The Annals of Statistics
2021, Vol. 49, No. 2, 1218–1238
https://doi.org/10.1214/20-AOS1999

ONLY CLOSED TESTING PROCEDURES ARE ADMISSIBLE FOR CONTROLLING FALSE DISCOVERY PROPORTIONS

BY JELLE J. GOEMAN[1], JESSE HEMERIK[2] AND ALDO SOLARI[3]

[1]*Department of Biomedical Data Sciences, Leiden University Medical Center, j.j.goeman@lumc.nl*
[2]*Oslo Centre for Biostatistics and Epidemiology, University of Oslo, and Biometris, Wageningen University & Research, jesse.hemerik@wur.nl*
[3]*Department of Economics, Management and Statistics, University of Milano-Bicocca, solari@stat.unipd.it*

We consider the class of all multiple testing methods controlling tail probabilities of the false discovery proportion, either for one random set or simultaneously for many such sets. This class encompasses methods controlling familywise error rate, generalized familywise error rate, false discovery exceedance, joint error rate, simultaneous control of all false discovery proportions, and others, as well as gene set testing in genomics and cluster inference in neuroimaging. We show that all such methods are either equivalent to a closed testing procedure, or are uniformly improved by one. Moreover, we show that a closed testing method is admissible if and only if all its local tests are admissible. This implies that, when designing methods, it is sufficient to restrict attention to closed testing. We demonstrate the practical usefulness of this design principle by obtaining more informative inferences from the method of higher criticism, and by constructing a uniform improvement of a recently proposed method.

1. Introduction. Closed testing (Marcus, Peritz and Gabriel (1976)) is a fundamental principle of familywise error rate (FWER) control in multiple hypothesis testing. Indeed, almost every known procedure controlling FWER has been shown to be a special case of closed testing, and many procedures have been explicitly constructed as such. This is natural from a theoretical perspective, as Sonnemann (1982, 2008), and Sonnemann and Finner (1988) have shown that closed testing is necessary for FWER control: every admissible procedure that controls FWER is a special case of closed testing. Romano, Shaikh and Wolf (2011) extended the results of Sonnemann and Finner, proving that from a FWER perspective not every closed testing procedure is admissible; only consonant procedures are. These results are valuable for designers of FWER controlling methods, who can rely exclusively on closed testing as a general design principle. Alternative design principles exist, such as the partitioning principle (Finner and Strassburger (2002)) and sequential rejection (Goeman and Solari (2010)), but these are equivalent to closed testing.

Rather than only for FWER control, Goeman and Solari (2011) showed that closed testing may also be used to obtain simultaneous confidence bounds for the false discovery proportion (FDP) of all subsets within a family of hypotheses. Used in this way, closed testing allows a form of post-selection inference. It allows users to look at the data prior to choosing thresholds and criteria for significance, while still keeping control of tail probabilities of the FDP. The approach of Goeman and Solari (2011) is equivalent to an earlier approach by Genovese and Wasserman (2004), Genovese and Wasserman (2006) that did not explicitly use closed testing. A natural question that arises is whether similar results to those of Sonnemann (1982), Sonnemann and Finner (1988) and Romano, Shaikh and Wolf (2011) also hold for this novel

Received February 2019; revised June 2020.
MSC2020 subject classifications. 62F03.
Key words and phrases. Familywise error rate, selective inference, simultaneous inference, multiple testing, higher criticism.

use of closed testing. When controlling FDP, is it sufficient to look only at closed testing-based methods? Which methods controlling FDP are admissible? These are the questions we will address in this paper.

2. Overview and main results. This paper has three main contributions. First, it presents a unification of methods and error rates, rewriting a wide range of diverse procedures as examples of a novel class. Within this class, we give first a necessary condition for admissibility, and then a sufficient one. We start with a birds-eye view of these main results.

Genovese and Wasserman (2004) and Goeman and Solari (2011) considered simultaneous control of FDP for all subsets of a testing problem. For a family of hypotheses of interest $(H_i)_{i\in I}$, these authors have proposed methods to find upper $(1-\alpha)$-confidence bounds $\mathbf{q}^I(S)$ for the FDP $\pi_0(S)$, for all $S \subseteq I$, that are simultaneous for all such S. This means that

$$\mathrm{P}\big(\pi_0(S) \le \mathbf{q}^I(S) \text{ for all } S \subseteq I\big) \ge 1-\alpha. \tag{1}$$

In this paper, we investigate the class of all methods controlling the error rate (1). In Section 3, we show that many procedures that seem to target control of other quantities than FDP at first sight are members of this general class. These include all methods with regular FWER control; FWER control of intersection hypotheses; k-FWER control; simultaneous k-FWER control; False Discovery Exceedance control; control of the Joint Error Rate; and methods constructing confidence intervals for the overall proportion of true (or false) hypotheses. Essentially, the procedures we can rewrite as a special case of (1) are all procedures that control a tail probability of the number or proportion of false discoveries from above, either for one random set or simultaneously for several such sets.

Broad though this class of methods may be, it turns out that we can make strong statements that are valid for the whole class. We focus on admissibility, and therefore on the existence of uniform improvements of methods controlling (1). The first central result of this paper is given in Section 7: given any method controlling (1), we can always construct a closed testing procedure that is either equivalent to the method we started with, or a uniform improvement of that method. Thus, our result implies that a necessary condition for admissibility is equivalence to a closed testing procedure. Moreover, we give an explicit construction of the improvement using the approach of Genovese and Wasserman (2004) and Goeman and Solari (2011). The second main result of this paper is a sufficient condition for admissibility. We show in Section 8 that a closed testing procedure is admissible if the local tests that define the closed testing procedure are admissible.

Taken together, these results give design principles for multiple testing procedures. To design admissible procedures it is sufficient to create a closed testing procedure with admissible local tests. To show admissibility for a procedure designed in a different way, it is sufficient to show that the procedure is equivalent to such a procedure. We will discuss practical implications of our results for researchers seeking to develop new methods. We do this by revisiting two testing procedures: Higher Criticism (Donoho and Jin (2004), Meinshausen and Rice (2006)) and the simultaneous FDP bounds of Katsevich and Ramdas (2020). In both cases, we do not only uniformly improve the inferential statements of the methods, we also extend their scope by deriving nontrivial bounds for the FDP of sets these methods did not initially target.

3. Inference on false discovery proportions. Assume that we have data $\mathbf{X}$ distributed according to some unknown probability distribution $\mathrm{P} \in \Omega$. About P we may formulate hypotheses of the form $H \subseteq \Omega$. Let the family of hypotheses of interest be $(H_i)_{i\in I}$, where $I \subseteq C \subseteq \mathbb{N}$ is finite. The set C, possibly infinite, is arbitrary here, but will become important in Section 6. Within the family I, let $I_0 = \{i \in I : \mathrm{P} \in H_i\}$ be the index set of the true hypotheses and $I_1 = I \setminus I_0$ the index set of the false hypotheses. We will make no further model

assumptions in this paper: any models, any test statistics, and any dependence structures will be allowed. Equalities and inequalities between random variables should be read as holding almost surely for all $\mathrm{P} \in \Omega$ unless otherwise stated. Proofs of all theorems, lemmas and propositions are in Section E in the Supplementary Material (Goeman, Hemerik and Solari (2020)). Throughout the paper, we will denote all random quantities in boldface. Upper case variables (except P) always refer to sets.

We will be studying procedures with FDP control. The FDP of a finite set S is given by

$$\pi_0(S) = \frac{|S \cap I_0|}{|S| \vee 1}.$$

We define a procedure with FDP control on I (i.e., on $(H_i)_{i \in I}$) as a random function $\mathbf{q}^I : 2^I \to [0, 1]$, where 2^I is the power set of I, such that for all $\mathrm{P} \in \Omega$ it satisfies (1).

It will be more convenient to use an equivalent representation that gives a simultaneous lower $(1-\alpha)$-confidence bound for $|S \cap I_1|$, the number of true discoveries. We say that a random function $\mathbf{d}^I : 2^I \to \mathbb{R}$ has a $(1-\alpha)$-*true discovery guarantee* on I if, for all $\mathrm{P} \in \Omega$,

$$\mathrm{P}\big(\mathbf{d}^I(S) \le |S \cap I_1| \text{ for all } S \subseteq I\big) \ge 1 - \alpha. \tag{2}$$

We will usually suppress the dependence on α when talking about true discovery guarantees. To see that the class of methods of FDP control and the class of procedures with a true discovery guarantee are equivalent, note that if $\mathbf{q}^I$ fulfils (1), then

$$\mathbf{d}^I(S) = \big(1 - \mathbf{q}^I(S)\big)|S|,$$

fulfils (2) and, if $\mathbf{d}^I$ fulfils (2), then

$$\mathbf{q}^I(S) = \frac{|S| - \mathbf{d}^I(S)}{|S| \vee 1}$$

fulfills (1). In the rest of the paper, we will focus on true discovery guarantee procedures, which are mathematically easier to work with than methods with FDP control, for example, because they automatically avoid issues with empty sets S. Without loss of generality, we may assume that $\mathbf{d}^I(S)$ takes integer values, and that $0 \le \mathbf{d}^I(S) \le |S|$. If $\mathbf{d}^I(S)$ is not integer, we may freely replace $\mathbf{d}^I(S)$ by $\lceil \mathbf{d}^I(S) \rceil$.

The class of FDP control (cf. true discovery guarantee) procedures encompasses seemingly diverse methods. Only few authors (Genovese and Wasserman (2006), Goeman and Solari (2011), Goeman et al. (2019), Blanchard, Neuvial and Roquain (2020)) have explicitly proposed procedures that target control of FDP for all sets S simultaneously as implied by (1). However, many other well-known types of multiple testing procedures turn out to be special cases of FDP control procedures, even if they were not directly formulated to control (1) or its equivalent. We will review these procedures briefly in the rest of this section in order to emphasize the wide range of applications of the results of this paper. We will reformulate such procedures in terms of $\mathbf{d}^I$.

Procedures that control FWER (e.g., Berk et al. (2013), Bretz et al. (2009), Westfall and Young (1993), Janson and Su (2016)) within the family defined by I are usually defined as producing a random set $\mathbf{K}$ (possibly empty) for which it is guaranteed that, for all $\mathrm{P} \in \Omega$, $\mathrm{P}(|\mathbf{K} \cap I_0| = 0) \ge 1 - \alpha$, a generalization, k-FWER (Guo and Rao (2010), Hommel and Hoffmann (1988), Lehmann and Romano (2005a), Romano and Shaikh (2006), Sarkar (2007), Finos and Farcomeni (2011)), makes sure that, for all $\mathrm{P} \in \Omega$,

$$\mathrm{P}\big(|\mathbf{K} \cap I_0| < k\big) \ge 1 - \alpha,$$

which reduces to regular FWER if $k = 1$ is chosen. It is easily seen that this is equivalent to requiring (2) if we take

$$\mathbf{d}^I(S) = \begin{cases} |S| - k + 1 & \text{if } S = \mathbf{K}, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Free additional statements may be obtained from (3) by direct logical implication. For example, if $\mathbf{d}^I(S) = |S| - k + 1$ then we may immediately set $\mathbf{d}^I(U) = |U| - k + 1$, if positive, for all $U \subseteq S$ without compromising (2). We will come back to such implications in Section 5.

Related to k-FWER are methods controlling False Discovery Exceedance (FDX), also known as γ-FDP, at level $\boldsymbol{\gamma}$ (Dudoit, van der Laan and Pollard (2004), Farcomeni (2009), Korn et al. (2004), Romano and Shaikh (2006), Sun et al. (2015), Delattre and Roquain (2015)). Such methods find a random set $\mathbf{K}$ (possibly empty) such that, for all $\mathrm{P} \in \Omega$,

$$\mathrm{P}\big(|\mathbf{K} \cap I_0| \leq \boldsymbol{\gamma}|\mathbf{K}|\big) \geq 1 - \alpha,$$

which is equivalent to (2) with

$$\mathbf{d}^I(S) = \begin{cases} \lceil (1 - \boldsymbol{\gamma})|S| \rceil & \text{if } S = \mathbf{K}, \\ 0 & \text{otherwise.} \end{cases}$$

In most methods controlling FDX the control level $\boldsymbol{\gamma}$ is fixed, but it may also be random as, for example, in the permutation-based method of Hemerik and Goeman (2018). Variants, such as kFDP (Guo, He and Sarkar (2014)), which allow a minimum number of false discoveries regardless of the size of $\mathbf{K}$, also fit (2).

Other methods allow γ to be chosen post-hoc by controlling FDX simultaneously over several values of γ. One way to achieve this is by control of the Joint Error Rate (JER). The JER (Blanchard, Neuvial and Roquain (2020)) constructs a sequence of $\mathbf{m} \geq 0$ distinct random sets $\mathbf{K}_1, \ldots, \mathbf{K}_{\mathbf{m}} \subseteq I$ and corresponding random bounds $\mathbf{k}_1, \ldots, \mathbf{k}_{\mathbf{m}}$, such that, for all $\mathrm{P} \in \Omega$,

$$\mathrm{P}\big(|\mathbf{K}_i \cap I_0| < \mathbf{k}_i \text{ for all } 1 \leq i \leq \mathbf{m}\big) \geq 1 - \alpha.$$

This is a special case of (2) if we set

$$\mathbf{d}^I(S) = \begin{cases} |\mathbf{K}_i| - \mathbf{k}_i + 1 & \text{if } S = \mathbf{K}_i \text{ for some } 1 \leq i \leq \mathbf{m}, \\ 0 & \text{otherwise.} \end{cases}$$

Joint error rate control may be used with nested sets (Blanchard, Neuvial and Roquain (2020)) or tree-structured sets (Durand et al. (2020)), and is meant to be combined with interpolation (see Section 5). Similar approaches were used by, for example, the permutation-based methods of Meinshausen (2006) and Hemerik, Solari and Goeman (2019). Also the approach of Katsevich and Ramdas (2020), discussed in detail in Section 11, can be seen as controlling JER with nested sets.

A different category of methods involves FWER control of many intersection hypotheses as, for example, used in gene set testing in genomics and in cluster inference in neuroimaging. In genomics, a collection of distinct sets $K_1, \ldots, K_m \subseteq I$ is given a priori, and the procedure generates corresponding random indicators $\mathbf{k}_1, \ldots, \mathbf{k}_m \in \{0, 1\}$ for detection of signal in the corresponding set. FWER is controlled over all statements made, that is, for all $\mathrm{P} \in \Omega$,

$$\mathrm{P}\big(|K_i \cap I_1| \geq \mathbf{k}_i \text{ for all } i = 1, \ldots, m\big) \geq 1 - \alpha. \tag{4}$$

This corresponds to (2) with

$$\mathbf{d}^I(S) = \begin{cases} \mathbf{k}_i & \text{if } S = K_i \text{ for some } 1 \leq i \leq m, \\ 0 & \text{otherwise.} \end{cases}$$

Examples of such methods include Meinshausen (2008), Goeman and Mansmann (2008), Goeman and Finos (2012), Meijer and Goeman (2015a), Meijer, Krebs and Goeman (2015), and Meijer and Goeman (2015b). In the latter two papers a connection with FDP control was already noted. In neuroimaging, cluster inference methods are similar except that in this case the sets $\mathbf{K}_1, \ldots, \mathbf{K}_\mathbf{m}$ and their number $\mathbf{m} \geq 0$ are random, and $k_i = 1$ for $i = 1, \ldots, \mathbf{m}$ is fixed (Poline and Mazoyer (1993)). FWER control (4) is guaranteed by Gaussian random field theory. Such control translates to a true discovery guarantee (2) in the same way.

In partial conjunction testing (Benjamini and Heller (2008), Wang and Owen (2019)), the hypothesis $H_0^{k/n} : |I_1| < k$ is tested for some $1 \leq k \leq n$. The requirement that $\boldsymbol{\delta}$, taking values in $\{0, 1\}$ is a valid test of $H_0^{k/n}$ is equivalent to (2) with

$$\mathbf{d}^I(S) = \begin{cases} \boldsymbol{\delta} k & \text{if } S = I, \\ 0 & \text{otherwise.} \end{cases}$$

Finally, related to partial conjunction methods are methods that aim to make one-sided confidence intervals for $\pi_0(I)$, the proportion of true null hypotheses in the testing problem as a whole (Meinshausen and Rice (2006), Ge and Li (2012)). Here, the requirement that $[0, \mathbf{u}]$ is a valid confidence interval for $\pi_0(I)$ is equivalent to demanding (2) with

$$\mathbf{d}^I(S) = \begin{cases} (1 - \mathbf{u})|I| & \text{if } S = I, \\ 0 & \text{otherwise.} \end{cases}$$

This listing of the different types of methods that may be written as true discovery guarantee methods is certainly not exhaustive, but a general pattern emerges. Any method controlling a $(1-\alpha)$-tail probability of the number or proportion of true discoveries (from below) or false discoveries (from above) either in one subset of I, or in several subsets simultaneously, are special cases of general discovery control procedures. The sets and bounds are all allowed to be random; only α must be fixed.

Writing procedures as true discovery guarantee procedures, even when the rewriting is trivial, may bring a new perspective to the use of the procedure. As proposed by Goeman and Solari (2011), procedures that fulfil (1) or (2) allow a different, flexible way of using multiple testing methods. In flexible multiple testing the user may look at the data before choosing post hoc one or several sets $S \subseteq I$ of interest, based on any desired criteria, and find their $\mathbf{d}^I(S)$. Regardless of this data peeking the bounds on the selected sets are simultaneously valid due to the simultaneity in (2). Writing procedures in this form, therefore, in principle opens the way to their use as post-selection inference methods (see Rosenblatt et al. (2018), Ebrahimpoor et al. (2020), for applications). Of course, this is only useful if the user has some real choice, that is, if $\mathbf{d}^I(S) \neq 0$ for a number of sets S. We will see in Section 5 how to get rid of some of the zeros in the definitions above.

4. True discovery guarantee using closed testing. A general way to construct true discovery guarantee procedures is provided by closed testing, introduced by Marcus, Peritz and Gabriel (1976) for FWER control. Genovese and Wasserman (2006) and Goeman and Solari (2011) adapted closed testing to make it usable for true discovery guarantee and FDP control. We will briefly review these methods here.

For every finite set $S \subseteq C$ we define a corresponding intersection hypothesis as $H_S = \bigcap_{i \in S} H_i$. This hypothesis is true if and only if all H_i, $i \in S$ are true. We have $H_\varnothing = \Omega$, which is always true. For every intersection hypothesis H_S we may choose a *local test* $\boldsymbol{\phi}_S$, taking values in $\{0, 1\}$, with 1 indicating rejection of H_S. This is a valid statistical test for H_S if it has the property that, for all $\mathrm{P} \in H_S$,

$$\mathrm{P}(\boldsymbol{\phi}_S = 1) \leq \alpha.$$

We always choose $\boldsymbol{\phi}_\varnothing = 0$ surely. Choosing a local test for every finite $S \subseteq C$ will yield a *suite of local tests* $\boldsymbol{\phi} = (\boldsymbol{\phi}_S)_{S \subseteq C, |S| < \infty}$. To deal with restricted combinations (Shaffer (1986)) efficiently, if present, we demand that identical hypotheses have identical tests: if for $U, V \subseteq C$ we have $H_U = H_V$, then $\boldsymbol{\phi}_U = \boldsymbol{\phi}_V$. If $H_U = \varnothing$ for some $U \subseteq C$, we may take $\boldsymbol{\phi}_U = 1$ surely.

From a suite of local tests we may obtain a true discovery guarantee procedure in two simple steps. First, we need to correct the tests for multiple testing. We define the *effective local test* within the family I by

$$\boldsymbol{\phi}_S^I = \min\{\boldsymbol{\phi}_U : S \subseteq U \subseteq I\}.$$

As shown by Marcus, Peritz and Gabriel (1976), the effective local tests have FWER control over all intersection hypotheses H_S, $S \subseteq I$, that is, for all $\mathrm{P} \in \Omega$,

$$\mathrm{P}(\boldsymbol{\phi}_S^I \leq |S \cap I_1| \text{ for all } S \subseteq I) \geq 1 - \alpha.$$

Next, we calculate $\mathbf{d}^I(S)$. We see that the procedure defined by $\mathbf{d}^I(S) = \boldsymbol{\phi}_S^I$ already fulfils (2). More recently, however, Goeman and Solari (2011) showed that closed testing may also be used for more powerful FDP control. For any suite of local tests $\boldsymbol{\phi}$, these authors defined the associated procedure

$$\mathbf{d}_{\boldsymbol{\phi}}^I(S) = \min_{U \in 2^S} \{|S \setminus U| : \boldsymbol{\phi}_U^I = 0\}, \tag{5}$$

and proved the true discovery guarantee. Note that the minimum is always defined since $\boldsymbol{\phi}_\varnothing^I = \boldsymbol{\phi}_\varnothing = 0$ surely.

An earlier general approach to developing true discovery guarantee procedures was developed, without reference to closed testing, by Genovese and Wasserman (2004), Genovese and Wasserman (2006). Starting from a suite of local tests, they proved coverage for the general true discovery guarantee procedure

$$\mathbf{g}_{\boldsymbol{\phi}}^I(S) = \min_{V \in 2^I} \{|S \setminus V| : \boldsymbol{\phi}_V = 0\}. \tag{6}$$

The difference between approaches (5) and (6) is that (5) uses a two-step approach, first correcting the local tests for multiple testing using the closed testing procedure, while (6) works directly on the local tests. In compensation, (5) only needs to look through the subsets of the set of interest S, while (6) looks through all subsets of the family I. The end result, however, is identical (Hemerik, Solari and Goeman (2019)):

LEMMA 1. $\mathbf{g}_\phi^I = \mathbf{d}_\phi^I$.

The expressions (5) and (6) are very useful for constructing true discovery guarantee procedures. Local tests tend to be easy to specify in most models, as each local test is a test of a single hypothesis, so that standard statistical test theory may be used. Given a suite of local tests, (5) or (6) takes care of the multiplicity. A computational problem remains: direct application of (5) or (6) takes exponential time. Often, however, shortcuts are available that allow faster computation (Goeman and Solari (2011), Goeman et al. (2019), Dobriban (2020)). We will see examples in Sections 10 and 11.

Comparing (5) and (6), the single step expression of Genovese and Wasserman (2006) is clearly more elegant. However, the link of (5) to closed testing is valuable because it connects true discovery guarantee procedures to the enormous literature on closed testing (see Henning and Westfall (), for an overview). The detour via effective local tests is often profitable in practice because expressions for $\mathbf{d}_{\boldsymbol{\phi}}^I(S)$ can be easier to derive through expressions for $\boldsymbol{\phi}_S^I$ (Hemerik and Goeman (2018), Goeman et al. (2019)).

5. Coherence and interpolation. By viewing methods in terms of true discovery guarantees, as we have done in Section 3, they are upgraded from making a confidence statement about discoveries in a limited number of sets $S \subseteq I$ to doing the same for all subsets of I. However, in the definitions of Section 3, most of these statements are the trivial $\mathbf{d}^I(S) = 0$. Often, however, some of the statements can be uniformly improved by a process called *interpolation*. In this section, we discuss interpolation and how it can improve true discovery guarantee procedures. We will define coherent procedures as procedures that cannot be improved by interpolation.

Let $\mathbf{d}^I$ be some true discovery guarantee procedure. We define the *interpolation* $\bar{\mathbf{d}}^I$ of $\mathbf{d}^I$ as

$$\bar{\mathbf{d}}^I(S) = \max_{U \in 2^I} \{\mathbf{d}^I(U) - |U \setminus S| + \mathbf{d}^I(S \setminus U)\}. \tag{7}$$

Interpolation was used in weaker versions or in specific cases by several authors (Blanchard, Neuvial and Roquain (2020), Genovese and Wasserman (2006), Meinshausen (2006), Durand et al. (2020)). Taking $U = S$, we see that $\bar{\mathbf{d}}^I(S) \geq \mathbf{d}^I(S)$. Moreover, the improvement from $\mathbf{d}^I$ to $\bar{\mathbf{d}}^I$ is for free, as noted in the following lemma.

LEMMA 2. *If $\mathbf{d}^I$ is a true discovery guarantee procedure then so is $\bar{\mathbf{d}}^I$.*

Intuitively, the rationale for interpolation is as follows. If $\mathbf{d}^I(U)$ is large, and S has so much overlap with U that the signal $\mathbf{d}^I(U)$ in U does not fit in $U \setminus S$, then the remaining signal must be in S. Since this reasoning follows by direct logical implication, it will not increase the occurrence of type I error: we can only make an erroneous statement about S if we had already made one about U. As an example, consider interpolation for k-FWER controlling procedures. The interpolated version of (3) is simply

$$\bar{\mathbf{d}}^I(S) = 0 \vee \big(|S \cap \mathbf{K}| - k + 1\big), \tag{8}$$

an expression that simplifies even further to $\bar{\mathbf{d}}^I(S) = |S \cap \mathbf{K}|$ with regular FWER when $k = 1$.

Interpolation is not necessarily a one-off process, and interpolated procedures may sometimes be further improved by another round of interpolation. We call a procedure *coherent* if it cannot be improved by interpolation, that is, if

$$\bar{\mathbf{d}}^I(S) = \mathbf{d}^I(S) \quad \text{for all } S \subseteq I. \tag{9}$$

We can characterize coherent procedures further with the following lemma.

LEMMA 3. *$\mathbf{d}^I$ is coherent if and only if for every disjoint $V, W \subseteq I$ we have*

$$\mathbf{d}^I(V) + \mathbf{d}^I(W) \leq \mathbf{d}^I(V \cup W) \leq \mathbf{d}^I(V) + |W|.$$

We intentionally use the same term *coherent* that was used by Sonnemann (1982) in the context of FWER control of intersection hypotheses. Looking only at FWER control of intersection hypotheses is equivalent to looking only at $\mathbb{1}\{\mathbf{d}^I(S) > 0\}$ for every S, where $\mathbb{1}\{\cdot\}$ denotes an indicator function. In that case (9) reduces to simply requiring that $U \subseteq V$ and $\mathbf{d}^I(U) > 0$ implies that $\mathbf{d}^I(V) > 0$, which is exactly Sonnemann's definition of coherence.

Methods that are created through closed testing are automatically coherent, as the following lemma claims.

LEMMA 4. *The procedure $\mathbf{d}^I_\phi$ is coherent.*

Since an incoherent procedure can always be replaced by a coherent procedure that is at least as good, we will restrict attention to coherent procedures for the rest of this paper.

6. Monotone procedures. The methods from the literature discussed in Sections 3 and 4 are usually not defined for a specific family I of hypotheses, but as generic procedures that can be used for any family, large or small. Researchers developing methods are usually not looking for good properties for a specific family at a specific scale $|I|$, but for methods that are generally applicable and have good properties whatever I.

We can embed the procedure $\mathbf{d}^I$ into a stack of procedures $\mathbf{d} = (\mathbf{d}^I)_{I \subseteq C, |I| < \infty}$, where we may have some maximal family $C \subseteq \mathbb{N}$. We will briefly call $\mathbf{d}$ a *monotone procedure* if it fulfills the three criteria below. In contrast, we call $\mathbf{d}^I$ for a specific I a *local procedure*, or a local member of $\mathbf{d}$.

1. *true discovery guarantee*: $\mathbf{d}^I$ is a true discovery guarantee procedure for every finite $I \subseteq C$;
2. *coherence*: $\mathbf{d}^I$ is coherent for every finite $I \subseteq C$;
3. *monotonicity*: $\mathbf{d}^I(S) \geq \mathbf{d}^J(S)$ for every finite $S \subseteq I \subseteq J \subseteq C$.

The first two criteria are no more than natural. We demand a true discovery guarantee for every member of the monotone procedure, and we demand coherence for every local member since otherwise we may always improve it by a coherent procedure. The monotonicity requirement relates local procedures at different scales to each other. It says that inference on the number of discoveries in a set S should never get better if we embed S in a larger family J rather than in a smaller family I. As the multiple testing problem gets larger, inference should get more difficult. This requirement relates closely to the "subsetting property" of Goeman and Solari (2014) and the monotonicity property of various FWER control procedures (e.g., Bretz et al. (2009), Goeman and Solari (2010)). It is a natural requirement, and the procedures cited in Section 3 generally adhere to it by construction.

There are a few notable exceptions to the rule that method designers tend to design monotone rather than local procedures. All the examples we are aware of are FWER-controlling procedures. Rosenblum, Liu and Yen (2014) proposed a local procedure for $|I| = 2$ hypotheses that optimizes the power for rejecting at least one of these. Their method is specific for the scale $|I|$ it was defined for; extensions to $|I| > 2$ do not exist (Rosset et al. (2018)). In another example, Rosset et al. (2018) developed methods that optimize the power for detecting at least one true effect for specific scales $|I|$ under an exchangeability assumption. These methods also have nonmonotone behavior.

We remark, however, that every coherent local $\mathbf{d}^I$ true discovery guarantee procedure may be trivially embedded in a monotone procedure with $C = I$ (or even $C = \mathbb{N}$) by setting

$$\mathbf{d}^J(S) = \begin{cases} \mathbf{d}^I(S) & \text{if } S \subseteq I, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

This embedding allows translation of properties of monotone procedures to properties of their local members. We will mostly be studying monotone procedures in this paper, but investigate implications for local procedures where appropriate.

Procedures created using closed testing are automatically monotone, as formalized in the following lemma.

LEMMA 5. *The procedure* $\mathbf{d}_{\boldsymbol{\phi}} = (\mathbf{d}^I_{\boldsymbol{\phi}})_{I \subseteq C, |I| < \infty}$ *is a monotone procedure.*

The property of primary interest to us is admissibility. Let us formally define admissibility for true discovery guarantee procedures. Recall that a statistical test $\boldsymbol{\delta}$ of a hypothesis H is uniformly improved by a statistical test $\tilde{\boldsymbol{\delta}}$ of the same hypothesis if (1.) $\tilde{\boldsymbol{\delta}} \geq \boldsymbol{\delta}$; and (2.) $\mathrm{P}(\tilde{\boldsymbol{\delta}} > \boldsymbol{\delta}) > 0$ for some $\mathrm{P} \in \Omega$. A statistical test is *admissible* if no test exists that uniformly improves it (Lehmann and Romano (2005b, Section 6.7). We call a suite of local tests $\boldsymbol{\phi}$

admissible if $\boldsymbol{\phi}_S$ is admissible for all finite $\varnothing \subset S \subseteq C$. We note that existence of admissible tests is not assured in all models, but that under a weak condition all tests that exhaust the α-level are admissible. We discuss these technical issues in Section A in the Supplementary Material, where we also motivate our definition of admissibility compared to alternatives in the literature.

Analogously to admissibility of single tests we define admissibility for true discovery guarantee procedures. A uniform improvement of a monotone procedure $\mathbf{d}$ is a monotone procedure $\tilde{\mathbf{d}}$ such that (1.) $\tilde{\mathbf{d}}^I(S) \geq \mathbf{d}^I(S)$ for all finite $S \subseteq I \subseteq C$; and (2.) $\mathrm{P}(\tilde{\mathbf{d}}^I(S) > \mathbf{d}^I(S)) > 0$ for some $\mathrm{P} \in \Omega$ and some finite $S \subseteq I \subseteq C$. A uniform improvement of a local procedure $\mathbf{d}^I$ is a local procedure $\tilde{\mathbf{d}}^I$ such that (1.) $\tilde{\mathbf{d}}^I(S) \geq \mathbf{d}^I(S)$ for all $S \subseteq I$; and (2.) $\mathrm{P}(\tilde{\mathbf{d}}^I(S) > \mathbf{d}^I(S)) > 0$ for some $\mathrm{P} \in \Omega$ and some $S \subseteq I$. We call a local or monotone procedure that cannot be uniformly improved admissible. If all local members of a monotone procedure are admissible, then the monotone procedure is admissible, but the converse is not necessarily true, as illustrated in Section B in the Supplementary Material.

7. All admissible procedures are closed testing procedures. Theorem 1, below, claims that every monotone true discovery guarantee procedure is either equivalent to a closed testing procedure or can be uniformly improved by one. We already know from Lemma 3 that every incoherent procedure can be uniformly improved by a coherent procedure. It follows that every procedure that is not equivalent to a closed testing procedure is inadmissible: the class of all closed testing procedures is essentially complete (Lehmann and Romano (2005b), Section 1.8) for procedures with a true discovery guarantee, and therefore for FDP control. This is the first main result of this paper.

THEOREM 1. *Let* $\mathbf{d}$ *be a monotone procedure. Then, for every finite* $S \subseteq C$,

$$\boldsymbol{\phi}_S = \mathbb{1}\{\mathbf{d}^S(S) > 0\}$$

is a valid local test of H_S. *For the suite* $\boldsymbol{\phi} = (\boldsymbol{\phi}_S)_{S \subseteq C, |S| < \infty}$ *we have, for all* $S \subseteq I \subseteq C$ *with* $|I| < \infty$,

$$\mathbf{d}^I_{\boldsymbol{\phi}}(S) \geq \mathbf{d}^I(S).$$

Coherence is necessary but not sufficient to guarantee admissibility. The procedure $\mathbf{d}^I_{\boldsymbol{\phi}}(S)$ implied by Theorem 1 may in some cases be truly a uniform improvement over the original, coherent $\mathbf{d}^I(S)$. To see a classical example in which a coherent procedure can uniformly improved by closed testing, think of Bonferroni. Combined with (8), Bonferroni is coherent. However, it is uniformly improved by Holm's procedure that follows from a well-known step-down argument that incorporates an estimate of $\pi_0(I)$ into the procedure. This stepping-down can be seen as a direct application of closed testing with the local test defined in Theorem 1. Step-down arguments are standard for FWER control and have been applied to several FDP controlling methods in the past (Blanchard, Neuvial and Roquain (2020), Goeman et al. (2019), Hemerik, Solari and Goeman (2019)).

It should be noted that in case of a monotone procedure, the local test $\boldsymbol{\phi}_S$ defined in Theorem 1 is truly local, in the sense that it uses only the information used by the restricted testing problem $\mathbf{d}^S$ about the hypotheses H_i, $i \in S$. For example, in a testing problem based on p-values, the local test would use only the p-values p_i, $i \in S$. In other testing problems, some global information may be used, for example, the overall estimate of σ^2 in a large one-way ANOVA, but still in such situations the local test is very natural: as a local test for H_S we use the test for discovery of signal in hypotheses H_i, $i \in S$, that we would use in the situation where the hypotheses H_i, $i \notin S$ are not of interest to us. Such a local test is implicitly defined by the local procedure $\mathbf{d}^S$.

The result of the theorem is formulated in terms of monotone procedures. It applies immediately to local procedures as well if we use the trivial embedding (10) of a local procedure into a monotone one. With this embedding we even have $\mathbf{d}^I_{\boldsymbol{\phi}}(S) = \mathbf{d}^I(S)$. This leads to the following corollary.

COROLLARY 1. *Let $\mathbf{d}^I$ be a coherent procedure. Then, for every $S \subseteq I$,*

$$\boldsymbol{\phi}_S = \mathbb{1}\{\mathbf{d}^I(S) > 0\}$$

is a valid local test of H_S. For the suite $\boldsymbol{\phi} = (\boldsymbol{\phi}_S)_{S \subseteq I}$ we have, for all $S \subseteq I$,

$$\mathbf{d}^I_{\boldsymbol{\phi}}(S) = \mathbf{d}^I(S).$$

Corollary 1 shows that every coherent true discovery guarantee procedure is equivalent to a closed testing procedure. It may possibly be uniformly improved by another closed testing procedure if the suite of local tests $\boldsymbol{\phi}$ is not admissible, as we shall see in the next section.

Corollary 1 also confirms the equivalence between the closed testing and partitioning principles for FWER control. This has been clear since Finner and Strassburger (2002) showed that closed testing procedures may be rewritten as partitioning procedures and that this sometimes uniformly improves them, while Sonnemann (1982) and Sonnemann and Finner (1988) had already shown that the family of closed testing procedure is complete for FWER control. However, since the result is important and, as far as we know, not explicitly stated in the literature we phrase it as a separate theorem.

THEOREM 2. *For every closed testing procedure there exists a partitioning procedure that rejects exactly the same hypotheses. For every partitioning procedure there exists a closed testing procedure that rejects exactly the same hypotheses.*

Since both closed testing and partitioning procedures may be written as sequential rejection procedures (Goeman and Solari (2010)), while it cannot improve upon them by Corollary 1 and Theorem 2, sequential rejection could be labelled a third equivalent principle.

8. All closed testing procedures are admissible. So far we have seen that a true discovery guarantee procedure may be uniformly improved by interpolation to coherent procedures, which in turn may be uniformly improved by closed testing procedures. Clearly, equivalence to a closed testing procedure is necessary for admissibility. Are all closed testing procedures admissible? In this section, we derive a simple condition for admissibility of monotone procedures that is both necessary and sufficient. We show that admissibility of the monotone procedure $\mathbf{d}_{\boldsymbol{\phi}}$ follows directly from admissibility of its local tests. This is the second main result of this paper.

THEOREM 3. *$\mathbf{d}_{\boldsymbol{\phi}}$ is admissible if and only if the suite $\boldsymbol{\phi}$ is admissible.*

We have already seen from Theorem 1 that only closed testing procedures are admissible. Theorem 3 says that all closed testing procedures are admissible, provided they fulfil the reasonable demand that they are built from admissible local tests. To check admissibility of the local tests, Section A in the Supplementary Material shows that under a weak assumption it is sufficient to check that the local tests exhaust the α-level. Theorem 3 thus makes it easy to guarantee admissibility of monotone procedures.

Unlike Theorem 1, the result of Theorem 3 does not immediately translate to local procedures: even if $\boldsymbol{\phi}$ is admissible, it may happen for some finite $I \subseteq C$ that $\mathbf{d}^I_{\boldsymbol{\phi}}$ can be uniformly improved by some other procedure $\mathbf{d}^I$. About such local improvements, we have the following proposition.

PROPOSITION 1. *If* $\mathbf{d}^I \geq \mathbf{d}^I_{\boldsymbol{\phi}}$ *is admissible, then there is an admissible* $\boldsymbol{\psi}$ *such that* $\mathbf{d}^I = \mathbf{d}^I_{\boldsymbol{\psi}}$ *and, for all* $S \subseteq I$, $\boldsymbol{\psi}_S \geq \boldsymbol{\phi}^I_S$.

Proposition 1 limits the available room for local improvements of admissible monotone procedures. Combining Proposition 1 and Theorem 3 we see that such improvements have to be admissible monotone procedures, and therefore closed testing procedures, themselves. The difference between $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$, if both are admissible, is that for every $S \subseteq I$, $\boldsymbol{\phi}_S$ uses only the local information in $\mathbf{d}^S_{\boldsymbol{\phi}}(S)$, but the same does not necessarily hold for $\boldsymbol{\psi}_S$.

In Section B in the Supplementary Material, we give an example of a local improvement of an admissible monotone procedure. Local improvements are also possible in case null hypotheses are composite, using the Partitioning Principle, as shown in Finner and Strassburger (2002), Examples 4.1–4.3, and Goeman and Solari (2010), Section 4. For many well-known procedures, for example, Holm's procedure under arbitrary dependence, we believe that local improvements do not exist. However, we have no general theory on the relationship between admissibility of a monotone procedure and admissibility of its local members. We leave this as an open problem.

9. Consonance and familywise error. Theorem 3 establishes a necessary and sufficient condition for admissibility of monotone true discovery guarantee procedures, and therefore of FDP-controlling procedures. At first sight, our results may seem at odds with those of Romano, Shaikh and Wolf (2011), who proved that for FWER control, which is a special case of the true discovery guarantee requirement, only consonant procedures are admissible. However, this seeming contradiction disappears when we realize that admissibility of a procedure as a true discovery guarantee procedure does not automatically imply admissibility as a FWER controlling procedure and vice versa.

We call a procedure $\mathbf{d}^I$ *consonant* if it has the property that for every $S \subseteq I$, $\mathbf{d}^I(S) > 0$ implies that for at least one $i \in S$ we have $\mathbf{d}^I(\{i\}) = 1$, almost surely for all $\mathrm{P} \in \Omega$. Conceptually, consonant procedures allow pinpointing of effects. If $\mathbf{d}^I(S) > 0$, signal has been detected somewhere in S. A consonant procedure in this case can always find at least one elementary hypothesis to pin the effect down on. This is a desirable property, as it can be unsatisfactory for a researcher to know that an effect exists but not where it can be found. However, Goeman et al. (2019) argued that for FDP control, nonconsonant procedures can be far more powerful in large-scale multiple testing procedures than consonant ones.

In Section C of the Supplementary Material, we go more deeply into the theory of consonant procedures in relation to admissibility of procedures as FWER controlling procedures. We extend the result of Romano, Shaikh and Wolf (2011), showing that admissible FWER controlling procedures must be closed testing procedures with consonant local tests, but also closed testing procedures with admissible local tests. Conversely, if the local tests are both admissible and consonant, then the resulting closed testing procedure is admissible.

10. Improving methods 1: Meinshausen and Rice (2006). Existing methods may be improved by embedding them in a closed testing procedure. We illustrate this with the method of Higher Criticism (Donoho and Jin (2004)), which defines a global test for the null hypothesis H_I, as follows. Let $I = \{1, \ldots, m\}$, and assume we have p-values $\mathbf{p}_1, \ldots, \mathbf{p}_m$, independent and stochastically larger than uniform under H_I. For this null hypothesis, Higher Criticism defines the test

$$\boldsymbol{\phi}_I = \mathbb{1}\left\{ \max_{k_0 \leq j \leq k_1} \frac{\sqrt{m}(j/m - \mathbf{p}_{(j)})}{\sqrt{\mathbf{p}_{(j)}(1 - \mathbf{p}_{(j)})}} \geq a_m \right\},$$

for suitably chosen k_0 and k_1, where $\mathbf{p}_{(1)} \le \cdots \le \mathbf{p}_{(m)}$ are the sorted p-values, and a_m is a suitably chosen critical value. Donoho and Jin (2004) proposed $a_m = (1+a)\sqrt{2\log\log(m)}$ for some $a > 0$, assuming large m. Several finite-m adjustments have been proposed (Hall and Jin (2010), Barnett and Lin (2014)). We will use $k_0 = 1$ and $k_1 = m$. Meinshausen and Rice (2006) improved upon Higher Criticism by showing that discoveries may also be counted, proving that

$$\mathbf{f}_I = \left\lceil \max_{t\in[0,1)} \frac{|\{i \in I\colon p_i \le t\}| - mt - a_m\sqrt{mt(1-t)}}{1-t} \right\rceil$$

is a $(1-\alpha)$-lower confidence bound for the number of false hypotheses $|I_1|$. We have $\boldsymbol{\phi}_I = \mathbb{1}\{\mathbf{f}_I > 0\}$, so $\mathbf{f}_I$ is consistent with the higher criticism test, and uniformly improves it as a true discovery guarantee procedure.

Can we improve $\mathbf{f}_I$ further? First, we can use (7) to interpolate, getting

$$\mathbf{d}^I(S) = \mathbf{f}_I - m + |S|. \tag{11}$$

The resulting method is consonant, and by Corollary 1 it is equivalent to a closed testing procedure with local tests $\boldsymbol{\psi}_S = \mathbb{1}\{|S| > m - \mathbf{f}_I\}$ for every $S \subseteq I$, where we note that $\boldsymbol{\psi}_I = \boldsymbol{\phi}_I$. The interpolated method improves upon $\mathbf{f}_I$ by giving nontrivial $\mathbf{d}^I(S)$ for $S \ne I$ with large $|S|$, but still has $\mathbf{d}^I(I) = \mathbf{f}_I$.

Further improvement is possible by noting that the suite $\boldsymbol{\psi}$ is not admissible. In fact, $\boldsymbol{\psi}$ is uniformly improved by $\boldsymbol{\phi}$, the suite of Higher Criticism local tests. This test is suggested by the recipe of Theorem 1 for improving methods. In Section E of the Supplementary Material we show that

$$\boldsymbol{\phi}_S \ge \boldsymbol{\psi}_S \quad \text{for all } S \subseteq I, \tag{12}$$

and that $\boldsymbol{\phi}_S$ uniformly improves $\boldsymbol{\psi}_S$ for $\varnothing \subset S \subset I$. It follows that $\mathbf{d}^I_{\boldsymbol{\phi}}$ uniformly improves $\mathbf{d}^I$, and that even $\mathbf{d}^I_{\boldsymbol{\phi}}(I)$ uniformly improves $\mathbf{f}_I$ as a confidence bound for $|I_1|$, as we shall see.

To solve the issue of computing $\mathbf{d}^I_{\boldsymbol{\phi}}$, we write (Gontscharuk, Landwehr and Finner (2016))

$$\boldsymbol{\phi}_S = \mathbb{1}\{\mathbf{p}_{(i:S)} \le l_{i:|S|} \text{ for at least one } i = 1,\dots,|S|\}, \tag{13}$$

where $\mathbf{p}_{(i:S)}$, for $1 \le i \le |S|$, is the ith smallest p-value among the multiset $\{p_i : i \in S\}$, and

$$l_{i:s} = \frac{2i + a_s^2 - \sqrt{(2i+a_s^2)^2 - 4i^2(s+a_s^2)/s}}{2(s+a_s^2)}. \tag{14}$$

Written like this, we see that $\boldsymbol{\phi}$ is similar to the Simes tests investigated by Goeman et al. (2019). For calculating $\mathbf{d}^I_{\boldsymbol{\phi}}(S)$, we can use a generalization of the algorithm presented in that paper, given as Lemma 6.

LEMMA 6. *If $\boldsymbol{\phi}_S$, $\varnothing \ne S \subseteq I$, is of the form (13), with $l_{i:m} \ge l_{i:n}$ for all $i \ge 1$ and $0 \le m \le n$, then*

$$\boldsymbol{\phi}^I_S = \mathbb{1}\{\mathbf{p}_{(i:S)} \le l_{i:\mathbf{h}_I} \text{ for at least one } i = 1,\dots,|S|\},$$

and

$$\mathbf{d}^I_{\boldsymbol{\phi}}(S) = \max_{1\le u\le|S|} 1 - u + |\{i \in S\colon \mathbf{p}_i \le l_{u:\mathbf{h}_I}\}|, \tag{15}$$

where

$$\mathbf{h}_I = \max\{s \in \{0,\dots,|I|\} : \mathbf{p}_{(|I|-s+i:I)} > l_{i:s}, \text{ for } i = 1,\dots,s\}.$$

The lemma offers calculation in quadratic time in the general case. For Higher Criticism, $\mathbf{h}_I$ can be calculated using bisection as in Goeman et al. (2019), reducing computation time even to $O(m \log(m))$. We give a condition for the use of bisection in Section F of the Supplementary Material.

To illustrate the new method $\mathbf{d}^I_{\boldsymbol{\phi}}(S)$ we used a simple simulation using settings by Donoho and Jin (2004). We used $|I| = m = 10^6$ independent one-sided z-tests. Of these, 10^3 were under the alternative, with a mean shift of $\sqrt{0.30 \log(m)} \approx 2.04$. We used $a = 1.08$ in the calculation of the critical value, which empirically gives good control of type I error for $m \approx 10^6$ and $\alpha = 0.05$. We used 10^4 replications. The power of Higher Criticism in this setting is 98.0%.

In this simulation, we found that $\mathbf{d}^I_{\boldsymbol{\phi}}(I)$ indeed improved Meinshausen and Rice's $\mathbf{f}_I$, although in this setting an improvement was found in only 2.2% of the realizations. More importantly, however, the new $\mathbf{d}^I_{\boldsymbol{\phi}}(S)$ also makes meaningful statements for $S \neq I$. Figure 1 (bottom) gives 20 realizations of $\mathbf{d}^I(\mathbf{K}_i)$ as a function of i, where $\mathbf{K}_i$ consists of the indices of the i hypotheses with smallest p-values, with ties broken arbitrarily, as well as the expected curve. Figure 1 (top) gives the estimate of $\mathrm{P}(\mathbf{d}^I_{\boldsymbol{\phi}}(\mathbf{K}_i) > 0)$. We see that, by embedding Higher Criticism into a closed testing procedure, even with this weak signal we may make much stronger statements than only pure detection. In about 88.3% of the realizations, we confidently detected signal within the set of 100 hypotheses with smallest p-values; in about 67.1% this signal was in the top 10, and in 38.3% of the realizations we even had a confident rejection of the single hypothesis with the smallest p-value. Substantial improve-

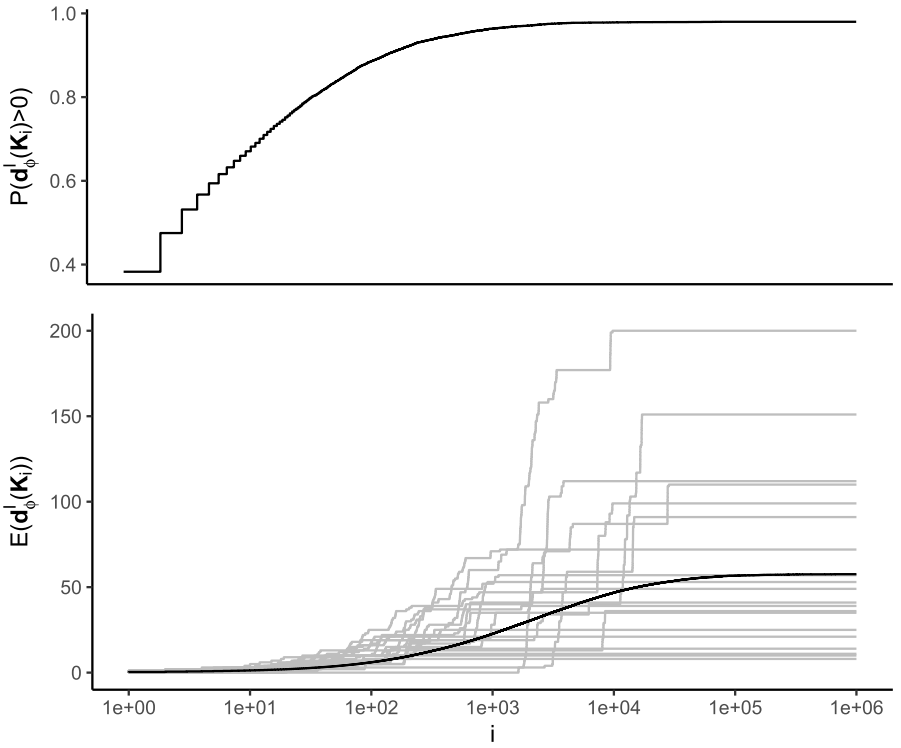

FIG. 1. *Top*: *estimated probability of detection among the* i *hypotheses with smallest* p*-values, using closed testing with the higher criticism local test for* $m = 10^6$ *independent* p*-values. Bottom*: *Lower confidence bounds on the number of true discoveries among the same sets. Grey lines are* 20 *individual realizations*; *the black curve is the average over* 10^4 *realizations.*

ment of $\mathbf{d}^I_{\boldsymbol{\phi}}$ over $\mathbf{d}^I$, defined in (11), is also clear since the latter gives $\mathbf{d}^I(S) = 0$ whenever $|S| \le m - \mathbf{f}_I \approx 0.9999 \times 10^6$.

The strict line drawn by Donoho and Jin (2004) between detectable and estimable effects is therefore, in our view, more like a gray zone. If we can detect effects, closed testing can count them. Though we may be unable to pinpoint effects, closed testing can close in on them.

11. Improving methods 2: Katsevich and Ramdas (2020). We chose as a second example a method recently proposed by Katsevich and Ramdas (2020). This elegant method (abbreviated K&R) allows users to choose a p-value cutoff for significance post hoc, and uses stochastic process arguments to control both FDP and FDR. We focus on the FDP control property here. We use the same setting as in the previous section of m p-values, independent and stochastically uniform under the null, and we use the same notation.

Katsevich and Ramdas showed that, if $\alpha \le 0.31$,

$$\mathrm{P}\big(|\mathbf{K}_i \cap I_0| \le c(1 + m\mathbf{p}_{(i)}) \text{ for all } 1 \le i \le m\big) \ge 1 - \alpha,$$

where $c = -\log(\alpha)/\log(1 - \log(\alpha))$, and $\mathbf{p}_{(i)}$ is the ith smallest p-value. For $\alpha = 0.05$ we have $c \approx 2.163$. As in Section 3, we can write this as a true discovery guarantee procedure on I by writing

$$\mathbf{d}^I(S) = \begin{cases} 0 \vee \lceil i - c(1 + m\mathbf{p}_{(i)}) \rceil & \text{if } S = \mathbf{K}_i \text{ for some } 1 \le i \le m, \\ 0 & \text{otherwise,} \end{cases} \tag{16}$$

where we round up to ensure that $\mathbf{d}^I(S)$ is always an integer.

Is the procedure (16) admissible, and if not, how can we improve it? We apply the results of this paper. First, we remark that the method as defined is not coherent. The interpolation of the procedure is given by

$$\mathbf{d}^I(S) = 0 \vee \max_{k=1,\dots,|S|} \lceil k - c(1 + m\mathbf{p}_{(k:S)}) \rceil, \tag{17}$$

taking $\mathbf{d}^I(\varnothing) = 0$ implicitly. The derivation of equation (17) is given in Section E of the Supplementary Material.

We note that interpolated method (17) makes nontrivial statements for sets S not of the form $\mathbf{K}_i$, and may even improve $\mathbf{d}^I(\mathbf{K}_i)$ for some i. It may be checked using Lemma 3 that the procedure (17) is coherent, so no further rounds of interpolation are needed. The K&R procedure was not developed for a specific scale m. Writing $|I|$ for m in (17) we have a procedure that is defined for general I, and it is easy to check that $(\mathbf{d}^I)_{I \subseteq \mathbb{N}, |I| < \infty}$ is monotone.

Next, we use Theorem 1 to embed the method in a closed testing procedure, which results in further improvement of the procedure. By the theorem, the local test for finite $\varnothing \ne S \subseteq \mathbb{N}$ is given by

$$\boldsymbol{\phi}_S = \mathbb{1}\big\{\mathbf{p}_{(i:S)} \le (i - c)/c|S| \text{ for at least one } i = 1, \dots, |S|\big\}.$$

We will construct the closed testing procedure based on this local test. We note that the local test is of the form assumed in Lemma 6, with

$$l_{i:s} = \frac{i - c}{cs} \tag{18}$$

if $s \ne 0$, and $l_{i:0} = 1$. By Theorem 1, the method (15) with (18) is everywhere at least as powerful as the interpolated method (17). In fact, it is a uniform improvement of that method as we shall see in the simulation experiment below.

TABLE 1
Values of c_s calculated by Monte Carlo integration (10^6 samples)

s	1	2	3	4	5	7	10	15	20	50	100	500	1000
c_s	0.95	1.38	1.55	1.64	1.71	1.78	1.84	1.90	1.92	1.98	2.00	2.01	2.02

The next question is whether the method defined by (15) with (18) is admissible, or whether it can be further improved. We can verify this using Theorem 3 by checking whether the local tests are admissible. It is immediately obvious that this is not the case. Taking, for example, $|S| = 1$, we see that at $\alpha = 0.05$ with $c \approx 2.163$ we have $\boldsymbol{\phi}_S = \mathbb{1}\{\mathbf{p}_{(1:S)} \leq (1-c)/c < 0\} = 0$, which is clearly not admissible. We may freely decrease c to $c_1 = 1/(1+\alpha) \approx 0.952$ to obtain the uniformly more powerful local test $\boldsymbol{\phi}_S = \mathbb{1}\{\mathbf{p}_{(1:S)} \leq \alpha\}$. We can use the same reasoning for $|S| = 2, 3, \ldots$, decreasing the value of c to the minimal value that guarantees type I error control. This value may easily be calculated numerically since the worst case distribution of $(\mathbf{p}_i)_{i \in S}$ under H_S is the independent uniform case. We obtain a new local test of the form (13) with

$$l_{i:s} = \frac{i - c_s}{c_s s}. \tag{19}$$

We tabulated the values of c_s (taking $\alpha = 0.05$) for some values of s in Table 1. Note that $c_s \leq c$ for all s. Since $l_{i:s}$ is monotone in c_s the new local test uniformly improves the old one. We note that with these choices of c_s the critical values $l_{i:s}$ cannot be further increased without destroying type I error control of the local tests, so we conclude that the resulting local tests are admissible, provided that the test $\mathbb{1}\{\mathbf{p}_i \leq \alpha\}$ is admissible as an α-level local test of H_i for all i and α. Assuming this, by Theorem 3 the resulting true discovery guarantee procedure is admissible. We note that, since c_s is increasing in s, (19) still fulfils the conditions of Lemma 6, so that the admissible method is still computable using Lemma 6.

We have started with the procedure of Katsevich and Ramdas (2020) and improved it uniformly in three steps: the method was first improved by interpolation. The resulting coherent method was further improved by embedding it in a closed testing procedure, and finally that closed testing procedure was improved to an admissible method by improving its local tests. This way we obtained a sequence of four methods, each uniformly improving the previous one. We will call them the *original* (16), *coherent* (17), *closed*, defined by (15) with (18), and *admissible* method, defined by (15) with (19). We performed a small simulation experiment to assess the relative improvement made with each of the three steps. We used $m = 1000$ hypotheses, of which m_0 were true, and $m_1 = m - m_0$ false. We sampled p-values independently. For true null hypotheses, we used $\mathbf{p}_i \sim \mathcal{U}(0,1)$. For false null hypotheses, we used $\mathbf{p}_i \sim \Phi^{-1}(-\gamma \mathbf{Z})$, where Φ is the standard normal distribution function, and $\mathbf{Z} \sim \mathcal{N}(0,1)$. We took values $m_1 = 8, 40, 200$ and $\gamma = 2, 2.5, 3$. A true discovery guarantee procedure gives exponentially many output values. We report only results for sets $\mathbf{K}_i$ of the i smallest p-values, as the original method did. Calculation for the closed and admissible methods was in quadratic time based on Lemma 6. We calculated $\mathbf{d}^I(\mathbf{K}_i)$, $i = 1, \ldots, m$ for the closed and admissible methods in less than 0.1 seconds on a standard PC.

The results are given in Figure 2, in terms of number of true discoveries $\mathbf{d}^I(\mathbf{K}_i)$ (top) and in terms of true discovery proportions $\mathbf{d}^I(\mathbf{K}_i)/i$ (TDP; bottom). For each setting and each method we report the average value of $\mathbf{d}^I(\mathbf{K}_i)/i$ over 10^4 simulations. Several things can be noticed about these simulation results.

The most important finding is that all three improvement steps can be substantial. The improvement from the original to the coherent procedure is perhaps largest. It is especially

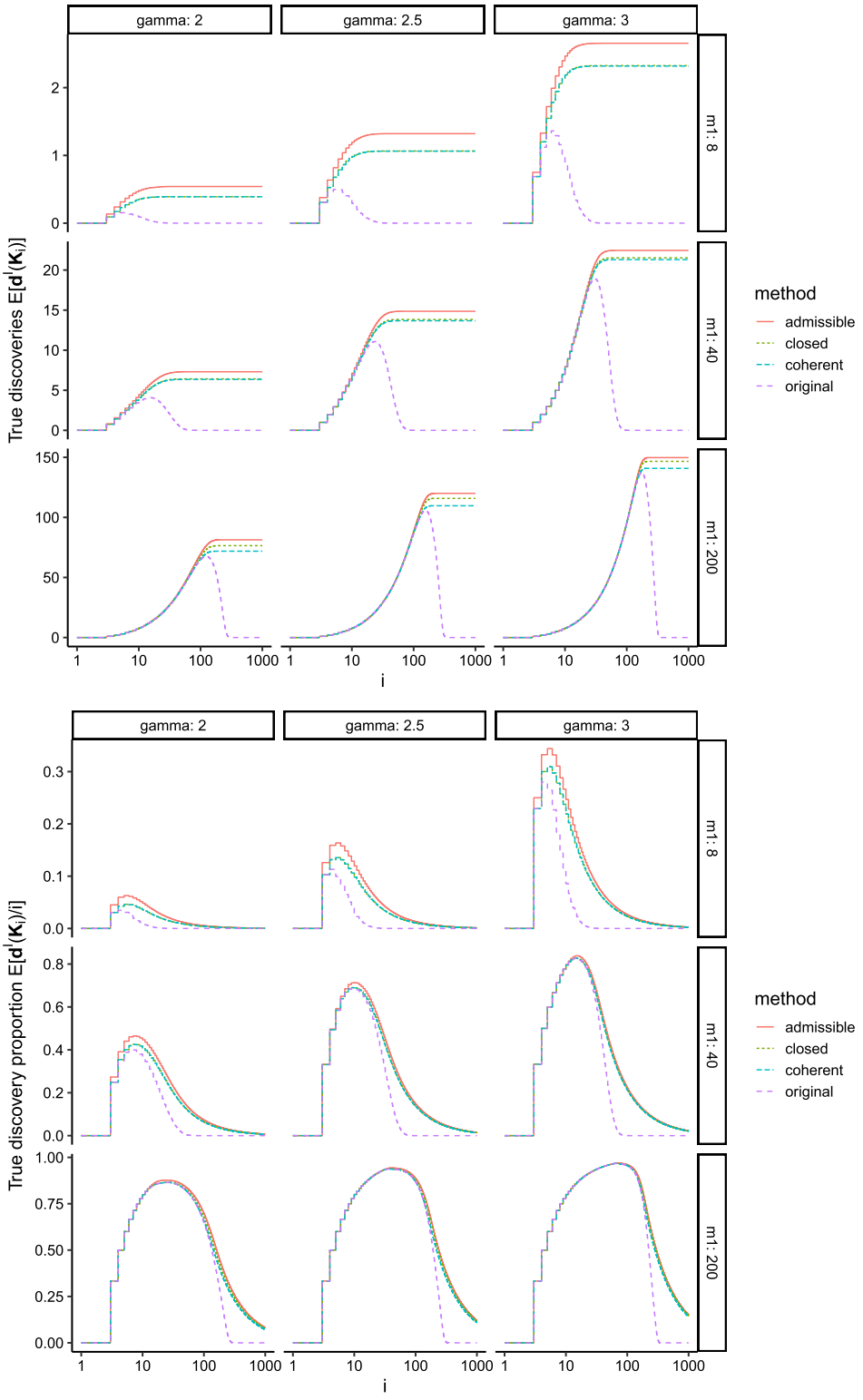

FIG. 2. *Lower confidence bound on the number of true discoveries* (*top*) *and true discovery proportion* (*bottom*) *among the* i *hypotheses with smallest* p*-values, relating to the method of Katsevich and Ramdas* (2020) *and its successive uniform improvements. Curves are the averages over* 10^4 *realizations. Since each method uniformly improves the previous one, any observed difference is automatically significant.*

noticeable for large rejected sets, where the original method may all too often give $\mathbf{d}^I(S) = 0$, especially if $|S| \gg m_1$. The peak of the TDP is improved if the TDP of the original method was low. The second improvement, from the coherent procedure to closed testing is substantial in terms of numbers of discoveries only if m_1 is large. This is natural because the improvement can be seen as a "step-down" argument, implicitly incorporating an estimate of m_1 into the procedure. Even with large m_1 it is the improvement is negligible on the TDP scale. The final improvement from the initial closed testing to the admissible procedure is clear throughout the figure. It is largest in terms of number of discoveries when m_1 is large, but largest in terms of TDP when m_1 is small.

Although the improvement from the coherent to the closed procedure seems the smallest one, we emphasize that closed testing was crucial for the construction of the admissible procedure. We also note that the method of K& R, as well as its improvements, make no useful FWER rejections at all: in Figure 2 we see that $\mathrm{E}[\mathbf{d}^I(\mathbf{K}_i)] \approx 0$ for $i \leq 2$. This phenomenon is analyzed in more detail in Section D of the Supplementary Material.

12. Discussion. We have studied the class of all methods controlling tail probabilities of false discovery proportions. This class encompasses very diverse methods, for example, familywise error control procedures, false discovery exceedance procedures, joint error rate controlling methods and cluster inference. We have shown that all such procedures can be written as methods simultaneously controlling false discovery proportions over all subsets of the family of hypotheses. This rewrite, trivial as it may be in some cases, is valuable in its own right, because it makes it possible to study methods jointly that seemed incomparable before, and takes a step in reducing the "plethora of error rates" lamented by Benjamini (2010). Moreover, methods that were constructed to give nontrivial error bounds for only a single random hypothesis set of interest, now give simultaneous error bounds for all such sets, allowing their use in flexible selective inference in the sense advocated by Goeman and Solari (2011).

We have formulated all such procedures in terms of a $(1 - \alpha)$-true discovery guarantee, that is, giving a $(1 - \alpha)$-lower confidence bound to the number of true discoveries in each set, because this representation is mathematically easier to work with. Also, by emphasizing true rather than false discoveries, it gives a valuable positive frame to the multiple testing problem. Otherwise, this change in representation is purely cosmetic; we may continue to speak of FDP control procedures.

Admissibility is a very weak requirement for statistical tests, as under a weak assumption all tests that exhaust their α-level are admissible. However, admissibility is not so easy to achieve for FDP control procedures. We have formulated a condition for admissibility of FDP control procedures that is both necessary and sufficient. All admissible FDP control procedures are closed testing procedures, and all closed testing procedures are admissible as FDP control procedures, provided they are well designed in the sense that all their local tests are admissible. Apparently, control of false discovery proportions and closed testing procedures are so closely tied together that the relationship seems almost tautological. Admissibility is closely tied to optimality. Since optimal methods must be admissible, and admissible methods must be closed testing procedures, we have shown that only closed testing procedures can be optimal.

This theoretical insight has great practical value for methods designers. It can be used to uniformly improve existing methods, as we have demonstrated on the methods of Meinshausen and Rice (2006) and Katsevich and Ramdas (2020). Given a procedure that controls FDP, we first make sure it is coherent. Next, we can explicitly construct the local tests implied by the procedure, and turn it into a closed testing procedure. To check admissibility, we now only need to check admissibility of the local tests. Each step may result in

substantive improvement, as we have shown in simulations. Alternatively, when designing a method we may start from a suite of local tests that has good power properties. The options are virtually unlimited here. The validity of the local test as an α-level test guarantees control of FDP. Correlations between test statistics, that often complicate multiple testing procedures, should be properly taken into account by the local test. Admissibility of the local tests guarantees admissibility of the resulting procedure. In both cases the computational problem remains that closed testing may require exponentially many tests, but this is the only remaining problem. Polynomial time shortcuts are possible. Ideally these are exact, as for K&R and higher criticism above, and admissibility is retained. If the full closed testing procedure is not computable for large testing problems, we may settle for an inadmissible but computable method, based on a conservative shortcut (e.g., Hemerik and Goeman (2018), Hemerik, Solari and Goeman (2019)). It may still be worthwhile to compare such a method to full closed testing in small-scale problems to see how much power is lost.

Concretely, in Lemma 6 have given an exact computational shortcut that can be used for closed testing with a wide range of local tests, for example, to the local tests implied by the False Discovery Rate controlling procedures of Blanchard and Roquain (2009), to other local tests implied by the Dvoretzky-Kiefer-Wolfowitz inequality (Genovese and Wasserman (2004), Meinshausen and Rice (2006)), to the local tests implied by second and higher order generalized Simes constants (Cai and Sarkar (2008), Gou and Tamhane (2014)), and to the local tests implied by the FDR controlling procedures of Benjamini and Liu (1999), and Romano and Shaikh ((2006), Equation 4.1). Using the lemma, computation time of closed testing is quadratic, even reducing to linearithmic in some cases.

We have defined admissibility in terms of simultaneous FDP control for all possible subsets of the family of hypotheses. In some cases we may not be interested in all of these sets as, for example, when targeting FWER control exclusively. Even when only interested in some of the subsets, we retain the result that admissible procedures must be closed testing procedures. We lose, however, the property that all such procedures are automatically admissible if they have admissible local tests. Additional criteria might come in, such as consonance in the case of familywise error control. Variants of consonance may be useful as well (Brannath and Bretz (2010)).

Our focus has been mostly on monotone procedures. Such procedures are defined for multiple testing problems on different scales simultaneously. Connecting between different scales, they have the property that adding more hypotheses to the multiple testing problem will never result in stronger inference for the hypotheses that were already there. This is an intuitively desirable property by itself, which prevents some paradoxes (Goeman and Solari (2014)). Monotone procedures have additional valuable properties: viewed as closed testing procedures, they have local tests that are truly local: the local test on S uses only the information that the corresponding local procedure $\mathbf{d}^S$ uses. Admissible monotone procedures, however, may sometimes be locally improved, and we have given an example of this. Such improvements, if admissible, must still be closed testing procedures with admissible local tests themselves.

We have restricted to finite testing problems. Extensions to countably infinite problems are of interest, for example, when considering online control (Javanmard and Montanari (2018)). The results of this paper may trivially be extended to allow infinite $|I|$ if we are willing to assume that $|I_1| < \infty$, so that $\mathbf{d}^I < \infty$. If $|I_1|$ is unbounded, care must be taken to scale $\mathbf{d}$ properly to keep it in the nontrivial range. This scaling adds some technical complexity, and is not assumption-free because $\mathbf{d}^I(S)$ scales with the unknown $|S \cap I_1|$. However, since most of the results of this paper compare methods that obviously require the same scaling, we conjecture that the optimality of closed testing will translate to FDP control in countable and even uncountable multiple testing problems. We leave this to future research.

Finally, we remark that we have only considered procedures that control tail probabilities of the false discovery proportion. These methods can also be used for bounding the median FDP (Goeman and Solari (2011)). However, if there is interest in the central tendency of FDP it is more common to bound the mean FDP, better known as False Discovery Rate (FDR). Given the close connection we have established between closed testing and FDP tail probabilities, it is likely that there is also a connection between closed testing and FDR control. Some connections have already been found between Simes-based closed testing and the procedure of Benjamini and Hochberg (1995) by Goeman et al. (2019). It is likely that there are more such connections. Any procedure that controls FDR, since FDR control implies weak FWER control, implies a local test and can therefore be used to construct a closed testing procedure. Conversely, if FDP is controlled with $(1-\alpha)$-confidence at level γ, then FDR is controlled at $\alpha(1-\gamma)+\gamma$, as Lehmann and Romano (2005a) have shown. More profound relationships may be found in the future.

Acknowledgements. This paper was inspired by many discussions during the workshop "Post-selection Inference and Multiple Testing" in Toulouse, February 2018, organized by Pierre Neuvial, Etienne Roquain and Gilles Blanchard. We thank the organizers and all participants, and especially Ruth Heller for asking the question that triggered this research project. We thank Jonathan Rosenblatt and the Israel Science Foundation for financing the computing equipment used for the simulations (grants 926/14 and 900/16). Jelle Goeman was supported by NWO VIDI grant 639.072.412.

SUPPLEMENTARY MATERIAL

Supplement to "Only closed testing procedures are admissible for controlling false discovery proportions" (DOI: 10.1214/20-AOS1999SUPP; .pdf). Supplementary information.

REFERENCES

BARNETT, I. J. and LIN, X. (2014). Analytical p-value calculation for the higher criticism test in finite-d problems. *Biometrika* **101** 964–970. MR3286929 https://doi.org/10.1093/biomet/asu033

BENJAMINI, Y. (2010). Simultaneous and selective inference: Current successes and future challenges. *Biom. J.* **52** 708–721. MR2758547 https://doi.org/10.1002/bimj.200900299

BENJAMINI, Y. and HELLER, R. (2008). Screening for partial conjunction hypotheses. *Biometrics* **64** 1215–1222. MR2522270 https://doi.org/10.1111/j.1541-0420.2007.00984.x

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

BENJAMINI, Y. and LIU, W. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence **82** 163–170. J. Statist. Plann. Inference, 1-2, Multiple comparisons (Tel Aviv, 1996). MR1736441 https://doi.org/10.1016/S0378-3758(99)00040-3

BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122 https://doi.org/10.1214/12-AOS1077

BLANCHARD, G., NEUVIAL, P. and ROQUAIN, E. (2020). Post hoc confidence bounds on false positives using reference families. *Ann. Statist.* **48** 1281–1303. MR4124323 https://doi.org/10.1214/19-AOS1847

BLANCHARD, G. and ROQUAIN, É. (2009). Adaptive false discovery rate control under independence and dependence. *J. Mach. Learn. Res.* **10** 2837–2871. MR2579914

BRANNATH, W. and BRETZ, F. (2010). Shortcuts for locally consonant closed test procedures. *J. Amer. Statist. Assoc.* **105** 660–669. MR2724850 https://doi.org/10.1198/jasa.2010.tm08127

BRETZ, F., MAURER, W., BRANNATH, W. and POSCH, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Stat. Med.* **28** 586–604. MR2655732 https://doi.org/10.1002/sim.3495

CAI, G. and SARKAR, S. K. (2008). Modified Simes' critical values under independence. *Statist. Probab. Lett.* **78** 1362–1368. MR2528334 https://doi.org/10.1016/j.spl.2007.12.018

DELATTRE, S. and ROQUAIN, E. (2015). New procedures controlling the false discovery proportion via Romano-Wolf's heuristic. *Ann. Statist.* **43** 1141–1177. MR3346700 https://doi.org/10.1214/14-AOS1302

DOBRIBAN, E. (2020). Fast closed testing for exchangeable local tests. *Biometrika* **107** 761–768. MR4138990 https://doi.org/10.1093/biomet/asz082

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195 https://doi.org/10.1214/009053604000000265

DUDOIT, S., VAN DER LAAN, M. J. and POLLARD, K. S. (2004). Multiple testing. I. Single-step procedures for control of general type I error rates. *Stat. Appl. Genet. Mol. Biol.* **3** Art. 13, 71. MR2101462 https://doi.org/10.2202/1544-6115.1040

DURAND, G., BLANCHARD, G., NEUVIAL, P. and ROQUAIN, E. (2020). Post hoc false positive control for structured hypotheses. *Scand. J. Stat.* **47** 1114–1148. https://doi.org/10.1111/sjos.12453

EBRAHIMPOOR, M., SPITALI, P., HETTNE, K., TSONAKA, R. and GOEMAN, J. (2020). Simultaneous enrichment analysis of all possible gene-sets: Unifying self-contained and competitive methods. *Brief. Bioinform.* **21** 1302–1312. https://doi.org/10.1093/bib/bbz074

FARCOMENI, A. (2009). Generalized augmentation to control the false discovery exceedance in multiple testing. *Scand. J. Stat.* **36** 501–517. MR2549707 https://doi.org/10.1111/j.1467-9469.2008.00633.x

FINNER, H. and STRASSBURGER, K. (2002). The partitioning principle: A powerful tool in multiple decision theory. *Ann. Statist.* **30** 1194–1213. MR1926174 https://doi.org/10.1214/aos/1031689023

FINOS, L. and FARCOMENI, A. (2011). k-FWER control without p-value adjustment, with application to detection of genetic determinants of multiple sclerosis in Italian twins. *Biometrics* **67** 174–181. MR2898829 https://doi.org/10.1111/j.1541-0420.2010.01443.x

GE, Y. and LI, X. (2012). Control of the false discovery proportion for independently tested null hypotheses. *J. Probab. Stat.* Art. ID 320425, 19. MR2934965 https://doi.org/10.1155/2012/320425

GENOVESE, C. and WASSERMAN, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32** 1035–1061. MR2065197 https://doi.org/10.1214/009053604000000283

GENOVESE, C. R. and WASSERMAN, L. (2006). Exceedance control of the false discovery proportion. *J. Amer. Statist. Assoc.* **101** 1408–1417. MR2279468 https://doi.org/10.1198/016214506000000339

GOEMAN, J. J. and FINOS, L. (2012). The inheritance procedure: Multiple testing of tree-structured hypotheses. *Stat. Appl. Genet. Mol. Biol.* **11** Art. 11, 20. MR2924200 https://doi.org/10.1515/1544-6115.1554

GOEMAN, J. J, HEMERIK, J. and SOLARI, A. (2021). Supplement to "Only closed testing procedures are admissible for controlling false discovery proportions." https://doi.org/10.1214/20-AOS1999SUPP

GOEMAN, J. J. and MANSMANN, U. (2008). Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics* **24** 537–544.

GOEMAN, J. J. and SOLARI, A. (2010). The sequential rejection principle of familywise error control. *Ann. Statist.* **38** 3782–3810. MR2766868 https://doi.org/10.1214/10-AOS829

GOEMAN, J. J. and SOLARI, A. (2011). Multiple testing for exploratory research. *Statist. Sci.* **26** 584–597. MR2951390 https://doi.org/10.1214/11-STS356

GOEMAN, J. J. and SOLARI, A. (2014). Multiple hypothesis testing in genomics. *Stat. Med.* **33** 1946–1978. MR3257576 https://doi.org/10.1002/sim.6082

GOEMAN, J. J., MEIJER, R. J., KREBS, T. J. P. and SOLARI, A. (2019). Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika* **106** 841–856. MR4046036 https://doi.org/10.1093/biomet/asz041

GONTSCHARUK, V., LANDWEHR, S. and FINNER, H. (2016). Goodness of fit tests in terms of local levels with special emphasis on higher criticism tests. *Bernoulli* **22** 1331–1363. MR3474818 https://doi.org/10.3150/14-BEJ694

GOU, J. and TAMHANE, A. C. (2014). On generalized Simes critical constants. *Biom. J.* **56** 1035–1054. MR3270109 https://doi.org/10.1002/bimj.201300258

GUO, W., HE, L. and SARKAR, S. K. (2014). Further results on controlling the false discovery proportion. *Ann. Statist.* **42** 1070–1101. MR3210996 https://doi.org/10.1214/14-AOS1214

GUO, W. and RAO, M. B. (2010). On stepwise control of the generalized familywise error rate. *Electron. J. Stat.* **4** 472–485. MR2657378 https://doi.org/10.1214/08-EJS320

HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38** 1686–1732. MR2662357 https://doi.org/10.1214/09-AOS764

HEMERIK, J. and GOEMAN, J. J. (2018). False discovery proportion estimation by permutations: Confidence for significance analysis of microarrays. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 137–155. MR3744715 https://doi.org/10.1111/rssb.12238

HEMERIK, J., SOLARI, A. and GOEMAN, J. J. (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika* **106** 635–649. MR3992394 https://doi.org/10.1093/biomet/asz021

HENNING, K. S. S. and WESTFALL, P. H. Closed testing in pharmaceutical research: Historical and recent developments. *Stat. Biopharm. Res.* **7** 126–147. https://doi.org/10.1080/19466315.2015.1004270

HOMMEL, G. and HOFFMANN, T. (1988). Controlled uncertainty. In *Multiple Hypothesenprüfung/Multiple Hypotheses Testing* 154–161. Springer, Berlin.

JANSON, L. and SU, W. (2016). Familywise error rate control via knockoffs. *Electron. J. Stat.* **10** 960–975. MR3486422 https://doi.org/10.1214/16-EJS1129

JAVANMARD, A. and MONTANARI, A. (2018). Online rules for control of false discovery rate and false discovery exceedance. *Ann. Statist.* **46** 526–554. MR3782376 https://doi.org/10.1214/17-AOS1559

KATSEVICH, E. and RAMDAS, A. (2020). Simultaneous high-probability bounds on the false discovery proportion in structured, regression, and online settings. *Ann. Statist.* **48** 3465–3487. https://doi.org/doi:10.1214/19-AOS1938

KORN, E. L., TROENDLE, J. F., MCSHANE, L. M. and SIMON, R. (2004). Controlling the number of false discoveries: Application to high-dimensional genomic data. *J. Statist. Plann. Inference* **124** 379–398. MR2080371 https://doi.org/10.1016/S0378-3758(03)00211-8

LEHMANN, E. L. and ROMANO, J. P. (2005a). Generalizations of the familywise error rate. *Ann. Statist.* **33** 1138–1154. MR2195631 https://doi.org/10.1214/009053605000000084

LEHMANN, E. L. and ROMANO, J. P. (2005b). *Testing Statistical Hypotheses*, 3rd ed. *Springer Texts in Statistics*. Springer, New York. MR2135927

MARCUS, R., PERITZ, E. and GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655–660. MR0468056 https://doi.org/10.1093/biomet/63.3.655

MEIJER, R. J. and GOEMAN, J. J. (2015a). Multiple testing of gene sets from gene ontology: Possibilities and pitfalls. *Brief. Bioinform.* **17** 808–818.

MEIJER, R. J. and GOEMAN, J. J. (2015b). A multiple testing method for hypotheses structured in a directed acyclic graph. *Biom. J.* **57** 123–143. MR3298222 https://doi.org/10.1002/bimj.201300253

MEIJER, R. J., KREBS, T. J. P. and GOEMAN, J. J. (2015). A region-based multiple testing method for hypotheses ordered in space or time. *Stat. Appl. Genet. Mol. Biol.* **14** 1–19. MR3305943 https://doi.org/10.1515/sagmb-2013-0075

MEINSHAUSEN, N. (2006). False discovery control for multiple tests of association under general dependence. *Scand. J. Stat.* **33** 227–237. MR2279639 https://doi.org/10.1111/j.1467-9469.2005.00488.x

MEINSHAUSEN, N. (2008). Hierarchical testing of variable importance. *Biometrika* **95** 265–278. MR2521583 https://doi.org/10.1093/biomet/asn007

MEINSHAUSEN, N. and RICE, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.* **34** 373–393. MR2275246 https://doi.org/10.1214/009053605000000741

POLINE, J.-B. and MAZOYER, B. M. (1993). Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise-ratio pixel clusters. *J. Cereb. Blood Flow Metab.* **13** 425–437.

ROMANO, J. P. and SHAIKH, A. M. (2006). Stepup procedures for control of generalizations of the familywise error rate. *Ann. Statist.* **34** 1850–1873. MR2283720 https://doi.org/10.1214/009053606000000461

ROMANO, J. P., SHAIKH, A. and WOLF, M. (2011). Consonance and the closure method in multiple testing. *Int. J. Biostat.* **7** Art. 12, 27. MR2775079 https://doi.org/10.2202/1557-4679.1300

ROSENBLATT, J. D., FINOS, L., WEEDA, W. D., SOLARI, A. and GOEMAN, J. J. (2018). All-resolutions inference for brain imaging. *NeuroImage* **181** 786–796. https://doi.org/10.1016/j.neuroimage.2018.07.060

ROSENBLUM, M., LIU, H. and YEN, E.-H. (2014). Optimal tests of treatment effects for the overall population and two subpopulations in randomized trials, using sparse linear programming. *J. Amer. Statist. Assoc.* **109** 1216–1228. MR3265692 https://doi.org/10.1080/01621459.2013.879063

ROSSET, S., HELLER, R., PAINSKY, A. and AHARONI, E. (2018). Optimal Procedures for Multiple Testing Problems. Preprint. Available at arXiv:1804.10256.

SARKAR, S. K. (2007). Stepup procedures controlling generalized FWER and generalized FDR. *Ann. Statist.* **35** 2405–2420. MR2382652 https://doi.org/10.1214/009053607000000398

SHAFFER, J. P. (1986). Modified sequentially rejective multiple test procedures. *J. Amer. Statist. Assoc.* **81** 826–831.

SONNEMANN, E. (1982). Allgemeine Lösungen Multipler Testprobleme. Universität Bern. Institut für Mathematische Statistik und Versicherungslehre.

SONNEMANN, E. (2008). General solutions to multiple testing problems. *Biom. J.* **50** 641–656. MR2542333 https://doi.org/10.1002/bimj.200810462

SONNEMANN, E. and FINNER, H. (1988). Vollständigkeitssätze für multiple Testprobleme. In *Multiple Hypothesenprüfung/Multiple Hypotheses Testing* 121–135. Springer, Berlin.

SUN, W., REICH, B. J., CAI, T. T., GUINDANI, M. and SCHWARTZMAN, A. (2015). False discovery control in large-scale spatial multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 59–83. MR3299399 https://doi.org/10.1111/rssb.12064

WANG, J. and OWEN, A. B. (2019). Admissibility in partial conjunction testing. *J. Amer. Statist. Assoc.* **114** 158–168. MR3941245 https://doi.org/10.1080/01621459.2017.1385465

WESTFALL, P. and YOUNG, S. (1993). *Resampling-Based Multiple Testing*. Wiley, New York.