Electronic Journal of Statistics
Vol. 20 (2026) 3178–3189
ISSN: 1935-7524
https://doi.org/10.1214/26-EJS2546

Carefree multiple testing with e-processes

Yury Tavyrikov[1], **Jelle J. Goeman**[2] **and Rianne de Heide**[3]

[1]*Vrije Universiteit Amsterdam,e-mail:* y.tavyrikov@vu.nl

[2]*Leiden University Medical Center,e-mail:* j.j.goeman@lumc.nl

[3]*University of Twente and Centrum Wiskunde & Informatica Amsterdam,e-mail:* r.deheide@utwente.nl

Abstract: E-processes enable hypothesis testing with ongoing data collection while maintaining Type I error control. However, when testing multiple hypotheses simultaneously, current e-value based multiple testing methods such as e-BH are not invariant to the order in which data are gathered for the different e-processes. This can lead to undesirable situations, e.g., where a hypothesis rejected at time t is no longer rejected at time $t+1$ after choosing to gather more data for one or more e-processes unrelated to that hypothesis. We argue for multiple testing methods working with suprema of e-processes. We provide an example to illustrate that e-BH does not control the FDR, at level α when applied to suprema of e-processes. From the same example we see that the FWER is not controlled with averaging, and also closed e-BH does not control the FDR. We show that adjusters can be used to ensure FDR-sup control with e-BH under arbitrary dependence.

MSC2020 subject classifications: Primary 62L99.
Keywords and phrases: e-values, e-BH, multiple testing, sequential testing.

Received July 2025.

The e-process is a recent innovative conceptual framework for hypothesis testing. An e-process is a nonnegative stochastic process $(E_t)_{t\geq 0}$, which starts at a value of $E_0 \leq 1$ and is continually updated over time as more information comes in. It does not grow beyond 1 in expectation if the null hypothesis H_0 is true, that is, $\mathbb{E}_P(E_\tau) \leq 1$ for every stopping time τ and for every $P \in H_0$. Therefore, we may reject H_0 when, at any time point t, we have $E_t \geq 1/\alpha$. Doing so will control Type I error, since, if H_0 is true, Ville's inequality ([11], Fact 6.4) guarantees that e-processes have the property that

$$P\left(\sup_{t\in\mathbb{N}} E_t \geq 1/\alpha\right) \leq \alpha. \tag{1}$$

In contrast to p-values, e-processes can be repeatedly evaluated until, hopefully, rejection follows for some t. This means that a researcher can continue gathering data as long as they like. Notably, where the definition of e-processes involves a stopping time, the Type I error control property (1) involves the supremum of the process. This implies that data gathering is carefree: the researcher never has to regret gathering more data. For example, suppose that the researcher decided to gather E_{t+1} before properly evaluating E_t. If $E_t \geq 1/\alpha$ but $E_{t+1} < 1/\alpha$, then they can still reject H_0 and appeal to (1) for Type I error control, effectively undoing the gathering of E_{t+1}, as long as α was fixed a priori (or even if it is not: see [9, 7]).

If more than one hypothesis is of interest, multiple testing considerations apply. There is by now a large literature on e-based multiple testing. Most notably, the e-BH procedure

arXiv: 2501.19360

[18] generalizes the Benjamini-Hochberg procedure for controlling the False Discovery Rate (FDR). However, this literature has so far almost exclusively focused on multiple testing with e-values, which are single instances of e-processes, i.e., random variables E for which $\mathbb{E}(E) \leq 1$ if H_0 is true. When applied to parallel e-processes $E_t^1, \ldots, E_t^K$, as we will show below, e-BH guarantees that

$$\sup_{t \in \mathbb{N}} \mathrm{FDR}\left(E_t^1, \ldots, E_t^K\right) \leq \alpha. \tag{2}$$

The supFDR (2) criterion reflects the fact that researchers may examine the e-process at arbitrary and data-dependent times. Since e-processes may trigger rejections earlier than the final value, the error rate must account for the worst moment over time. This is analogous to anytime-valid Type I control for single hypotheses.

Unfortunately, data gathering under the guarantee (2) is not carefree. Let us illustrate this with an example. For $K = 2$ e-values, e-BH rejects any hypothesis for which the e-value exceeds $2/\alpha$, and both hypotheses if both e-values exceed $1/\alpha$. Now suppose that at time t we have $1/\alpha \leq E_t^1 < 2/\alpha$ and $E_t^2 < 1/\alpha$. Gathering more data for both e-processes, however, the situation could reverse, and $E_{t+1}^1 < 1/\alpha$ and $1/\alpha \leq E_{t+1}^2 < 2/\alpha$. Now, the researcher regrets gathering more data for E^1, since (2) does not support any rejection now, but would have supported rejection of both null hypotheses if the researcher would have stopped E^1, taking $E_{t+1}^1 = E_t^1$. The lack of the carefree property complicates experimental design because collecting more data for one or more hypotheses can cause previously rejected hypotheses—those for which more data has been gathered, or even others—to lose their rejected status. This means that researchers may wish to consider carefully, in real time, for which hypotheses to gather more data, adding data where it seems most promising. As a result, the process becomes volatile and harder to manage. Carefree procedures avoid this problem by ensuring that collecting more data never leads to losing discoveries.

For carefree data gathering, the researcher must always be allowed to revert any e-process, for which they regret gathering more data, back to an earlier time point. The optimal reversion is to the maximum so far, so a carefree procedure should guarantee

$$\sup_{s \in \mathbb{N}} \mathrm{FDR}(\max_{t \leq s} E_t^1, \ldots, \max_{t \leq s} E_t^K) \leq \alpha. \tag{3}$$

We argue that (3), the *FDR-sup*, is the proper FDR criterion for e-processes. In this paper, we investigate how to construct procedures controlling (3).

1. Real-data illustration: why the carefree property matters

To ground the discussion in an applied setting, we illustrate the practical consequences of non-carefree multiple testing on a real genomic survival dataset. The key phenomenon is *volatility of discoveries*: when evidence is gathered sequentially for many hypotheses, a multiple testing procedure that is not carefree may reject a hypothesis at some intermediate time point and later fail to reject it after additional data collection, even if the later data collection is not targeted at that specific hypothesis. This behavior complicates experimental design and reporting, because discoveries become sensitive to how and when interim analyses are performed.

We use the `breastCancerNKI` microarray dataset derived from the NKI-295 breast cancer study [14], which contains gene expression measurements together with the distant

metastasis-free survival time and the event indicator. Each gene defines a separate hypothesis of no association between its expression level and survival, resulting in a large-scale multiple testing problem.

To mimic sequential acquisition of information, we partition the $n = 295$ patients into $K = 100$ disjoint groups of (approximately) equal size using a random split. The randomization is employed solely to avoid introducing any systematic ordering or selection effects; the qualitative behavior described below is robust and appears consistently across different partitions. At step $k = 1, \dots, K$, we reveal the next patient group and update the evidence. Thus, step k corresponds to an interim analysis based on the first k partitions.

For each gene and at each step k, we fit a Cox proportional hazards model of survival on the gene expression using only the patients in partition k and compute the score test p-value. We convert this p-value to an e-value using the calibrator

$$e(p) = \frac{1 - p + p \log p}{p(\log p)^2}, \qquad p \in (0, 1),$$

which yields a valid e-value for any null p-value [11]. Letting $p_{g,k}$ denote the p-value for gene g at step k, we define

$$e_{g,k} = e(p_{g,k}), \qquad E_{g,k} = \prod_{j=1}^{k} e_{g,j}.$$

This construction yields a nonnegative process $(E_{g,k})_{k \geq 1}$ that forms an e-process under the null hypothesis, and increases when successive partitions provide evidence against the null.

At each step k, we apply e-BH at level $\alpha = 0.05$ to the collection of current e-process values $\{E_{g,k}\}_g$, producing a rejection set $\mathcal{R}_k$. We visualize the rejections as a binary matrix $Z \in \{0, 1\}^{G \times K}$ with entries

$$Z_{g,k} = \mathbf{1}\{g \in \mathcal{R}_k\},$$

where G is the number of genes. Figure 1 plots this matrix as a black–white heatmap across steps (restricted to genes exhibiting non-monotone rejection behavior).

For many genes the rejection status is not monotone in k. Some genes are not rejected initially, become rejected at intermediate steps when the accumulated evidence temporarily crosses the e-BH threshold, and then cease to be rejected after additional partitions are incorporated. In other words, discoveries can appear and later disappear purely due to continued data collection. This is precisely the type of "regret" that carefree multiple testing aims to eliminate: the researcher would have preferred to stop (or revert) earlier for some hypotheses, but the standard stopped e-BH analysis at later times no longer supports the earlier discovery.

This real-data behavior motivates the FDR-sup perspective: to guarantee carefree data collection, inference must be stable under continued sampling and optional stopping. A natural way to enforce accept-to-reject behavior is to base inference on running maxima of the underlying e-processes, ensuring that once a hypothesis attains strong evidence, it remains eligible for rejection. However, as shown in Proposition 3.2, naive application of e-BH to running maxima does not in general control the FDR under arbitrary dependence, which motivates the adjusted running-maximum approach developed in Section 3.

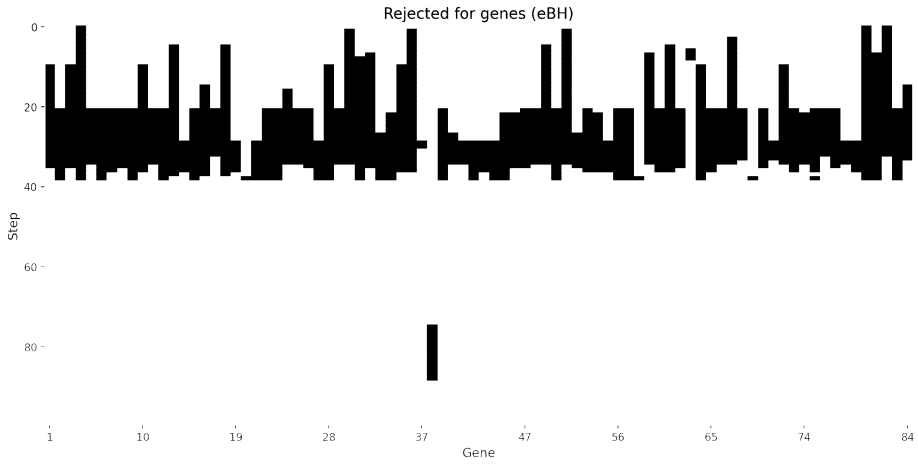

Fig 1. *Real-data illustration of volatility without carefree multiple testing. Heatmap of e-BH rejection indicators over time on the* `breastCancerNKI` *dataset. The vertical axis corresponds to sequential analysis steps* $k = 1, \ldots, K$ *with* $K = 100$*, where each step reveals one additional patient partition. Each column corresponds to a gene (restricted here to genes exhibiting non-monotone rejection behavior), and black indicates rejection of the corresponding null hypothesis at that step. The figure illustrates that, for many genes, rejection can occur at intermediate steps and subsequently disappear after additional data are incorporated, highlighting the lack of the carefree property when inference is not based on running maxima.*

2. FDR control with e-processes

2.1. The e-BH procedure

Let $H_1, \ldots, H_K$ be K hypotheses, denote the set of their indices with $\mathcal{K} = \{1, \ldots, K\}$, and let P^* be the true data-generating probability measure. For each $k \in \mathcal{K}$, H_k is called *true* if $P^* \in H_k$. Let $\mathcal{N} \subseteq \mathcal{K}$ be the set of indices of true null hypotheses and let K_0 be its cardinality. For each $k \in \mathcal{K}$, H_k is associated with an e-value E^k.

When there is only one observation of the e-value for each hypothesis, it has been shown [18] that the classical Benjamini-Hochberg (BH) procedure can be modified to accommodate e-values. The most powerful property of this *e-BH* procedure, is that it provides FDR control for any dependence structure between the e-values, in contrast with the classical BH procedure, which requires the assumption of positive regression dependence on a subset (PRDS). Both the classical BH and also the Benjamini-Yekutieli (BY) procedures can be recovered as special cases of the e-BH procedure.

2.2. e-BH with e-processes

Formally, a non-negative process E is called an e-process when $\mathbb{E}[E_\tau] \leq 1$ for any stopping time τ, where the expectation is taken under any distribution in the null hypothesis.

We consider the setting in which all e-processes are adapted to a common filtration $(\mathcal{F}_t)_{t \geq 0}$, as this is required to guarantee joint validity under stopping. This setting corresponds to many practical situations, and has been discussed in detail by [16].

For each $k \in \mathcal{K}$, let H_k be associated with an e-process $(E^k_t)_{t \geq 0}$, and let $\tau_1, \tau_2, \ldots, \tau_K$ be their arbitrary stopping times. Let $S_k = E^k_{\tau_k}$ be the stopped e-processes.

Procedure 2.1 (Naive e-BH with e-processes)**.** Let $S_{[k]}$ for $k \in \mathcal{K}$ be the k-th order statistic of $S_1, \ldots, S_K$, sorted from the largest to the smallest, such that $S_{[1]}$ is the largest stopped e-process. The stopped e-BH procedure $\mathcal{G}(\alpha) : [0, \infty]^K \to 2^{\mathcal{K}}$ then rejects hypotheses with the largest k^*_{se} stopped e-processes, where

$$k^*_{se} = \max \left\{ k \in \mathcal{K} : S_{[k]} \geq \frac{1}{\alpha} \frac{K}{k} \right\}.$$

Fact. The stopped e-BH procedure applied to arbitrarily dependent e-processes, adapted to a common filtration), has FDR at most $K_0\alpha/K$.

Since a stopped e-process has expectation under the null bounded by one, i.e. $\mathbb{E}S_k = \mathbb{E}E^k_{\tau_k} \leq 1$, the set of stopped e-processes $\{S_k\}_{k\in\mathcal{K}}$ is a set of e-variables. Thus, this procedure reduces to the standard e-BH procedure on the stopped e-processes $\{S_k\}_{k\in\mathcal{K}}$, and the FDR guarantee [18] carries over.

2.3. Related work

Online multiple testing has been considered both classically with p-values [4] and e-values [5, 6, 19], but online is here to be understood in the dimension of the hypothesis space: time points represent the addition of an hypothesis. Until recent, the only paper that considered multiple testing with e-values where data can be added sequentially, was [19], where they introduce the *doubly sequential* framework. For p-values there is work on interim-analyses and group-sequential multiple testing, e.g. [12] on FDR; but to the best of our knowledge there is no work on any-time valid methods. Between uploading the present paper to arxiv and the submission, two notable developments appeared. [16] provide a rigorous setup for using e-BH on stopped e-processes. They raise the important issue that one needs to be careful about the filtrations used for the stopped e-processes. In the present paper we circumvented the issue by assuming a common underlying filtration, but it would be good to investigate the interplay between their result and ours. Secondly, [21] recently proposed a novel necessary and sufficient principle for FDR-controlling methods. Their principle gives rise to a substantial improvement of e-BH, the resulting procedure is called *closed e-BH*, and we added a subsection to show that our central proposition holds for closed e-BH as well. Lastly, [8] study closed testing for familywise error rate (FWER) control, both in the static and sequential setting.

3. Multiple testing with the running maximum of an e-process

We have argued that multiple-testing procedures for e-processes should possesses the carefree property that sequential data gathering will not lead to regrets, and that this requires the procedure to work with the running maximum of the e-process: $M^k_t = \max_{1\leq s\leq t} E^k_s$.

It is straightforward to observe that in such a procedure, the set of rejected hypotheses forms a non-decreasing set. That is, if a hypothesis is rejected at a given point in time, it will remain rejected in all subsequent time points. It therefore automatically possesses the accept-to-reject property, as introduced in [6]: a multiple testing method that can change earlier non-rejections into a rejection at a later point in time, but not the other way around. As noted before, in that work the sequential aspect was in terms of adding hypotheses rather than data points. We argue however that also in the latter context, the accept-to-reject principle is a desirable feature of any multiple testing method.

Procedure 3.1 (Running maximum e-BH procedure)**.** Let $M_t^{[k]}$ for $k \in \mathcal{K}$ be the k-th order statistic of $M_t^1, \ldots, M_t^K$, sorted from the largest to the smallest so that $M_t^{[1]}$ is the largest maximum of an e-process. The stopped e-BH procedure $\mathcal{G}(\alpha) : [0, \infty]^K \to 2^{\mathcal{K}}$ then rejects hypotheses with the largest k_m^* maxima of e-processes, where

$$k_m^* = \max \left\{ k \in \mathcal{K} : M_t^{[k]} \geq \frac{1}{\alpha} \frac{K}{k} \right\}.$$

Since the running maximum of an e-process is not an e-process itself (see [1] for more on this), we cannot directly deduce FDR control for this procedure. However, in the case that all e-processes are independent, it is straightforward to see that this procedure ensures FDR control. This is because, in essence, the inverse of the running maximum of the e-processes $1/M_k$ is a valid p-process [1]. Thus, applying this procedure to the running maximum is equivalent to applying the classical BH procedure.

However, in the case of arbitrary dependence between e-processes, the e-BH procedure loses its FDR control, as asserted by the following proposition.

Proposition 3.2. *The running maximum e-BH procedure applied to arbitrarily dependent e-processes does not provide FDR control at level* α.

3.1. Proof of Proposition 3.2

We present a counterexample that demonstrates how the e-BH procedure, when applied to the running maxima of e-processes, fails to control the FDR at the nominal level α. The FDR is evaluated numerically.

Fix $\alpha \in (0, 1/2]$. Consider two true null hypotheses with corresponding e-processes E_t^1 and E_t^2, defined as:

$$E_t^1 = X_0^1 \prod_{s=1}^{t} e_s^1, \quad E_t^2 = X_0^2 \prod_{s=1}^{t} e_s^2,$$

where the initial values (X_0^1, X_0^2) are distributed as:

$$(X_0^1, X_0^2) \sim \begin{cases} (0,0), & \text{with probability } 1 - 2\alpha, \\ \left(\frac{1}{2\alpha}, \frac{1}{2\alpha}\right), & \text{with probability } 2\alpha, \end{cases}$$

and the increments (e_s^1, e_s^2) are distributed as:

$$\begin{cases} \left(\frac{1}{2}, \frac{1}{2}\right), & \text{with probability } \frac{1}{3}, \\ \left(2, \frac{1}{2}\right), & \text{with probability } \frac{1}{3}, \\ \left(\frac{1}{2}, 2\right), & \text{with probability } \frac{1}{3}. \end{cases}$$

The variables X_0^1, X_0^2, and all e_s^1, e_s^2 are independent.

It is straightforward to verify that the expected values of E_t^1 and E_t^2 are equal to 1 at all times, i.e., $\mathbb{E}[E_t^1] = \mathbb{E}[E_t^2] = 1$, which ensures that both E_t^1 and E_t^2 are valid e-processes.

However, the processes E_t^1 and E_t^2 are not independent because the random variables e_s^1 and e_s^2, as well as X_0^1, X_0^2, are dependent.

Since the null hypotheses are true, the FDR-sup is equal to the probability of rejecting at least one of them. This corresponds to the event:

$$M_t^1 \geq \frac{2}{\alpha}, \quad M_t^2 \geq \frac{2}{\alpha}, \quad \text{or both } M_t^1, M_t^2 \geq \frac{1}{\alpha}, \tag{4}$$

where $M_t^k = \max_{1 \leq s \leq t} E_s^k$ as above. To confirm the validity of this example, we performed simulations of the described e-processes (the code is provided in the Appendix B).

The results of the simulations show that, in this scenario, the FDR of the e-BH procedure applied to the running maxima is approximately 1.08α. In our simulations, we used $M = 10^6$ Monte Carlo iterations. Therefore, based on the standard error formula $\sqrt{p(1-p)/M}$, the estimated standard deviation of the error is approximately 0.00023 for $\alpha = 0.05$. This demonstrates that when using running maxima, additional corrections or normalizations are necessary to ensure proper FDR control.

3.2. Closed e-BH

Recently, [21] proposed a novel necessary and sufficient principle for FDR-controlling methods. The principle yields a procedure that offers substantial improvements over standard e-BH.

Let $\mathbf{E} \coloneqq (\mathbf{e}_S)_{S \in 2^{[m]}}$ be an *e-collection*: a family of nonnegative random variables satisfying $\mathbb{E}_{\mathrm{P}}(\mathbf{e}_{N_{\mathrm{P}}}) \leq 1$ for all $\mathrm{P} \in H_0$. That is, $\mathbf{e}_{N_{\mathrm{P}}}$ is an e-value for each $\mathrm{P} \in H_0$, where N_{P} denotes the index set of true nulls. The set $\mathcal{R}$ of candidate rejection sets R for the FDR at level α is defined as

$$\mathcal{R}_\alpha^{\mathrm{FDR}}(\mathbf{E}) \coloneqq \left\{ R \in 2^{[m]} : \mathbf{e}_S \geq \frac{|R \cap S|}{\alpha(|R| \vee 1)} \quad \forall S \in 2^{[m]} \right\}. \tag{5}$$

In their *e-Closure Principle* (Theorem 6 of [21]), the authors show that any procedure whose rejection sets lie in (5) controls the FDR at level α. As a novel FDR controlling method for e-values under arbitrary dependence, [21] introduce the *closed eBH* procedure $\overline{\mathrm{eBH}}$, in which the e-collection is composed of the arithmetic averages of the e-values in the corresponding subset:

$$\mathbf{e}_S = \frac{1}{|S|} \sum_{i \in S} \mathbf{e}_i \tag{6}$$

The authors prove that $\overline{\mathrm{eBH}}$ uniformly improves upon standard e-BH. In particular, the rejection set produced by standard e-BH is always contained in the rejection sets of $\overline{\mathrm{eBH}}$. Therefore, the proof of Proposition 3.2 extends directly to $\overline{\mathrm{eBH}}$.

Corollary 3.3. *Applying* $\overline{\mathrm{eBH}}$ *to the running maxima of arbitrarily dependent e-processes does not control the FDR at level* α.

Note that Proposition 3.2 already shows that the standard e-BH fails to control the FDR when applied to running maxima. Since closed e-BH uniformly improves e-BH by enlarging the rejection set ([21]), any violation of FDR control for e-BH immediately extends to closed e-BH, because the counterexample is all-null. Hence the counterexample from Proposition 3.2 is inherited, establishing the corollary.

3.3. FWER

Wang [17] recently proved that, under arbitrary dependence, the only admissible way of merging e-values is via a weighted arithmetic average. More precisely, any e-merging function that outputs a valid global e-value for all joint distributions and cannot be uniformly improved must be of the form $\sum_{k=1}^{K} w_k E^k$ with nonnegative weights summing to one; all other merging rules are inadmissible. This result generalizes earlier work of [15] on symmetric merging functions and rigorously justifies why averaging is the canonical operation for combining arbitrary e-values. In particular, in the FWER context, if we want a valid global e-value under arbitrary dependence, we are essentially forced to use (possibly weighted) arithmetic averaging of the individual e-values.

We see from the events for which one or both hypotheses are rejected in e-BH, stated Equation 4, that averaging those e-values will also always lead to rejection. Since e-values are always positive, if either one of them is larger than $2/\alpha$, the average will exceed $1/\alpha$, and if both of them are larger than $1/\alpha$ the average is trivially larger than $1/\alpha$ as well. Therefore, Proposition 3.2 extends to FWER as well.

Corollary 3.4. *The procedure of averaging the running maxima of arbitrarily dependent e-processes does not control the FWER at level α.*

4. Adjusting running maxima

Since Proposition 3.2 shows that e-BH does not control the FDR-sup, one way forward is to use *adjusters*, which can transform the running maximum process such that it becomes an e-process. Adjusters were introduced in the context of option pricing [2] and game theoretic probability [3, 13], ensuring against loss of capital or evidence, and in [1] as means to combine evidence across filtrations.

An *admissible adjuster* [1] is an increasing function $\mathrm{A} : [1, \infty] \to [0, \infty]$ if and only if it is right-continuous, $\mathrm{A}(\infty) = \lim_{E \to \infty} \mathrm{A}(E) = \infty$, and $\int_1^\infty \mathrm{A}(E)/E^2 \,\mathrm{d}\, E = 1$.

Intuitively, the running maximum of an e-process is not itself an e-process because it can only increase, which introduces an upward bias: once a large value is attained, it is never undone. As a result, the key validity property of e-processes—namely that the expectation at arbitrary stopping times is bounded by one under the null—no longer holds. An adjuster corrects this bias by applying a carefully calibrated, slowly growing transformation to the running maximum. Importantly, this transformation does not restore the supermartingale property (which the running maximum cannot satisfy), nor does the adjusted process start at one. Instead, the defining integral condition of an adjuster guarantees that, under the null hypothesis, the adjusted running maximum satisfies $\mathbb{E}[A(M_\tau)] \leq 1$ for all stopping times τ. In this sense, an adjuster acts as a compensating mechanism that restores the essential expectation control required for valid optional stopping, thereby making the carefree use of running maxima statistically sound.

Examples of adjusters can be found in the aforementioned papers, where they are named *lookback adjuster, capital calibrator, martingale calibrator*, or *adjuster*. If $M_t^k = \sup_{s \leq t} E_s$ is the running maximum of an e-process for a family of null hypotheses H_0, then $(\mathrm{A}(M_t^k))_{t \geq 0}$ is also an e-process for H_0. Since the adjusted process is non-decreasing, it will never result in regretting having collected more evidence.

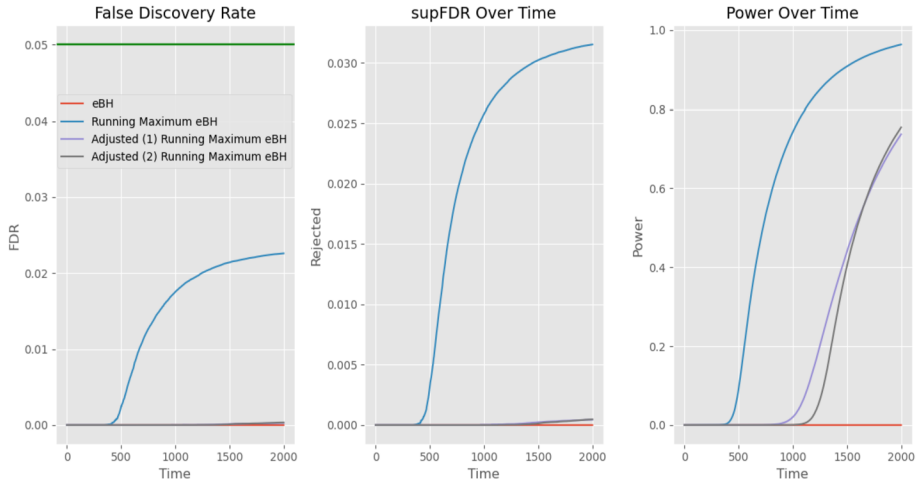

Fig 2. *Performance comparison of multiple testing methods (red line - the standard e-BH, blue line - running maximum e-BH, purple line - adjusted running maximum e-BH using A_1 from (7), black line - adjusted running maximum e-BH using A_2 from (7), green line - the confidence threshold). Plots: (1) FDR over time shows control below the threshold, (2) supFDR indicates worst-time FDR and (3) power - the proportion of rejected false hypotheses increases over time.*

4.1. FDR-sup control with the adjusted running maximum

Theorem 4.1. *The e-BH procedure applied to the adjusted running maxima of e-processes controls the FDR-sup at level $K_0\alpha/K$.*

One way to prove this is to follow the classical proof of [18], and is detailed in Appendix A.

4.2. Example: e-BH with the adjusted running maximum

Two examples of admissible adjusters are

$$\mathrm{A}_1(E) = \frac{E - 1 - \log E}{\log^2 E}, \qquad \mathrm{A}_2(E) = \sqrt{E} - 1. \tag{7}$$

In Figure 2 we see an example of running e-BH on the running maximum of an e-process adjusted by (7). The code to generate these plots can be found in Appendix B.

We conducted simulations to evaluate the performance of three procedures: standard e-BH, running maximum e-BH, and adjusted running maximum e-BH. The simulation involves 200 hypotheses, where the null hypothesis (H_0) assumes a normal distribution with mean $\mu = 0$, and the alternative hypothesis (H_1) assumes $\mu = 0.1$. For each hypothesis, data is generated sequentially, creating a time series of observations.

The simulation includes dependencies among the hypotheses, introduced via a fixed randomly generated correlation matrix for the multivariate normal distributions for $X_t^1, X_t^2, \ldots, X_t^K$, which are independent over time. To construct the e-process for each hypothesis, we compute the likelihood ratio at each time step as follows:

$$E_t^k(X^t) = \prod_{i=1}^{t} \frac{f_1(X_i^k)}{f_0(X_i^k)} = \frac{e^{\frac{1}{2}\sum_{i=1}^{t}(X_i^k - 0.1)^2}}{e^{\frac{1}{2}\sum_{i=1}^{t}(X_i^k)^2}}.$$

Each simulation generates $T = 2000$ observations for each hypothesis, and the performance metrics (e.g., FDR, power, number of rejections) are averaged over $M = 10000$ repetitions. This setup allows us to compare the efficiency and FDR control of the three methods under a dependent data structure.

In this example, we see that standard e-BH on the running maximum does not reach level α. E-BH is known to be very conservative [18, 10, 20], which is why the methods seems to control FDR, even though Proposition 3.2 says this is not the case in general. We also see that the adjusted maximum processes remain below level α as expected, and, unfortunately, we see an appreciable loss of power. Power could be boosted by using techniques such as conditional calibration [10] or stochastic rounding [20]. Still, the power loss of adjusted processes remains an important drawback of this way of controlling the FDR-sup.

5. Conclusion

We have argued that multiple testing procedures based on e-processes should be carefree, meaning that the researcher should never lose out on any rejections because they chose to gather more data for one or more e-processes. If procedures are not carefree, experimental design becomes highly complex, since the decision when to gather more data for which e-processes becomes consequential. Carefree multiple testing procedures must operate on the running maxima of e-processes. Such running maxima are not themselves e-processes, but can be turned into e-processes by adjusters. Adjustment is costly, however, in terms of power.

Appendix A: Proofs of FDR-sup control with adjusted e-BH

A.1. Proof of Theorem 4.1

In one way to prove that the e-BH procedure with e-processes controls the FDR-sup, an adjuster is needed to make the last step of the classical proof by [18] work. For each $s \in \mathbb{N}$, let

$$M_s^j := \sup_{t \le s} E_t^k, \qquad k \in \mathcal{K},$$

and let

$$\mathcal{R}_s := \mathcal{R}_{\text{e-BH}}\big(A(M_s^1), \dots, A(M_s^K)\big)$$

denote the e-BH rejection set based on the adjusted running maxima.

By the self-consistency property of e-BH, if $k \in \mathcal{R}_s$, then

$$A(M_s^k) \ge \frac{K}{\alpha |\mathcal{R}_s|}.$$

Consequently, for every $k \in \mathcal{N}$,

$$\frac{\mathbf{1}\{k \in \mathcal{R}_s\}}{|\mathcal{R}_s| \vee 1} \le \frac{\alpha}{K} A(M_s^k) \mathbf{1}\{k \in \mathcal{R}_s\} \le \frac{\alpha}{K} A(M_s^k). \tag{8}$$

It follows that, for every $s \in \mathbb{N}$,

$$\text{FDR}_s = \sum_{k \in \mathcal{N}} \mathbb{E}\left[\frac{\mathbf{1}\{k \in \mathcal{R}_s\}}{|\mathcal{R}_s| \vee 1}\right]$$

$$\leq \frac{\alpha}{K} \sum_{k \in \mathcal{N}} \mathbb{E}\left[A(M_s^k)\right] \tag{9}$$

$$\leq \frac{K_0 \alpha}{K}, \tag{10}$$

where the final inequality follows because $(A(M_s^k))_{s \geq 0}$ is an e-process under every true null hypothesis. Taking the supremum over $s \in \mathbb{N}$ gives

$$\sup_{s \in \mathbb{N}} \mathrm{FDR}_s \leq \frac{K_0 \alpha}{K},$$

which proves the result.

Appendix B: Code

B.1. Code for the proof of Proposition 3.2

The code can be found at the following link: https://github.com/tavyrikov/runmax_eBH/blob/main/counterexpample.py

B.2. Code for generating Figure 2

The code can be found at the following link: https://github.com/tavyrikov/runmax_eBH/blob/main/simulation.py

Acknowledgments

We thank Aaditya Ramdas and Nick Koning for helpful comments on an earlier version of this paper.

Funding

Rianne de Heide's work was supported by NWO Veni grant number VI.Veni.222.018.

References

[1] Choe, Y. J. and Ramdas, A. (2024). Combining Evidence Across Filtrations. arXiv: 2402.09698.

[2] Dawid, A. P., de Rooij, S., Grunwald, P., Koolen, W. M., Shafer, G., Shen, A., Vereshchagin, N. and Vovk, V. (2011). Probability-Free Pricing of Adjusted American Lookbacks. arXiv:1108.4113. https://doi.org/10.48550/arXiv.1108.4113.

[3] Dawid, A. P., de Rooij, S., Shafer, G., Shen, A., Vereshchagin, N. and Vovk, V. (2011). Insuring against Loss of Evidence in Game-Theoretic Probability. *Statistics & Probability Letters* **81** 157–162. https://doi.org/10.1016/j.spl.2010.10.013. MR2740080

[4] FISCHER, L., BOFILL ROIG, M. and BRANNATH, W. (2024). The Online Closure Principle. *The Annals of Statistics* **52**. https://doi.org/10.1214/24-AOS2370. MR4744197

[5] FISCHER, L. and RAMDAS, A. (2024). Online Closed Testing with E-Values. arXiv:2407.15733.

[6] FISCHER, L. and RAMDAS, A. (2024). An Online Generalization of the E-BH Procedure. arXiv:2407.20683. https://doi.org/10.48550/arXiv.2407.20683.

[7] GRÜNWALD, P. D. (2024). Beyond Neyman–Pearson: E-values enable hypothesis testing with a data-driven alpha. *Proceedings of the National Academy of Sciences* **121** e2302098121. MR4826968

[8] HARTOG, W. and LEI, L. (2025). Family-wise Error Rate Control with E-values.

[9] KONING, N. W. (2024). Post-Hoc α Hypothesis Testing and the Post-hoc p-Value. arXiv:2312.08040. https://doi.org/10.48550/arXiv.2312.08040.

[10] LEE, J. and REN, Z. (2024). Boosting E-BH via Conditional Calibration. arXiv:2404.17562. https://doi.org/10.48550/arXiv.2404.17562.

[11] RAMDAS, A. and WANG, R. (2024). Hypothesis Testing with E-Values. arXiv:2410.23614. https://doi.org/10.48550/arXiv.2410.23614.

[12] SARKAR, S. K., CHEN, A., HE, L. and GUO, W. (2019). Group sequential BH and its adaptive versions controlling the FDR. *Journal of Statistical Planning and Inference* **199** 219–235. https://doi.org/10.1016/j.jspi.2018.07.001. MR3857824

[13] SHAFER, G., SHEN, A., VERESHCHAGIN, N. and VOVK, V. (2011). Test Martingales, Bayes Factors and p-Values. *Statistical Science* **26** 84–101. https://doi.org/10.1214/10-STS347. MR2849911

[14] VAN DE VIJVER, M. J., HE, Y. D., VAN'T VEER, L. J., DAI, H., HART, A. A., VOSKUIL, D. W., SCHREIBER, G. J., PETERSE, J. L., ROBERTS, C., MARTON, M. J. et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* **347** 1999–2009.

[15] VOVK, V. and WANG, R. (2021). E-values: Calibration, combination and applications. *The Annals of Statistics* **49** 1736–1754. MR4298879

[16] WANG, H., DANDAPANTHULA, S. and RAMDAS, A. (2025). Anytime-valid FDR control with the stopped e-BH procedure. MR4943236

[17] WANG, R. (2025). The only admissible way of merging arbitrary e-values. *Biometrika* **112** asaf020. MR4926277

[18] WANG, R. and RAMDAS, A. (2022). False Discovery Rate Control with E-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **84** 822–852. https://doi.org/10.1111/rssb.12489. MR4460577

[19] XU, Z. and RAMDAS, A. (2023). Online Multiple Testing with E-Values. arXiv:2311.06412.

[20] XU, Z. and RAMDAS, A. (2024). More Powerful Multiple Testing under Dependence via Randomization. arXiv:2305.11126. https://doi.org/10.48550/arXiv.2305.11126.

[21] XU, Z., SOLARI, A., FISCHER, L., DE HEIDE, R., RAMDAS, A. and GOEMAN, J. (2025). Bringing closure to false discovery rate control: A general principle for multiple testing. *arXiv preprint* arXiv:2509.02517.