q\in\mathcal{Q}} R_2(\delta, Q) = \sup_{Q\in\mathcal{Q}_1} R_2(\delta, Q) = M - \inf_{Q\in\mathcal{Q}_1} \sum_{m\in\mathcal{M}} \pi_{\delta_m}(Q)$. The optimality condition on the MDR amounts therefore to maximizing $\sum_{m\in\mathcal{M}} \pi_{\delta_m}(Q_{m1})$. Interestingly, if we had standardized the loss function $L_2(a, Q)$ to take values in $[0, 1]$ via division by $|\mathcal{M}_1(Q)| = \theta(Q)^t 1$, the minimax justification does not carry through!

For *strong* FWER control, we seek a compound MDF, $\delta^* \in \mathcal{D}$, with $R_0(\delta^*, Q^*) \le \alpha$ *whatever* the true, but unknown, probability law Q^* of X is, and with $\sum_{m\in\mathcal{M}} \pi_{\delta_m^*}(Q_{m1})$ large, possibly maximal, among all $\delta \in \mathcal{D}$ satisfying $R_0(\delta, Q^*) \le \alpha$. For (strong) FDR-control, a threshold $q^* \in (0, 1)$ is specified and we seek a compound MDF, $\delta^* \in \mathcal{D}$, such that, *whatever* Q^* is, $R_1(\delta^*, Q^*) \le q^*$, and with $\sum_{m\in\mathcal{M}} \pi_{\delta_m^*}(Q_{m1})$ large, possibly maximal, among all $\delta \in \mathcal{D}$ satisfying $R_1(\delta, Q^*) \le q^*$. For discussion of weak and strong control, refer to [6, 7]. Discussion of optimality in multiple testing can be found in [25] where maximin optimality results are established for some step-down and step-up MTPs.

3. Revisiting MP tests and *p*-value statistics. An MDF $\delta \in \mathcal{D}$ whose mth component δ_m depends only on (X_m, U_m) for every $m \in \mathcal{M}$ is called simple; otherwise, it is compound. The subclass of simple MDFs, denoted by $\mathcal{D}_0$, will be our initial focus in searching for an optimal weak FWER-controlling MDF. The resulting optimal MDF will then anchor our search for strong FWER- and FDR-controlling compound MDFs. Before implementing this program, we introduce the unifying concept of decision processes.

3.1. *Decision processes*, *ROC functions*, *p-value statistics*. First, a brief review. Let $X : (\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{B})$ and $Q = \mathbf{P}X^{-1}$. Based on X, consider testing the pair of hypotheses $H_0 : Q = Q_0$ versus $H_1 : Q = Q_1$, where Q_0 and Q_1 are two probability measures on $(\mathcal{X}, \mathcal{B})$. Let q_0 and q_1 be versions of the densities of Q_0 and Q_1 with respect to some fixed dominating measure ν, for example, $\nu = Q_0 + Q_1$. Recall that a test or decision function is a $\delta : (\mathcal{X}, \mathcal{B}) \to ([0, 1], \sigma[0, 1])$, with $\sigma[0, 1]$ the Borel sigma-field on $[0, 1]$. Given $X = x$, $\delta(x)$ is the probability of deciding in favor of H_1. Its size is $\alpha_\delta = E_{Q_0}\delta(X)$; it is of level $\alpha \in [0, 1]$ if $\alpha_\delta \le \alpha$. Its power is $\pi_\delta = E_{Q_1}\delta(X)$. δ^* is most powerful (MP) of level α if $\alpha_{\delta^*} \le \alpha$ and for all δ with $\alpha_\delta \le \alpha$, we have $\pi_{\delta^*} \ge \pi_\delta$.

DEFINITION 3.1. A collection $\Delta = \{\delta_\eta : \eta \in [0, 1]\}$ of test functions such that, a.e. $[Q]$, $\delta_0(x) = 0$, $\delta_1(x) = 1$ and $\eta \mapsto \delta_\eta(x)$ is nondecreasing and right-continuous, is a decision process. Its *size* function is $A_\Delta : [0, 1] \to [0, 1]$ and its *power* function is $\rho_\Delta : [0, 1] \to [0, 1]$, where $A_\Delta(\eta) = \alpha_{\delta_\eta} = E_{Q_0}\delta_\eta(X)$ and

$\rho_\Delta(\eta) = \pi_{\delta_\eta} = E_{Q_1}\delta_\eta(X)$. Its *receiver operating characteristic* (ROC) curve is $\mathrm{ROC}(\Delta) \equiv \mathrm{Graph}\{(A_\Delta(\eta), \rho_\Delta(\eta)) : \eta \in [0,1]\}$. If $A_\Delta(\eta) = \eta$ for all $\eta \in [0,1]$, $\eta \mapsto \rho_\Delta(\eta)$ is the *ROC function* of Δ.

The use of the phrase *power function* in Definition 3.1 is atypical since we are not viewing this as a function of a parameter as is the usual meaning of this phrase. However, for lack of a better name, we shall adopt this terminology. In the sequel, δ_η and $\delta(\eta)$ will be used interchangeably to also represent $\delta(\cdot;\eta)$.

Let $L : (\mathcal{X}, \mathcal{B}) \to (\Re_+, \sigma(\Re_+))$ be a version of the likelihood ratio function: $L(x) = q_1(x)/q_0(x)$ a.e. $[\nu]$. Let $G_0(\cdot)$ and $G_1(\cdot)$ be the distribution functions of $L(X)$ when $\mathcal{L}(X) = Q_0$ and $\mathcal{L}(X) = Q_1$, where $\mathcal{L}(X)$ is probability measure of X. For a monotone nondecreasing right-continuous function $M(\cdot)$ from $\Re$ into $\Re$, let $M^{-1}(r) = \inf\{x \in \Re : M(x) \geq r\}$ and $\Delta M(r) = M(r) - M(r-)$. By the Neyman–Pearson fundamental lemma [27], the MP test function of level η for testing H_0 versus H_1 is

$$\delta^*(X;\eta) \equiv \delta^*_\eta = I\{L(X) > c(\eta)\} + \gamma(\eta) I\{L(X) = c(\eta)\}, \tag{3.1}$$

where $c(\eta) = G_0^{-1}(1-\eta)$ and $\gamma(\eta) = (G_0(c(\eta)) - (1-\eta))/\Delta G_0(c(\eta))$. Let $U \sim U(0,1)$ be independent of X. Redefine δ^* via $\delta^{**}_\eta \equiv \delta^{**}(X,U;\eta) = I\{U \leq \delta^*(X;\eta)\}$, which is nonrandomized w.r.t. (X,U). In essence, with the aid of an auxiliary randomizer U, the MP test could always be made nonrandomized. The decision process formed from these MP tests, given by

$$\Delta^* = \{\delta^*_\eta : \eta \in [0,1]\} = \{\delta^{**}_\eta : \eta \in [0,1]\}, \tag{3.2}$$

is called the most powerful (MP) decision process. The power (at $Q = Q_1$) of the MP test δ^*_η or δ^{**}_η is

$$\rho_{\Delta^*}(\eta) \equiv \pi_{\delta^*_\eta} = \pi_{\delta^{**}_\eta} = 1 - G_1(c(\eta)) + \gamma(\eta)\Delta G_1(c(\eta)). \tag{3.3}$$

It is well known [24] that $\pi_{\delta^*_\eta} < 1$ implies $\alpha_{\delta^*_\eta} = \eta$. We denote by A_{Δ^*} and ρ_{Δ^*} the size and power functions of Δ^*. If $\pi_{\delta^*_\eta} < 1$ for all $\eta < 1$, then $\eta \mapsto \rho_{\Delta^*}(\eta)$ is the ROC function of Δ^*. We present below some important properties of this function.

Before stating the proposition, we reiterate that all formal proofs of propositions, theorems, lemmas and corollaries are in the supplemental article [28].

PROPOSITION 3.1. *The function $\rho_{\Delta^*} : [0,1] \to [0,1]$ in (3.3) is concave, continuous and nondecreasing. Furthermore, $\rho_{\Delta^*}(\eta) \geq \eta$ and it is strictly increasing on the set $\mathcal{N}_< \equiv \{\eta \in [0,1] : \rho_{\Delta^*}(\eta) < 1\}$.*

DEFINITION 3.2. Let $\Delta = \{\delta_\eta : \eta \in [0,1]\}$ be a decision process, where $\delta_\eta : (\mathcal{X} \times [0,1], \mathcal{B} \otimes \sigma[0,1]) \to (\{0,1\}, \sigma\{0,1\})$. Its (randomized) p-value statistic is $S_\Delta : (\mathcal{X} \times [0,1], \mathcal{B} \otimes \sigma[0,1]) \to ([0,1], \sigma[0,1])$ with $S_\Delta(x,u) = \inf\{\eta \in [0,1] : \delta_\eta(x,u) = 1\}$.

When $\forall(\eta, x, u): \delta_\eta(x, u) = \delta_\eta(x)$, then $S_\Delta(X, U)$ is the usual p-value statistic. See also [4] for a more specialized definition of a randomized p-value statistic. We refer the reader to [18] for properties of this p-value statistic and its use in existing FDR-controlling procedures.

PROPOSITION 3.2. *Let* $\Delta = \{\delta_\eta : \eta \in [0, 1]\}$ *be a decision process with p-value statistic* S_Δ. *Then, for all* $s \in [0, 1]$, $H_0(s) \equiv \mathbf{P}_{Q_0}(S_\Delta \le s) = A_\Delta(s)$ *and* $H_1(s) \equiv \mathbf{P}_{Q_1}(S_\Delta \le s) = \pi_{\delta(s)} = \rho_\Delta(s)$. *Consequently,* $S_\Delta \sim U[0, 1]$ *under* $\mathcal{L}(X) = Q_0$ *if and only if* $\forall \eta \in [0, 1]: A_\Delta(\eta) = \eta$.

4. Optimal *weak* FWER control. Return now to the multiple decision problem in Section 2. We extend the notion of decision processes to the multiple decision setting.

DEFINITION 4.1. A collection $\mathbf{\Delta} = (\Delta_m : m \in \mathcal{M})$, where $\Delta_m = (\delta_m(\eta) : \eta \in [0, 1])$ is a decision process on $(\mathcal{X} \times [0, 1]^M, \mathcal{B} \otimes \sigma[0, 1]^M)$, is a multiple decision process (MDP). It is simple if each Δ_m is simple; otherwise, it is compound. When simple its multiple decision size function is $\mathbf{A}_{\mathbf{\Delta}} = (A_{\Delta_m} : m \in \mathcal{M})$ and its multiple decision ROC function is $\boldsymbol{\rho}_{\mathbf{\Delta}} = (\rho_{\Delta_m} : m \in \mathcal{M})$, where A_{Δ_m} and ρ_{Δ_m} are the size and ROC functions of Δ_m.

4.1. *Optimization problem.* Let $\mathbf{\Delta}$ be a simple MDP. Then, a multiple decision size vector $\boldsymbol{\eta} = (\eta_m : m \in \mathcal{M}) \in \mathcal{N} \equiv [0, 1]^M$ determines from $\mathbf{\Delta}$ an MDF $\delta_{\mathbf{\Delta}}(\boldsymbol{\eta}) = (\delta_m(\eta_m) : m \in \mathcal{M}) \in \mathcal{D}_0$. For this MDF, $\mathrm{FWER}(\delta_{\mathbf{\Delta}}(\boldsymbol{\eta})) = 1 - \prod_{m \in \mathcal{M}}[1 - A_{\Delta_m}(\eta_m)]$ and $R_2(\delta_{\mathbf{\Delta}}(\boldsymbol{\eta}), Q_1) = M - \sum_{m \in \mathcal{M}} \rho_{\Delta_m}(\eta_m)$ for $Q_1 \in \mathcal{Q}_1$. Fix an FWER-threshold $\alpha \in (0, 1)$. Suppose there exists a multiple decision size vector $\boldsymbol{\eta}^*_{\mathbf{\Delta}}(\alpha) \in \mathcal{N}$ such that

$$\boldsymbol{\eta}^*_{\mathbf{\Delta}}(\alpha) = \arg\max_{\boldsymbol{\eta} \in \mathcal{N}} \Big\{ \sum_{m \in \mathcal{M}} \rho_{\Delta_m}(\eta_m) : \prod_{m \in \mathcal{M}} [1 - A_{\Delta_m}(\eta_m)] \ge 1 - \alpha \Big\}.$$

Then, $\mathbf{A}_\Delta(\boldsymbol{\eta}^*_\Delta(\alpha)) = (A_{\Delta_m}(\eta^*_{\mathbf{\Delta},m}(\alpha)) : m \in \mathcal{M})$ is the optimal multiple decision size vector for weak FWER control at α associated with the simple MDP $\mathbf{\Delta}$. The associated optimal simple MDF is $\delta_{\mathbf{\Delta}}(\mathbf{A}_\Delta(\boldsymbol{\eta}^*_{\mathbf{\Delta}}(\alpha)))$.

But, since H_{m0} and H_{m1} are both simple, then there exists a simple most powerful MDP, $\mathbf{\Delta}^* = (\Delta^*_m : m \in \mathcal{M})$, where $\Delta^*_m = (\delta^*_m(\eta) : \eta \in [0, 1])$ with $\delta^*_m(\eta)$ being the simple Neyman–Pearson MP test function of size η for H_{m0} versus H_{m1}. Consider the simple MDF obtained from $\mathbf{\Delta}^*$ given by $(\delta^*_m(A_{\Delta_m}(\boldsymbol{\eta}^*_{\mathbf{\Delta},m}(\alpha))) : m \in \mathcal{M})$. This will satisfy the FWER constraint, and by virtue of the MP property of each $\delta^*_m(A_{\Delta_m}(\boldsymbol{\eta}^*_{\mathbf{\Delta},m}(\alpha)))$ for each $m \in \mathcal{M}$,

$$\sum_{m \in \mathcal{M}} \rho_{\Delta^*_m}(A_{\Delta_m}(\boldsymbol{\eta}^*_{\mathbf{\Delta},m}(\alpha))) \ge \sum_{m \in \mathcal{M}} \rho_{\Delta_m}(A_{\Delta_m}(\boldsymbol{\eta}^*_{\mathbf{\Delta},m}(\alpha))).$$

Thus, in searching for the optimal weak FWER-controlling simple MDF, it suffices to restrict to the simple most powerful MDP $\boldsymbol{\Delta}^*$. Without loss of generality (wlog), we may assume $A_{\Delta_m^*}(\eta) = \eta$ for $m \in \mathcal{M}$ and $\eta \in [0, 1]$. The optimization problem reduces to finding a $\boldsymbol{\eta}^*_{\boldsymbol{\Delta}^*}(\alpha) \in \mathcal{N}$ satisfying

$$\boldsymbol{\eta}^*_{\boldsymbol{\Delta}^*}(\alpha) = \arg\max_{\boldsymbol{\eta}\in\mathcal{N}} \Big\{ \sum_{m\in\mathcal{M}} \rho_{\Delta_m^*}(\eta_m) : \prod_{m\in\mathcal{M}} (1-\eta_m) \geq 1-\alpha \Big\}. \tag{4.1}$$

The optimal *weak* FWER-controlling simple MDF is then

$$\delta^*_W(\alpha) \equiv \big(\delta^*_m(\eta^*_{\boldsymbol{\Delta}^*,m}(\alpha)) : m \in \mathcal{M}\big). \tag{4.2}$$

Two well-known and conventional choices for the size vector $\boldsymbol{\eta} = (\eta_m : m \in \mathcal{M})$ which satisfy the weak FWER constraint are the Šidák sizes $\eta_m = \eta_m(\alpha) = 1 - (1-\alpha)^{1/M}$ and the Bonferroni-adjusted sizes $\eta_m = \eta_m(\alpha) = \alpha/M$. The former requires the independence Condition (I) and is sharp, the latter is conservative but does not require Condition (I). Both ignore possible differences in power traits of the individual test functions.

4.2. *Existence and uniqueness of optimal size vector.* We establish the existence of an optimal multiple decision size vector for weak FWER control within the class $\mathcal{D}_0$. As pointed out in Section 4.1, it suffices to look for the optimal weak FWER-controlling simple MDF by starting with the most powerful simple MDP $\boldsymbol{\Delta}^* = (\Delta_m^* : m \in \mathcal{M})$. For brevity, $\rho_m \equiv \rho_{\Delta_m^*}$ and $A_m(\eta) \equiv A_{\Delta_m^*}(\eta) = \eta$. Recall that $\mathcal{N} = [0, 1]^M$, the multiple decision size space. In a nutshell, the existence of an optimal multiple decision size vector for weak FWER control exploits convexity properties of relevant subsets of $\mathcal{N}$. This is formalized by establishing a sequence of propositions which are presented below. For $\alpha \in [0, 1]$, define the weak FWER constraint set

$$C_\alpha = \begin{cases} \Big\{ \boldsymbol{\eta} \in \mathcal{N} : \sum_{m\in\mathcal{M}} \log(1-\eta_m) \geq \log(1-\alpha) \Big\}, & \text{if } \alpha < 1, \\ \mathcal{N}, & \text{if } \alpha = 1. \end{cases} \tag{4.3}$$

PROPOSITION 4.1. *C_α satisfies* (i) $\boldsymbol{\eta} = \mathbf{0} \in C_\alpha$; (ii) $(\mathbf{0}, \alpha_m) \in C_\alpha$ *for all* $m \in \mathcal{M}$, *where* $(\mathbf{0}, \alpha_m)$ *is the zero-vector with the* m*th element replaced by* α; *and* (iii) *it is convex and closed.*

PROPOSITION 4.2. *For $\boldsymbol{\eta}_0 \in \mathcal{N}$ let $U(\boldsymbol{\eta}_0) = \{\boldsymbol{\eta} \in \mathcal{N} : \eta_m \geq \eta_{0m}, \forall m \in \mathcal{M}\}$, the upper set of η_0, and let $UB(C_\alpha) = \{\boldsymbol{\eta} \in \mathcal{N} : C_\alpha \cap U(\boldsymbol{\eta}) = \{\boldsymbol{\eta}\}\}$, the upper boundary set of C_α. Then, for all $\alpha \in [0, 1)$, $UB(C_\alpha) = \{\boldsymbol{\eta} \in \mathcal{N} : \sum_{m\in\mathcal{M}} \log(1 - \eta_m) = \log(1-\alpha)\}$.*

PROPOSITION 4.3. *Let $\mathcal{N}_b \equiv \{\boldsymbol{\eta} \in \mathcal{N} : \sum_{m\in\mathcal{M}} \rho_m(\eta_m) \geq Mb\}$ for $b \in [0, 1]$. Then $\{\mathcal{N}_b : b \in [0, 1]\}$ satisfies* (i) $\boldsymbol{\eta} = \mathbf{1} \in \mathcal{N}_b$, (ii) *it is closed and convex, and* (iii) $\mathcal{N} = \mathcal{N}_0 \supseteq \mathcal{N}_{b_1} \supseteq \mathcal{N}_{b_2}$ *for* $0 \leq b_1 \leq b_2 \leq 1$.

PROPOSITION 4.4. *Let* $B_\alpha = \{b \in [0,1] : \mathcal{N}_b \cap C_\alpha \neq \varnothing\}$ *for* $\alpha \in [0,1)$ *and let* $b_\alpha^* = \sup B_\alpha$. *Then* $B_\alpha = [0, b_\alpha^*]$.

Building on these intermediate results, the existence of an optimal weak FWER-controlling multiple decision size vector is obtained.

THEOREM 4.1 (Existence). *Let* $\alpha \in [0,1)$. *Then* $C_\alpha \cap \mathcal{N}_{b_\alpha^*} \neq \varnothing$. *Furthermore,* $\eta \in \mathcal{N}$ *is a weak FWER-*α *optimal multiple decision size vector if and only if* $\eta \in C_\alpha \cap \mathcal{N}_{b_\alpha^*}$.

Theorem 4.1 guarantees existence of an optimal weak FWER multiple decision size vector, but it does not address whether the solution is unique. We present a result on this issue in the following theorem.

THEOREM 4.2 (Uniqueness). *Let* $\alpha \in [0,1)$ *and define* $C_\alpha(m) = \{\eta_m \in [0,1] : \boldsymbol{\eta} \in C_\alpha\}$, *called the* m*th section of* C_α. *If, for all* $m \in \mathcal{M}$, *the mapping* $\eta_m \mapsto \rho_m(\eta_m)$ *is strictly increasing on* $C_\alpha(m)$, *then the optimal weak FWER-*α *multiple decision size vector is unique and it is the* $\boldsymbol{\eta}^*$ *satisfying* $C_\alpha \cap \mathcal{N}_{b_\alpha^*} = \{\boldsymbol{\eta}^*\}$.

It is easy to see that a sufficient condition for uniqueness of the optimal size vector is that, for all $m \in \mathcal{M}$, $\eta_m \in [0, \sup C_\alpha(\eta_m)) \Rightarrow \rho_m(\eta_m) < 1$. Nonuniqueness may occur with nonregular families of densities, for example, uniform or shifted exponential, where the power of the MP test may equal one even though its size is still less than one. It occurs if the decision processes in the MDP do *not* satisfy the condition that $\forall \eta \in [0,1], \forall m \in \mathcal{M}, A_m(\eta) = \eta$, which is the case with discrete data or when using nonparametric rank-based test functions with randomization not permitted.

4.3. *Finding optimal size vector.* Generally, without differentiability of the ROC functions as in the case with discrete distributions, linear or nonlinear programming methods are needed to obtain the optimal solution. In the case, however, where the ROC functions are twice-differentiable, the optimal size vector is in a more explicit form.

THEOREM 4.3. *Let* $\boldsymbol{\Delta}^* = (\Delta_m^*, m \in \mathcal{M})$ *be the MP MDP, and assume that the ROC functions* $\eta_m \mapsto \rho_m(\eta_m)$ *are strictly increasing and twice-differentiable with first and second derivatives* ρ_m' *and* ρ_m'', *respectively. Given* $\alpha \in (0,1)$, *the optimal weak FWER-*α *multiple decision size vector* $\boldsymbol{\eta}^* \equiv \boldsymbol{\eta}_{\boldsymbol{\Delta}^*}^*(\alpha) = (\eta_m^*(\alpha), m \in \mathcal{M})$ *is the* $\boldsymbol{\eta} \in \mathcal{N}$ *satisfying* (i) *for some* $\lambda \in \Re_+$, $\forall m \in \mathcal{M}$, $\rho_m'(\eta_m)(1-\eta_m) = \lambda$ *and* (ii) $\sum_{m \in \mathcal{M}} \log(1-\eta_m) = \log(1-\alpha)$.

A question arises as to whether the optimal sizes are monotonic in α. Such a property is desirable since it will imply that if at FWER size α_1 we have

$\delta_m(\eta_m(\alpha_1)) = 1$, then at an FWER size α_2 with $\alpha_2 > \alpha_1$, we will also have $\delta_m(\eta_m(\alpha_2)) = 1$. This property will also be critical in proving a martingale property needed for the development of the FDR-controlling procedure. This issue is the content of the following proposition.

PROPOSITION 4.5. *Assume the conditions of Theorem 4.3. Then, for each $m \in \mathcal{M}$, the mapping $\alpha \mapsto \eta_m^*(\alpha)$ is nondecreasing and continuous.*

4.4. *Gaussian example for weak FWER control.* For $m \in \mathcal{M}$, let $X_m \sim N(\mu_m, \sigma_{m0}^2)$, where the μ_m's are unknown and σ_{m0}^2's are known. Consider the multiple hypotheses testing problem $H_{m0}: \mu_m = \mu_{m0}$ and $H_{m1}: \mu_m = \mu_{m1}$ with $\mu_{m0} < \mu_{m1}$ for $m \in \mathcal{M}$. The MP test of size η_m for H_{m0} versus H_{m1} is $\delta_m^*(X_m; \eta_m) \equiv \delta_m^*(\eta_m) = I\{X_m \geq \mu_{m0} + \sigma_{m0}\Phi^{-1}(1-\eta_m)\}$, where $\Phi(\cdot)$ and $\Phi^{-1}(\cdot)$ are the cumulative distribution and quantile functions, respectively, of a standard normal variable. The mth effect size is $\gamma_m = (\mu_{m1} - \mu_{m0})/\sigma_{m0}$, and the ROC function of the decision process $\Delta_m^* = (\delta_m^*(\eta_m) : \eta_m \in [0,1])$ is $\rho_m(\eta_m) \equiv \rho_m(\eta_m; \gamma_m) = \Phi(\gamma_m - \Phi^{-1}(1-\eta_m))$, clearly twice-differentiable with respect to η_m. With $\phi(\cdot)$ the standard normal density function,

$$(\rho_m)'(\eta_m) = \frac{\phi(\gamma_m - \Phi^{-1}(1-\eta_m))}{\phi(\Phi^{-1}(1-\eta_m))}.$$

For fixed $\alpha \in (0,1)$ and γ_m's, consider the mappings $d \mapsto \eta_m(d)$, $m \in \mathcal{M}$, defined implicitly by the equation

$$\frac{\phi(\gamma_m - \Phi^{-1}(1-\eta_m))}{\phi(\Phi^{-1}(1-\eta_m))}(1-\eta_m) - d = 0. \tag{4.4}$$

The optimal value of d, denoted by d^*, solves the equation

$$\sum_{m\in\mathcal{M}} \log\bigl(1-\eta_m(d)\bigr) - \log(1-\alpha) = 0. \tag{4.5}$$

The optimal sizes of the M MP tests are then $\eta_m(d^*)$, $m \in \mathcal{M}$. An `R` [19] implementation of this numerical problem first defines $v_m = 1 - \Phi^{-1}(1-\eta_m)$, so condition (4.4) amounts to solving for $v_m = v_m(d)$ the equation

$$\log\Phi(v_m) + \gamma_m v_m - \log(d) - \gamma_m^2/2 = 0. \tag{4.6}$$

We utilized a Newton–Raphson iteration in solving for v_m's in (4.6) and the `uniroot` routine in the `R` Library to solve for d in (4.5). Upon obtaining $v_m(d)$'s, the $\eta_m(d)$'s are computed via $\eta_m(d) = 1 - \Phi(v_m(d))$.

Figure 1 demonstrates the optimal sizes when $M = 2{,}000$ and for uniformly distributed effect sizes. Observe from the second panel that when the effect size is small, which converts to low power, then the optimal size for the test is also small, but also note that when the effect size is large, which converts to high power, then

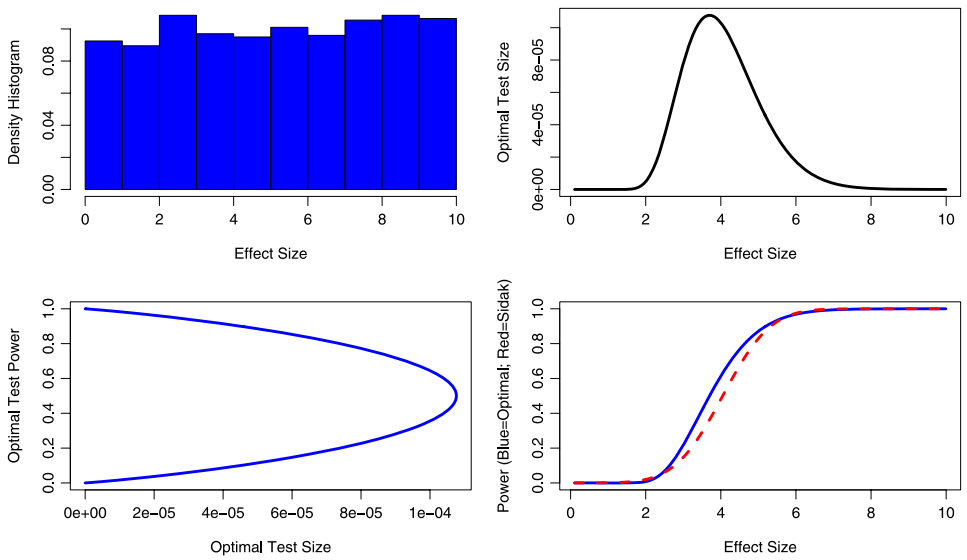

FIG. 1. *Optimal test sizes and powers for* 2,000 *MP tests of hypotheses under normality when the effect sizes were generated from a* uniform$[0.1, 10]$ *distribution. Panel four shows the powers for both the optimal* [*solid black*] *and the Šidák* [*dashed red*] *tests with respect to effect sizes.*

the optimal test size is also small. For the tests with moderate effect sizes or power, then the optimal sizes are higher. This behavior could also be seen by looking at the third panel in the figure which shows the achieved power of the tests at the optimal sizes.

The efficiency of the optimal procedure relative to the Šidák procedure was measured via the ratio (multiplied by 100) of the average power over the M tests, defined by $\sum_{m\in\mathcal{M}} \rho_m(\eta_m)/M$, of the optimal procedure and the average power of the Šidák procedure. The fourth panel in Figure 1 depicts the powers of the resulting tests versus the effect size for both procedures (solid blue = optimal; dashed red = Šidák). For these uniformly-generated effect sizes, the efficiency of the optimal procedure over the Šidák is 103.5%. This efficiency is affected by the vector of effect sizes. For instance, when we change the effect sizes in Figure 1 to be generated from a uniform over $[0.1, 2]$, then the efficiency jumps to 181.7%, though it should also be pointed out that since the effect sizes are small, then the overall powers of both procedures are also small.

4.5. *A size-investing strategy.* In the preceding Gaussian example, as well as in other situations we examined, for example, with exponential and Bernoulli distributions, we observed the phenomenon where, among the M tests, those with low powers (small effect sizes) and those with high powers (large effect sizes) are allocated relatively small sizes in the weak FWER-controlling optimal procedure. The tests with larger sizes are those with moderate powers or effect sizes.

This is a *size-investing strategy* in the multiple hypotheses testing problem, and it has intuitive content. With the overall goal of making more real discoveries while controlling the proportion of false discoveries for a pre-specified, usually small, overall size α, the optimal procedure dictates that not much size should be accorded those tests with either very low or very high powers. The former case will not lead to any discoveries anyway if the size that could be allocated is small, while the latter case will lead to discoveries even if the test sizes are made small. Thus, there is more to be gained by investing larger sizes on those tests that are of moderate powers, and an appropriate tweaking of their test sizes according to condition (i) in Theorem 4.3 improves the ability to achieve more real discoveries. However, this phenomenon is dependent on the magnitude of the overall size. If this overall size is made larger, more leeway ensues to the extent that it may then be more beneficial to allocate more size to those with low powers since those tests with moderate powers, when they had small sizes, may now have larger powers because of the consequent increase in their sizes. The precise and crucial determinant of where the differential sizes should be allocated are the rates of change of the ROC functions, with some size-attenuation. Interesting discussions of size and weight allocation strategies can also be found in [49], where the size allocation was related to the "α-spending" function of [22], in [14] which deals with α-investing in sequential procedures that control expected false discoveries, and in [16, 29] which discuss optimal weights for the p-values.

A tangential real-life manifestation of this strategy occurred during the 2008 American presidential election, with the total resources (financial, manpower, etc.) available to the candidates analogous to the overall size in the multiple testing problem. In the waning days of the campaign, the major candidates, then-Senator Barack Obama of the Democratic Party and Senator John McCain of the Republican Party, focused their campaign efforts, in terms of allocating their financial and manpower resources, in the "battleground states" of North Carolina, Virginia and Pennsylvania, while basically ignoring the "in-the-bag states" of South Carolina, then expected to vote for McCain, and California, then expected to vote for Obama. Also, by virtue of the deep resources of the Obama campaign, it was able to allocate more resources *even* in states that traditionally voted Republican, whereas the McCain campaign, with a relatively smaller war chest, had to "drop" some states (e.g., Michigan) in their campaign. The behaviors of the two camps somehow mirror the size-investing strategy with proper accounting of each campaign's overall resources.

5. Restrictions, extensions and connections.

5.1. *On the restriction to $\mathcal{D}_0$.* The optimization problem for weak FWER control could be construed as limited since we restricted to the subclass $\mathcal{D}_0$ thus leading to an optimal weak FWER-controlling procedure that is still simple. In [42, 45], it was demonstrated that performance is enhanced via compound MDFs.

Examples of compound MDFs are the *estimated* optimal discovery procedure (ODP) in [42, 43], the FDR-controlling procedure in [1], and the oracle-based adaptive MDFs in [45].

Could we immediately start from compound MDFs in the search for an optimal weak FWER-controlling compound MDF? Let us suppose that $\boldsymbol{\delta} = (\delta_m : m \in \mathcal{M})$ is a compound MDF, so δ_m depends on (X, U) and not only on (X_m, U_m). For such an MDF, we have

$$R_0(\boldsymbol{\delta}, Q) = \mathbf{P}_Q\left\{\bigcup_{m\in\mathcal{M}_0(Q)} [\delta_m(X, U) = 1]\right\}. \tag{5.1}$$

Now, even if the independence Condition (I) holds, $(\delta_m(X, U) : m \in \mathcal{M}_0(Q))$ need not be an independent collection. As such no closed-form *exact* expression for $R_0(\boldsymbol{\delta}, Q)$ need exist. The right-hand side in (5.1) could be Bonferroni-bounded by

$$\mathrm{EFP}(\boldsymbol{\delta}, Q) \equiv \sum_{m\in\mathcal{M}_0(Q)} \alpha_{\delta_m}(Q), \tag{5.2}$$

called the expected number of false positives in [42]. Alternatively, if a generalized positive quadrant dependence (PQD) condition holds, with

$$\mathbf{P}_Q\left\{\bigcap_{m\in\mathcal{M}_0(Q)} [\delta_m(X, U) = 0]\right\} \geq \prod_{m\in\mathcal{M}_0(Q)} \mathbf{P}_Q\{\delta_m(X, U) = 0\},$$

then the right-hand side in (5.1) could be upper-bounded by

$$\mathrm{PQD}(\boldsymbol{\delta}, Q) \equiv 1 - \prod_{m\in\mathcal{M}_0(Q)} [1 - \alpha_{\delta_m}(Q)], \tag{5.3}$$

where $\alpha_{\delta_m}(Q) = E_Q\delta_m(X, U)$, the size of δ_m when $m \in \mathcal{M}_0(Q)$. For this compound MDF, its MDR is $R_2(\boldsymbol{\delta}, Q) = \sum_{m\in\mathcal{M}_1(Q)}[1 - \pi_{\delta_m}(Q)]$, where $\pi_{\delta_m}(Q) = E_Q\delta_m(X, U)$ is the power of δ_m when $m \in \mathcal{M}_1(Q)$.

An optimization approach could proceed by putting an upper threshold $\alpha \in (0, 1)$ on either (5.2) or (5.3), and then finding the $\boldsymbol{\delta}$ that minimizes $R_2(\boldsymbol{\delta}, Q)$, or equivalently, maximizes $\mathrm{ETP}(\boldsymbol{\delta}, Q) \equiv \sum_{m\in\mathcal{M}_1(Q)} \pi_{\delta_m}(Q)$, the latter quantity referred to as the expected number of true positives in [42]. The MDFs in [38] and [42] were both obtained through this program. The MDF in [38] is

$$\boldsymbol{\delta}_{\mathrm{SPJ}}(\alpha) = \arg\max_{\boldsymbol{\delta}\in\mathcal{D}_0}\{\mathrm{ETP}(\boldsymbol{\delta}, Q_1) : \mathrm{EFP}(\boldsymbol{\delta}, Q_0) \leq \alpha\}, \tag{5.4}$$

where $Q_0 \in \mathcal{Q}_0$ and $Q_1 \in \mathcal{Q}_1$; whereas the optimal discovery procedure (ODP) in [42] is

$$\boldsymbol{\delta}_{\mathrm{STO}}(\alpha; Q) = \arg\max_{\boldsymbol{\delta}\in\mathcal{D}}\{\mathrm{ETP}(\boldsymbol{\delta}, Q) : \mathrm{EFP}(\boldsymbol{\delta}, Q) \leq \alpha\}, \tag{5.5}$$

where Q is the true probability measure of X. The use of EFP as type I error measure in [42] enabled a calculus of variations optimization to obtain the ODP.

This has a particularly interesting structure when we utilize as its input the vector of p-value statistics $(S_m^*(x_m, u_m) : m \in \mathcal{M})$ from the MP MDP $\boldsymbol{\Delta}^* = (\Delta_m^* : m \in \mathcal{M})$ with multiple decision size function $\mathbf{A}_{\boldsymbol{\Delta}^*}^* = \{(A_m^*(\eta) : \eta \in [0,1]) : m \in \mathcal{M}\}$ and multiple decision ROC function $\boldsymbol{\rho}_{\boldsymbol{\Delta}^*}^* = \{(\rho_m^*(\eta) : \eta \in [0,1]) : m \in \mathcal{M}\}$ and with $A_m^*(\cdot)$ and $\rho_m^*(\cdot)$ both differentiable with derivatives $(A_m^*)'(\cdot)$ and $(\rho_m^*)'(\cdot)$. The significance thresholding function $\mathcal{S} : ([0,1], \sigma[0,1]) \to (\Re, \sigma(\Re))$ utilized in the ODP becomes

$$\mathcal{S}(s; Q) = \frac{\sum_{m \in \mathcal{M}_1(Q)} (\rho_m^*)'(s)}{\sum_{m \in \mathcal{M}_0(Q)} (A_m^*)'(s)}, \tag{5.6}$$

a consequence of Lemma 2 in [42] and Proposition 3.2. The ODP $\boldsymbol{\delta}_{\mathrm{STO}} = (\delta_{m,\mathrm{STO}} : m \in \mathcal{M})$ has a single-thresholding structure with components

$$\delta_{m,\mathrm{STO}}(S_m^*(x_m, u_m); Q) = I\{\mathcal{S}(S_m^*(x_m, u_m); Q) \geq \lambda\}, \qquad m \in \mathcal{M},$$

where $\lambda \in [0, \infty)$ is chosen so the size constraint on $\mathrm{EFP}(\boldsymbol{\delta}_{\mathrm{STO}}(\alpha; Q), Q)$ is approximately satisfied. Observe that each of these components is still of simple-type, unless λ is determined in a data-dependent manner using the full data (x, u). Note also that $\boldsymbol{\delta}_{\mathrm{STO}}$ was derived under complete knowledge of the unknown Q, or more specifically, the sets $\mathcal{M}_0(Q)$ and $\mathcal{M}_1(Q)$, as can be seen in (5.6), hence is referred to as an oracle MDF. For the simple null versus simple alternative hypotheses case, the size functions $A_m^*(\cdot)$'s and the ROC functions $\rho_m^*(\cdot)$'s will be known, but with composite hypotheses they will be unknown. To implement $\boldsymbol{\delta}_{\mathrm{STO}}$, it was proposed in [42, 43] that these unknown quantities, sets, functions, or significance thresholding function, be estimated using the data (x, u). This will make the estimated ODP of compound type. But note that through this plug-in approach the exact optimality property of the ODP need not anymore hold for the estimated version; see also [13, 45]. In contrast, $\boldsymbol{\delta}_{\mathrm{SPJ}}$ is determined only by the two classes of extreme probability measures, $\mathcal{Q}_0$ and $\mathcal{Q}_1$, so the marginal probability measures, Q_m's, are completely known, and not by the unknown true probability measure Q governing X. This fact was criticized in [42] as a "potentially problematic optimality" criterion. More importantly, it should be recognized that both $\boldsymbol{\delta}_{\mathrm{SPJ}}$ and $\boldsymbol{\delta}_{\mathrm{STO}}$ need not be the optimal weak or strong FWER- or FDR-controlling MDFs since the Bonferroni upper bound for $R_0(\boldsymbol{\delta}, Q)$ utilized in their derivations is hardly a sharp upper bound.

The criticism leveled against $\boldsymbol{\delta}_{\mathrm{SPJ}}$ could also be invoked against our optimal weak FWER-controlling procedure since we also relied on a criterion determined only by the extreme classes $\mathcal{Q}_0$ and $\mathcal{Q}_1$. However, note that each component of the optimal weak FWER-controlling multiple decision size vector, and consequently each component of $\boldsymbol{\delta}_W^*(\alpha)$, uses all of the Q_{m0}'s and Q_{m1}'s, analogously to the ODP, though the MDF $\boldsymbol{\delta}_W^*(\alpha)$ is still neither adaptive nor compound. Our development of this simple MDF, which is optimal in the class $\mathcal{D}_0$, is a prelude to our development of adaptive and compound MDFs *strongly*-controlling FWER

and FDR. The MDF $\boldsymbol{\delta}_W^*(\alpha)$ will be the anchor for these FWER and FDR strongly-controlling compound MDFs. These new MDFs are discussed in Section 6 for strong FWER-control and in Section 7 for FDR control. Our approach to obtaining these strongly-controlling MDFs is indirect, whereas that in [42] is direct. There is also an intrinsic difference in the problems considered since our focus is on the type I error risk functions R_0 and R_1, whereas in [38, 42] the simpler type I error metric of EFP was utilized. Looking forward, though our starting point is the optimal weak FWER-controlling simple MDF $\boldsymbol{\delta}_W^*(\alpha)$, there is confidence in the viability of our indirect approach to generate good MDFs since we will establish later that both the sequential Šidák procedure and the BH procedure are special cases of our new MDFs under exchangeability.

5.2. *Families with MLR property.* The initial simplification to the simple null versus simple alternative hypotheses for each $m \in \mathcal{M}$ could be perceived as a limitation because of the need to know the Q_{m1}'s to determine the ROC functions. However, this approach, which was also implemented in [29, 38, 42], is natural and historically-justified by the Neyman–Pearson framework. We surmise that in this multiple decision problem, the solution to the simple null versus simple alternative hypotheses setting will play a prominent role in solving the composite hypotheses setting, since it appears that for an MDF to possess optimality, it will require knowledge, either in exact, approximate, or estimated forms, of the alternative hypotheses distributions. We touch on this aspect in the presence of the monotone likelihood ratio (MLR) property; see [24].

Suppose that for each $m \in \mathcal{M}$, the density function q_m belongs to a one-dimensional parametric family $\mathcal{F}_m = \{q_m(\cdot;\xi_m) : \xi_m \in \Gamma_m \subset \Re\}$ which possesses the MLR property. A typical pair of hypotheses to be tested would be $H_{m0}^* : \xi_m \leq \xi_{m0}$ versus $H_{m1}^* : \xi_m > \xi_{m0}$, where ξ_{m0} is known. With the MLR property, a uniformly most powerful (UMP) test function $\delta_m(X_m, U_m; \eta_m)$ of size η_m exists, with this UMP test identical to the MP test of size η_m for the simple null hypothesis $H_{m0} : \xi_m = \xi_{m0}$ versus the simple alternative hypothesis $H_{m1} : \xi_m = \xi_{m1}$, with $\xi_{m1} > \xi_{m0}$. When dealing with the single-pair hypothesis testing problem, recall that exact knowledge of the value of ξ_1 is not necessary since the critical constants of the size-η MP test for $H_0 : \xi = \xi_0$ versus $H_1 : \xi = \xi_1$ can be made independent of ξ_1. In contrast, for the multiple decision problem, to determine the optimal size allocations for each of the M MP tests, the powers of the tests at the ξ_{m1}'s are required, hence the need to know the values of the ξ_{m1}'s. When M is large, such information may not be so forthcoming. The default procedure is the simplistic approach of simply assuming that the (Q_{m0}, Q_{m1}) is invariant in m, which is the exchangeable setting. However, this exchangeable assumption is most likely wrong as a consequence of varied effect sizes or different test functions utilized. See, for instance, [11] for real situations where exchangeability do not hold. We propose two possible solutions to this dilemma.

The first approach is to solicit from the scientific investigator the values of the ξ_{m1}'s for which the powers are of most interest. Such values may coincide with those that are scientifically different from the ξ_{m0}'s. Such elicitation, which may not be very feasible in practice if M is large, but which may be made possible by forming subclasses or clusters of the M genes as in [11], amounts to specifying effect sizes. Formation of such clusters must be made in close consultation with the investigator, or perhaps guided by the result of a preliminary cluster analysis using data independent of that used in the decision functions. For the specified ξ_{m1}'s, the ROC functions in the determination of the optimal weak FWER-controlling multiple size vector become $\rho_m(\eta) = \pi_{\delta_m^*(\eta)}(\xi_{m1})$ for $m \in \mathcal{M}$, where $\delta_m^*(\eta)$ is the simple MP test of size η for testing $H_{m0}: \xi_m = \xi_{m0}$ versus $H_{m1}: \xi_m = \xi_{m1}$, and $\pi_{\delta_m^*(\eta)}(\xi_{m1})$ is the power of $\delta_m^*(\eta)$ (at $\xi_m = \xi_{m1}$). In the clustered situation with $\mathcal{M} = \biguplus_{k=1}^K \mathcal{M}_k$, we may denote by $\bar{\rho}_k(\eta)$ and ζ_k, respectively, the common ROC function and size for the decision functions in cluster $\mathcal{M}_k$. Under second-order differentiability of $\bar{\rho}_k(\eta)$'s, by Theorem 4.3, the optimal weak FWER-α controlling multiple size vector $\zeta(\alpha) = (\zeta_1(\alpha), \zeta_2(\alpha), \ldots, \zeta_K(\alpha))$ is the $\zeta = (\zeta_1, \zeta_2, \ldots, \zeta_K)$ that solves the set of equations $\forall k = 1, 2, \ldots, K: \bar{\rho}_k'(\zeta_k)(1 - \zeta_k) = \lambda$ for some $\lambda \in \Re_+$ with $\sum_{k=1}^K |\mathcal{M}_k| \log(1 - \zeta_k) = \log(1 - \alpha)$.

The second approach, analogous to those in [21, 30, 42, 43, 45, 49] is to estimate or approximate the underlying values of the ξ_m's either using the observed data x, possibly via shrinkage-type estimators, or through the use of prior information which could be informed by external covariates as in [13]. Addressing this same restriction of requiring knowledge of the simple null and simple alternative hypotheses and advocating this second approach, [29], page 679, stated: "although leading to oracle procedures, it can be used in practice as soon as the null and alternative distributions are estimated or guessed reasonably accurately from independent data." By "independent data" is meant in [29] as data different from that used in performing the actual tests. However, such external data need not always be used for estimating or imputing the unknown parameters. For example, suppose that for each $m \in \mathcal{M}$, data x_m could be partitioned into (v_m, w_m). We may then use $\tilde{\xi}_m(v_m) = \max\{\xi_{m0}, \hat{\xi}_m(v_m)\}$, where $\hat{\xi}_m(v_m)$ is the maximum likelihood estimate of ξ_m based on v_m, and proceed as in the preceding paragraph with ξ_{m1} set to $\tilde{\xi}_m(v_m)$ for each $m \in \mathcal{M}$, and with the component data w_m used in the test functions. The resulting MDF will be of an adaptive type, possibly also compound as in [45] if shrinkage estimators are used for estimating the ξ_m's using the v_m components. Observe that if for some $m_0 \in \mathcal{M}$, $\tilde{\xi}_m(v_{m_0})$ and $\xi_{m_0 0}$ are very close or identical, then a relatively small size will be allocated to the MP test for component m_0. This amounts to downgrading the testing problem for this component, a fact of importance since a criticism of multiple hypotheses testing, especially when using FDR, is that an unscrupulous investigator may keep adding irrelevant genes. When using the adaptive MDF arising from the optimal multiple decision size vector, this investigator's strategy will backfire since the adaptive MDF will

automatically downgrade the irrelevant genes. This second approach still requires deeper study. For instance, there is the issue of how to partition each x_m into the v_m and w_m components. Furthermore, the impact of a misspecified ξ_{m1}, possibly arising from the estimation procedure, needs to be ascertained.

5.3. *Connections to p-value statistics*. Proposition 3.2 indicates that the ROC function $\eta \mapsto \rho_m(\eta)$ is differentiable if and only if the distribution function of the p-value statistic $S_m(X_m, U_m)$ under $H_{m1}: Q_m = Q_{m1}$ is differentiable. In this case, $\rho'_m(\cdot)$ coincides with $h_m(\cdot)$, the density function of $S_m(X_m, U_m)$ under $H_{m1}: Q_m = Q_{m1}$. Condition (i) in Theorem 4.3 is equivalent to the constancy in m of $h_m(\eta_m)(1-\eta_m)$. This is surprising since it indicates that it is *not* enough to simply find the sizes that maximize these $h_m(\cdot)$'s, as dictated by the Neyman–Pearson lemma when dealing with a single pair of null and alternative hypotheses. Rather, in the multiple hypotheses testing scenario, there is attenuation in that larger sizes incur penalties. Condition (i) in Theorem 4.3 governs the interactions among the M tests regarding their size allocations to achieve the best overall result, in terms of overall type II error, among themselves.

The optimal weak FWER-controlling MDF can be converted to a procedure based on the p-value statistics. If $\boldsymbol{\eta}^*(\alpha) = (\eta^*_m(\alpha), m \in \mathcal{M})$ is the optimal weak FWER-α multiple decision size vector and $(S_m(x_m, u_m), m \in \mathcal{M})$ is the vector of computed p-value statistics, the decision based on data $(x, u) = ((x_m, u_m), m \in \mathcal{M})$ is $\delta^*(x, u) = (I\{S_m(x_m, u_m) \le \eta^*_m(\alpha)\}, m \in \mathcal{M})$, an MDF based on weighted p-values. This is related to the approach in several papers using weighted p-values such as [16, 21, 29, 30, 46]. In our case, the weights are tied-in to the optimal sizes.

6. *Strong* FWER control. Let $\boldsymbol{\Delta}^* = (\Delta^*_m, m \in \mathcal{M})$ be the MP MDP with $\Delta^*_m = (\delta^*_m(\eta) : \eta \in [0, 1])$ the MP decision process for $H_{m0}: Q_m = Q_{m0}$ versus $H_{m1}: Q_m = Q_{m1}$ based on (X_m, U_m). Wlog, assume that the size function $A_m(\cdot)$ of Δ^*_m satisfies $A_m(\eta) = \eta$. Define $\boldsymbol{\eta}: [0, 1] \to [0, 1]^M$ such that $\boldsymbol{\eta}(\alpha) = (\eta_m(\alpha), m \in \mathcal{M})$ is the optimal weak FWER-controlling multiple decision size vector at level α. Assume that each component of this mapping is nondecreasing and continuous, which is the case when the ROC functions of $\boldsymbol{\Delta}^*$ are twice-differentiable as established in Proposition 4.5.

For a weak FWER threshold of $\alpha \in [0, 1]$, the optimal MDF in $\mathcal{D}_0$ is $\boldsymbol{\delta}^*_W(\alpha) = (\delta^*_m(\eta_m(\alpha)), m \in \mathcal{M})$, as given in (4.2). Associated with this MDF is the *generalized* multiple decision p-value statistic $\mathbf{W} = (W_m, m \in \mathcal{M})$, where

$$W_m \equiv W_m(X_m, U_m) = \inf\{\alpha \in [0, 1] : \delta^*_m(\eta_m(\alpha)) = 1\}. \tag{6.1}$$

The $w_m = W_m(x_m, u_m)$ is the smallest weak FWER size leading to rejection of H_{m0} when using $\boldsymbol{\delta}^*_W(\alpha)$ given data $(x, u) = ((x_m, u_m), m \in \mathcal{M})$. The usual p-value statistic S_m [see (3.2)] for δ^*_m is related to W_m via

$$\forall m \in \mathcal{M} : S_m(X_m, U_m) = \eta_m(W_m(X_m, U_m)). \tag{6.2}$$

Now, a lá [42, 45], suppose an Oracle knows Q, the true underlying probability measure of X. For the MDF $\boldsymbol{\delta}^*_W(\alpha)$, its FWER is

$$R_0(\boldsymbol{\delta}^*_W(\alpha), Q) = 1 - \prod_{m\in\mathcal{M}} [1-\eta_m(\alpha)]^{1-\theta_m(Q)}.$$

This is nondecreasing and continuous in α since the mappings $\alpha \mapsto \eta_m(\alpha)$ for each $m \in \mathcal{M}$ are nondecreasing and continuous. If the Oracle desires to control this type I error rate at a value $q^* \in [0,1]$ and also minimize the MDR given by $R_2(\boldsymbol{\delta}^*_W(\alpha), Q) = |\mathcal{M}_1(Q)| - \sum_{m\in\mathcal{M}_1(Q)} \rho_m(\eta_m(\alpha))$, where $\rho_m(\eta_m(\alpha))$ is the power of $\delta^*_m(\eta_m(\alpha))$, then she should choose the largest $\alpha \in [0,1]$ such that $R_0(\boldsymbol{\delta}^*_W(\alpha), Q) = q^*$. Owing to the continuity and nondecreasing properties of $R_0(\boldsymbol{\delta}^*_W(\alpha), Q)$ in α, the Oracle's optimal α could also be expressed via

$$\alpha^\dagger(q^*; Q) = \inf\Bigg\{\alpha \in [0,1]\colon \prod_{m\in\mathcal{M}} [1-\eta_m(\alpha)]^{1-\theta_m(Q)} < 1-q^*\Bigg\}.$$

However, there is no Oracle and Q is not known, else there is no multiple decision problem. Thus, $\alpha^\dagger(q^*; Q)$ is not observable. A natural idea is to estimate the unknown $\theta_m(Q)$, the state of the mth pair of hypotheses. An intuitive and simple estimator of $\theta_m(Q)$ for a fixed value of α is

$$\widehat{\theta}_m(Q) = \delta^*_m(\eta_m(\alpha)-) \equiv \delta^*_m(X_m, U_m; \eta_m(\alpha)-). \tag{6.3}$$

In turn, we obtain a *step-down* estimator $\alpha^\dagger(q^*) \equiv \alpha^\dagger(X, U; q^*)$ of the Oracle-based $\alpha^\dagger(q^*; Q)$ given by

$$\alpha^\dagger(q^*) = \inf\Bigg\{\alpha \in [0,1]\colon \prod_{m\in\mathcal{M}} [1-\eta_m(\alpha)]^{1-\delta^*_m(\eta_m(\alpha)-)} < 1-q^*\Bigg\}. \tag{6.4}$$

This determines a compound MDF $\boldsymbol{\delta}^*_S(q^*) \equiv \boldsymbol{\delta}^*_S(X, U; q^*) \in \mathcal{D}$, where

$$\boldsymbol{\delta}^*_S(q^*) = \big(\delta^*_m(\eta_m(\alpha^\dagger(q^*))), m \in \mathcal{M}\big). \tag{6.5}$$

By virtue of the optimal choice of the $\eta_m(\alpha)$'s and the use of the MP tests, we expect $\boldsymbol{\delta}^*_S(q^*)$ to possess excellent, if not optimal, MDR-properties. By taking the infimum over the weak FWER-size α coupled with the estimation of $\theta_m(Q)$ by $\delta^*_m(\eta_m(\alpha)-)$ in (6.4), there occurs an adaptive downweighting of components whose H_{m0}'s are most likely correct as dictated by the data (x, u). Theorem 6.1 below establishes that $\boldsymbol{\delta}^*_S(q^*)$ in (6.5) does strongly control the FWER.

THEOREM 6.1. *Let* $q^* \in [0,1]$. *Then*, $\forall Q \in \mathcal{Q}$, $R_0(\delta^*_S(q^*), Q) \le q^*$.

Next, we reexpress $\boldsymbol{\delta}^*_S(q^*)$ in terms of the generalized p-value statistic $\mathbf{W}$. This is achieved by defining the random variable

$$J^\dagger(q^*) = \max\Bigg\{j \in \mathcal{M}\colon \prod_{m=i}^{M} \big[1-\eta_{(m)}\big(W_{(i)}\big)\big] \ge 1-q^*, i = 1, 2, \ldots, j\Bigg\}.$$

Since $\alpha^\dagger(q^*) \in [W_{(J^\dagger(q^*))}, W_{(J^\dagger(q^*)+1)})$, then

$$\boldsymbol{\delta}_S^*(q^*) = \big(\delta_m^*\big(\eta_m\big(W_{(J^\dagger(q^*))}\big)\big), m \in \mathcal{M}\big).$$

The next result shows that the sequential step-down Šidák MDF, which strongly controls FWER, is a special case of $\boldsymbol{\delta}_S^*(q^*)$ under exchangeability.

PROPOSITION 6.1. *If the M ROC functions are identical, then $\boldsymbol{\delta}_S^*(q^*)$ coincides with the sequential Šidák step-down FWER-controlling MDF.*

7. *Strong* FDR control. Assume the same framework as in Section 6. Our idea in obtaining an FDR-controlling MDF builds on the development of the BH MDF, specifically the rationale of Theorem 2 in [1]. Let $q^* \in [0, 1]$ be the desired FDR threshold and Q be the underlying probability measure of X. We introduce two stochastic processes: $\mathbf{T}_0 = \{T_0(\alpha; Q) : \alpha \in [0, 1]\}$ and $\mathbf{T} = \{T(\alpha) : \alpha \in [0, 1]\}$, where

$$T_0(\alpha; Q) = \sum_{m \in \mathcal{M}_0(Q)} \delta_m^*(\eta_m(\alpha)) \quad \text{and} \quad T(\alpha) = \sum_{m \in \mathcal{M}} \delta_m^*(\eta_m(\alpha)).$$

For the MDF $\delta_W^*(\alpha)$, its FDR is

$$R_1(\delta_W^*(\alpha), Q) = E_Q\left\{\frac{T_0(\alpha; Q)}{T(\alpha)} I\{T(\alpha) > 0\}\right\}.$$

By the definition of the generalized p-value statistics W_m's in (6.1), we have for $\alpha \in [W_{(m)}, W_{(m+1)})$ that $T(\alpha) = m$, whereas

$$E_Q\{T_0(\alpha; Q)\} = \sum_{m \in \mathcal{M}} \big(1 - \theta_m(Q)\big)\eta_m(\alpha) \le \sum_{m \in \mathcal{M}} \eta_m(\alpha). \tag{7.1}$$

Focus now on an $\alpha \in [W_{(m)}, W_{(m+1)})$. If $\sum_{j \in \mathcal{M}} \eta_j(W_{(m)}) \le mq^*$, then the best α in this interval will be the largest value satisfying $\sum_{j \in \mathcal{M}} \eta_j(\alpha) \le mq^*$, since by increasing α, the MDR decreases as argued in the development of $\delta_S^*(q^*)$ in Section 6. This motivates our definition of $\alpha^*(q^*) = \alpha^*(X, U; q^*)$ as the *step-up* estimator

$$\alpha^*(q^*) = \sup\left\{\alpha \in [0, 1] : \sum_{m \in \mathcal{M}} \eta_m(\alpha) \le q^* \sum_{m \in \mathcal{M}} \delta_m^*(\eta_m(\alpha))\right\}. \tag{7.2}$$

This induces a compound MDF $\boldsymbol{\delta}_F^*(q^*) \equiv \boldsymbol{\delta}_F^*(X, U; q^*) \in \mathcal{D}$ given by

$$\boldsymbol{\delta}_F^*(q^*) = \big(\delta_m^*(\eta_m(\alpha^*(q^*))), m \in \mathcal{M}\big). \tag{7.3}$$

Theorem 7.1 establishes that $\boldsymbol{\delta}_F^*(q^*)$ does control the FDR at q^*. Interestingly, the proof of this theorem, which can be found in [28], employs a reverse martingale argument.

THEOREM 7.1. *Let* $q^* \in [0,1]$. *If,* $\forall Q \in \mathcal{Q} \setminus \{Q_0\}$ *and* $\forall \alpha \in (0,1)$, $|\mathcal{M}_0(Q)| \max_{m\in\mathcal{M}_0(Q)} \eta_m(\alpha) \le \sum_{m\in\mathcal{M}} \eta_m(\alpha)$, *then* $R_1(\boldsymbol{\delta}_F^*(q^*), Q) \le q^*$ *for* $\forall Q \in \mathcal{Q}$.

Some remarks are in order regarding the condition in Theorem 7.1. Clearly, the Šidák multiple decision size vector, which is the optimal multiple decision size vector when the ROC functions are identical, always satisfies this condition. When not in this exchangeable setting, this condition induces some control on the differences of the ROC functions. The next proposition establishes that the BH procedure is a special case of $\boldsymbol{\delta}_F^*(q^*)$ under exchangeability.

PROPOSITION 7.1. *If the ROC functions are identical, then* $\boldsymbol{\delta}_F^*(q^*)$ *is the FDR-*q^* *controlling MDF in* [1].

Examination of the proof of Proposition 7.1 as presented in [28] shows that the BH MDF $\boldsymbol{\delta}^{\mathrm{BH}}(q^*)$ coincides with the Šidák-size based MDF $\boldsymbol{\delta}^{S}(q^*)$. The martingale proof for Theorem 7.1 thus carries over to establishing FDR control by $\boldsymbol{\delta}^{\mathrm{BH}}(q^*)$. We mention that a martingale-based proof of FDR control by $\boldsymbol{\delta}^{\mathrm{BH}}(q^*)$ has also been presented in [44].

We also provide an alternative form of $\boldsymbol{\delta}_F^*(q^*)$ in terms of the generalized p-value statistics W_m's, a form analogous to the conventional formulation of the BH procedure. Define

$$J^*(q^*) \equiv J^*(X,U;q^*) = \max\left\{ m \in \mathcal{M} : \sum_{j\in\mathcal{M}} \eta_j\big(W_{(m)}\big) \le q^* m \right\}. \tag{7.4}$$

Then, it is easy to see that $\boldsymbol{\delta}_F^*(q^*)$ rejects $H_{(m)0}$ for $m \in \{1,2,\ldots,J^*(q^*)\}$ and accepts $H_{(m)0}$ for $m \in \{J^*(q^*)+1, J^*(q^*)+2,\ldots,M\}$.

Finally, let us examine further the generalized p-value statistics W_m's. Focusing on $W_{(1)}$, under Q_0, we have that, for $a \in (0,1)$,

$$\mathbf{P}_{Q_0}\big(W_{(1)} > a\big) = \mathbf{P}_{Q_0}\left\{ \bigcap_{m\in\mathcal{M}} [\delta_m^*(\eta_m(a)) = 0] \right\} = \prod_{m\in\mathcal{M}} [1-\eta_m(a)] = 1-a,$$

the second equality obtained by using the independence of the δ_m^*'s under Q_0. Thus, $W_{(1)}$ is standard uniform when all null hypotheses are correct. Using this uniformity result and Lemma D.2 presented in [28] dealing with lower and upper bounds of $\eta_\bullet$ for $\eta \in UB(C_\alpha)$, we obtain in Proposition 7.2 presented below a lower bound for $R_1(\delta_F^*(q^*), Q_0)$, the FDR when all the null hypotheses are correct.

PROPOSITION 7.2. $\forall q^* \in [0,1]$, $1-(1-q^*/M)^M \le R_1(\delta_F^*(q^*), Q_0) \le q^*$.

8. A modest simulation. We compared through computer simulations the performances of $\boldsymbol{\delta}_F^*$ and $\boldsymbol{\delta}^{\mathrm{BH}}$ in terms of FDR and MDR. The simulation model utilized is similar to the Gaussian example illustrating the optimal weak FWER-controlling procedure in Section 4.4. In this model, the observables are $X_m \sim N(\mu_m, 1)$ for each $m \in \mathcal{M}$, which are independent of each other. The mth pair of hypotheses is $H_{m0}: \mu_m \leq 0$ versus $H_{m1}: \mu_m > 0$. The UMP size-η_m test is $\delta_m^*(X_m; \eta_m) = I\{X_m > \Phi^{-1}(1-\eta_m)\}$. The true values of the means μ_m's are $\mu_m = \xi_m \theta_m, m \in \mathcal{M}$, with $\theta_m \sim \mathrm{Ber}(p)$ and effect sizes $\xi_m \sim |N(\nu, 1)|$, again independently generated from each other. The parameter combinations were induced by taking $M \in \{20, 50, 100\}$, $p \in \{0.1, 0.2, 0.4\}$ and $\nu \in \{1, 2, 4\}$. The FDR-threshold utilized were $q^* \in \{0.05, 0.10\}$. Since the computational implementation of $\boldsymbol{\delta}_F^*$ takes time, for each combination of (q^*, M, ν, p), we limited our simulations to 1,000 replications. The simulated FDR and MDR* were the averages of the false discovery proportions, $L_1(a, Q)$'s, and the standardized missed discovery proportions, $L_2(a, Q)/|\mathcal{M}_1(Q)|$, over the 1,000 replications. We used this standardized MDR since, for each replicate, a Q is generated, hence $|\mathcal{M}_1(Q)|$ differs over the replications. In essence, we are comparing the averages of $R_2(\delta_F^*, Q)/|\mathcal{M}_1(Q)|$ and $R_2(\delta^{\mathrm{BH}}, Q)/|\mathcal{M}_1(Q)|$, where the averaging is with respect to the mechanism generating the Q's over the simulation replications.

We only report results for $q^* = 0.10$ in Table 1 since results for $q^* = 0.05$ lead to similar conclusions. From this table, we observe that both $\boldsymbol{\delta}_F^*$ and $\boldsymbol{\delta}^{\mathrm{BH}}$ fulfill the FDR-constraint, and in a conservative manner, which is expected from theory. More importantly, the MDR-performance of $\boldsymbol{\delta}_F^*$ is better compared to that of $\boldsymbol{\delta}^{\mathrm{BH}}$, with this dominance holding for all twenty-seven parameter combinations. Observe that as M is increased with (ν, p) remaining the same, there is an increase in their MDR*'s; whereas, when ν is increased, which increases the effect sizes, their MDR*'s decrease. Interestingly, the impact of a change of value in p, the proportion of true alternative hypotheses, did not necessarily translate into a monotone change in their MDR*'s, especially when $M = 20$, though for the larger M-values, the change in MDR* appears monotonically decreasing.

It may appear from this simulation study that the standardized improvement of δ_F^* over δ^{BH} is minuscule. However, note that when translated to overall number of discoveries, when M is large, δ_F^* will lead to many more discoveries than δ^{BH} while still maintaining desired FDR control. Such an increase in the number of discoveries may have important practical implications, such as enlarging the number of genes to be explored in consequent studies. This may translate to enhanced chances of discovering crucial and important genes without sacrificing the type I error rate.

9. Summary and concluding remarks. This paper provides some resolution on the role of the individual powers of test or decision functions, more appropriately their ROC functions, in multiple hypotheses testing problems. The importance and relevance of these problems have arisen because of the proliferation of

TABLE 1

Comparison of the false discovery rate (FDR) and standardized missed discovery rate (MDR) performance of MDFs δ_F^* and δ^{BH} under a variety of simulation parameters. This table is for $q^* = 0.10$. The FDR and MDR* are in percentages. The number of replications is* 1,000

	q^*	M	v	p	δ_F^*-FDR	δ_F^*-MDR*	δ^{BH}-FDR	δ^{BH}-MDR*
1	0.1	20	1	0.1	8.03	70.80	8.43	72.64
2	0.1	20	1	0.2	7.55	79.64	8.77	81.99
3	0.1	20	1	0.4	6.05	77.47	6.65	80.30
4	0.1	20	2	0.1	7.70	54.42	8.43	55.80
5	0.1	20	2	0.2	7.39	56.32	7.59	57.31
6	0.1	20	2	0.4	6.47	47.82	6.21	49.38
7	0.1	20	4	0.1	9.14	8.62	9.48	10.30
8	0.1	20	4	0.2	7.80	7.34	6.97	9.20
9	0.1	20	4	0.4	6.15	3.58	5.65	5.53
10	0.1	50	1	0.1	8.83	84.87	9.26	87.05
11	0.1	50	1	0.2	7.11	83.49	7.14	86.65
12	0.1	50	1	0.4	6.45	78.91	6.42	82.30
13	0.1	50	2	0.1	8.36	63.36	8.99	65.04
14	0.1	50	2	0.2	8.74	57.30	8.73	58.93
15	0.1	50	2	0.4	5.80	48.71	5.93	50.21
16	0.1	50	4	0.1	8.84	10.28	8.93	12.09
17	0.1	50	4	0.2	7.93	6.91	7.81	8.79
18	0.1	50	4	0.4	6.34	3.40	6.07	5.68
19	0.1	100	1	0.1	9.14	87.10	9.02	90.02
20	0.1	100	1	0.2	8.21	84.05	8.78	87.38
21	0.1	100	1	0.4	5.92	80.12	5.88	83.73
22	0.1	100	2	0.1	9.79	66.10	9.24	67.93
23	0.1	100	2	0.2	7.68	58.25	7.94	59.93
24	0.1	100	2	0.4	5.74	49.29	6.10	50.90
25	0.1	100	4	0.1	8.37	10.44	8.62	12.36
26	0.1	100	4	0.2	7.72	5.93	7.81	8.22
27	0.1	100	4	0.4	5.69	3.80	6.14	5.72

high-dimensional "large M, small n" data sets in the natural, medical, physical, economic and social sciences. Such data sets are being created or generated due to advances in high-throughput technology, the latter fueled by speedy developments in computer technology and miniaturization.

Almost a century ago, Neyman and Pearson demonstrated the need to take into account the power function and the alternative hypothesis configuration when seeking an optimal test procedure in single-pair hypothesis testing. Their work led to a divorce from the then-existing significance or p-value approach. Currently, many multiple hypotheses testing procedures, epitomized by the Šidák procedures for weak and strong FWER control and by the Benjamini–Hochberg (BH) proce-

dure for FDR control, are based on the p-values of the individual tests and do not consider differences in the power traits of the individual tests. They are appropriate in so-called exchangeable settings wherein power characteristics of the individual tests are identical. Such settings, however, are more the exception than the rule, since nonidentical power characteristics easily arise due to differences in the effect sizes, the dispersion parameters, or the test functions that are employed.

This paper examined whether differences in power characteristics of the individual tests could be exploited to improve on existing procedures for FWER and FDR control. Procedures were developed under the historically most fundamental scenario where the null and the alternative hypotheses are simple. First, an optimal MDF within the class of simple MDFs was shown to exist for weak FWER control. This MDF is better than the Šidák weak FWER-controlling MDF, though the latter is a special case of the optimal MDF under exchangeability. Optimality also informs us of an optimal size-investing strategy. Second, by using this optimal, though still restricted, MDF as an anchor, a compound MDF strongly controlling FWER was obtained. The sequential Šidák MDF is a special case of this MDF under exchangeability. Third, we developed a compound MDF that controls FDR. The BH procedure obtains from this MDF under exchangeability. By construction, these new MDFs have smaller MDRs relative to those that did not exploit power differences. The improvement was demonstrated through a modest simulation study by comparing the new FDR-controlling MDF and the BH MDF.

Though the proposed MDFs do improve on existing ones, we could not claim that they are optimal among *all* compound MDFs for strong FWER or FDR control. This question of global optimality is a difficult and elusive one. So far none of the existing compound MDFs, such as the estimated ODP in [42], could claim global optimality. In our case, the possible drawback is that in constructing the new MDFs, we started with the class of simple MDFs. The resulting MDFs are indeed compound, but establishing global optimality is not transparent. A question even arise as to whether there truly exists an optimal MDF among all compound MDFs that, say, control FDR. One thing certain about our MDFs is that they do control FWER or FDR. This is in contrast to some MDFs that are obtained from oracle MDFs via plugging-in of estimates for unknown quantities. Even though the oracle MDF, which are unimplementable, satisfies the type I error rate control, the plug-in step will usually invalidate such control. See [45] where optimality was in an asymptotic sense and with the type I error rate being the mFDR, as well as [13, 29] for more discussions on these issues.

A natural layer to add in the decision-theoretic formulation of the problem is a Bayesian layer where a prior measure is specified on the unknown probability measure Q or, alternatively, on $\theta(Q)$. There is a possibility that through this Bayesian approach, one may be able to obtain a characterization of the class of optimal MDFs controlling type I error rates, or when the two types of error rates are combined, for example, via a weighted linear combination. The papers [10, 11,

26, 33] which employ Bayes or empirical Bayes approaches are highly relevant on this front.

Finally, we mention that there are still other aspects of the multiple decision problem not dealt with in this paper. First is the extension to situations with composite null and alternative hypotheses. We indicated some ideas in Section 5.2 for distributional models possessing the MLR property, but further and more extensive studies are needed. Second are possible dependencies among the components in $(X_m, m \in \mathcal{M}_0(Q))$. We have assumed that this is an independent collection, but it is certainly of theoretical and applied relevance to examine dependent settings. Potential results in such scenarios will extend those in [2, 31, 32]. In these composite hypotheses and dependent data settings, we expect that resampling-based ideas and approaches, such as those in [47, 48], will be central.

Acknowledgments. The first author is grateful to Dr. James Berger for facilitating his sabbatical leave visit at the Statistical and Applied Mathematical Sciences Institute (SAMSI) during Fall 2008 as this afforded him quality time for generating ideas relevant to this project. As such this work was partially supported by the National Science Foundation (NSF) under Grant DMS-0635449 to SAMSI. However, any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation. He is also grateful to Prof. Odd Aalen and Prof. Bo Lindqvist for facilitating his visits to the University of Oslo and the Norwegian University of Science and Technology (NTNU) which led to critical ideas for this project. The authors are highly grateful to the two reviewers, Associate Editor and the Editors for their comments, suggestions and criticisms. Special thanks to Prof. Sanat Sarkar and Prof. Lan Wang for a careful reading of an earlier version of the manuscript, and thank the following for comments or for pointing out references: Prof. J. Lynch, Dr. A. McLain, Prof. G. Rempala, Prof. J. Sethuraman, Prof. G. Taraldsen, Prof. A. Vidyashankar, Prof. L. Wasserman and Prof. P. Westfall. We also thank Dr. M. Peña for discussions about microarrays.

SUPPLEMENTARY MATERIAL

Supplement to "Power-Enhanced Multiple Decision Functions Controlling Family-Wise Error and False Discovery Rates" (DOI: 10.1214/10-AOS844SUPP; .pdf). The proofs of lemmas, propositions, theorems and corollaries are provided in this supplemental article [28].

REFERENCES

[1] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

[2] BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

[3] BONFERRONI, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Publ. R. Instit. Super. Sci. Econ. Commere. Firenze* **8** 1–62.
[4] COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London. MR0370837
[5] DUDOIT, S., GILBERT, H. N. and VAN DER LAAN, M. (2007). Resampling-based empirical Bayes multiple testing procedures for controlling generalized tail probability and expected value error rates: Focus on the false discovery rate and simulation study. Technical report, Univ. California, Berkeley.
[6] DUDOIT, S., SHAFFER, J. P. and BOLDRICK, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.* **18** 71–103. MR1997066
[7] DUDOIT, S. and VAN DER LAAN, M. J. (2008). *Multiple Testing Procedures With Applications to Genomics*. Springer, New York. MR2373771
[8] EFRON, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104. MR2054289
[9] EFRON, B. (2007). Size, power and false discovery rates. *Ann. Statist.* **35** 1351–1377. MR2351089
[10] EFRON, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23** 1–22. MR2431866
[11] EFRON, B. (2008). Simultaneous inference: When should hypothesis testing problems be combined? *Ann. Appl. Statist.* **2** 197–223. MR2415600
[12] EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571
[13] FERKINGSTAD, E., FRIGESSI, A., RUE, H., THORLEIFSSON, G. and KONG, A. (2008). Unsupervised empirical Bayesian multiple testing with external covariates. *Ann. Appl. Statist.* **2** 714–735. MR2524353
[14] FOSTER, D. P. and STINE, R. A. (2008). α-investing: A procedure for sequential control of expected false discoveries. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 429–444. MR2424761
[15] GENOVESE, C. and WASSERMAN, L. (2002). Operating characteristic and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 499–517. MR1924303
[16] GENOVESE, C. R., ROEDER, K. and WASSERMAN, L. (2006). False discovery control with p-value weighting. *Biometrika* **93** 509–524. MR2261439
[17] GUINDANI, M., MULLER, P. and ZHANG, S. (2009). A Bayesian discovery procedure. *J. Roy. Statist. Soc. Ser. B* **71** 905–925.
[18] HABIGER, J. and PEÑA, E. A. (2010). Randomized P-values and nonparametric procedures in multiple testing. *J. Nonparametr. Stat.* 1–22. DOI: 10.1080/10485252.2010.482154.
[19] IHAKA, R. and GENTLEMAN, R. (1996). R: A language for data analysis and graphics. *J. Comput. Graph. Statist.* **5** 299–314.
[20] JIN, J. and CAI, T. T. (2007). Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* **102** 495–506. MR2325113
[21] KANG, G., YE, K., LIU, N., ALLISON, D. and GAO, G. (2009). Weighted multiple hypothesis testing procedures. *Stat. Appl. Genet. Mol. Biol.* **8** 1–21.
[22] LAN, K. K. G. and DEMETS, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70** 659–663. MR0725380
[23] LANGAAS, M., LINDQVIST, B. H. and FERKINGSTAD, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 555–572. MR2168204
[24] LEHMANN, E. L. (1997). *Testing Statistical Hypotheses*, 2nd ed. Springer, New York. MR1481711
[25] LEHMANN, E. L., ROMANO, J. P. and SHAFFER, J. P. (2005). On optimality of stepdown and stepup multiple test procedures. *Ann. Statist.* **33** 1084–1108. MR2195629

[26] MÜLLER, P., PARMIGIANI, G., ROBERT, C. and ROUSSEAU, J. (2004). Optimal sample size for multiple testing: The case of gene expression microarrays. *J. Amer. Statist. Assoc.* **99** 990–1001. MR2109489

[27] NEYMAN, J. and PEARSON, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Ser. A* **231** 289–337.

[28] PEÑA, E., HABIGER, J. and WU, W. (2010). Supplement to "Power-enhanced multiple decision functions controlling family-wise error and false discovery rates." DOI: 10.1214/10-AOS844SUPP.

[29] ROQUAIN, E. and VAN DE WIEL, M. A. (2009). Optimal weighting for false discovery rate control. *Electron. J. Stat.* **3** 678–711. MR2521216

[30] RUBIN, D., DUDOIT, S. and VAN DER LAAN, M. (2006). A method to increase the power of multiple testing procedures through sample splitting. *Stat. Appl. Genet. Mol. Biol.* **5** Art. 19, 20 pp. (electronic). MR2240850

[31] SARKAR, S. K. (1998). Some probability inequalities for ordered MTP_2 random variables: A proof of the Simes conjecture. *Ann. Statist.* **26** 494–504. MR1626047

[32] SARKAR, S. K. (2008). Generalizing Simes' test and Hochberg's stepup procedure. *Ann. Statist.* **36** 337–363. MR2387974

[33] SARKAR, S. K., ZHOU, T. and GHOSH, D. (2008). A general decision theoretic formulation of procedures controlling FDR and FNR from a Bayesian perspective. *Statist. Sinica* **18** 925–945. MR2440399

[34] SCHWEDER, T. and SPJØTVOLL, E. (1982). Plots of P-values to evaluate many tests simultaneously. *Biometrika* **69** 493–502.

[35] SCOTT, J. and BERGER, J. (2006). An exploration of aspects of Bayesian multiple testing. *J. Statist. Plann. Inference* **136** 2144–2162. MR2235051

[36] ŠIDÁK, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J. Amer. Statist. Assoc.* **62** 626–633. MR0216666

[37] SORIĆ, B. (1989). Statistical "discoveries" and effect-size estimation. *J. Amer. Statist. Assoc.* **84** 608–610.

[38] SPJØTVOLL, E. (1972). On the optimality of some multiple comparison procedures. *Ann. Math. Statist.* **43** 398–411. MR0301871

[39] STEVENSON, R. L. (1886). *The Strange Case of Dr Jekyll and Mr Hyde*, 1st ed. Longmans, Green and Co., London.

[40] STOREY, J. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 479–498. MR1924302

[41] STOREY, J. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.* **31** 2012–2035. MR2036398

[42] STOREY, J. (2007). The optimal discovery procedure: A new approach to simultaneous significance testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 347–368. MR2323757

[43] STOREY, J., DAI, J. and LEEK, J. (2007). The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics* **8** 414–432.

[44] STOREY, J. D., TAYLOR, J. E. and SIEGMUND, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 187–205. MR2035766

[45] SUN, W. and CAI, T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102** 901–912. MR2411657

[46] WASSERMAN, L. and ROEDER, K. (2006). Weighted hypothesis testing. Technical report, Carnegie-Mellon Univ. Available at http://arxiv.org/abs/math.ST/0604172.

[47] WESTFALL, P. and TROENDLE, J. (2008). Multiple testing with minimal assumptions. *Biom. J.* **50** 1–11. MR2526520

[48] WESTFALL, P. and YOUNG, S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. Wiley, New York.

[49] WESTFALL, P. H., KRISHEN, A. and YOUNG, S. S. (1998). Using prior information to allocate significance levels for multiple endpoints. *Stat. Med.* **17** 2107–2119.

E. PEÑA
W. WU
DEPARTMENT OF STATISTICS
UNIVERSITY OF SOUTH CAROLINA
COLUMBIA, SOUTH CAROLINA 29208
USA
E-MAIL: pena@stat.sc.edu
wu26@mailbox.sc.edu
URL: http://www.stat.sc.edu/~pena

J. D. HABIGER
DEPARTMENT OF STATISTICS
OKLAHOMA STATE UNIVERSITY
301-G MSCS BLDG
STILLWATER, OKLAHOMA 74078
USA
E-MAIL: jhabige@okstate.edu