Optimal and maximin procedures for multiple testing problems

Saharon Rosset\textsuperscript{1} | Ruth Heller\textsuperscript{1} | Amichai Painsky\textsuperscript{2} | Ehud Aharoni\textsuperscript{3}

\textsuperscript{1}Department of Statistics and Operations Research, Tel-Aviv University, Tel-Aviv, Israel
\textsuperscript{2}Department of Industrial Engineering, Tel-Aviv University, Tel-Aviv, Israel
\textsuperscript{3}IBM Haifa Research Labs, Haifa, Israel

Abstract
Multiple testing problems (MTPs) are a staple of modern statistical analysis. The fundamental objective of MTPs is to reject as many false null hypotheses as possible (that is, maximize some notion of power), subject to controlling an overall measure of false discovery, like family-wise error rate (FWER) or false discovery rate (FDR). In this paper we provide generalizations to MTPs of the optimal Neyman-Pearson test for a single hypothesis. We show that for simple hypotheses, for both FWER and FDR and relevant notions of power, finding the optimal multiple testing procedure can be formulated as infinite dimensional binary programs and can in principle be solved for any number of hypotheses. We also characterize maximin rules for complex alternatives, and demonstrate that such rules can be found in practice, leading to improved practical procedures compared to existing alternatives that guarantee strong error control on the entire parameter space. We demonstrate the usefulness of these novel rules for identifying which studies contain signal in numerical experiments as well as in application to clinical trials with multiple studies. In various settings, the increase in power from using optimal and maximin procedures can range from 15\% to more than 100\%.

KEYWORDS
FDR, FWER, infinite linear programming, multiple comparisons, optimal testing, strong control

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. Journal of the Royal Statistical Society: Series B (Statistical Methodology) published by John Wiley & Sons Ltd on behalf of Royal Statistical Society.
1 | INTRODUCTION

Hypothesis testing is a fundamental component of scientific research. Having methods for assuring validity of results and controlling false discoveries is critical, to avoid a situation where ‘most published research is false’ as has often been claimed in recent years Ioannidis (2005) and Simmons et al. (2011). However, it is arguably just as critical that these methods not be over-conservative, since this limits the ability of scientists to make true discoveries. Methods that control false discovery while being as liberal as possible yield higher statistical power, and facilitate more discoveries, thus are highly desirable for scientific use.

Major advances have been made since the late 90s of methods for control of false discovery rate (FDR) in lieu of more traditional and conservative measures like familywise error rate (FWER). The Benjamini-Hochberg (BH) suite of FDR control methods gained tremendous influence due to their ability to increase power and make more discoveries while provably controlling a relevant measure of false discovery. However, such methods are not appropriate when false discoveries cannot be tolerated, as, for example, is the case in clinical trials with multiple studies/endpoints/subgroups. Hence the development of new and powerful MTPs for controlling either FDR or FWER remains of prime importance.

The problem of developing ‘most powerful’ MTPs for these problems have thus far been addressed in the literature to a limited extent and mostly under major simplifying assumptions (Dobriban et al., 2015; Lehmann et al., 2005; Spjotvoll, 1972; Storey, 2007; Sun & Cai, 2007; Westfall et al., 1998). For example, although we know that BH guarantees FDR control, it is not the ‘most powerful’ method for controlling it.

In a classic hypothesis testing problem, we are given null and alternative hypotheses, and we wish to find good statistical tests for this problem. A good test is expected to be valid and have the desired probability of rejection under the null model, while being powerful and having a high probability of rejection under the alternative. When the hypotheses are both simple and fully specify the distribution of the data, the Neyman-Pearson (NP) Lemma characterizes the most powerful test at every given level, as rejecting for high values of the likelihood ratio.

This ‘most powerful test’ problem can be viewed as an optimization problem, where every point in sample space has to be assigned to reject or non-reject regions, in a manner that maximizes the expected rejection under the alternative distribution, subject to a constraint on its expectation under the null. When the sample space is infinite (such as a Euclidean space), this is an infinite dimensional binary optimization problem, whose optimal solution happens to have the simple structure characterized by NP.

When moving from testing a single hypothesis to multiple testing scenarios, several complications are added. First, there is no longer a single universally accepted definition of false discovery. Given a rejection policy, denote the (random) number of rejected hypotheses by $R$, and the number of falsely rejected hypotheses (true nulls) by $V$. Two commonly used measures of false discovery, which we denote generically by $Err$, are:

$$FWER: \Pr(V > 0); \quad FDR: \mathbb{E}\left(\frac{V}{R}; R > 0\right).$$

The problem is especially challenging since we require strong $Err$ control, that is, $Err \leq \alpha$ for all possible configurations of null and non-null hypotheses and parameter values. This is the commonly used error control requirement for MTPs, and it expresses the fact that the true configuration is unknown.
Second, there is no longer a single notion of power. For example, we may seek a test which maximizes the expected number of rejections if all nulls are false, or one which maximizes our chance of correctly rejecting a single false null, or we may want to maximize the number of true rejections under some (prior, estimated or known) distribution on the percentage of false nulls, as in the Bayesian approach to FDR (Efron, 2010; Genovese & Wasserman, 2002). The chosen definition should capture the true ‘scientific’ goal of the testing procedure and the type of discoveries we wish to make.

However, once we choose a false discovery criterion and a power criterion, we can write the problem of finding the optimal test as an optimization problem.

Our first challenge in this paper is to develop theory and algorithms for such Optimal Multiple Testing (OMT) procedures for simple hypotheses, and investigate the implications that our results have on design of practical multiple testing procedures, and we address this in Section 2. Our main result is that the goal of finding OMT procedures is attainable in principle for any exchangeable MTP. The key observation towards this result is that like in single hypothesis testing, the key test statistic is the likelihood ratio of each test, and the optimal policy is a step-down procedure on this statistic. We prove optimality and derive the resulting algorithm using Lagrange theory. We apply our algorithms to find OMT procedures for testing three normal means with strong FWER or FDR control. The resulting OMT procedures are much more powerful than relevant alternatives.

However, in realistic testing problems the assumption of simple hypotheses is usually not realistic. In particular, accounting for complexity of alternative hypotheses (potential discoveries) is critical, in assuring that strong control is guaranteed for any combination of relevant alternatives. In the classic results, the NP theory is expanded to notions of optimality for complex or one-sided alternatives, like uniformly most powerful (UMP) tests or maximin tests maximizing the minimal power among the relevant alternatives (Lehmann & Romano, 2005).

To address the challenge of optimally dealing with complex alternatives in MTPs, we define in Section 3 a maximin formulation of multiple testing with complex alternatives. For this setting we derive sufficient conditions for existence of maximin solutions, and formulate the resulting algorithm. We demonstrate that these sufficient conditions hold in interesting examples, such as controlling FWER or FDR for testing normal means, allowing us to find maximin procedures for complex alternatives. In Section 4 we demonstrate the utility of maximin OMT procedures with FWER control in subgroup analysis of randomized clinical trials.

It is important to note that the maximin formulation relaxes the requirement of exchangeability — the actual distribution can be different for each of the testing problems where the null does not hold. When the sufficient conditions hold, the maximin solution is valid across all possible combinations of such alternatives, that is, it provides strong control of the error rate. Moreover, the maximin solutions are fundamentally different than existing alternatives. For example, all procedures for FWER control are closed testing procedures (Goeman et al., 2020), but the vast majority of existing methods can be expressed as applying the closure principle (Marcus et al., 1976) to global tests, while our approach cannot, with the desirable result of significant power increase while still guaranteeing strong FWER control.

Table 1 presents the taxonomy of optimality definitions and our original solutions in the context of MTPs.

The optimal rejection regions presented in Sections 2 and 3 can have counter-intuitive shapes, in particular display non-monotone behaviors, where smaller p-values can lead to fewer rejections. In Section S7 we discuss and interpret these interesting behaviors noting that they are typical for low-power settings, but tend to vanish as power increases. We also offer in Section S7 an extension
which enforces properly defined monotonicity and demonstrate that we can still solve the maximin problem formulation with the added constraints, and that the loss of power from enforcing it is typically small.

Optimality results for MTPs available in the literature typically start from seeking the optimal procedure for a restricted class of decision rules (e.g. single step procedures), and within this restricted class they provide procedures with optimal properties, see Spjotvoll (1972), Westfall et al. (1998), Lehmann et al. (2005) and Dobriban et al. (2015) for results for FWER control and Genovese and Wasserman (2002), Storey (2007), Sun and Cai (2007), Arias-Castro and Chen (2017) and Finner et al. (2009) for results for FDR control. For the specific setting covered by the ‘two-group model’, where the test statistics are generated from a mixture distribution of null and non-null densities, Sun and Cai (2007) and Xie et al. (2011) provide optimal procedures for marginal FDR control, and we provide the optimal procedure for FDR control in Heller and Rosset (2021). A notable exception which considers all possible decision rules, like our approach, is Rosenblum et al. (2014), which used a discrete optimization approximation (as compared to our infinite optimization approach) to estimate the OMT procedure for two independent hypotheses with normal distributions. In Section 5 we compare and contrast our optimality results with the above previous work.

Our key contributions, OMT solutions for exchangeable MTPs and maximin solutions for non-exchangeable MTPs (with or without enforcing monotonicity of the rejection region), lead to novel MTPs with improved power compared to existing alternatives. The new formulations are conceptually important, since they are the first (as far as we know) that do not restrict the class of decision rules in which to search for the optimal solution. However, the computational complexity of the resulting algorithms is high, and therefore the dimension (number of hypotheses $K$) of problems that can be solved is limited with current algorithms. We demonstrate here solutions for up to $K = 3$ hypotheses. We show that this is already useful for important applications like those in Section 4. In Section 6 we present some directions for extensions and practical applications, including dealing with high dimensional problems and complex dependence structures.

## 2 OMT PROCEDURES FOR EXCHANGEABLE SIMPLE HYPOTHESES

### 2.1 Problem formulation and notation

We start by assuming that we have $K > 1$ identical simple hypothesis testing problems, and that we have a sample for testing each of the $K$ null hypotheses. For the $i$th hypothesis testing problem, using data vector $\vec{X}_i$, let $\mathcal{L}_0(\vec{X}_i)$ and $\mathcal{L}_1(\vec{X}_i)$ be the null and non-null likelihood, and $\Lambda(\vec{X}_i) = \frac{\mathcal{L}_1(\vec{X}_i)}{\mathcal{L}_0(\vec{X}_i)}$ the likelihood ratio (LR).

### TABLE 1 Optimality definitions in single testing, parallels for multiple testing problems (MTPs), and our contributions

| Single hypothesis | Simple alternative | Complex alternative |
|-------------------|--------------------|---------------------|
| Most powerful (Neyman-Pearson) | UMP/maximin (Lehmann & Romano, 2005) |
| OMT | maximin |
| (Section 2) | (Section 3) |
We denote the true states of all $K$ tests by the fixed (yet unknown) vector $\tilde{h} \in \{0, 1\}^K$, where, for the $k$th hypothesis testing problem, $h_k = 1$ if the alternative hypothesis is true and $h_k = 0$ if the null hypothesis is true. Let $L_{\tilde{h}}(\tilde{X})$ be the likelihood of $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_K)$ when the true configuration is $\tilde{h}$, and $\Lambda(\tilde{X}) = (\Lambda(\tilde{X}_1), \ldots, \Lambda(\tilde{X}_K))$ be the vector of likelihood ratios. We assume throughout this manuscript that the joint likelihood $L_{\tilde{h}}$ is well defined under all possible true and false null combinations $\tilde{h}$, and similarly for the likelihood ratios $\Lambda(\tilde{X})$, meaning there is an appropriately defined dominating measure and proper densities on the entire support of $\tilde{X}$.

We denote by $\tilde{h}_L$, $0 \leq L \leq K$ the special configuration with the first $L$ nulls being false, and the rest true:

$$\tilde{h}_L = (1, 1, \ldots, 1, 0, ..., 0)_L^{K-L}.$$  

Let $S_K$ be the permutation group for $K$ elements. For $\sigma \in S_K$, $\sigma(\tilde{X}_1, \ldots, \tilde{X}_K)$ and $\sigma(\tilde{h})$ are the corresponding permutations of the data vector $\tilde{X}$ and hypothesis state vector $\tilde{h}$, respectively. We denote by $\sigma_{ij} \in S_K$ the permutation that interchanges entries $i$ and $j$ only. For example, $\sigma_{12}(\tilde{X}) = (\tilde{X}_2, \tilde{X}_1, \tilde{X}_3, \ldots, \tilde{X}_K)$, $\sigma_{12}(\tilde{h}) = (h_2, h_1, h_3, \ldots, h_K)$, and $\Lambda(\sigma_{12}(\tilde{X})) = (\Lambda(\tilde{X}_2), \Lambda(\tilde{X}_1), \Lambda(\tilde{X}_3), \ldots, \Lambda(\tilde{X}_K))$.

If $\tilde{X}_1, \ldots, \tilde{X}_K$ are mutually independent, the likelihood of $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_K)$ is $L_{\tilde{h}}(\tilde{X}) = \prod_{k=1}^{K} L_{h_k}(\tilde{X}_k)$. Consider two coordinates $i$ and $j$ such that $\Lambda(\tilde{X}_i) > \Lambda(\tilde{X}_j)$ and $h_i \geq h_j$. Then $L_{\tilde{h}}(\tilde{X}) \geq L_{\sigma_{ij}(\tilde{h})}(\tilde{X})$ with strict inequality if and only if $h_i = 1$ and $h_j = 0$, since then

$$\frac{L_{\tilde{h}}(\tilde{X})}{L_{\sigma_{ij}(\tilde{h})}(\tilde{X})} = \frac{L_1(\tilde{X}_1)L_0(\tilde{X}_j)}{L_0(\tilde{X}_i)L_1(\tilde{X}_j)} = \frac{\Lambda(\tilde{X}_i)}{\Lambda(\tilde{X}_j)} > 1.$$  

It turns out that this order relation between the LR$s$ and the joint likelihood is enough in order to prove that the OMT procedure will reject first the largest LR statistics.

We shall prove that the OMT procedure rejects first the largest LR statistics in a slightly more general setting, where $\tilde{X}_1, \ldots, \tilde{X}_K$ need not be mutually independent but they need to satisfy the following two conditions (which are clearly satisfied if $\tilde{X}_1, \ldots, \tilde{X}_K$ are mutually independent).

**Assumption 1** \( \tilde{h} \) - **Exchangeability**: The $K$ tests are $\tilde{h}$-exchangeable if

$$L_{\tilde{h}}(\tilde{X}) = L_{\sigma(\tilde{h})}[\sigma(\tilde{X})], \quad \forall \tilde{X}, \sigma \in S_K.$$  

As a consequence $\tilde{X}_i$ and $\tilde{X}_j$ necessarily have the same distribution if $h_i = h_j$.

**Assumption 2** **Arrangement increasing**: The likelihood function $L_{\tilde{h}}(\cdot)$ is arrangement increasing if, when the following holds for some pair of indexes $i \neq j$ and their corresponding data:

$$\left(\Lambda(\tilde{X}_i) - \Lambda(\tilde{X}_j)\right)(h_i - h_j) \leq 0,$$

then by interchanging entries $i$ and $j$ in the vector $\tilde{X}$ to form the vector $\sigma_{ij}(\tilde{X})$, we have the relation:

$$L_{\tilde{h}}(\tilde{X}) \leq L_{\tilde{h}}(\sigma_{ij}(\tilde{X})).$$
In words, improving the individual likelihood ratios for false null hypotheses (at the expense of true null hypotheses) improves the overall likelihood of the data.

For independent exchangeable hypothesis tests, the arrangement increasing property always holds as was demonstrated above. For non-independent exchangeable tests it is typically mild and requires that the joint likelihood is consistent with the marginal likelihoods in terms of ordering. In Section S2 we provide examples for which $\vec{X}_1, \ldots, \vec{X}_K$ are not mutually independent but Assumption 2 is satisfied.

A testing policy has to make a decision which hypotheses are rejected at each possible data vector with the binary decision function $\vec{D} : \mathcal{X} \rightarrow \{0, 1\}^K$, where $\mathcal{X}$ is the range of $\vec{X}$. Given the $\vec{h}$-exchangeability assumption, we limit our interest to decision functions that are symmetric, that is, the decision commutes with the permutation of the data (for example, for $K = 2$ we have $D_1(\vec{X}_1, \vec{X}_2) = D_2(\vec{X}_2, \vec{X}_1)$).

**Definition 1** A decision function $\vec{D} : \mathcal{X} \rightarrow \{0, 1\}^K$ is symmetric if

$$\sigma(\vec{D}(\vec{X})) = \vec{D}(\sigma(\vec{X})), \quad \forall \vec{X} \in \mathcal{X}, \sigma \in S_K.$$ 

To formulate the OMT problem as an optimization problem we need to select the false discovery criterion we wish to control, and the power function we wish to optimize. For power many notions can be considered, as discussed in Section 1. In this paper we limit our consideration to the following options that serve well our purpose of demonstrating the utility of our approach:

$$\Pi_{\text{any}}(\vec{D}) = \mathbb{P}_{\vec{h}_K}(R(\vec{D})(\vec{X}) > 0)$$

$$\Pi_L(\vec{D}) = \mathbb{E}_{\vec{h}_L}[D_1(\vec{X}) + \cdots + D_L(\vec{X})]/L, \quad 1 \leq L \leq K.$$ 

In words, $\Pi_{\text{any}}$ is the probability of making any discoveries if all alternatives are true, and it was discussed for example in Lehmann et al. (2005) and Bittman et al. (2009). $\Pi_L$ is the average power (also known as total power, Westfall & Krishen, 2001), and it seeks to maximize the expected number of true rejections given that $L$ nulls are false. Note that although calculated assuming the first $L$ nulls are false, due to the $\vec{h}$-exchangeability assumption and symmetry requirement on $\vec{D}$, the value of $\Pi_L$ would be the same if the expectation is calculated relative to any other configuration of $L$ false nulls.

Given a selected power measure $\Pi$ and false discovery measure to control $Err$, we can write the OMT problem of finding the optimal test subject to strong control as an infinite dimensional binary program, where the optimization is over the value of the function $\vec{D}$ at every point in the cube:

$$\max_{\vec{D} : \mathcal{X} \rightarrow \{0, 1\}^K, \text{symmetric}} \Pi(\vec{D})$$

$$\text{s.t. } Err_{\vec{h}_L}(\vec{D}) \leq \alpha, \quad 0 \leq L < K.$$ 

We denote the optimal solution to this problem (assuming it exists) by $\vec{D}^*$. We have only $K$ constraints and not $2^K - 1$ due to $\vec{h}$-exchangeability and symmetry.

---

1 Given the definitions that follow, we can in fact prove that OMT procedures are indeed symmetric without assuming this form, if we allow randomized policies. For simplicity we choose to state this as a requirement here.
Several aspects of this optimization problem appear to make it exceedingly difficult to solve:

1. The optimization is over an infinite number of variables
2. This is an integer (binary) optimization problem, which can be impractical to solve even in finite dimensional cases.

2.2 The ordering of OMT procedures by their likelihood ratio statistics

Our first key theoretical result is that an optimal solution to Problem (3) necessarily rejects first the largest likelihood ratio statistics. We formally define the LR-ordering property:

**Definition 2** A decision function $\bar{D} : \mathcal{X} \rightarrow \{0, 1\}^K$ is LR-ordered if it always prioritizes rejecting hypotheses with a higher LR in favor of a hypotheses with a lower LR:

$$\Lambda(\bar{X}_i) \geq \Lambda(\bar{X}_j) \Rightarrow D_i(\bar{X}) \geq D_j(\bar{X}).$$

With this definition we can formalize the advantage of LR-ordered policies in our setting of interest:

**Theorem 1** Assume the following:

(a) Assumptions 1–2 hold
(b) The power criterion $\Pi$ is $\Pi_{\text{any}}$ or $\Pi_L$, $1 < L \leq K$, and the error criterion $\text{Err}$ is either FDR or FWER

Assume also we are given a decision policy $\bar{D}$ which is symmetric. Then there is an improved symmetric policy $\bar{E}$ such that:

(a) $\bar{E}$ is LR-ordered
(b) $\text{Err}_{\bar{E}}(\bar{X}) \leq \text{Err}_{\bar{D}}(\bar{X}), \quad 0 \leq L < K$
(c) $\Pi(\bar{E}) \geq \Pi(\bar{D})$

Note that if $\bar{D}$ is already LR-ordered then we can simply set $\bar{E} = \bar{D}$.

Proofs are supplied in Section S1 in the supplementary material.

This theorem and its proof demonstrate that any policy can be improved by making it LR-ordered. Hence we can conclude that to solve Problem (3) it is sufficient to search over LR-ordered policies, as formalized in the next corollary, which is the main result we need in order to find the OMT policy:

**Corollary 1** Under the assumptions of Theorem 1, if we find an optimal solution of Problem (3) in the class of LR-ordered symmetric policies, it is globally optimal.

Formally, if we find a symmetric solution $\bar{D}^*$ such that:

(a) $\bar{D}^*$ is LR-ordered
(b) $\bar{D}^*$ complies with the constraints of Problem (3)
(c) \( \Pi(\vec{D}^*) \geq \Pi(\vec{D}) \) for any LR-ordered symmetric policy \( \vec{D} \) that complies with the constraints of Problem (3).

Then \( \vec{D}^* \) is an optimal solution of Problem (3).

**Proof.** Assume we have a policy \( \vec{D}^* \) with the required properties. Now assume by negation that \( \vec{D} \) is a different policy which complies with the constraints of Problem (3) and has higher power. Then by Theorem 1 we can improve \( \vec{D} \) with an LR-ordered policy \( \vec{E} \). This contradicts optimality of \( \vec{D}^* \) in the LR-ordered family.

We note that the LR-ordering property holds more generally. One important generalization is to \( K \) simple non-exchangeable hypotheses, if \( \vec{X}_1, \ldots, \vec{X}_K \) are mutually independent, see Section S3 for details. It also applies to other power and error criteria discussed in Section 7, with proofs along the same lines.

Since Corollary 1 states that rejections are based on the LR statistics, we can in fact limit the definition of \( \vec{D} \) to consider only the ordered LR statistics. The \( p \)-value of a LR statistic is necessarily a non-increasing function of the LR, so without loss of generality we move to consider the LR \( p \)-values: for realized LR \( \Lambda_i \) for the \( i \) hypothesis testing problem, \( u_i = \int I\{\vec{X}_i : \frac{L_i(\vec{X}_i)}{L_0(\vec{X}_i)} \geq \Lambda_i\} L_0(\vec{X}_i) d\vec{X}_i \). The null density of \( u_i \) is uniform on \( (0, 1) \), and we denote its alternative density by \( g(\cdot) \).

Our MTP thus formulated is to test the following family of \( K \) hypotheses:

\[
H_{0k} : \quad u_k \sim U(0, 1) \quad \text{and} \quad H_{Ak} : \quad u_k \sim g.
\]

Given the symmetry of \( \vec{D} \) and the fact that \( u_1, \ldots, u_K \) are sufficient statistics for the MTP, we limit the definition of \( \vec{D} \) to consider only the ‘lower corner’ set \( Q = \{u : 0 \leq u_1 \leq u_2 \leq \cdots \leq u_K \leq 1\} \), and extend it to \([0, 1]^K\) through the symmetry. Throughout our discussion we limit our attention to functions \( \vec{D} \) that are Lebesgue measurable on \( Q \) or \([0, 1]^K\) (note that \( \vec{D} \) is also bounded by definition and so integrable).

The LR-ordering result in Corollary 1 implies that on the ‘lower corner’ set \( Q \), \( \vec{D}^* \) can be characterized via \( \kappa^* (\vec{u}) = \max \{k \leq K : D^*_k (\vec{u}) = 1\} \), the ‘last and largest’ \( p \)-value rejected by \( \vec{D}^* \) at \( \vec{u} \in Q \). Given this solution on \( Q \) we can extend it to \([0, 1]^K\) using the symmetry property:

\[
\vec{D}^*(\vec{u}) = \sigma_{\vec{u}}^{-1} \left( \vec{D}^* (\sigma_{\vec{u}}(\vec{u})) \right),
\]

where \( \sigma_{\vec{u}} \) is the sorting permutation for \( \vec{u} \), so that \( \sigma_{\vec{u}}(\vec{u}) \in Q \) is the order statistic of \( \vec{u} \).

### 2.3 The Linear representation of the objectives and constraints

Once we limit our discussion to functions \( \vec{D} \) that have this structure, we can simplify the mathematical description of the objective and constraints of our optimization problem. For this purpose, we add the following notation. Let \( f_{\vec{h}} \) denote the joint density of the \( K \) \( p \)-values when the true configuration is \( \vec{h} \), for example, \( f_{\vec{h}_L} \) is the density for the configuration with the first \( L \) nulls being false.

Taking into account \( \vec{h} \)-exchangeability and symmetry, the L-expected power can be written as a linear functional of \( \vec{D} \) on \( Q \):
\[
\Pi_L(\vec{D}) = \int_{[0,1]^k} f_L(\vec{\bar{u}}) \sum_{k=1}^{L} D_k(\vec{\bar{u}}) d\vec{\bar{u}} / L = \\
= L!(K - L)! \int_Q \sum_{i \in \binom{K}{L}} f_i(\vec{\bar{u}}) \sum_{k \in i} D_k(\vec{\bar{u}}) d\vec{\bar{u}} / L,
\]

where \( i \) indexes the set of all subsets of size \( L \); the notation \( k \in i \) is shorthand that the \( k \)th null is set to false by the \( i \)th configuration; \( f_i(\cdot) \) is the density under the configuration of \( L \) false nulls indexed by \( i \). The first equality follows since without loss of generality we may assume the first \( L \) hypotheses are the non-null hypotheses. The second equality follows by expressing the integral on the ‘lower corner’ set \( Q \).

The LR-ordering property of the optimal solution (Corollary 1) is sufficient to simplify \( \Pi_{\text{any}} \) to a linear functional too:

\[
\Pi_{\text{any}}(\vec{D}) = K! \int_Q D_1(\vec{\bar{u}}) f_L(\vec{\bar{u}}) d\vec{\bar{u}}.
\]

Moving to the constraints, symmetry and \( \vec{h} \)-exchangeability allow us to write the constraints of Problem (3) in the form:

\[
FWER_L(\vec{D}) = \int_{\vec{\bar{u}} \in [0,1]^k} \mathbb{1}\{ V(\vec{D}(\vec{\bar{u}})) > 0 \} f_L(\vec{\bar{u}}) d\vec{\bar{u}} \leq \alpha, \quad 0 \leq L < K.
\]

A similar expression can be written for FDR. By Corollary 1, we can rewrite these \( K \) constraints as linear functionals of the decision function \( \vec{D} \) on \( Q \) for both FWER and FDR:

\[
FWER_L(\vec{D}) = L!(K - L)! \int_Q \sum_k D_k(\vec{\bar{u}}) \sum_{i \in \binom{K}{L}} \mathbb{1}\{ V(\vec{\bar{u}}) > 0 \} f_i(\vec{\bar{u}}) d\vec{\bar{u}};
\]

\[
FDR_L(\vec{D}) = L!(K - L)! \int_Q \sum_k D_k(\vec{\bar{u}}) \sum_{i \in \binom{K}{L}} r_{ki} f_i(\vec{\bar{u}}) d\vec{\bar{u}},
\]

where \( \vec{i}_{\min} \) is the minimal element not in the \( i \)’th configuration of false nulls (that is, the true null hypothesis with the smallest index), and \( r_{ki} \) is the difference in false discovery proportion (FDP) if we reject the \( k \) versus \( k - 1 \) smallest \( p \)-values, that is,

\[
r_{ki} = \frac{|\{1, \ldots, k\} \cap \vec{i}^c|}{k} - \frac{|\{1, \ldots, k - 1\} \cap \vec{i}^c|}{k - 1},
\]

where \( \vec{i}^c \) denotes the actual set of true nulls in the configuration indexed by \( i \) (and we also assume 0/0 = 0). See Section S4 for the derivation of \( FDR_L(\vec{D}) \).

Taking all of these together we conclude that for any combination of objective of the form \( \Pi_{\text{any}} \), \( \Pi_L \), and strong control of FWER or FDR, we can rewrite Problem (3) as an infinite linear binary program on the set \( Q \), with a linear objective and \( K \) linear constraints.
2.4 The OMT procedures

Given our results in Section 2.3, we aim to solve the following optimization problem:

$$\max_{\tilde{D} : Q \rightarrow \{0,1\}^K} \int_Q \left( \sum_{i=1}^{K} a_i(\tilde{u}) D_i(\tilde{u}) \right) d\tilde{u}$$

s.t. $$\int_Q \left( \sum_{i=1}^{K} b_{L,i}(\tilde{u}) D_i(\tilde{u}) \right) d\tilde{u} \leq \alpha, \quad 0 \leq L < K,$$  

(8)

where $$a_i, b_{L,i}, i = 1, \ldots, K, L = 0, \ldots, K - 1$$ are fixed non-negative real functions over $$Q$$, and are linear combinations of the density functions $$\left\{ f^k_h : h \in \{0,1\}^K \right\}$$, with non-negative coefficients that depend on the specific choice of $$\Pi, Err$$, as shown in Equations (4)–(7).

From the LR-ordering property (Corollary 1) it follows that for $$\tilde{u} \in Q$$ we need to consider only policies in the range $$D = \{(0, \ldots, 0), (1, 0, \ldots, 0), \ldots, (1, \ldots, 1, 0), (1, \ldots, 1)\}$$, instead of $$\{0,1\}^K$$.

The Lagrangian is

$$L(\tilde{D}, \mu) = \int_Q \left( \sum_{i=1}^{K} a_i(\tilde{u}) D_i(\tilde{u}) \right) d\tilde{u} + \sum_{L=0}^{K-1} \mu_L \times \left\{ \alpha - \int_Q \left( \sum_{i=1}^{K} b_{L,i}(\tilde{u}) D_i(\tilde{u}) \right) d\tilde{u} \right\}$$

$$= \sum_{L=0}^{K-1} \mu_L \times \alpha + \int_Q \left\{ \sum_{i=1}^{K} D_i(\tilde{u}) \left( a_i(\tilde{u}) - \sum_{L=0}^{K-1} \mu_L \times b_{L,i}(\tilde{u}) \right) \right\} d\tilde{u}$$  

(9)

for $$\mu = (\mu_0, \ldots, \mu_{K-1})$$ with nonnegative entries (i.e., $$\mu_L \geq 0, L = 0, \ldots, K - 1$$).

Let $$f^*$$ denote the optimal (maximum) value of Problem (8). Clearly,

$$f^* \leq \max_{\tilde{D} : Q \rightarrow D} L(\tilde{D}, \mu).$$

Let $$R^k_\mu(\tilde{u}) = a_k(\tilde{u}) - \sum_{L=0}^{K-1} \mu_L \times b_{L,k}(\tilde{u})$$. It is easy to see that for every $$\tilde{u},$$

$$\max_{\tilde{D}(\tilde{u}) \in D} \sum_{k=1}^{K} D_k(\tilde{u}) R^k_\mu(\tilde{u}) = \max \left\{ \max_{i=1,\ldots,K} \sum_{k=1}^{l} R^k_i(\tilde{u}), 0 \right\}$$

achieved for the following policy:

$$D^\mu_i(\tilde{u}) = \prod_{i=1}^{K} \left\{ \sum_{k=1}^{l} R^k_i(\tilde{u}) > 0 \right\}$$  

(10)

$$D^\mu_i(\tilde{u}) = 1 \left\{ \tilde{D}^\mu_{i-1}(\tilde{u}) = 1 \right\} \times \max_{i=1,\ldots,K} \left\{ \sum_{k=1}^{l} R^k_i(\tilde{u}) > 0 \right\}, \quad i = 2, \ldots, K.$$  

(11)

It thus follows that

$$\tilde{D}^\mu(\tilde{u}) = \arg \max_{\tilde{D} : Q \rightarrow D} L(\tilde{D}, \mu).$$
Proposition 1 If \( \exists \mu^* > \bar{0} \) such that \( \tilde{D}^{\mu^*} : \mathbb{R}^K \to D \) satisfies for \( L = 0, \ldots, K - 1 \):

\[
\begin{align*}
&\left( \mu_L^* \geq 0 \text{ and } \int_Q \left( \sum_{i=1}^K b_{L,i}(\bar{u})D_i^{\mu^*}(\bar{u}) \right) \, d\bar{u} = \alpha \right) \text{ or } \\
&\left( \mu_L^* = 0 \text{ and } \int_Q \left( \sum_{i=1}^K b_{L,i}(\bar{u})D_i^{\mu^*}(\bar{u}) \right) \, d\bar{u} \leq \alpha \right),
\end{align*}
\]

then \( \tilde{D}^{\mu^*} \) is an optimal policy for Problem (8).

Proof. \( L(\tilde{D}^{\mu^*}, \mu^*) = \int_Q \left( \sum_{i=1}^K a_i(\bar{u})D_i^{\mu^*}(\bar{u}) \right) \, d\bar{u} \), so \( L(\tilde{D}^{\mu^*}, \mu^*) \leq f^* \) since \( \tilde{D}^{\mu^*} \) is feasible (for \( L \in \{0, \ldots, K - 1\} \) and \( \mu_L^* \geq 0 \) satisfies \( \int_Q \left( \sum_{i=1}^K b_{L,i}(\bar{u})D_i^{\mu^*}(\bar{u}) \right) \, d\bar{u} \leq \alpha \)). On the other hand, \( f^* \leq L(\tilde{D}^{\mu^*}, \mu^*) \forall \mu > \bar{0} \). Therefore, \( f^* = L(\tilde{D}^{\mu^*}, \mu^*) \) and the optimal solution is achieved at \( \tilde{D}^{\mu^*} \).

Thus, a complete algorithm for solving our OMT problems involves:

(a) An approach for searching the space \((\mathbb{R}^+ \cup \{0\})^K\) of possible \( \mu \) vectors for a solution \( \mu^* \).

(b) An approach for (exact or numerical) integration, to calculate

\[
\int_Q \left( \sum_{i=1}^K b_{L,i}(\bar{u})D_i^{\mu^*}(\bar{u}) \right) \, d\bar{u}
\]

for any \( \mu \) vector and assess the error relative to the conditions on \( \mu^* \).

In Section S5 we demonstrate a detailed derivation of the formulas and resulting algorithm for a specific instance of the general problem: Maximizing \( \Pi_3 \) for \( K = 3 \) independent tests under FDR control.

A solution exists under mild ‘non-redundancy’ conditions. We make the following assumption:

Assumption 3 The set of density functions \( \left\{ f_{\bar{h}} : \bar{h} \in \{0, 1\}^K \right\} \) is non-redundant, that is, any non-trivial linear combination of the \( 2^K \) density functions is non-zero almost everywhere on \([0, 1]^K\):

\[
\sum_{\bar{h}} \gamma_{\bar{h}} f_{\bar{h}}(\bar{u}) \neq 0 \text{ almost everywhere for any fixed vector } \gamma_{\bar{h}} \in \mathbb{R}, \sum_{\bar{h}} |\gamma_{\bar{h}}| > 0.
\]

This assumption is mild given the highly non-linear nature of the functions \( f_{\bar{h}} \) in typical applications (as in our examples below).

Proposition 2 Under Assumption 3, the optimal solution \( \mu^* \) exists.

2.5 Example applications: \( K = 3 \) independent normal means

We now utilize our results to illustrate the potential power gain from using OMT procedures as well as the potential insight gained from examining the optimal solutions.
We consider tests of the form $H_0^k : X_k \sim N(0, 1) \text{ vs } H_A^k : X_k \sim N(\theta, 1)$ for $k = 1, 2, 3$ with $\theta < 0$, where all test statistics are independent. The power functions we consider are $\Pi_{\theta, 3}(\vec{D})$, the average power when all three nulls are false, and $\Pi_{\theta, \text{any}}(\vec{D})$, the probability of making at least one rejection when all three nulls are false, which we term minimal power.

For strong FWER control, we demonstrate the potential power gain over Hommel’s procedure (Hommel, 1988), which is the closed-testing procedure (Marcus et al., 1976) that applies the Simes test (Simes, 1986) for each intersection hypothesis. This procedure is popular for independent test statistics since it is uniformly more powerful than the sequentially rejective Bonferroni procedure suggested by Holm (Holm, 1979), as well as from the step-up Hochberg (Hochberg, 1988) procedure. In Lehmann et al. (2005) the step-down Sidak procedure (which performs at each step Sidak’s test (Sidak, 1967) for independent test statistics) was shown to maximize certain aspects of power,2 among all monotone rejection policies (see definition in Section 5). For $\alpha = 0.05$, the difference of this policy from Bonferroni-Holm is negligible (less than $10^{-5}$ in average or minimal power for our experiments below), thus Hommel’s procedure is at least as powerful as the step-down Sidak procedure.

Table 2 shows the power comparison for various values of $\theta$. The power of Hommel’s procedure is smaller than the power of the OMT policy for $\Pi_{\theta, 3}$ by more than 10%. However, the average power of the OMT policy for $\Pi_{\theta, \text{any}}$ can be lower than that of Hommel’s procedure. Of course, the minimal power of the OMT policy for $\Pi_{\theta, \text{any}}$ is much higher than that of Hommel’s procedure.

The OMT policy for $\Pi_{\theta, \text{any}}$ rejects only the minimal $p$-value, since there is no gain in the objective function for rejecting more than one hypothesis. Interestingly, for a large range of $\theta$s the only tight constraint for the OMT policy is the global null constraint. The optimal global test statistic is $\sum_{i=1}^{K} \Phi^{-1}(u_i) / \sqrt{K}$. Therefore, the level $\alpha$ OMT policy for $K$ false nulls when the only tight constraint is the global null constraint is to reject the hypothesis with minimal $p$-value if $\sum_{i=1}^{K} \Phi^{-1}(u_i) / \sqrt{K} < z_{\alpha}$, where $z_{\alpha}$ is the $\alpha$th quantile of the standard normal distribution. For $K = 3$, this is the OMT policy for $\theta > -0.75$ or $\theta < -1.6$, but the OMT rejection region is smaller for $\theta \in (-1.6, -0.75)$ since both the global null constraint and the constraint of FWER control when there is one false null are tight.

Interestingly, for $K = 2$ the OMT policy for $\Pi_{\theta, \text{any}}$ is to reject the hypothesis with minimal $p$-value if $\sum_{i=1}^{2} \Phi^{-1}(u_i) / \sqrt{2} < z_{\alpha}$ for any $\theta < 0$. This was also noted in Rosenblum (2014) in a similar setting for two hypotheses. However, for $K = 3$, such a policy is no longer valid for all $\theta < 0$, since the FWER when there is one false null will be inflated for $\theta \in (-1.6, -0.75)$.

---

2The specific aspect of power considered in Lehmann et al. (2005) was the following: the minimal probability of at least one rejection, among all configurations with $K$ exchangeable non-null hypotheses with signal $\theta \geq \epsilon$. 

---

**Table 2** Average power (columns 2–4) and minimal power (columns 5–7) when all null hypotheses are false, for different discovery policies with strong family-wise error rate (FWER) control at level 0.05

| $\theta$ | $\Pi_{\theta, 3}$ | $\Pi_{\theta, \text{any}}$ |
| --- | --- | --- |
| | Hommel’s procedure | OMT policy | OMT policy | Hommel’s procedure | OMT policy | OMT policy |
| --- | --- | --- | --- | --- | --- | --- |
| $-0.5$ | 0.056 | 0.111 | 0.073 | 0.151 | 0.194 | 0.218 |
| $-1.33$ | 0.254 | 0.363 | 0.247 | 0.524 | 0.660 | 0.742 |
| $-2$ | 0.554 | 0.634 | 0.323 | 0.848 | 0.928 | 0.968 |
For analysis of the OMT policy with FDR control in this setting, see Section S6. An interesting finding is that when the signal is weak, the OMT policy with FDR control is to reject all hypotheses if the optimal global null test is rejected at the nominal FDR level (Proposition S6.1).

3 | BEYOND SIMPLE HYPOTHESES: DEALING WITH COMPLEX ALTERNATIVES

In practical multiple testing scenarios, it is often more realistic to assume that there is no specific known alternative distribution, but that there is a family of relevant alternatives indexed by a parameter \( \theta \in \Theta_A \) (Lehmann & Romano, 2005). Hence, it is important to expand our results to dealing with complex alternatives. Requiring strong control under a range of alternatives translates to requiring that the constraints in Problem (3) hold for every alternative distribution (note that in multiple testing, unlike the single hypothesis case, the constraints do depend on the alternative). We limit our setting of interest to cases that have the monotone likelihood ratio property (Lehmann & Romano, 2005), i.e., for which the ratio of alternative to null density, \( \Lambda(\tilde{X}) = \mathcal{L}_\theta(\tilde{X}) / \mathcal{L}_{\theta_0}(\tilde{X}) \), is decreasing in a properly chosen statistic \( T(\tilde{X}) \) for any \( \theta < \theta_0 \). Therefore, the \( p \)-value for each hypothesis is uniquely defined based on the cumulative distribution of \( T(\tilde{X}) \) (or \( \Lambda(\tilde{X}) \)) under the null, regardless of the specific alternative value of the parameter \( \theta < \theta_0 \). This property holds for large classes of distributions, including exponential families (Lehmann & Romano, 2005).

We define some additional notation. As before, assume each test \( k \) deals with a single parameter \( \theta_k \), with \( H_{0k} : \theta_k = 0 \). For a vector of potential alternatives \( \tilde{\theta} \in \Theta_A^K \), a vector of \( p \)-values \( \tilde{u} \in [0, 1]^K \), and a configuration of hypotheses \( \tilde{h} \in \{0, 1\}^K \), denote the density by \( f_{\tilde{h}, \tilde{\theta}}(\tilde{u}) \), and correspondingly the error measure \( Err_{\tilde{h}, \tilde{\theta}} \). The power of the policy \( \tilde{D} \) is \( \Pi_{\tilde{\theta}}(\tilde{D}) \) when the parameter is \( \tilde{\theta} \) (and the power can be any of \( \Pi_{\tilde{\theta}, \text{any}} \), \( \Pi_{\tilde{\theta}, \text{L}} \) as before). In case \( \theta_1 = \theta_2 = \cdots = \theta_K = \theta \) we use the scalar notations \( \Pi_{\theta}(\tilde{D}), Err_{\tilde{h}, \tilde{\theta}} \).

We consider two approaches to defining the objective in the complex alternative case:

(a) **Single objective.** Assume we have a specific alternative that is of special interest, denote it \( \theta_0 \), and wish to optimize the power for this selected alternative, while maintaining validity for all considered alternatives:

\[
\max_{\tilde{D} : [0, 1]^K \to \{0, 1\}^K} \Pi_{\theta_0}(\tilde{D}) \\
\text{s.t.} \quad Err_{\tilde{h}, \tilde{\theta}}(\tilde{D}) \leq \alpha, \quad \forall \tilde{h} \in \{0, 1\}^K, \quad \tilde{\theta} \in \Theta_A^K. \tag{13}
\]

(b) **Maximin.** In this case, we aim to maximize the minimal power among all alternatives of interest \( \tilde{\theta} \in \Theta_B^K \subseteq \Theta_A^K \), under the same set of constraints:

\[
\max_{\tilde{D} : [0, 1]^K \to \{0, 1\}^K} \min_{\tilde{\theta} \in \Theta_A^K} \Pi_{\tilde{\theta}}(\tilde{D}) \\
\text{s.t.} \quad Err_{\tilde{h}, \tilde{\theta}}(\tilde{D}) \leq \alpha, \quad \forall \tilde{h} \in \{0, 1\}^K, \quad \tilde{\theta} \in \Theta_A^K. \tag{14}
\]

Note that we do not assume that the actual parameter is the same for all alternatives, but the range of potential alternatives is the same. Hence the symmetry requirement on the resulting regions is still applicable.
These optimization problems now have, in addition to an infinite number of variables, also an infinite number of integral constraints (assuming $\Theta_A$ is an infinite set).

We are unable to offer guarantees on existence and sparsity of the optimal solutions to the problems we pose, as we have in the simple hypotheses case. Instead, we offer an approach that assumes existence of an optimal solution that can be characterized using a single value of the parameter. If the assumption holds, our proposed approach is able to find this OMT solution and, importantly, confirm its optimality.

Let $\vec{D}^*(\theta_0, \theta_A)$ be the optimal solution of the optimization problem that uses fixed parameters $\theta_0$ in the objective and $\theta_A \in \Theta_A$ in the constraints:

$$\vec{D}^*(\theta_0, \theta_A) = \arg \max_{\vec{D} : [0,1]^K \rightarrow \{0,1\}^K} \Pi_{\theta_0}(\vec{D})$$

s.t. $Err_{h,\theta_A}(\vec{D}) \leq \alpha, \forall h \in \{0,1\}^K$. \hspace{1cm} (15)

The following result states a sufficient condition for an optimal solution to Problem (13).

**Proposition 3** Assume that we find a parameter value $\theta_A \in \Theta_A$ such that the solution $\vec{D}^*(\theta_0, \theta_A)$ controls $Err$ at level $\alpha$ at all parameter values $\theta \in \Theta_A^K$:

$$Err_{h,\theta}(\vec{D}^*(\theta_0, \theta_A)) \leq \alpha, \forall h \in \{0,1\}^K, \theta \in \Theta_A^K,$$

then $\vec{D}^*(\theta_0, \theta_A)$ is the optimal solution to the complex alternative Problem (13).

The following corollary simplifies the use of this result for finding $\theta_A$.

**Corollary 2** If $\theta_A$ in Proposition 3 exists, then we have:

$$\Pi_{\theta_0}(\vec{D}^*(\theta_0, \theta_A)) \leq \Pi_{\theta_0}(\vec{D}^*(\theta_0, \theta)) \forall \theta \in \Theta_A.$$

In words: the power of the optimal solution for constraints at $\theta_A$ is minimal among all optimal solutions $\vec{D}^*(\theta_0, \theta)$.

With this corollary, we have a simple policy for trying to solve Problem (13):

(a) Search over $\Theta_A$ to find $\theta_A = \arg \min_{\theta} \Pi_{\theta_0}(\vec{D}^*(\theta_0, \theta))$.
(b) Check whether the control condition in Proposition 3 holds.

This approach requires solving problems of the form Problem (15), which are instances of Problem (8), where the parameter $\theta$ of the density function can be different in the power objective and in the constraints. It is straightforward to confirm that the results in Section 2 hold unchanged in this case, hence we can use the same ideas and algorithms to solve the current problems.

Next, we derive a similar sufficient condition for existence of a maximin solution, and corresponding approach for finding it.

**Theorem 2** Assume that we can find two values $\theta_0 \in \Theta_B, \theta_A \in \Theta_A$ such that:

(a) $\vec{D}^*(\theta_0, \theta_A)$ is the optimal solution of the single objective Problem (13) at $\theta_0$. 

(b) The power of this solution at other values is higher:

$$\Pi_{\theta_0} \left( \bar{D}^*(\theta_0, \theta_A) \right) \leq \Pi_{\bar{\theta}} \left( \bar{D}^*(\theta_0, \theta_A) \right) \quad \forall \bar{\theta} \in \Theta^K_B.$$ 

Then $$\bar{D}^*(\theta_0, \theta_A)$$ is the solution to the maximin Problem (14).

The usefulness of this last result is not immediately evident, since the conditions seem harsh. As we show next, it can be practically useful when the problem is such that $$\Theta_B$$ has a minimal element ('closest alternative'), and there exists inherent monotonicity in the problem such that when $$\theta_0$$ is taken as the closest alternative, the conditions hold.

3.1 Example: Testing independent normal means

Assume we are testing $$K$$ independent normal means with variance 1, with $$\Theta_B = (-\infty, \theta_0]$$ and $$\Theta_A = (-\infty, 0]$$. We seek the maximin rejection policy with the objective function of average power for $$K = 3$$ false nulls, $$\Pi_{\theta_0, K}$$. By solving the optimization problem for a single $$\theta < 0$$ constraint at a time, we compute series of rejection policies $$\bar{D}^*(\theta_0, \theta_A)$$. We identify the value of $$\theta$$ with minimal power, $$\theta_A$$, so that $$\Pi_{\theta_0, K}(\bar{D}^*(\theta_0, \theta_A)) \leq \Pi_{\theta_0, K}(\bar{D}^*(\theta_0, \theta))$$ for all $$\theta < 0$$. Once we find this, we can check if we have:

$$\Pi_{\bar{\theta}, K}(\bar{D}^*(\theta_0, \theta_A)) \geq \Pi_{\theta_0, K}(\bar{D}^*(\theta_0, \theta_A)), \quad \forall \bar{\theta} \in \Theta^K_B,$$

and

$$Err_{\bar{h}, \bar{\theta}} (\bar{D}^*(\theta_0, \theta_A)) \leq \alpha, \quad \forall \bar{h} \in \{0, 1\}^K, \bar{\theta} \in \Theta_A^K,$$

in which case by Theorem 2, the computed solution is the maximin solution for $$\Theta^K_B$$. This turned out to be the case for all $$\theta_0$$ values in the examples considered below.

We examine three independent normal means with strong FWER control, using $$\theta_0 = -2$$, and considering average power $$\Pi_{\theta_0, 3}$$. This maximin rejection policy will be used in subgroup analysis in Section 4. We examine its power at a range of $$\bar{\theta}$$ values, and compare it to Hommel’s procedure, as well as to the following closed-testing procedure (Marcus et al., 1976), which is commonly applied to subgroup analysis: for each intersection hypothesis, the sum of z-scores for the hypotheses in the intersection is the test statistic (Stouffer et al., 1949). We refer to the resulting test as closed-Stouffer. Figure 1 shows the power of the three procedures for a range of $$\bar{\theta} = (\theta_1, \theta_2, -2)$$ values, where $$(\theta_1, \theta_2) \in [-2, 0)^2$$. The maximin policy has better power than closed-Stouffer for all values of $$\bar{\theta}$$; while the power gap is fairly small at $$(2, -2, -2)$$, the gap increases as $$\theta_1$$ and $$\theta_2$$ approach zero. Hommel’s (closed testing of Simes) procedure has a small power advantage over the maximin rejection policy at $$\theta_1 = \theta_2 = 0$$, but it is much less powerful than maximin for negative values of $$\theta_1$$ and $$\theta_2$$ Overall, the maximin rejection policy has better power properties than Hommel’s procedure and closed-Stouffer even when one or two of the coordinates have weaker signal than $$-2$$. This suggests that in applications where the test statistics are Gaussian and independent, the maximin rejection policy may be a useful alternative to the Hommel and closed-Stouffer procedures. We describe such an application in the next section Section 4.

Figure 2 shows the rejection regions for the three procedures. For a 2-dimensional display, we selected slices of the 3-dimensional rejection region that are fixed by the minimum $$p$$-value.
FIGURE 1  Power of the maximin (solid), closed-Stouffer (dotted) and Hommel’s (dashed) rejection policies for normal test statistics at \( \vec{\theta} = (\theta_1, \theta_2, -2) \). The maximin procedure is optimized for average power at \( \theta_0 = -2 \).

\[
u_1 = 1.05e - 03 \quad u_1 = 1.66e - 02 \quad u_1 = 4.88e - 02 \quad u_1 = 5.63e - 02
\]

FIGURE 2  For fixed values of the minimum \( p \)-value, displayed are 2-dimensional slices of the following three-dimensional rejection regions for strong family-wise error rate control at level 0.05: Hommel’s procedure (row 1); closed-Stouffer (row 2); maximin rejection policy for optimizing the average power at \( \Theta_B = (-\infty, -2] \) and \( \Theta_A = (-\infty, 0] \) (row 3). In green: reject all three hypotheses; in red: reject exactly two hypotheses; in blue: reject only one hypothesis. The last panel in the first two rows is empty since Hommel’s procedure and closed-Stouffer makes no rejections if all \( p \)-values are greater than 0.05. The third panel in the first row contains only a single grid point on the diagonal where all three hypotheses are rejected (which is hard to see). For each panel, the rejection region is in the top right quadrant of the partition of the plane by the point \((u_1, u_1)\) [Colour figure can be viewed at wileyonlinelibrary.com]
We show the slices with a small minimum $p$-value, with the minimum $p$-value having the Bonferroni threshold (i.e., 0.05/3), with a minimal $p$-value slightly below the nominal level, and with a minimal $p$-value above the nominal level. The boundaries between one, two, or three rejections are necessarily parallel to the axes for Hommel’s procedure but not parallel to the axes for the closed-Stouffer or the maximin rejection policy. Therefore, the latter two policies are not monotone policies (Definition 3), and the non-monotonicity is manifest in the sloping decision boundaries. Interestingly, for the maximin rejection policy, rejections of hypotheses with $p$-values greater than the nominal level are possible: if the smallest $p$-value is 0.0563, there is a fairly large region where only the smallest $p$-value is rejected, but if the two smallest $p$-values are about the same, for a fairly large range of the maximal $p$-value the two smallest $p$-values are rejected, and there is even a small region near the diagonal (in green color) where all three hypotheses are rejected.

4 | APPLICATION TO SUBGROUP ANALYSIS IN THE COCHRANE LIBRARY

The Cochrane database of systematic reviews (CDSR) is the leading resource for systematic reviews in health care. Each review typically includes several outcomes, and for each outcome there may be several subgroups for a subgroup analysis. The outcomes measure the effectiveness or adverse effects of an intervention. The subgroups may differ by patient characteristics (e.g. males and females), by study (e.g. studies in different geographic locations), or by intervention used (Higgins & Green, 2011).

The statistical analysis in these reviews includes forest plots and a fixed effect or random effects meta-analysis, with the underlying assumption that the estimated effect sizes are normally distributed and independent across subgroups. Thus, the set-up is that of inference on $K$ independent normal means.

We considered all the updated reviews up to 2017 in all domains. For subgroup analysis, we considered outcomes that satisfied the following criteria: the outcome was a comparison of means; the number of participants in each comparison group was at least 30; there were at least three subgroups. For simplicity, if the outcome had more than three subgroups we only considered the first three, in order to have $K = 3$ subgroup hypotheses for each outcome. The number of outcomes that passed our selection criteria was 791.

For each outcome, we applied the following three procedures: the maximin procedure with $\Theta_B = (-\infty, -2]$ and $\Theta_A = (-\infty, 0]$ for optimizing the average power ($\Pi_3$), Hommel, and closed-Stouffer. The 2-dimensional slices of their rejection policies are depicted in Figure 2.

Table 3 summarizes the average number of discoveries, and the number of outcomes with at least one discovery, for each procedure. The cross-tabulations in Table 4 show that for every outcome in which closed-Stouffer makes discoveries, the maximin rejection policy makes}

|                  | Maximin | Hommel | Closed-Stouffer |
|------------------|---------|--------|-----------------|
| Average number of discoveries | 1.207   | 1.191  | 1.137           |
| Fraction with at least one discovery | 0.650   | 0.630  | 0.574           |
discoveries as well. There are 74 outcomes for which exactly one of the two procedures, maximin and Hommel, make discoveries, and in 45 of these outcomes, it is maximin that makes the discoveries.

Interestingly, there are two outcomes in which the maximin rejection policy makes exactly two discoveries, yet the other two procedures make no discoveries. Their p-values are as follows: (0.024, 0.026, 0.482); (0.042, 0.044, 0.323).

## 5 RELATED WORK ON OPTIMAL AND MAXIMIN PROCEDURES

### 5.1 Previous work on optimal FWER control

The simplest class of FWER controlling procedures is that of single-step procedures, where the decision whether to reject a hypothesis is only based on the test statistic (or p-value) for that hypothesis. For the weighted Bonferroni procedures, weights to maximize the average power have been considered, e.g., in Spjotvoll (1972), Westfall et al. (1998) and Dobriban et al. (2015).

Optimality results are also available for a more general class of FWER controlling procedures, which requires the selection rules to be monotone.

**Definition 3** (Lehmann et al., 2005). A decision rule $\mathbf{D} : [0, 1]^K \rightarrow \{0, 1\}^K$ is said to be monotone if $u_i' \leq u_i$ for $D_i(\mathbf{u}) = 1$ but $u_i' > u_i$ for $D_i(\mathbf{u}) = 0$ implies that $\mathbf{D}(\mathbf{u}) = \mathbf{D}(\mathbf{u}')$.

In words, if the value of the rejected p-values is decreased, and the value of the non-rejected p-values is increased, the set of rejections remains unchanged.

If restricted to monotone decision rules, the optimal procedure is in the family of stepwise procedures (Lehmann et al., 2005).³

The restriction to monotone decision rules excludes closed testing procedures (Marcus et al., 1976) that are based on combined test statistics (e.g. based on the sum of the z-scores) for testing the intersection hypotheses. Such procedures can have better power than stepwise procedures, unless there is a single strong signal among a group of otherwise null or very weak signals (in which case step-down tests are best), see for example, Lehmacher et al. (1991) and Bittman et al. (2009).

³The decision on whether to reject in stepwise procedures depends on the rank of the p-value: step-down procedures begin by looking at whether the most significant p-value should be rejected; step-up procedures begin by looking at the least significant p-value.
A direction that is most similar to ours, of pursuing optimal power with strong FWER control, with no restriction on the form of regions generated, was explored in Rosenblum et al. (2014), and optimal rejection regions were presented for $K = 2$ which are clearly not in the family of monotone selection rules. While the derivation of optimal monotone selection rules in Lehmann et al. (2005) is relatively easy, the derivation of the optimal rules as suggested in Rosenblum et al. (2014) is computationally difficult. Their optimization technique is based on discrete approximation of the relevant probabilities, which is computationally feasible with two hypotheses, but may be infeasible for more hypotheses. They leave the extension to more than two hypotheses for future research.

Our objective of maximizing power with strong FWER control is similar to that of Rosenblum et al. (2014). However, we address the optimization of the continuous problem in a general framework, which is different from their approach. From the equations of the optimal solution, we demonstrate how we can gain insight into the nature of the rejection region. We demonstrate for $K = 3$ hypotheses the significantly higher power that can be obtained over the stepwise procedures of Lehmann et al. (2005), and we show that the optimal rejection regions are not monotone.

### 5.2 Previous work on optimal FDR control

The best known FDR controlling procedure is the Benjamini-Hochberg (BH) procedure (Benjamini & Hochberg, 1995), which has been shown to perform nearly optimally for various loss functions assuming the hypotheses are exchangeable, when the fraction of null hypotheses is close to one (Genovese & Wasserman, 2002). Asymptotically, as the number of hypotheses grows to infinity, Arias-Castro and Chen (2017) showed that the BH procedure is optimal in some sense; Finner et al. (2009) derived an asymptotically optimal rejection curve under some restrictions on the possible rejections.

An important line of work in recent years concerns control of the marginal FDR, $\mathbb{E}(V)/\mathbb{E}(R)$, and the positive FDR, $\mathbb{E}(V/R|R > 0)$, where the expectations are over $\vec{h}$ and $\vec{u}$, assuming the Bayesian setting that the hypotheses come from the two-group model (Efron, 2008; Storey, 2007; Sun & Cai, 2007). The OMT policy for the two-group model is provided in Sun and Cai (2007) for marginal FDR control, and in Heller and Rosset (2021) for FDR and positive FDR control.

Our objective of maximizing power with strong FDR control stands apart from this line of work in an important way: our strong FDR control guarantee applies to any true state vector $\vec{h}$, in particular the realized one, and it is non-asymptotic. Moreover, we do not assume knowledge of the percentage of true null hypotheses. We present the optimization of the continuous problem, and show from the equations of the optimal solution how we can gain insight into the nature of the rejection region. We demonstrate for $K = 3$ hypotheses the significantly higher power that we can obtain over the BH procedure as well as over the procedure of Solari and Goeman (2017) which provides a small but uniform improvement over the BH procedure. We further show that the optimal rejection region is not monotone, but that monotonicity constraints can be enforced in order to receive an OMT policy that is more well behaved, in the sense that a vector of smaller realized $p$-values cannot result in fewer rejections (see details in Section S7).

### 5.3 Previous work on optimal maximin procedures

We are not aware of much related work. We note a line of work under the name Generalized Neyman Pearson (GNP), which deals with finding minimax tests for single hypotheses when
the null and alternative are allowed to be complex. In this setting, the simple NP solution no longer holds, but extensions using convexity and duality arguments allow asserting the existence of minimax-optimal solutions in certain cases (Cvitanic & Karatzas, 2001; Rudloff & Karatzas, 2010). Our maximin OMT problems are more complex because of the structure and number of constraints, and our policy in Section 3 of deriving testable sufficient conditions rather than theoretical guarantees reflects that.

6 APPLICATIONS AND EXTENSIONS

This work is focused on setting a fundamental framework for extending optimality results to multiple testing, and for deriving optimal policies. Our development has focused on establishing solvability of OMT problems, and solving relatively low dimensional instances numerically, up to $K = 3$. In some modern applications, $K$ can be in hundreds, thousands or even millions (like in Genome Wide Association Studies), while in others, like in clinical trials (Dmitrienko & D'Agostino, 2018), $K$ that is bigger than one but small is still the norm. Either by directly applying the tools developed here, or by extending them to different problem formulations, a number of important extensions of practical interest can be pursued. We review some of these here, including extensions we have studied in follow-up work and others we plan to pursue in the future.

The first direction is to focus on applications to clinical trials, where a small number of tests $K$ is typically performed and where optimizing power (equivalently, decreasing sample size to achieve equal power) can be especially critical for reducing costs. In follow-up work (Heller et al., 2021), we have focused on the application to clinical trials, examining how multiple testing in clinical trials can be framed within our framework and demonstrating the potential benefits from following this route. In this framework, it is also possible to efficiently deal with general dependence structure and non-exchangeability, as we discuss.

For large scale inference problems (typically, $K$ at least few thousands), the two-group model, first introduced by Efron et al. (2001), is often assumed. Using this model, inference is typically based on methods that control the mFDR (Efron, 2010; Sun & Cai, 2007). Our infinite-dimensional formulation of the hypothesis testing problem led us to find the OMT policy with FDR and positive FDR (Storey, 2003) control within the two-group model framework in a companion paper (Heller & Rosset, 2021). By assuming the two-group model, the number of error constraints is reduced to one, since the error is the expectation over all possible configurations $\tilde{h}$. For example, for FDR control the single constraint is $\mathbb{E}_{\tilde{h}, \tilde{u}} \left( \frac{V}{R+0} \right) \leq \alpha$. Therefore, with computational shortcuts, we were able to find the OMT policies for practically any $K$.

To address feasibility of solution for larger $K$ within our current framework of strong control, we need to consider the computational complexity of numerical solution, and in particular its dependence on $K$. There are three components to the computation:

(a) Searching in parameter space for the vector $\mu^*$ of $K$ Lagrange multipliers which solve the problem.
(b) For each set of multipliers considered, performing numerical integration over the set $Q$ in the $K$-dimensional hypercube.
(c) For each evaluation of the integrand in the integration, calculating the coefficients in Equations (6) and (7).

The complexity of the first two items depends on the specific algorithms used for search and integration, of which there is a large variety (Press et al., 2007), and identifying the best approaches
The expected number of true discoveries if all null hypotheses are false, for the following rejection policies: the maximin procedure with $\Theta_B = (-\infty, -2]$ and $\Theta_A = (-\infty, 0]$ for optimizing the average power ($\Pi_3$) at level $0.05/N$ for three normal means; Hommel’s procedure at level $0.05/N$ for three normal means.

| $N \times 3$ | Grouped maximin | Grouped Hommel |
|--------------|-----------------|----------------|
| 3            | 1.90            | 1.66           |
| 30           | 8.37            | 5.92           |
| 300          | 29.26           | 17.20          |

for our type of problems is a topic for future research. For the third item — calculation of coefficients for the linear constraints — we can make some progress. The representation in Equations (6) and (7) appears to be exponential in $K$, however it is easy to see that these coefficients can be calculated in complexity $O(K^2)$ using a dynamic programming approach, for independent hypotheses (details in Section S8). Hence by combining state of the art approaches for parameter optimization and numerical integration, with efficient calculation of the coefficients at each integration point, problems of dimension higher than $K = 3$ can be solved exactly and efficiently.

It seems quite clear, however, that to go to dimensions in the thousands or higher, approximations would be required. One direction for such approximations is the use of hierarchical controlling procedures, where hypotheses are divided to groups, within each group an optimal testing procedure is employed, and the results are summed up using group-aggregation techniques. For example, for an MTP with $N \times K$ hypotheses, if we have optimal rejection policies for $K$ hypotheses, we can adjust the level of testing within each of $N$ groups of $K$ hypotheses in order to solve the bigger problem with the same error guarantee. Specifically, for FWER control at level $\alpha$, we can apply the optimal rejection policy at level $\alpha/N$ for each group of $K$ hypotheses, and this procedure will clearly be far more powerful than standard Bonferroni-type procedures on the $N \times K$ hypotheses $p$-values. Table 5 shows the advantage over using Hommel’s procedure for each group (which is uniformly more powerful than using Bonferroni): as $N \times K$ increases (for $K = 3$), the advantage of using the maximin policy on three hypotheses increases. We leave important questions regarding the best strategies for grouping hypotheses, and using less naive approaches than Bonferroni for combining the groups, to follow-up work.

7 | SUMMARY

We present a complete mathematical treatment of OMT procedures for multiple testing of exchangeable simple hypotheses, with demonstration of the resulting solutions for $K = 3$ hypotheses and their power advantage over existing alternatives. In Section 3 we expand the results to the case of complex alternatives, which is the relevant setup in most practical applications, and offer sufficient conditions for a maximin solution, which controls false discovery at all alternatives, while maximizing the minimal power for the range of relevant ones. We emphasize that this maximin solution does not require that all alternatives are the same, and in that sense it also relaxes the exchangeability requirement from the simple hypotheses solution. Critically, we demonstrate that these sufficient conditions hold for testing independent normal means, and in Section 4 use this for subgroup analysis, generating more discoveries than Hommel’s and closed-Stouffer’s approaches.

As far as we know, this is the first work that shows that the objective and constraints are linear in the decision function, thus enabling the computation of optimal rejection regions for FDR and
FWER control. Similar steps can be followed to establish that other objectives that are of interest with FDR control, such as expected weighted loss minimization (Sun & Cai, 2007), are also linear in the decision function. We leave for future research examination of the utility of OMT policies for other power functions, which can also be solved within our framework. For example, we can use linear combinations of the functions we propose here: let \( w_L = \mathbb{P}(\sum_{k=1}^{K} h_k = L) \) be the prior probability that exactly \( L \) null hypotheses are false; then the resulting objective \( \sum_{k=1}^{K} \prod_{l} (\vec{D}) \times w_L \) can be optimally solved within our existing framework. As another example, we can consider incorporating group structure or covariate information into the objective using weights (see, e.g., Roquain and van de Wiel (2009) and Durand (2019) for multiple testing with weights). We also leave for future research consideration of other error measures (see Table 1 in Benjamini (2010) for a list of measures which can be viewed as generalization of the FWER and FDR).

For FWER control, hierarchical procedures that base the decision for a hypothesis on the values of all test statistics have been advocated on an intuitive basis for the setting of non-sparse signals, but without the justification by an optimality theory (Lehmacher et al., 1991; Marcus et al., 1976). Lehmann et al. (2005) developed an optimality theory for procedures restricted to be monotone (which exclude procedures suggested in Lehmacher et al., 1991; Marcus et al., 1976). Under this requirement, optimal testing procedures can take simple forms like being limited to ‘step-down’ rules (Lehmann et al., 2005), and these can be derived relatively easily, with no need for complex methodology we develop here. However, the OMT regions we obtain for all problems we consider are not monotone according to this very strict definition. The reward is significantly higher power than step-down procedures can supply for such problems (Table 2).

We may still ask what constitutes a ‘reasonable’ test, and what happens when optimal tests do not comply with reasonableness expectations. This issue has been raised in other contexts in Perlman and Wu (1999). The form of monotonicity that appears to be uniformly desired is what we term ‘weak monotonicity’ in Section S7, implying that smaller \( p \)-values cannot result in fewer rejections. Surprisingly, some of our derived optimal procedures do not comply with this seemingly sensible requirement, including FDR regions in Figure S1, and more pronounced, maximin regions in Figure S2. To accommodate this requirement, we offer in Section S7 a formulation which adds it to the optimization problem, show that it can be solved, but at some additional costs, of increased computational complexity, and decreased power. In the examples we tested, the power loss is minimal, but more research is required to understand whether it can be more substantial in other cases.

**ACKNOWLEDGEMENTS**

The authors thank Yoav Benjamini, Marina Bogomolov, Thorsten Dickhaus, Jelle Goeman, Abba Krieger, Aaditya Ramdas, Ryan Tibshirani, and Wenguang Sun for useful discussions and comments on a previous version of the manuscript. The authors also thank Ian Pan and Orestis Panagiotou for providing the processed CDSR data. This research was supported by Israeli Science Foundation grant 2180/20.

**ORCID**

Saharon Rosset \( \text{https://orcid.org/0000-0002-4458-9545} \)

Ruth Heller \( \text{https://orcid.org/0000-0003-4329-1015} \)

---

4The step-down Sidak procedure considered in (Lehmann et al., 2005), which performs at each step Sidak’s test, is the same as Bonferroni-Holm up to the fifth decimal place, and hence less powerful than Hommel’s procedure, in the settings considered in Table 2.
REFERENCES
Arias-Castro, E. & Chen, S. (2017) Distribution-free multiple testing. *Electronic Journal of Statistics*, 11, 1983–2001.
Benjamini, Y. (2010) Simultaneous and selective inference: current successes and future challenges. *Biometrical Journal*, 52(6, SI), 708–721.
Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society series B - statistical methodology*, 57(1), 289–300.
Bittman, R., Romano, J., Vallarino, C. & Wolf, M. (2009) Optimal testing of multiple hypotheses with common effect direction. *Biometrika*, 96(2), 399–410.
Cvitanic, J. & Karatzas, I. (2001) Generalized neyman-pearson lemma via convex duality. *Bernoulli*, 7(1), 79–97.
Dmitrienko, A. & D'Agostino, Sr. R.B. (2018) Multiplicity considerations in clinical trials. *New England Journal of Medicine*, 378(22), 2115–2122.
Dobriban, E., Fortney, K., Kim, S.K. & Owen, A.B. (2015) Optimal multiple testing under a gaussian prior on the effect sizes. *Biometrika*, 102(4), 753–766.
Duran G. (2019) Adaptive p-value weighting with power optimality. *Electronic Journal of Statistics*, 13(2), 3336–3385.
Efron, B. (2008) Microarrays, empirical bayes and the two-groups model. *Statistical Science*, 23(1), 1–22.
Efron, B. (2010) *Large-scale inference*. Cambridge: Cambridge University Press.
Efron, B., Tibshirani, R., Storey, J. & Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456), 1151–1160.
Finner, H., Dickhaus, T. & Roters, M. (2009) On the false discovery rate and an asymptotically optimal rejection curve. *The Annals of Statistics*, 37, 596–618.
Genovese, C. & Wasserman, L. (2002) Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society series B - statistical methodology*, 64(3), 499–517.
Goeman, J., Hemerik, J. & Solari, A. (2020) Only closed testing procedures are admissible for controlling false discovery proportions. *Annals of Statistics*, 49(2), 1218–1238.
Heller, R. & Rosset, S. (2021) Optimal control of false discovery criteria in the two-group model. *Journal of the Royal Statistical Society, Series B*, to appear.
Heller, R., Krieger, A. & Rosset, S. (2021) Optimal multiple testing and design in clinical trials. arXiv:2104.01346.
Higgins, J. & Green, S. (2011) *The cochrane hand book for systematic reviews of interventions version 5.1*. Hoboken: Wiley.
Hochberg, Y. (1988) A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4), 800–802.
Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
Hommel, G. (1988) A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, 75(2), 383.
Ioannidis, J. (2005) Why most published research findings are false. *PLOS Medicine*, 2(8), 696–701.
Lehacher, W., Wassmer, G. & Reitmeir, P. (1991) Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics*, 47(2), 511–521.
Lehmann, E. & Romano, J. (2005) *Testing statistical hypotheses*. Berlin: Springer.
Lehmann, E., Romano, J. & Shaffer, J. (2005) On optimality of stepdown and stepup multiple test procedures. *The Annals of Statistics*, 33(3), 1084–1108.
Marcus, R., Peritz, E. & Gabriel, K. (1976) Closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3), 655–660.
Perlman, M. & Wu, L. (1999) The emperor’s new tests. *Statistical Science*, 14(4), 355–369.
Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (2007) *Numerical recipes 3rd edition: the art of scientific computing*, 3rd edn. New York, NY, USA: Cambridge University Press.
Roquain, E. & van de Wiel, M. (2009) Optimal weighting for false discovery rate control. *Electronic Journal of Statistics*, 3, 679–711.
Rosenblum, M. (2014) Uniformly most powerful tests for simultaneously detecting a treatment effect in the overall population and at least one subpopulation. *Journal of Statistical Planning and Inference*, 155, 107–116.
Rosenblum, M., Liu, H. & Yen, E. (2014) Optimal tests of treatment effects for the overall population and two subpopulations in randomized trials, using sparse linear programming. *Journal of the American Statistical Association*, 109(507), 1216–1228.
Rudloff, B. & Karatzas, I. (2010) Testing composite hypotheses via convex duality. *Bernoulli*, 16(4), 1224–1239.

Sidak, Z. (1967) Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62, 626–633.

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

Simmons, J., Nelson, L. & Simonsohn, U. (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366.

Solari, A. & Goeman, J.J. (2017) Minimally adaptive bh: a tiny but uniform improvement of the procedure of benjamini and hochberg. *Biometrical Journal*, 59(4), 776–780.

Spjotvoll, E. (1972) On the optimality of some multiple comparison procedures. *The Annals of Mathematical Statistics*, 43(2), 398–411.

Storey, J. (2003) The positive false discovery rate: a bayesian interpretation and the q-value. *Annals of Statistics*, 31(6), 2013–2035.

Storey, J.D. (2007) The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society Series B - Statistical Methodology*, 69(3), 347–368.

Stouffer, S., Suchman, E., Devinney, L. & Star, S. (1949) *The American soldier: adjustment during army life*, volume 1. Princeton: Princeton University Press.

Sun, W. & Cai, T.T. (2007) Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479), 901–912.

Westfall, P. & Krishen, A. (2001) Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *Journal of Statistical Planning and Inference*, 99(1), 25–40.

Westfall, P., Krishen, A. & Young, S. (1998) Using prior information to allocate significance levels for multiple endpoints. *Statistics in Medicine*, 17(18), 2107–2119.

Xie, J., Cai, T., Maris, J. & Li, H. (2011) Optimal false discovery rate control for dependent data. *Statistics and its Interface*, 4, 417–430.

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of the article at the publisher’s website.

**How to cite this article:** Rosset, S., Heller, R., Painsky, A. & Aharoni, E. (2022) Optimal and maximin procedures for multiple testing problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(4), 1105–1128. Available from: [https://doi.org/10.1111/rssb.12507](https://doi.org/10.1111/rssb.12507)