E-values as unnormalized weights in multiple testing

Ruodu Wang* Aaditya Ramdas†

May 17, 2022

Abstract

Most standard weighted multiple testing methods require the weights to deterministically add up to the number of hypotheses being tested (equivalently, the average weight is unity). We show that this normalization is not required when the weights are not constants, but are themselves e-values obtained from independent data. This could result in a massive increase in power, especially if the non-null hypotheses have e-values much larger than one. More broadly, we study how to combine an e-value and a p-value, and design multiple testing procedures where both e-values and p-values are available for every hypothesis (or one of them is available for an implied hypothesis). For false discovery rate (FDR) control, analogous to the Benjamini-Hochberg procedure with p-values (p-BH) and the recent e-BH procedure for e-values, we propose the ep-BH and the pe-BH procedures, which have valid FDR guarantee under different dependence assumptions. The procedures are designed based on several admissible combining functions for p/e-values. The method can be directly applied to family-wise error rate control problems. We also collect several miscellaneous results, such as a tiny but uniform improvement of e-BH, a soft-rank permutation e-value, and the use of e-values as masks in interactive multiple testing.

Keywords: multiple testing, FDR, p-values, betting scores, supermartingales

Contents

*Department of Statistics and Actuarial Science, University of Waterloo. E-mail: wang@uwaterloo.ca.
†Departments of Statistics and Machine Learning, Carnegie Mellon University. E-mail: aramdas@cmu.edu.

1 Introduction

Procedures for controlling the false discovery rate (FDR) have been extensively studied since the seminal work of Benjamini and Hochberg [3] for p-values. Recently, the corresponding FDR procedure for e-values, termed the e-BH procedure, is studied by Wang and Ramdas [31] and further by Xu et al. [33]. Similarly to the literature, we use "e-value" as an abstract umbrella term which encompasses betting scores, likelihood ratios, and stopped supermartingales, which appear in the recent literature, e.g., Shafer [21], Vovk and Wang [29], Grünwald et al. [12], Wasserman et al. [32] and Howard et al. [13, 14].

The main purpose of this note is to collect several advances on the e-BH procedure as well as its generalizations. In particular, we will design procedures for settings in which we have both p-values and e-values, perhaps from different sources.

One key insight of this paper, which is methodologically new and practically meaningful, appears in the context of weighted multiple testing with p-values. Usually, the weighting of p-values requires the weights to deterministically sum to the number of hypotheses; hence none of the weights can be larger than the number of hypotheses being tested, and if any one of the weights is very large then all others must be small, so that they can sum to a fixed constant. Since larger weights are better for power, both of these previous implications hurt power. Our insight here is that the following:

> Independent e-values can be directly used as weights for p-values in all standard multiple testing procedures, without needing to normalize them in any way. This can lead to huge increases in power relative to standard weighted procedures.

We first describe the basic setting. Our terminology is the same as Wang and Ramdas [31]. Let $H_1, \dots, H_K$ be K hypotheses, and write $\mathcal{K} = \{1, \dots, K\}$. Let the true (unknown) data-generating probability measure be denoted by $\mathbb{P}$. For each $k \in \mathcal{K}$, it is useful to think of hypothesis H_k as implicitly defining a set of joint probability measures, and H_k is called a true null hypothesis if $\mathbb{P} \in H_k$. A *p-variable* P for a hypothesis H is a random variable that satisfies $Q(P \le \alpha) \le \alpha$ for all $\alpha \in (0, 1)$ and all $Q \in H$. In other words, a p-variable is stochastically larger than U[0, 1], often truncated at 1. An *e-variable* E for a hypothesis H is a $[0, \infty]$-valued random variable satisfying $\mathbb{E}^Q[E] \le 1$ for all $Q \in H$. E-variables are often obtained from stopping an *e-process* $(E_t)_{t \ge 0}$, which is a nonnegative stochastic process adapted to a pre-specified filtration such that $\mathbb{E}^Q[E_\tau] \le 1$ for any stopping time τ and any $Q \in H$ (in case H is a singleton, an example of e-process is a supermartingale with initial value 1). Let $\mathcal{N} \subseteq \mathcal{K}$ be the set of indices of true null hypotheses, which is unknown to the decision maker, and K_0 be the number of true null hypotheses, thus the cardinality of $\mathcal{N}$.

Two settings of testing multiple hypotheses were considered by Wang and Ramdas [31]:

1. For each $k \in \mathcal{K}$, P_k is a p-variable for H_k, and the p-value p_k is its realization.
2. For each $k \in \mathcal{K}$, E_k is an e-variable for H_k, and the e-value e_k is its realization.

In this paper we will consider the setting where both P_k and E_k are available for each H_k.

Since we are testing whether $\mathbb{P} \in H_k$ for each k, we will only use the following condition: If $k \in \mathcal{N}$, then P_k is a p-variable for $\{\mathbb{P}\}$, and E_k is an e-variable for $\{\mathbb{P}\}$. There is no restrictions of P_k and E_k if they $k \notin \mathcal{N}$. We will omit $\mathbb{P}$ in the statements (by simply calling them p-variables and e-variables) and the expectations.

A testing procedure $\mathcal{D}$ is a mapping that produces a subset of $\mathcal{K}$ representing the indices of rejected hypotheses based on observed p-values, e-values, or a combination of two as the input. We tacitly require that all testing procedures are Borel functions. The terms "p-values/e-values" refer to both the random variables and their realized values; these should be clear from the context.

Let $\mathcal{D}$ be a testing procedure. The rejected hypotheses by $\mathcal{D}$ are called discoveries. We write $F_\mathcal{D} := |\mathcal{D} \cap \mathcal{N}|$ as the number of true null hypotheses that are rejected (i.e., false discoveries), and $R_\mathcal{D} := |\mathcal{D}|$ as the total number of discoveries. The value of interest is $F_\mathcal{D}/R_\mathcal{D}$, called the false discovery proportion (FDP), which is the ratio of the number of false discoveries to that of all claimed discoveries,

with the convention $0/0 = 0$ (i.e., FDP is 0 if there is no discovery). Benjamini and Hochberg [3] proposed to control FDR, which is the expected value of FDP, that is, $\mathrm{FDR}_{\mathcal{D}} := \mathbb{E}[F_{\mathcal{D}}/R_{\mathcal{D}}]$, where the expected value is taken under the true probability. Other ways of controlling the false discovery other than the FDR are studied by, for instance, Genovese and Wasserman [9, 10] and Goeman and Solari [11] with p-values, and Vovk and Wang [27] with e-values.

Another related error measurement, which is particularly relevant for testing the global null, is the family-wise error rate (FWER), defined as $\mathrm{FWER}_{\mathcal{D}} = \mathbb{P}(F_{\mathcal{D}} \geq 1)$. FWER is equal to FDR if all hypotheses are true null.

In the remainder of the paper, we will study discovery procedures that use both e-values and p-values for each hypothesis. After briefly reviewing some concepts in Section 2, we obtain several admissible combiners of an e-value and a p-value (Section 3) and propose the e-weighted p-BH and p-weighted e-BH procedures, with their FDR guarantee shown under difficult dependence assumptions (Section 4). Some other extensions on e-BH and e-values are also obtained and collected in Section 5.

2 Recap: PRDS, p-BH, and e-BH

To discuss the dependence structure among p-values and e-values, we rely on the notion of *positive regression dependence on a subset (PRDS)* of Benjamini and Yekutieli [4], flipped when imposed on e-values. A set $A \subseteq \mathbb{R}^K$ is said to be *decreasing* (resp. *increasing*) if $\mathbf{x} \in A$ implies $\mathbf{y} \in A$ for all $\mathbf{y} \leq \mathbf{x}$ (resp. all $\mathbf{y} \geq \mathbf{x}$). All terms "increasing" and "decreasing" are in the non-strict sense, and inequalities should be interpreted component-wise when applied to vectors.

1. A random vector $\mathbf{P}$ of p-values satisfies PRDS if for any null index $k \in \mathcal{N}$ and increasing set $A \subseteq \mathbb{R}^K$, the function $x \mapsto \mathbb{P}(\mathbf{P} \in A \mid P_k \leq x)$ is increasing on $[0, 1]$.

2. A random vector $\mathbf{E}$ of e-values satisfies PRDS if for any null index $k \in \mathcal{N}$ and decreasing set $A \subseteq \mathbb{R}^K$, the function $x \mapsto \mathbb{P}(\mathbf{E} \in A \mid E_k \geq x)$ is decreasing on $[0, \infty)$.

We use the version of PRDS in Finner et al. [8, Section 4] (also see Barber and Ramdas [2]) which is slightly weaker than the original one used in Benjamini and Yekutieli [4].

If the null p-values (e-values) are mutually independent and independent of the non-null p-values (e-values), then PRDS holds; as such, PRDS is a generalization of independence. Moreover, increasing individual transforms do not affect the PRDS property. Hence, the PRDS property is preserved when moving from p-values to e-values using calibrators, or vice versa by inversion.

Next, we briefly describe the p-BH and e-BH procedures. For $k \in \mathcal{K}$, let $p_{(k)}$ be the k-th order statistics of the observed p-values $p_1, \dots, p_K$, from the smallest to the largest, and $e_{[k]}$ be the k-th order statistic of the observed e-values $e_1, \dots, e_K$, from the largest to the smallest. Let $\alpha \in (0, 1)$ be the target FDR level.

The *p-BH procedure* (Benjamini and Hochberg [3]) uses observed p-values as input and rejects all hypotheses with the smallest k_p^* p-values, where

$$k_p^* = \max\left\{k \in \mathcal{K} : \frac{K p_{(k)}}{k} \leq \alpha\right\}, \tag{1}$$

with the convention $\max(\varnothing) = 0$. The *e-BH procedure* (Wang and Ramdas [31]) uses observed e-values as input and rejects all hypotheses with the largest k_e^* e-values, where

$$k_e^* = \max\left\{k \in \mathcal{K} : \frac{k e_{[k]}}{K} \geq \frac{1}{\alpha}\right\}. \tag{2}$$

An equivalent way to describe the e-BH procedure is to apply the p-BH procedure to $(e_1^{-1}, \dots, e_K^{-1})$.

It is known from Benjamini and Hochberg [3] and Benjamini and Yekutieli [4] that for PRDS p-values, the p-BH procedure at level α has FDR at most $K_0\alpha/K$, and for arbitrarily dependent

p-values, the p-BH procedure at level α has FDR at most $\ell_K K_0 \alpha / K$, where $\ell_K = \sum_{k=1}^{K} k^{-1} \approx \log K$. As for the FDR of the e-BH procedure, Wang and Ramdas [31] showed a surprising property that the base e-BH procedure controls FDR at level α even under unknown arbitrary dependence between the e-values.

The *weighted p-BH* and *weighted e-BH* procedures are obtained by applying the p-BH procedure and the e-BH procedure to $(P_1/w_1, \dots, P_K/w_K)$ and $(w_1 E_1, \dots, w_k E_K)$, respectively, where $(w_1, \dots, w_K) \in [0, \infty)^K$ is a pre-specified vector of weights. A key requirement for the FDR guarantee of these procedures is $\sum_{k=1}^{K} w_k = K$, that is, the weights have an average of 1; see Ramdas et al. [19] and Wang and Ramdas [31]. Therefore, if the weights are obtained from preliminary experiments, classical thinking suggests that they need to be normalized so that their average is 1. Later, we will see if the weights are obtained from e-values independent of $\mathbf{P}$ (or $\mathbf{E}$), then this normalization is not needed, and this can improve power substantially.

3 Combining a p-value and an e-value

One of the main objective of the paper is to design and understand procedures when both p-values and e-values are available. For this purpose, we first look at the single-hypothesis setting, in which case we drop the subscripts and use P for a p-variable and and E for an e-variable.

We briefly review calibration between a p-value and an e-value as developed previously by Shafer et al. [24] and Shafer and Vovk [23, Chapter 11.5], amongst other sources. Denote by $\overline{\mathbb{R}}_+ = [0, \infty]$. First, an e-variable E can be converted to a p-variable $P = 1/E$ (its validity follows from Markov's inequality). Further, the function $f : e \mapsto (1/e) \wedge 1$ is the unique admissible e/p calibrator (Vovk and Wang [29, Proposition 2.2]).

A p-variable P can also be converted to an e-variable, but there are many admissible choices; one simple example is to set $E = P^{-1/2} - 1$. More generally, we speak of p/e calibrators. Note that small p-values correspond to large e-values, which represent stronger evidence against a null hypothesis. A p/e calibrator is a decreasing function $h : [0, 1] \to \overline{\mathbb{R}}_+$ satisfying $\int_0^1 h(u) \, \mathrm{d}u \le 1$. Clearly, $h(P)$ is an e-variable for any p-variable P. Vovk and Wang [29, Proposition 2.1] shows that the set $\mathcal{C}^{\mathrm{p/e}}$ of all admissible p/e calibrators is given by

$$\mathcal{C}^{\mathrm{p/e}} = \left\{ h : [0,1] \to \overline{\mathbb{R}}_+ \mid h(0) = \infty, \int_0^1 h(u) \, \mathrm{d}u = 1, \ h \text{ is decreasing and upper semicontinuous} \right\}.$$

In the above statements, admissibility of a calibrator (or a combiner below) means that it cannot be improved strictly, where an improvement means obtaining a larger e-value or a smaller p-value.

Combining several p-values or e-values to form a new p-value or e-value is the main topic of Vovk and Wang [28, 29] and Vovk et al. [30]. For the objective of this paper, we need to combine a p-value P and an e-value E, first in a single-hypothesis testing problem. There are four cases to consider:

(i) If P and E are independent, how should we combine them to form an e-variable?

(ii) If P and E are independent, how should we combine them to form a p-variable?

(iii) If P and E are arbitrarily dependent, how should we combine them to form an e-variable?

(iv) If P and E are arbitrarily dependent, how should we combine them to form a p-variable?

We use the following terminology, similar to that of Vovk and Wang [29]. A function $f : [0, 1] \times \overline{\mathbb{R}}_+ \to \overline{\mathbb{R}}_+$ is called an i-pe/e combiner if $f(P, E)$ is an e-variable for all independent p-variable P and e-variable E, and $(p, e) \mapsto f(p, e)$ is decreasing in p and increasing in e. Similarly, we define i-pe/p, pe/p, and pe/e combiners, where "i" indicates independence, and "p" and "e" are self-explanatory. Note that if the output is a p-value, the combiner is increasing in p and decreasing in e.

We provide four natural answers to the above four questions, some relying on an admissible calibrator $h \in \mathcal{C}^{\mathrm{p/e}}$:

(i) Return $h(P)E$ by using the function $\Pi_h(p,e) := h(p)e$. The convention here is $0 \times \infty = \infty$.

(ii) Return P/E, capped at 1, by using the function $Q(p,e) := (p/e) \wedge 1$.

(iii) Return $\lambda h(P) + (1-\lambda)E$ by using the function $M_h^\lambda(p,e) := \lambda h(p) + (1-\lambda)e$ for some $\lambda \in (0,1)$.

(iv) Return $2\min(P, 1/E)$, capped at 1, by using the function $B(p,e) := (2(p \wedge e^{-1})) \wedge 1$.

The notation chosen for these functions is due to the initials of (i) product (but we avoid P which is reserved for p-variables); (ii) quotient; (iii) mean; (iv) Bonferroni correction on p and e^{-1}.

Note that Π_h and M_h^λ depend on h whereas Q and B do not. For the function M_h^λ, it may be convenient to choose $\lambda = 1/2$, so that $M_h^\lambda(P,E)$ is the arithmetic average of two e-variables $h(P)$ and E. As shown by Vovk and Wang [29, Proposition 3.1], the arithmetic average essentially dominates, in a natural sense, all symmetric e-merging function. In our context, $\lambda = 1/2$ has no special role, since the positions of $h(P)$ and E are not symmetric.

Theorem 3.1. *For $h \in \mathcal{C}^{\mathrm{p/e}}$ and $\lambda \in (0,1)$, Π_h is an admissible i-pe/e combiner, Q is an admissible i-pe/p combiner, M_h^λ is an admissible pe/e combiner, and B is an admissible pe/p combiner.*

Proof. Let P be a p-variable and E be an e-variable. They are assumed independent in (i) and (ii) below. For a fixed $e \in [0,\infty)$, we will frequently rely on a specific distribution F_e of e-variables given by, for $X \sim F_e$, $\mathbb{P}(X = e) = 1/e = 1 - \mathbb{P}(X = 0)$ if $e \ge 1$ and $\mathbb{P}(X = e) = 1/2 = \mathbb{P}(X = 2-e)$ if $e < 1$. It is clear that $\mathbb{E}[X] = 1$.

(i) We have $\mathbb{E}[h(P)E] \le 1$ since $h(P)$ is an e-variable independent of E. Hence, Π_h is an i-pe/e combiner. To show its admissibility, suppose for the purpose of contradiction that an i-pe/e combiner f satisfies $f \ge \Pi_h$ and $f(p,e) > \Pi_h(p,e)$ for some $(p,e) \in [0,1] \times \overline{\mathbb{R}}_+$. Clearly, $e \in [0,\infty)$. Since h is upper semicontinuous and $q \mapsto f(q,e)$ is decreasing, there exists $p' < p$ such that $f(q,e) \ge f(p,e) > \Pi_h(p',e) \ge \Pi_h(q,e)$ for all $q \in [p',p]$. Take $P \sim \mathrm{U}[0,1]$ and $E \sim F_e$. Since $\mathbb{E}[\Pi_h(P,E)] = 1$, $f \ge \Pi_h$, and $f(q,e) > \Phi_h(q,e)$ for all $q \in [p',p]$, we have $\mathbb{E}[f(P,E)] > \mathbb{E}[\Pi_h(P,E)] = 1$ which means that $f(P,E)$ is not an e-variable, contradicting the fact that f is an i-pe/e combiner. This contradiction shows that Π_h is admissible.

(ii) For $\alpha \in (0,1)$, we have $\mathbb{P}(Q(P,E) \le \alpha) = \mathbb{P}(P \le \alpha E) = \mathbb{E}[\mathbb{P}(P \le \alpha E | E)] \le \mathbb{E}[\alpha E] \le \alpha$. Therefore, Q is an i-pe/p combiner. To show its admissibility, suppose for the purpose of contradiction that an i-pe/p combiner f satisfies $f \le Q$ and $f(p,e) < (p/e) \wedge 1$ for some $(p,e) \in [0,1] \times \overline{\mathbb{R}}_+$. Since $a \mapsto f(p,a)$ is decreasing, we can assume $e \in [p,\infty)$ by replacing e with p if $e < p$. Take $P \sim \mathrm{U}[0,1]$ and $E \sim F_e$. Since $q \mapsto f(q,e)$ is increasing, there exists $p' < p$ such that $f(q,e) \le f(p,e) < p'/e$ for all $q \in [0,p]$. This gives $\mathbb{P}(f(P,e) \le p'/e) \ge \mathbb{P}(P \le p) = p$. For $\alpha = p'/e \in (0,1)$, if $e \ge 1$, then

$$\mathbb{P}(f(P,E) \le \alpha) = \mathbb{P}(f(P,e) \le p'/e)e^{-1} + \mathbb{P}(f(P,0) \le \alpha)(1 - e^{-1}) \ge p/e > \alpha.$$

If $e < 1$, then

$$\begin{aligned}\mathbb{P}(f(P,E) \le \alpha) &= \frac{1}{2}\mathbb{P}(f(P,e) \le p'/e) + \frac{1}{2}\mathbb{P}(f(P,2-e) \le \alpha) \\ &\ge \frac{1}{2}p + \frac{1}{2}\mathbb{P}(P \le \alpha(2-e)) \\ &= \frac{1}{2}\left(p + (2-e)p'/e\right) = \frac{1}{2}(p - p') + p'/e > \alpha.\end{aligned}$$

Hence, $f(P,E)$ is not a p-variable, and this contradicts the fact that f is an i-pe/p combiner. This contradiction shows that Q is admissible.

(iii) Note that the weighted average of two arbitrary e-variables is an e-variable, and hence M_h^λ is a pe/e combiner. Its admissibility follows essentially the same proof as in part (i), which we shall not repeat.

(iv) Since $1/E$ is a p-variable, the Bonferroni combination of P and $1/E$, $2\min(P, 1/E)$, is a p-variable, and hence B is a pe/p combiner. To show its admissibility, suppose for the purpose of contradiction that a pe/p combiner f satisfies $f \le B$ and $f(p,e) < (2(p \wedge e^{-1}) \wedge 1)$ for some $(p,e) \in [0,1] \times \overline{\mathbb{R}}_+$. By monotonicity of f, we can decrease e to $1/p$ or increase p to $e^{-1} \wedge (1/2)$, and this does not change the value of $(2(p \wedge e^{-1}) \wedge 1)$. Hence, we can assume that $f(p, 1/p) < 2p$ for some $p \in (0, 1/2]$ by noting that $f(p, 1/p) < 2p$ automatically holds for $p > 1/2$ because $B \le 1$. Since $q \mapsto f(q, e)$ is increasing, there exists $p' < p$ such that $f(q, 1/p) \le f(p, 1/p) < 2p'$ for all $q \in [0, p]$. Take $P \sim \mathrm{U}[0,1]$ and define $E = (1/p')\mathbb{1}_{\{P \in [p, p+p']\}}$. It is clear that $\mathbb{E}[E] = 1$. We have

$$\mathbb{P}(f(P,E) \le 2p') = \mathbb{P}(P \in [0, p+p']) = p + p' > 2p'.$$

Hence, $f(P,E)$ is not a p-variable, and this contradicts the fact that f is a pe/p combiner. This contradiction shows that B is admissible. □

Certainly, there are other useful combiners than the ones studied in Theorem 3.1, but we will mainly focus on the latter choices, as they are sufficient for the development of our theory as well as most applications.

4 Using e-values and p-values as weights in multiple testing

In this section, we move back to multiple testing by considering the setting where each hypothesis is associated with a p-value and an e-value. The FDR or FWER guarantee of testing procedures in this setting depends on the dependence amongst the e-values, the dependence amongst the p-values, and the dependence between the p-values and e-values.

We are mostly interested in the case when the vector $\mathbf{P} = (P_1, \dots, P_K)$ of p-values and the vector $\mathbf{E} = (E_1, \dots, E_K)$ of e-values are independent, and each of $\mathbf{P}$ and $\mathbf{E}$ may have dependent components. This situation is the most interesting as it represents the case where we accumulate statistical evidence by combining results from two stages of independent experiments.

We will omit discussions on the cases where $\mathbf{P}$ and $\mathbf{E}$ are not independent, but these cases can be handled analogously with the pe/e and pe/p combiners in Theorem 3.1, although the test power will likely be weak due to the conservative dependence assumption.

4.1 The e-weighted BH procedure (ep-BH)

Inspired by the combiners in Theorem 3.1, there are at least two ways to utilize both p-values and e-values in the setting where $\mathbf{P}$ and $\mathbf{E}$ are independent. Let $\alpha \in (0,1)$ be a specified target FDR level.

1. For $k \in \mathcal{K}$, compute $P_k^* = P_k/E_k$ by applying the i-pe/p merger Q, and then supply $(P_1^*, \dots, P_K^*)$ to the p-BH procedure at level α. This will be called the *ep-BH procedure*, or more fully, the e-weighted p-BH procedure.

2. Choose $h \in \mathcal{C}^{\mathrm{p/e}}$. For $k \in \mathcal{K}$, compute $E_k^* = h(P_k)E_k$ by applying the i-pe/e merger Π_h, and then supply $(E_1^*, \dots, E_K^*)$ to the e-BH procedure at level α. This will be called the *pe-BH procedure*, or more fully, the p-weighted e-BH procedure.

We note that the first procedure always dominates the second one. The e-BH procedure with input $(e_1, \dots, e_K)$ is equivalent to the p-BH procedure with input $(1/e_1, \dots, 1/e_K)$. Hence, the second procedure can be seen as applying the p-BH procedure to $(1/E_1^*, \dots, 1/E_K^*)$. Note that if $h(p) > 1/p$ for even a single $p \in (0,1)$, then h is not a p/e calibrator. Indeed, for $P \sim \mathrm{U}[0,1]$, by decreasing monotonicity of h, we get $\mathbb{E}[h(P)] \ge \mathbb{E}[h(P)\mathbb{1}_{\{P<p\}}] > \mathbb{E}[\mathbb{1}_{\{P<p\}}/p] = \mathbb{P}(P < p)/p = 1$, which

contradicts the fact that $h(P)$ is an e-value. Therefore, $p \mapsto 1/p$ is an upper bound for all p/e calibrators h, and this implies, for each $k \in \mathcal{K}$,

$$\frac{1}{E_k^*} = \frac{1}{h(P_k)E_k} \geq \frac{P_k}{E_k} = P_k^*.$$

Hence, the second procedure (pe-BH) is dominated by the first procedure (ep-BH). However, there is a caveat: the pe-BH procedure relies on feeding e-values to the e-BH procedure, and hence the dependence structure within $\mathbf{P}$ and within $\mathbf{E}$ can be arbitrary. On the other hand, the ep-BH procedure relies on feeding p-values to the p-BH procedure, and some dependence assumptions are needed (see Theorem 4.2).

Our main proposal is the ep-BH procedure, as it dominates the pe-BH procedure when both are valid, but we will soon see a very particular instance where the latter may be preferable.

If $(P_k)_{k\in\mathcal{N}}$ are independent of $(E_k)_{k\in\mathcal{N}}$, then regardless of the other dependence structures assumed, the e-values can be employed as *weights* in the BH-procedure, while maintaining FDR control. The remarkable fact is the following: the weighted-BH procedure requires using weights that are normalized, meaning that they sum to K. However, when the weights are e-values, no such normalization is required, so a potentially huge gain in power is possible by using unnormalized weights that are (hopefully) much larger than one. There are two further amendments that may be relevant in a real application of the ep-BH procedure, which we will omit in the main text of the paper.

1. If one is excessively worried that some e-values may be arbitrarily close to zero, one can simply replace the original e-values by $\widetilde{E}_k = (1 + E_k)/2$. This will make sure that, if a p-value turns out to be very strong in the second-stage experiment, it will not be thrown away due to a very small weight obtained in the first-stage experiment.

2. If some of the e-values are missing, one can safely set them to 1. Therefore, we do not require a full weight vector $\mathbf{E}$ for all hypotheses. This may represent the situation where some hypotheses have preceding experiments or preliminary data analysis whereas the others do not.

Remark 4.1. In the ep-BH procedure, e-values are used as weights for the p-values. Intuitively, if $E_k > 1$, then there is some evidence against H_k being a null, and we have $P_k/E_k < P_k$ (assuming $P_k \neq 0$); that is, the weight strengthens the signal of P_k in this case. Conversely, if $E_k < 1$, then there is no evidence against H_k being a null, and we have $P_k/E_k > P_k$. The above interpretation of e-values as weights is quite natural. The situation for the pe-BH procedure, where p-values are used as weights for the e-values, is somewhat different. For simplicity, suppose that we use the calibrator $h \in \mathcal{C}^{\mathrm{p/e}}$ given by $h(p) = p^{-1/2} - 1$. It is clear that $h(p) > 1$ if and only if $p < 1/4$. Hence, the signal of the e-value E_k will be strengthened in case $P_k < 1/4$. This is not surprising as observing a p-value in $(0.25, 1)$ generally does not indicate evidence against the null. Other choices of $h \in \mathcal{C}^{\mathrm{p/e}}$ lead to different thresholds, and this is consistent with the fact that there is no universal agreement on which moderate values of a p-value should be considered as carrying some (weak) evidence against the null.

The FDR guarantee of the pe-BH and the ep-BH procedures is summarized in Theorem 4.2 below, and we first make some immediate observations. The FDR of the pe-BH procedure follows by combining the fact that $E_1^*, \dots, E_K^*$ are e-values for $H_1, \dots, H_K$ due to Theorem 3.1 and the FDR guarantee $\alpha K_0/K$ of the e-BH procedure in Wang and Ramdas [31, Theorem 2]. This requires no dependence assumption on either $\mathbf{E}$ or $\mathbf{P}$. On the other hand, the ep-BH procedure requires some dependence assumption, such as PRDS, similarly to the p-BH procedure. Note that the PRDS of $\mathbf{P}$ does not imply that of $\mathbf{P}^* = (P_1^*, \dots, P_K^*)$, and hence some arguments are needed to establish FDR guarantee.

Theorem 4.2. *Suppose that* $\mathbf{P}$ *is independent of* $\mathbf{E}$*. The pe-BH procedure has FDR at most* $\alpha K_0/K$*. If* $\mathbf{P}$ *satisfies PRDS, then the ep-BH procedure has FDR at most* $\alpha K_0/K$*.*

Proof. The statement on the pe-BH procedure has been discussed above. We will only show the statement on the ep-BH procedure. Let $\mathcal{D}$ be the ep-BH procedure at level α. Since $\mathbf{E}$ is independent of $\mathbf{P}$, conditional on the event $\mathbf{E} = \mathbf{e} = (e_1, \dots, e_K) \in \mathbb{R}_+^K$, the ep-BH procedure becomes a weighted p-BH procedure with weight vector $\mathbf{e}$ applied to the PRDS p-values $\mathbf{P}$. Using existing results on the FDR of the weighted p-BH procedure (e.g., Ramdas et al. [19, Theorem 1]), we get

$$\mathbb{E}\left[\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}} \mid \mathbf{E} = \mathbf{e}\right] \le \frac{1}{K}\sum_{k\in\mathcal{N}} e_k \alpha.$$

Hence,

$$\mathbb{E}\left[\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}} \mid \mathbf{E}\right]\right] \le \frac{\alpha}{K}\mathbb{E}\left[\sum_{k\in\mathcal{N}} E_k\right] \le \frac{K_0}{K}\alpha,$$

and this gives the desired FDR control of the ep-BH procedure. □

By Theorem 4.2, if $\mathbf{P}$ satisfies PRDS, then the ep-BH procedure is valid, and it should be the better choice than the pe-BH procedure which is dominated. However, if there is no dependence information of $\mathbf{P}$, then $\mathbf{P}^*$ is arbitrarily dependent, and one may need to apply the p-BH procedure with the BY correction in Benjamini and Yekutieli [4]. In this case, the pe-BH and the ep-BH procedures do not dominate each other. In particular, one needs to compare the inputs

$$\frac{1}{h(P_k)E_k} \quad \text{and} \quad \frac{\ell_K P_k}{E_k}, \quad \text{where } \ell_K = \sum_{k=1}^{K}\frac{1}{k} \approx \log K.$$

This corresponds to the same trade-off between the p-BH procedure with BY correction and the e-BH procedure, as discussed in Wang and Ramdas [31], where one needs to compare $h(P_k)$ and $(\ell_K P_k)^{-1}$.

4.2 Combining e-values and p-values obtained from different hypotheses

We now consider the situation where $\mathbf{E}$ represents e-values for hypotheses $H_1^e, \dots, H_K^e$ and $\mathbf{P}$ represents p-values for hypotheses $H_1^p, \dots, H_K^p$, and the two set of hypotheses may not be equal. In this case, we can still apply the ep-BH and pe-BH procedures, and we are effectively testing $H_k = H_k^e \cap H_k^p$, $k \in \mathcal{K}$. The reason is that if E_k is an e-variable for H_k^e, then it is also one for any subset of H_k^e. The case of p-variables is similar.

In practical testing problems, we may encounter situations where

(a) $H_k^p \subseteq H_k^e$ for all $k \in \mathcal{K}$, or

(b) $H_k^e \subseteq H_k^p$ for all $k \in \mathcal{K}$.

By combining $\mathbf{P}$ and $\mathbf{E}$, we are testing $H_1^p, \dots, H_K^p$ in case (a) , and we are testing $H_1^e, \dots, H_K^e$ in case (b). This represents the situation where one set of experiments is conducted on looser hypotheses and the other set on stricter hypotheses.

As a concrete example, consider dose-based trials, where for drug k one hospital might run a high-dose study, another might run a low-dose study. Typically, we assume that if a drug k is not effective at high dose, then it also not effective at low dose. The combined pe-BH or ep-BH procedure will be effectively testing that the high-dose drugs are not effective.

If we can produce both e-values and p-values for hypotheses in both settings, then we recommend producing e-values for the high-dose trials and using them as weights for p-values from the low-dose trials. (As we discussed above, it seems more natural to use $\mathbf{E}$ as weights for testing $\mathbf{P}$, although this can be easily reverted to suit needs in different applications.)

4.3 The e-weighted Bonferroni procedure for FWER or PFER control

We next briefly discuss FWER control using a combination of e-values and p-values. Given input K p-values, the Bonferroni procedure at level $\alpha \in (0,1)$ rejects all hypotheses with a p-value less or equal to α/K. For a given level α, the *pe-Bonferroni procedure* is defined by applying the Bonferroni procedure to $(P_1^*, \dots, P_K^*) = (P_1/E_1, \dots, P_K/E_K)$. The per-family error rate (PFER) of a procedure $\mathcal{D}$ is defined as $\mathrm{PFER}_{\mathcal{D}} = \mathbb{E}[F_{\mathcal{D}}]$.

Proposition 4.3. *Suppose that* $\mathbf{P}$ *is independent of* $\mathbf{E}$. *The pe-Bonferroni procedure at level* α *has FWER and PFER at most* $\alpha K_0/K$.

Proof. Let $\mathcal{D}$ be the pe-Bonferroni procedure applied at level α. The PFER statement follows from

$$\mathrm{PFER}_{\mathcal{D}} = \mathbb{E}[F_{\mathcal{D}}] \le \mathbb{E}\left[\sum_{k\in\mathcal{N}} \mathbb{1}_{\{P_k^* \le \alpha/K\}}\right] \le \mathbb{E}\left[\sum_{k\in\mathcal{N}} \mathbb{E}[\mathbb{1}_{\{P_k \le \alpha E_k/K\}} \mid E_k]\right] \le \mathbb{E}\left[\sum_{k\in\mathcal{N}} \frac{\alpha E_k}{K}\right] \le \frac{K_0}{K}\alpha.$$

By definition, FWER is smaller than or equal to PFER, and hence the pe-Bonferroni procedure has FWER at most $K_0\alpha/K$. □

This result does not require assumptions on the dependence within $\mathbf{P}$ or within $\mathbf{E}$. The above proof, and essentially many of the earlier proofs as well, make the following point: the utility of e-values as random weights is quite far-reaching in multiple testing. It is clear that the same idea also applies to online multiple testing [18, 26], as well many other structured settings [19, 20].

5 Further extensions on e-BH and e-values

In this section we provide a few technical extensions on improving the e-BH procedure, constructing permutation e-values, and using e-values as masks.

5.1 A tiny but uniform improvement of e-BH

In this section, we propose a tiny but uniform improvement of the e-BH procedure, inspired by Solari and Goeman [25]. We remark that, similarly to the situation of Solari and Goeman [25], this improvement is negligible for large values of K and it may only be practically interesting for small K such as $K \le 10$. We mainly focus on the case without boosting (see Wang and Ramdas [31] for e-value boosting).

First, choose an e-merging function $F : [0,\infty]^K \to [0,\infty]$ in the sense of Vovk and Wang [29], i.e., F satisfies that $F(E_1, \dots, E_K)$ is an e-variable for any e-variables $E_1, \dots, E_K$. By Proposition 3.1 of Vovk and Wang [29], the arithmetic average

$$M : (e_1, \dots, e_K) \mapsto \frac{1}{K}\sum_{k=1}^{K} e_k$$

is the "best" symmetric e-merging function, in the sense that it is uniformly more powerful than any other symmetric e-merging functions. We allow for a general choice of F other than M as it will be useful for the discussion later on boosted e-values.

With a chosen e-merging function F and a level $\alpha \in (0,1)$, the improved e-BH procedure, denoted by $\mathcal{D}^F(\alpha)$, is designed as follows. We first test the global null $\bigcap_{k=1}^K H_k$ via the rejection condition $F(e_1, \dots, e_K) \ge 1/\alpha$, which has a type-I error of at most α, and if the global null is rejected, we then apply the e-BH procedure at level $\alpha' = K\alpha/(K-1)$. In other words,

1. if $F(e_1, \dots, e_K) < 1/\alpha$, then $\mathcal{D}^F(\alpha) = \varnothing$;

2. if $F(e_1, \ldots, e_K) \geq 1/\alpha$, then $\mathcal{D}^F(\alpha) = \mathcal{D}(\alpha')$ where $\alpha' = K\alpha/(K-1)$ and $\mathcal{D}(\alpha')$ is the e-BH procedure at level α'.

The next proposition shows that by choosing $F = M$, the resulting improved BH procedure dominates the base BH procedure.

Proposition 5.1. *The improved e-BH procedure $\mathcal{D}^F(\alpha)$ applied to arbitrary e-values has FDR at most α. In case $F = M$, $\mathcal{D}^M(\alpha)$ dominates the e-BH procedure $\mathcal{D}(\alpha)$, that is, $\mathcal{D}(\alpha) \subseteq \mathcal{D}^M(\alpha)$.*

Proof. The first statement on FDR can be shown in a similar way as Solari and Goeman [25]. Let A be the the event that $F(e_1, \ldots, e_K) \geq 1/\alpha$, treated as random. If $K_0 < K$, then, by using Theorem 5.1 of Wang and Ramdas [31],

$$\mathbb{E}\left[\frac{F_{\mathcal{D}^F(\alpha)}}{R_{\mathcal{D}^F(\alpha)}}\right] = \mathbb{E}\left[\frac{F_{\mathcal{D}(\alpha')}}{R_{\mathcal{D}(\alpha')}}\mathbb{1}_A\right] + \mathbb{E}\left[\frac{F_{\varnothing}}{R_{\varnothing}}(1-\mathbb{1}_A)\right] = \mathbb{E}\left[\frac{F_{\mathcal{D}(\alpha')}}{R_{\mathcal{D}(\alpha')}}\mathbb{1}_A\right] \leq \mathbb{E}\left[\frac{F_{\mathcal{D}(\alpha')}}{R_{\mathcal{D}(\alpha')}}\right] \leq \frac{K_0}{K}\alpha' \leq \alpha.$$

If $K_0 = K$, then the FDR of $\mathcal{D}^F(\alpha)$ is at most the probability $\mathbb{P}(A)$ of rejecting the global null via $F(e_1, \ldots, e_K) \geq 1/\alpha$. In this case, $\mathbb{P}(A) \leq \alpha$ by Markov's inequality and the fact that F is an e-merging function. Hence, in either case, the FDR of $\mathcal{D}^F(\alpha)$ is at most α.

To show the second statement on dominance, let

$$S : (e_1, \ldots, e_K) \mapsto \max_{k=1,\ldots,K} \frac{k e_{[k]}}{K}. \tag{3}$$

The function S is an e-merging function and it is dominated by M on $[0,\infty]^K$; see Section 6 of Vovk and Wang [29]. Note that by definition, $S(e_1, \ldots, e_K) < 1/\alpha$ implies $\mathcal{D}(\alpha) = \varnothing$. Therefore, if $M(e_1, \ldots, e_K) < 1/\alpha$, then $\mathcal{D}(\alpha) = \varnothing = \mathcal{D}^M(\alpha)$. Moreover, since $\alpha < \alpha'$, we always have $\mathcal{D}(\alpha) \subseteq \mathcal{D}(\alpha')$. Hence, $\mathcal{D}(\alpha) \subseteq \mathcal{D}^M(\alpha)$. □

Next, we briefly discuss the case of boosted e-values. The arithmetic average of boosted e-values is not necessarily a valid e-value, so one must be a bit more careful. Nevertheless, it turns out that we can use the function S in (3) on the boosted e-values. The new procedure can be described as the following steps.

1. Boost the raw e-values with level α.
2. If $S(e'_1, \ldots, e'_K) < 1/\alpha$ where $e'_1, \ldots, e'_K$ is the boosted e-values in step 1, then return $\varnothing$.
3. Else: boost the raw e-values with level $\alpha' = K\alpha/(K-1)$.
4. Return the discoveries by applying the base e-BH procedure to the boosted e-values in step 3.

This new procedure dominates the e-BH procedure, and it has FDR at most α. To show these two statements, it suffices to note that the probability of rejecting the global null test $S(e'_1, \ldots, e'_K) \geq 1/\alpha$ is at most α since the e-BH procedure has FDR at most α by Theorem 5.1 of Wang and Ramdas [31]; the rest of the proof is similar to that of Proposition 5.1.

5.2 The soft-rank permutation e-test

In this section, we introduce a new way of generating e-values, in addition to stopped e-processes (or test supermartingales), likelihood ratios, universal inference and p/e calibrations.

There is a large class of classical and modern testing procedures that use some form of Monte-Carlo sampling in order to produce test statistics that are exchangeable under the null, and use the rank of the original test statistic as a corresponding p-value. Examples include the permutation test, the conditional randomization test (Candes et al. [5]) and conformal prediction (Shafer and Vovk [22]). Then one can construct a natural e-value as well, as we show next. To keep the notation simple,

consider for now a single hypothesis and let L_0 be the test statistic calculated from the original data. Let $L_1, \ldots, L_B$ be B statistics that are constructed to be exchangeable with L_0 under the null, and $r > 0$ be a prespecified constant. Define $L_* = \min_{b=0,\ldots,B} L_b$ to be the smallest of the $(B+1)$ test statistics. For $b = 0, \ldots, B$, define the transformed statistic

$$R_b = \frac{\exp(rL_b) - \exp(rL_*)}{r}.$$

This transformation is performed to ensure that R_b is nonnegative while the ordering amongst the test statistics is preserved. The limiting case of $r = 0$ yields $R_b = L_b - L_*$. (Note that the L_i's are not assumed to be positive. If they were, we could choose $R_b = L_b/L_*$, or simply set $R_b = L_b$.) Most importantly, $\{R_b\}_{b=0}^B$ are also exchangeable under the null. Now, define

$$E := (B+1)\frac{R_0}{\sum_{i=0}^B R_i}.$$

It is not hard to check that E is an e-variable, since under the null $\mathbb{E}[R_0 | \sum_{b=0}^B R_i] = \sum_{b=0}^B R_b/(B+1)$. Contrasting this with the usual p-variable, we find that

$$P := \frac{\sum_{b=0}^B \mathbb{1}_{\{L_b/L_0 \geq 1\}}}{B+1} = \frac{\sum_{b=0}^B \mathbb{1}_{\{R_b/R_0 \geq 1\}}}{B+1} \leq \frac{\sum_{b=0}^B R_b/R_0}{B+1} = 1/E.$$

Since the P value is the rank of L_0 amongst $L_0, \ldots, L_B$, we see that $1/E$ can be seen as a smoothed notion of rank, or 'soft-rank' for short (much like the 'soft-max' is achieved by exponentiation). Hence, we call E as the soft-rank e-value and $1/E$ as the soft-rank p-value. When these are thresholded at α (meaning we reject if $E \geq 1/\alpha$), it yields level α test, we call the resulting method the soft-rank permutation test (or soft-rank conditional randomization test, and so on). Since the direct p-value P is always smaller than the soft-rank p-value $1/E$, there is apparently no advantage to using the latter for testing single hypothesis.

However, the situation changes when dealing with multiple such p-variables that are arbitrarily dependent, as may happen in application areas like neuroscience or genetics. In this case, applying the p-BH procedure to the p-values with dependence correction may be less powerful than applying the e-BH procedure to the e-values (because the $\approx \log K$ correction factor could outweight the benefit of using p-values). However it appears hard to compare their powers in general. When strong signals are expected, then it is beneficial to using e-values since $1/E \approx P$ if $R_b \ll R_0$ for almost all $b \neq 0$. One simple example of a turning point in comparing the powers of p-BH to e-BH occurs, for example, when $R_0 \approx 2R_{[1]} \approx 3R_{[2]} \approx 4R_{[3]} \ldots$, where $R_{[b]}$ is the b-th largest value amongst $R_1, \ldots, R_B$. This results in P being to equal $1/(K+1)$ while we have $1/E \approx \ell_K/(B+1)$, so the benefit of using e-BH would then be washed out.

In summary, if handling arbitrary dependence is a concern in practice and ranks of exchangeable statistics are being used as p-values, then switching to e-values may be a competitive alternative. Further examination of this special case, including a practical way to choose the free parameter r, is left to future work.

5.3 E-values as masks in interactive multiple testing

Since the original work by Lei and Fithian [15], the idea of "masking" p-values in order to enable user interaction during a multiple testing procedure has received much attention [16, 6, 7].

The broad idea is simple yet effective: split each p-value into two parts, $f(P)$ and $g(P)$, such that $\mathbb{E}[g(P)|f(P)] \leq 1$ under the null. Initially, one starts with the candidate discovery set being all hypotheses, and the user only sees $f(P_i)$ for every i. Then, the user "peels off" hypotheses one by one from the discovery set, each time observing the corresponding $g(P_i)$. One uses $f(P)$ for interactively selecting a set of discoveries, while using $g(P)$ to provide inferential statements on the selected set.

The original AdaPT algorithm used a very particular choice of f and g: they set $f(P) = \min(p, 1-p)$ and $g(p) = 2\mathbb{1}_{\{p>1/2\}}$. Later, Lei et al. [16] generalize this construction to show that for any monotonic function g that integrates to at most one, meaning that $\int_0^1 g(u)\,\mathrm{d}u \leq 1$, one can construct a function f such that $\mathbb{E}[g(P)|f(P)] \leq 1$ for null p-values. The astute reader will recognize that this condition on g is exactly the definition of a p/e calibrator. (One caveat in this interpretation is that calibrators h are monotonically decreasing, while the aforementioned papers choose g to be increasing, but this is minor because the information content in $h(P)$ and $g(P)$ is often identical and there is typically a one-to-one transformation from one to the other.)

The takeaway message is the following:

> the *interaction* in interactive multiple testing is enabled by the "gap" between a p-value and the corresponding calibrated e-value.

To elaborate, when we calibrate a p-value into an e-value using a p/e calibrator, the calibration process is slightly lossy — if we then convert the obtained e-value back to a p-value using an e/p calibrator (such as $e \mapsto 1/e$), we do not obtain the original p-value; we obtain a larger p-value which is thus less powerful. This "gap" between p-values and the corresponding calibrated e-value $g(P)$ is precisely what is exploited by these interactive procedures: it is what we have denoted $f(P)$, and it is revealed to the users at the start of the procedure. This extra revealed information $f(P)$ is essentially what distinguishes the above methods from their noninteractive counterparts which appeared earlier in the literature, namely the knockoffs procedure by Barber and Candès [1] and the ordered testing method by Li and Barber [17].

6 Conclusion

In this paper, selection procedures for multiple hypotheses testing are studied in the setting where both, or one of, e-value and p-value are available to each hypothesis. To obtain valid statistical guarantee, a few admissible ways to combine a p-value and an e-value into a p-value or an e-value are proposed (Theorem 3.1). Under the assumption that the p-values are independent of the e-values, these combination methods lead to the pe-BH and ep-BH procedures for FDR control (Theorem 4.2), as well as the pe-Bonferroni procedure for FWER control (Proposition 4.3). These procedures reveal the important fact that, unlike most standard weighted multiple testing methods which require the weights on the hypotheses to sum to one, such normalization is not required when the weights are e-values obtained from independent experiments. In case the some of the non-null hypotheses have e-values much larger than one from the first set of experiments, the new procedure can significantly boost the power of testing from the second set of experiments which produce p-values. We also collect several extensions on e-BH and e-values, including an improvement of e-BH, a soft-rank permutation e-value, and the use of e-values as masks.

References

[1] Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

[2] Rina Foygel Barber and Aaditya Ramdas. The p-filter: multilayer false discovery rate control for grouped hypotheses. *Journal of the Royal Statistical Society Series B*, 79(4):1247–1268, 2017.

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

[4] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29:1165–1188, 2001.

[5] Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection. *arXiv preprint arXiv:1610.02351*, 2016.

[6] Boyan Duan, Aaditya Ramdas, Sivaraman Balakrishnan, and Larry Wasserman. Interactive martingale tests for the global null. *Electronic Journal of Statistics*, 14(2):4489–4551, 2020.

[7] Boyan Duan, Aaditya Ramdas, and Larry Wasserman. Familywise error rate control by interactive unmasking. In *International Conference on Machine Learning*, pages 2720–2729. PMLR, 2020.

[8] Helmut Finner, Thorsten Dickhaus, and Markus Roters. On the false discovery rate and an asymptotically optimal rejection curve. *The Annals of Statistics*, 37(2):596–618, 2009.

[9] Christopher R. Genovese and Larry Wasserman. A stochastic process approach to false discovery control. *The Annals of Statistics*, 32:1035–1061, 2004.

[10] Christopher R. Genovese and Larry Wasserman. Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101:1408–1417, 2006.

[11] Jelle J. Goeman and Aldo Solari. Multiple testing for exploratory research. *Statistical Science*, 26:584–597, 2011. Correction: 28:464.

[12] Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. Technical Report arXiv:1906.07801, 2020.

[13] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

[14] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

[15] Lihua Lei and William Fithian. Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4): 649–679, 2018.

[16] Lihua Lei, Aaditya Ramdas, and William Fithian. A general interactive framework for false discovery rate control under structural constraints. *Biometrika*, 108(2):253–267, 2021.

[17] Ang Li and Rina Foygel Barber. Accumulation tests for fdr control in ordered hypothesis testing. *Journal of the American Statistical Association*, 112(518):837–849, 2017.

[18] Aaditya Ramdas, Fanny Yang, Martin J Wainwright, and Michael I Jordan. Online control of the false discovery rate with decaying memory. *Advances in neural information processing systems*, 30, 2017.

[19] Aaditya Ramdas, Rina F. Barber, Martin J. Wainwright, and Michael I. Jordan. A unified treatment of multiple testing with prior knowledge using the p-filter. *The Annals of Statistics*, 47:2790–2821, 2019.

[20] Aaditya Ramdas, Jianbo Chen, Martin J Wainwright, and Michael I Jordan. A sequential algorithm for false discovery rate control on directed acyclic graphs. *Biometrika*, 106(1):69–86, 2019.

[21] Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):407–431, 2021.

[22] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(Mar):371–421, 2008.

[23] Glenn Shafer and Vladimir Vovk. *Game-Theoretic Foundations for Probability and Finance*, volume 455. John Wiley & Sons, 2019.

[24] Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84–101, 2011.

[25] Aldo Solari and Jelle J Goeman. Minimally adaptive BH: A tiny but uniform improvement of the procedure of Benjamini and Hochberg. *Biometrical Journal*, 59(4):776–780, 2017.

[26] Jinjin Tian and Aaditya Ramdas. Online control of the familywise error rate. *Statistical Methods in Medical Research*, 30(4):976–993, 2021.

[27] Vladimir Vovk and Ruodu Wang. Confidence and discoveries with e-values. *arXiv preprint arXiv:1912.13292*, 2019.

[28] Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107(4):791–808, 2020.

[29] Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

[30] Vladimir Vovk, Bin Wang, and Ruodu Wang. Admissible ways of merging p-values under arbitrary dependence. *The Annals of Statistics*, 50(1):351–375, 2022.

[31] Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B*, 2021.

[32] Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.

[33] Ziyu Xu, Ruodu Wang, and Aaditya Ramdas. A unified framework for bandit multiple testing. In *Advances in Neural Information Processing Systems*, volume 34, pages 16833–16845, 2021.