Electronic Journal of Statistics
Vol. 18 (2024) 2292–2338
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2253

Post-selection inference for e-value based confidence intervals

Ziyu Xu[1], **Ruodu Wang**[3] **and Aaditya Ramdas**[1,2]

[1]*Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, USA, e-mail:* xzy@cmu.edu; aramdas@cmu.edu

[2]*Machine Learning Department, Carnegie Mellon University, Pittsburgh, USA*

[3]*Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada, e-mail:* wang@uwaterloo.ca

Abstract: Suppose that one can construct a valid $(1-\delta)$-confidence interval (CI) for each of K parameters of potential interest. If a data analyst uses an arbitrary data-dependent criterion to select some subset S of parameters, then the aforementioned CIs for the selected parameters are no longer valid due to selection bias. We design a new method to adjust the intervals in order to control the false coverage rate (FCR). The main established method is the "BY procedure" by Benjamini and Yekutieli (JASA, 2005). The BY guarantees require certain restrictions on the selection criterion and on the dependence between the CIs. We propose a new simple method which, in contrast, is valid under any dependence structure between the original CIs, and any (unknown) selection criterion, but which only applies to a special, yet broad, class of CIs that we call e-CIs. To elaborate, our procedure simply reports $(1-\delta|S|/K)$-CIs for the selected parameters, and we prove that it controls the FCR at δ for confidence intervals that implicitly invert e-values; examples include those constructed via supermartingale methods, via universal inference, or via Chernoff-style bounds, among others. The e-BY procedure is admissible, and recovers the BY procedure as a special case via a particular calibrator. Our work also has implications for post-selection inference in sequential settings, since it applies at stopping times, to continuously-monitored confidence sequences, and under bandit sampling. We demonstrate the efficacy of our procedure using numerical simulations and real A/B testing data from Twitter.

MSC2020 subject classifications: Primary 60G10, 62M07.
Keywords and phrases: Post-selection inference, e-value, e-confidence interval, false coverage rate.

Received May 2023.

Contents

arXiv: 2203.12572

1. Introduction

One of the most classical problems in statistics is the problem of parameter estimation e.g. estimating the mean of a distribution. The time-tested solution is to produce a confidence interval (CI) in the parameter space that covers the true parameter with high probability. However, many scientists are not simply interested in estimating a single parameter — they may have many *potentially interesting* parameters to estimate concurrently.

For example, a scientist might be experimenting with K vaccines for a strain of virus. She may only be interested in reporting CIs for the vaccines that the data suggest to be effective. However, she does not know which vaccines may be effective until after she looks at the data. A reasonable thing she might do is to report the usual 95% CIs for the vaccines where the CI is positive (i.e., the entire interval for the treatment effect is above zero). Consequently, she is using the *same data* to both select the parameters she wishes to estimate, and construct

estimates of these parameters. As a result, the reported (uncorrected) CIs do not provide valid statistical coverage. To see this, consider a scenario where none of the vaccines reduce mortality. Any reported CI will have a coverage probability of 0%, since the CI for a vaccine is reported only if it is positive, meaning that it excludes the true effect of zero.

The above example illustrates the issue of selection bias for post-selection inference, as prominently noted by Benjamini and Yekutieli [4]. Motivated by this example, let us quickly introduce the formal problem setup.

Let $\mathcal{P}$ denote the universe of all possible distributions for the data. Let $\vartheta : \mathcal{P} \to \Theta$ denote a functional (or parameter) that maps distributions to parameter values lying in some set Θ. A scientist observes some data, $\mathbf{X} = (X_1, \ldots, X_K)$ that is drawn from an unknown distribution, $\mathtt{P}^*$, and we are potentially interested in the values of K of its functionals[1] $\vartheta_1, \ldots, \vartheta_K$, but our interest in them depends on the unknown parameter values $\boldsymbol{\theta}^* := (\theta_1^*, \ldots, \theta_K^*)$, where $\theta_i^* := \vartheta_i(\mathtt{P}^*)$.

Let $\mathbb{E}_{\mathtt{P}}$ and $\mathbb{P}_{\mathtt{P}}$ denote the expectation and probability under a distribution $\mathtt{P}$, respectively, although we sometimes suppress the subscript in the case of $\mathtt{P}^*$ for simplicity. For each $i \in [K]$, we assume that from X_i, the scientist can construct a marginal $(1-\alpha)$-confidence interval for θ_i^*, for any desired level $\alpha \in [0,1]$, which is a set $C_i(\alpha) \subseteq \Theta$ such that $\mathbb{P}(\theta_i^* \notin C_i(\alpha)) \leq \alpha$, or more explicitly, for any $\mathtt{P} \in \mathcal{P}$, we would have $\mathbb{P}_{\mathtt{P}}(\vartheta_i(\mathtt{P}) \notin C_i(\alpha)) \leq \alpha$.

The scientist uses the data $\mathbf{X}$ to select a subset of "interesting" parameters, $\mathcal{S} \subseteq [K]$, using some potentially complex data-dependent selection rule $\mathfrak{S} : \mathbf{X} \mapsto \mathcal{S}$. The scientist must then devise confidence levels for the CI of each selected parameter, $\{\alpha_i\}_{i \in \mathcal{S}}$, that *can depend on the data* $\mathbf{X}$. The *false coverage proportion* (FCP) and *false coverage rate* (FCR) of such a procedure are:

$$\mathrm{FCP} := \frac{\sum_{i \in \mathcal{S}} \mathbf{I}\left\{\theta_i^* \notin C_i(\alpha_i)\right\}}{|\mathcal{S}| \vee 1}, \qquad \mathrm{FCR} := \mathbb{E}\left[\mathrm{FCP}\right],$$

where $a \vee b = \max(a, b)$. Our goal is to design a method for choosing $\{\alpha_i\}_{i \in \mathcal{S}}$ which guarantees $\mathrm{FCR} \leq \delta$ for a predefined level $\delta \in [0,1]$ provided by the scientists in advance, *regardless of what the selection rule* $\mathfrak{S}$ *is, and in particular even if the rule is unknown and we only observe the selected set* $\mathcal{S} = \mathfrak{S}(\mathbf{X})$. Figure 1 illustrates the setup of this post-selection inference problem.

We consider FCR over other error metrics for two main reasons. First, constructing a "conditional CI" which provides a coverage probability guarantee conditional on the selected set, $\mathcal{S}$, requires knowing the selection *rule* $\mathfrak{S}$ beforehand — while the scientist may sometimes be willing to provide this before seeing the data (and stick to it after seeing the data), this requirement still limits the usage of any such method because a scientist may wish to explore the data informally before deciding on a set $\mathcal{S}$. Our e-BY method will overcome this limitation. Another option is to obtain a simultaneous coverage guarantee

[1] Technically these functionals could each lie in different sets Θ but this complicates notation, and we will anyway not explicitly need these sets later in the paper. Note that each θ_i need not be bijective. For example, ϑ_i could capture the median of a distribution.

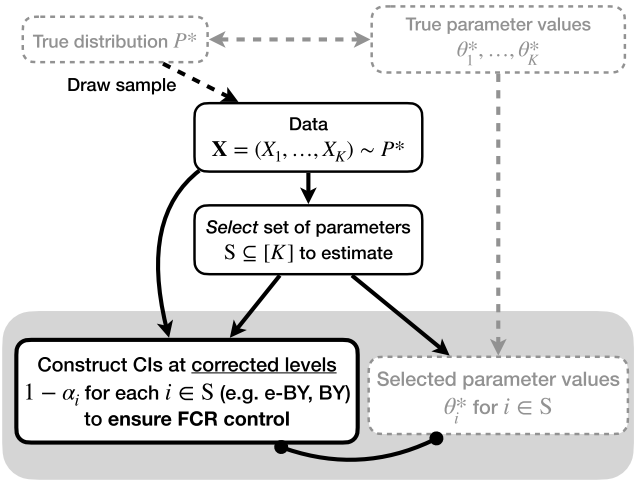

FIG 1. *The post-selection inference with confidence intervals (CI) problem. We must choose corrected confidence levels* $(1-\alpha_i)$ *for the marginal CI constructed for each* θ_i^* *that is selected. The goal is to correct for the bias introduced by the selection rule and still provide a coverage guarantee for the resulting CIs. The coverage metric that we are interested in is the expected proportion of selected parameters that are not covered by their respective CIs, that is, the false coverage rate (FCR). Our corrected confidence level must guarantee that the FCR is below some fixed level of error* $\delta \in [0, 1]$.

over all parameters, but this usually requires an overly conservative Bonferroni correction. We discuss these other metrics in more detail in Appendix F. With that said, we will focus our attention on FCR control in this work.

Our primary point of comparison is the BY procedure that was proposed in the same aforementioned paper by Benjamini and Yekutieli [4]. The BY procedure's choice of $\{\alpha_i\}_{i \in \mathcal{S}}$ and resulting guarantees depend upon assumptions (or knowledge) of the dependence structure in $\mathbf{X}$ and the selection algorithm $\mathfrak{S}$. Under certain restrictions (omitted here for brevity) on $\mathfrak{S}$, the BY procedure sets $\alpha_i = \delta|\mathcal{S}|/K$ to ensure that the FCR is controlled at level δ for mutually independent $X_1, \ldots X_K$. However, when no such assumptions can be made (i.e., under arbitrary dependence and an unknown selection rule) the BY procedure sets $\alpha_i = \delta|\mathcal{S}|/(K\ell_K)$, where $\ell_K := \sum_{i=1}^{K} i^{-1} \approx \log K$ is the Kth harmonic number. Clearly, the BY procedure produces much more conservative CIs when no assumptions can be made about dependence or selection.

In this paper, we introduce the *e-BY procedure*, which achieves the best of both worlds: it requires no restrictions on the dependence structure or selection rule and produces CIs with larger error levels $\alpha_i = \delta|\mathcal{S}|/K$ (yielding tighter CIs) and ensures FCR $\leq \delta$ *without any requirements on the dependence structure or selection rule.* However, its applicability is restricted to a smaller, but still quite broad, class of CIs. We refer to this class of intervals as "e-confidence intervals" (e-CIs) and we introduce them below.

E-confidence interval (e-CI) A crucial concept to defining an e-CI is the *e-value* [30, 24, 12, 21], a concept recently formulated in the hypothesis testing literature as an alternative to the classical p-value.

Definition 1. A nonnegative random variable E is called an *e-value* w.r.t. a set of distributions $\mathcal{Q}$, if $\sup_{\mathtt{P}\in\mathcal{Q}} \mathbb{E}_{\mathtt{P}}[E] \leq 1$.

Above, $\mathcal{Q}$ represents the null hypothesis being tested and E quantifies evidence against the null (large e-values are more evidence). As a simple example, we note that if $\mathcal{Q} = \{Q\}$ is a singleton, then the only admissible e-values are likelihood ratios of the form dR/dQ for some alternative distribution R [21].

Given a particular functional ϑ of interest, let $\mathcal{P}_\theta := \{\mathtt{P} \in \mathcal{P} : \vartheta(\mathtt{P}) = \theta\}$ denote the set of all distributions with functional value θ.

Definition 2. Let $\{E(\theta)\}_{\theta\in\Theta}$ be a family of e-values such that for each $\theta \in \Theta$, $E(\theta)$ is an e-value w.r.t. $\mathcal{P}_\theta$. For any $\alpha \in [0, 1]$, we define the $(1-\alpha)$-*e-confidence interval (e-CI)* as follows:

$$C(\alpha) = \left\{\theta \in \Theta : E(\theta) < \frac{1}{\alpha}\right\}. \tag{1}$$

Proposition 1. *Every e-CI is a valid CI.*

Proof. For any $\mathtt{P} \in \mathcal{P}_\theta$, we have $\mathbb{P}_{\mathtt{P}}(\theta \notin C(\alpha)) = \mathbb{P}_{\mathtt{P}}(E(\theta) \geq 1/\alpha) \leq \alpha$, where the last inequality follows by Markov's inequality, since $\mathbb{E}_{\mathtt{P}}[E(\theta)] \leq 1$ because $E(\theta)$ is an e-value with respect to any $P \in \mathcal{P}_\theta$. □

Our contributions Our key contributions are twofold. First, we study the aforementioned novel "e-CI" subclass of CIs. We show that a reasonably large array of existing CIs are already e-CIs. This new categorization sets the stage for our second main contribution: the e-BY procedure for FCR control (Definition 6). The e-BY procedure shows that one can attain tighter CIs than the BY procedure when no assumptions are made on the dependence structure or selection rule. The e-BY procedure is quite straightforward — it simply sets the confidence level of the CI for each selected parameter to be $1-\delta|\mathcal{S}|/K$, and this is sufficient to guarantee FCR $\leq \delta$.

On the other hand, the BY procedure *requires knowledge of all counterfactual selections one would make on different data*. The BY procedure relies on quantities concerning these counterfactual selections (that cannot be derived without knowing the selection rule beforehand), and has to pay a costly correction factor (approximately logarithmic in the number of initial parameters, K) on the confidence levels when the selection rule is unknown. The BY procedure has such limits *even when the data* $X_1, \ldots, X_K$ *are mutually independent*. Thus, BY has no guarantees when parameters are selected by the scientist in an ad hoc fashion, which is often the case (e.g., exploratory data analysis, selection based on domain knowledge). In addition, the BY procedure — even in optimal conditions of having independent data and a known selection rule — cannot produce confidence levels that are tighter than those produced by e-BY. As a

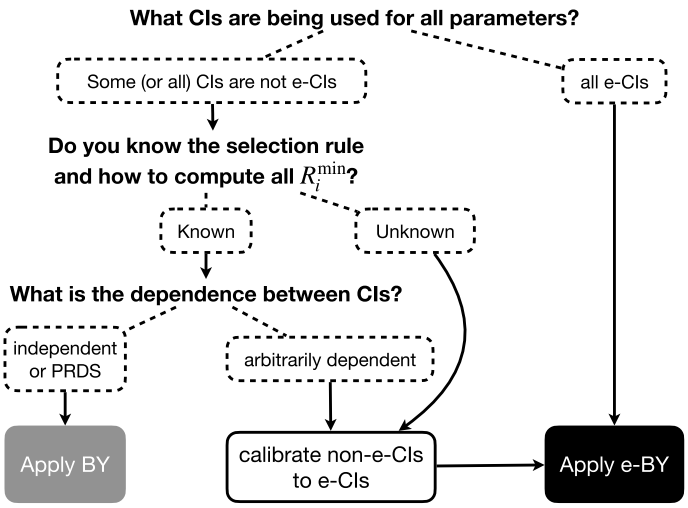

FIG 2. *A flowchart for deciding when to use e-BY vs. BY. The only case where the BY procedure should be employed is when strong assumptions are made upon both the selection rule and the type of data dependence, while the e-BY procedure is a uniform improvement over the BY procedure in all other cases. If e-CIs can only be constructed for some parameters, and standard CIs for others, we can calibrate standard CIs to e-CIs. In fact, calibration reduces the BY procedure to a special case of e-BY (Section 2.3).*

result, the e-BY procedure dominates the BY procedure when e-CIs are available or when one cannot make assumptions about either the selection rule or dependence structure. We elaborate on these points below:

D1. *Selection rule must be known for the BY procedure to yield the tightest possible confidence intervals.* Under independence (and a more general condition of PRDS — see Benjamini and Yekutieli [3, Section 1.3]), the BY procedure guarantees FCR control when the ith CI has a confidence level of $1-\delta|\mathcal{S}|/K$ only when the selection rule satisfies certain properties. When the selection rule is not guaranteed to satisfy such properties, the BY procedure can only provide FCR control level δ by yielding CIs at a confidence level of $1-\delta|\mathcal{S}|/(K\ell_K)$. On the contrary, the e-BY procedure can provide the same level of FCR control by producing CIs at a confidence level of $1-\delta|\mathcal{S}|/K$ i.e. the smallest level that can be achieved by the BY procedure even when the data is independent and the selection rule satisfies certain conditions.

D2. *CIs produced by the BY procedure are larger when the CIs are arbitrarily dependent.* Even if the selection rule $\mathfrak{S}$ is known, the BY procedure produces more conservative intervals than the e-BY procedure when the CIs for each parameter, $(C_1,\ldots,C_K)$, can be arbitrarily dependent. The BY procedure must correct the confidence levels to be $1-\delta|\mathcal{S}|/(K\ell_K)$ while the e-BY procedure maintains the same confidence levels of $1-\delta|\mathcal{S}|/K$. Arbitrarily dependent data can arise when we make multiple measure-

ments of a single unit of experimentation, for example, measuring different gene expression levels in a single cell. Dependence is also prevalent in settings where data are collected sequentially — we explore this usage further in Section 7, where we analyze both procedures using real data collected by an information technology company (Twitter) for the purposes of product testing.

D3. *The e-BY procedure applies to e-CIs.* Of course, the e-BY procedure is beneficial solely when e-CIs can be constructed for the parameters of interest. Although "e-CI" is a new term, many existing CIs are e-CIs. These include non-asymptotic CIs based on Chernoff-type inequalities such as Hoeffding (Appendix E), Bernstein, and empirical Bernstein CIs [14, 5, 19], universal inference CIs [34] (Section 2.1), and CIs formed by stopping confidence sequences [22, 15, 16, 35, 33] (Section 2.2). Thus, finding powerful e-CIs is not a major limitation in many applications. Further, we will demonstrate in Section 2.3 that we can *always* construct an e-CI from a CI, and one can use specific calibrators to recover the BY procedure (Section 3.2). Hence, if one has both CIs and e-CIs, calibrating the CIs to e-CIs and applying e-BY dominates the BY procedure.

We summarize the tradeoffs between the e-BY and BY procedures in Table 1.

TABLE 1

Tradeoffs between the e-BY procedure in this paper and the BY procedure [4]. The confidence levels of the e-BY procedure are not penalized for certain selection rules or for different types of dependencies between CIs. However, the e-BY procedure only being applicable to e-CIs, a special class of CIs derived from e-values [30, 24, 12, 21].

Procedure	Type of CI	Selection rule	Dependence
BY	general	penalized for unknown $\mathcal{S}$ (D1)	penalized (D2)
e-BY	e-CIs (D3)	no penalty	no penalty

Outline Since the e-BY procedure applies only to e-CIs, we describe several useful examples of e-CIs in Section 2. In Section 3, we prove that e-BY controls FCR under any selection rule and dependence structure, and contrast it with the BY procedure. To complement our validity results, we also show that the e-BY procedure is unimprovable in some notions — Section 4 shows that the e-BY procedure has sharp control of the FCR and Section 5 shows that e-BY is admissible among a general class of e-CI reporting procedures. Lastly, we compare the empirical behavior of the e-BY and BY procedures on simulations in Section 6 and real user metric data from A/B testing experiments at Twitter in Section 7.

2. Confidence intervals obtained via e-values (e-CIs)

E-CIs are are referred to "warranty sets" by Shafer [24] in the context of game-theoretic statistics, and Vovk and Wang [31] defined a similar notion under the name of an *e-confidence region*, although their version allows for thresholds smaller than 1. Recent work has shown that e-CIs are pertinent to or arise

naturally in many settings. For example, universal inference [34] allows us to construct e-CIs for estimation of certain parameters in nonparametric and high-dimensional settings where the only requirement is having access to the likelihood functions for each null distribution. For sequential testing and estimation problems, e-processes, the sequential counterpart of e-values, play a central role in constructing time-uniform versions of CIs, known as confidence sequences [22, 16, 35, 33]. Further, many classical statistical tools already construct e-CIs, since they are implicitly using e-values or e-processes under the hood e.g. likelihood ratio tests and Chernoff methods [15]. Hence, limiting ourselves to e-CIs does not restrict the regime we may perform efficient inference in. In some cases, e-CIs might be the best tool we have for the job. For example, when estimating the mean of bounded random variables, the e-CIs in Waudby-Smith and Ramdas [35] are empirically among the tightest known CIs for sequential estimation of the mean; see also [20]. On the other hand, when we are in a setting where e-CIs are not the default choice of CI, one can *calibrate* any CI into an e-CI. In fact, we can demonstrate that the BY procedure is a special case of the e-BY procedure through calibration. We elaborate on a couple of the aforementioned examples below.

2.1. E-CIs from universal inference

A particularly useful method for constructing e-CIs is through universal inference [34]. Universal inference is a generalization of the standard likelihood ratio test that allows for testing against composite null hypotheses. Hence, e-CIs formed through universal inference only require access to likelihood functions (or a function that upper bounds the likelihood functions) for each possible distribution.

The simplest form of the universal inference e-CI can be derived from the split likelihood ratio test of Wasserman, Ramdas and Balakrishnan [34]. Assume that we are given $A_1, \dots, A_{2n}$ i.i.d. samples. We first split this sample into two datasets D_0 and D_1 — for simplicity we can assume that the two datasets are of equal sizes of n samples, although they do not have to be. Let $\mathtt{P}_1 \in \mathcal{P}$ be some distribution that we choose based solely on D_1. In essence, $\mathtt{P}_1$ is a guess of the "most likely alternative" and chosen to make the resulting e-value as large as possible. Consequently, the more accurately $\mathtt{P}_1$ models the true distribution, the tighter the resulting e-CI will be. Let $\mathcal{L}_0(\mathtt{P})$ denote the likelihood of D_0 under a distribution, $\mathtt{P}$. We define the universal inference e-value as follows:

$$E_{\mathrm{UI}}(\theta) \coloneqq \frac{\mathcal{L}_0(\mathtt{P}_1)}{\operatorname{argmax}_{\mathtt{P}_0 \in \mathcal{P}_\theta} \mathcal{L}_0(\mathtt{P}_0)}. \tag{2}$$

Proposition 2 (Split universal inference e-CI [34])**.** *For any parameter $\theta \in \Theta$, $E_{\mathrm{UI}}(\theta)$ is an e-value w.r.t. to θ for any choice of procedure that derives $\mathtt{P}_1$ from D_1. Consequently, the following set is a valid e-CI associated with E_{UI}:*

$$C_{\mathrm{UI}}(\alpha) \coloneqq \left\{ \theta : \frac{\mathcal{L}_0(\mathtt{P}_1)}{\operatorname{argmax}_{\mathtt{P}_0 \in \mathcal{P}_\theta} \mathcal{L}_0(\mathtt{P}_0)} < \frac{1}{\alpha} \right\}.$$

As a result, universal inference ensures that a nontrivial e-CI exists in any situation where the likelihood (or an upper bound on it) is known. In many settings, universal inference remains the only method for deriving nontrivial tests and CIs e.g. estimating the number of mixtures in a Gaussian mixture model when the dimension is greater than 1.

2.2. E-CIs from stopped confidence sequences

In the sequential setting, we assume that we receive samples of data, $A_1, A_2, \ldots$, in a stream. The goal is to produce a CI that is valid when the number of samples is data dependent, i.e., the user has the option to continuously monitor the data and continue or stop sampling based on the values observed so far. A typical sampling strategy is for a scientist to stop sampling as soon as she has evidence to reject the null hypothesis. Ramdas et al. [21] showed that any admissible sequence of CIs that is valid under adaptive sampling must be an e-CI.

Denote the sigma-algebra formed at each time step that contains all the random samples seen so far as $\mathcal{F}_t \coloneqq \sigma(\{A_i\}_{i \le t})$, with $(\mathcal{F}_t)_{t\in\mathbb{N}}$ being the corresponding filtration. A random variable $\tau \in \mathbb{N}$ is a *stopping time* if $\mathbf{I}\{\tau = t\}$ is measurable w.r.t. $\mathcal{F}_t$. This means that whether τ "stops" at time $t \in \mathbb{N}$ can only depend on $(A_1, \ldots A_t)$. Since samples arrive one at a time, we can consider a sequence of intervals $(C^t(\alpha))$, where $C^t(\alpha)$ is the CI we construct after collecting t samples. However, a sequence of $(1-\alpha)$-CIs cannot guarantee that the CI at a stopping time, $C^\tau(\alpha)$, is also a $(1-\alpha)$-CI. In fact, Howard et al. [16] provide simulations showing that the coverage probability can be drastically less than $(1-\alpha)$. Thus, we want to strengthen the definition of a standard CI by providing a coverage guarantee at stopping times, as opposed to only fixed sample sizes.

Definition 3. A $(1-\alpha)$-*confidence sequence* for a functional ϑ is a sequence of intervals $(C^t(\alpha))_{t\in\mathbb{N}}$ such that $C^\tau(\alpha)$ is an $(1-\alpha)$-CI for any stopping time τ. It also has the following equivalent definition [21, Lemma 2]: for any $\mathsf{P} \in \mathcal{P}$ and $\alpha \in [0,1]$, we have

$$\mathbb{P}_{\mathsf{P}}(\vartheta(\mathsf{P}) \in C^t(\alpha) \text{ for all } t \in \mathbb{N}) \ge 1-\alpha.$$

The formulation of a CS in the above definitions emphasizes the *time-uniform* coverage a CS provides — a CS ensures the probability θ^* is in the CS at every time step is high. One way to construct such an object is using a family of *e-processes*, which are the sequential versions of the e-values.

Definition 4. Let the index set $\mathbb{I}$ be $\mathbb{N}$ or $(0,\infty)$. An *e-process* w.r.t. to some filtration $(\mathcal{F}_t)_{t\in\mathbb{I}}$ and set of distributions $\mathcal{Q}$ is defined as a sequence of random variables $(E^t)_{t\in\mathbb{I}}$ which are all nonnegative under any $\mathsf{P} \in \mathcal{Q}$ and satisfy $\sup_{\mathsf{P}\in\mathcal{Q}} \mathbb{E}_{\mathsf{P}}[E^\tau] \le 1$, i.e., E^τ is an e-value w.r.t. $\mathcal{Q}$, for all stopping times τ w.r.t. $(\mathcal{F}_t)$.

For recent literature relevant to e-processes, see [12, 21, 15, 16]. E-processes are a superset of nonnegative supermartingales (one can see nonnegative supermartingales are e-processes as a result of the optional stopping theorem).

Thus, a sequence of intervals constructed through applying the formula in (1) to a family of e-processes $(E^t(\theta))_{t\in\mathbb{N}}$ is a CS. For example, the following is a Hoeffding-esque e-process for testing the mean of a bounded random variable [14, 35]:

$$E^t_{\text{Hoef}}(\theta) \coloneqq \exp\left(\sum_{i=1}^{t} \lambda_i(A_i - \theta) - \frac{\lambda_i^2}{8}\right),$$

where (λ_t) is predictable w.r.t. $(\mathcal{F}_t)$, i.e., λ_t is measurable w.r.t. $\mathcal{F}_{t-1}$ for all $t \in \mathbb{N}$. We can define a two-sided e-CI based on $(E^t_{\text{Hoef}}(\theta))$ as follows:

$$C^t_{\text{Hoef}}(\alpha) \coloneqq \left(\frac{\sum_{i=1}^{t} \lambda_i A_i}{\sum_{i=1}^{t} \lambda_i} \pm \frac{\log(2/\alpha) + \frac{1}{8}\sum_{i=1}^{t} \lambda_i^2}{\sum_{i=1}^{t} \lambda_i}\right).$$

Notice that setting $\lambda_t = \sqrt{\log(2/\alpha)/n}$ recovers the classic Hoeffding CI [14] for the specific sample size of $n \in \mathbb{N}$ and fixed $\alpha \in [0, 1]$. We discuss how to optimize these parameters in the post-selection inference setting where there are multiple potential values of α in Appendix E.

Another implication of the above formulation is that the e-process view allows one to extend the coverage guarantee of existing CIs to be valid sequentially. This coverage improvement is not only applicable to Hoeffding's, but also a wide range of Chernoff type inequalities [15, 16, 33]. Similarly, the universal inference e-CI formulated in (2) can also be extended to a sequential form, where P_1 may be recomputed at each time step based on the previously observed samples. For bounded random variables, one can construct e-processes that directly correspond to the wealth process of a sequence of fair gambles [35]. Since e-CIs are necessary for sequential inference, the e-BY procedure is a strict improvement over the BY procedure in settings that allow for adaptive sampling and stopping such as in multi-armed bandits and the A/B testing setting we discuss in Section 7.

2.3. E-CIs constructed from CIs through calibration

Although the definition of an e-CI is much more specific than a general CI, we will present a method for calibrating arbitrary marginal CIs to e-CIs using a method for p-value to e-value calibration [29, 23, 26, 25, 30].

Definition 5. A *calibrator* is a nonincreasing, upper semicontinuous function $f : [0, 1] \to [0, \infty]$ where $\int_0^1 f(x)\, dx \le 1$.

Define f^{-1} to be the dual of the calibrator f, i.e., $f^{-1}(x) \coloneqq \sup\{p : f(p) \ge x\}$. When f is invertible, f^{-1} is the inverse of f. Using f^{-1}, we can convert any CI to an e-CI. Also we define two properties about any set-valued CI function $C : [0, 1] \mapsto 2^{\Theta}$. Let C be *nonincreasing* if, for any $\alpha, \beta \in [0, 1]$, $\alpha \le \beta$ implies

that $C(\alpha) \supseteq C(\beta)$. Further, define *continuous from below* to be the property that for any $\alpha \in [0,1]$, $C(\alpha) = \bigcup_{\beta>\alpha} C(\beta)$.

Theorem 1. *Let $C : [0,1] \mapsto 2^{\Theta}$ be a nonincreasing function that is continuous from below such that $C(\alpha)$ produces a $(1-\alpha)$-CI, and f be a calibrator (with dual calibrator f^{-1}). Then, the following set is a $(1-\alpha)$-e-CI:*

$$C^{\text{cal}}(\alpha) = C\left(f^{-1}\left(\frac{1}{\alpha}\right)\right).$$

We provide a proof of Theorem 1 in Appendix A.1. As a consequence of this theorem, we can construct a nontrivial e-CI from any nontrivial CI and a dual calibrator. Of course, this prompts the question: when should one use e-CIs with the e-BY procedure instead of CIs with the BY procedure? Generally speaking, the calibrated e-CI will be looser than the original CI, since $f(x) = 1/x$ is not a calibrator. However, we show in Section 3 that applying e-BY to calibrated e-CIs is actually equivalent to applying the BY procedure to the original CIs when the selection rule is unknown or the original CIs are dependent. Thus, calibrating CIs to e-CIs creates e-CIs across all parameters, in the case where some parameters initially do not have e-CIs.

3. The e-BY procedure

Now, we formally define the e-BY procedure as follows.

Definition 6. The *e-BY procedure* at level $\delta \in [0,1]$ sets $\alpha_i = \delta|\mathcal{S}|/K$ for each $i \in \mathcal{S}$.

We show that a FCR bound can be proven quite simply given the fact that e-CIs are constructed for each selected parameter.

Theorem 2. *Let $C_i(\alpha)$ be a $(1-\alpha)$-e-CI for each $i \in [K]$ and $\alpha \in [0,1]$. Then, the e-BY procedure in Definition 6 ensures* FCR $\leq \delta$ *for any $\delta \in (0,1)$ under any dependence structure between $X_1, \ldots, X_K$, and any selection rule $\mathfrak{S}$.*

Proof. We directly show an upper bound for the FCR as follows:

$$\begin{aligned}
\text{FCR} &= \mathbb{E}\left[\frac{\sum_{i\in\mathcal{S}} \mathbf{I}\{\theta_i^* \notin C_i(\delta|\mathcal{S}|/K)\}}{|\mathcal{S}| \vee 1}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i\in[K]} \mathbf{I}\{E_i(\theta_i^*)|\mathcal{S}|\delta/K > 1\} \cdot \mathbf{I}\{i \in S\}}{|\mathcal{S}| \vee 1}\right] \\
&\leq \sum_{i\in[K]} \mathbb{E}\left[\frac{E_i(\theta_i^*)|\mathcal{S}|\delta}{K(|\mathcal{S}| \vee 1)}\right] = \sum_{i\in[K]} \frac{\delta}{K}\mathbb{E}\left[E_i(\theta_i^*) \cdot \frac{|\mathcal{S}|}{|\mathcal{S}| \vee 1}\right] \leq \delta,
\end{aligned}$$

where $\mathbf{I}\{\cdot\}$ is the indicator function. The first inequality is because $\mathbf{I}\{x > 1\} \leq x$ for all $x \geq 0$. The second inequality is a result of the definition of the e-value

for θ_i^* having its expectation under P^* be upper bounded by 1. This achieves our desired bound. □

Remark 1. FCR control of the e-BY procedure implies FDR control of the e-BH procedure [32], while the converse is not true. The arbitrary nature of $\mathcal{S}$ is unique to the post-selection inference problem (as e-BH can only reject "self-consistent" [6] sets of hypotheses), and defining the concept of e-CIs is key to achieving FCR control. We expand on the relationship between these procedures in Appendix B and visualize their relationships in Figure 6. We also introduce a directional variant of e-BH that uses the FCR control of the e-BY procedure to achieve directional FDR control in Appendix B.3.

False coverage control for data-dependent δ An interesting result of this proof is that we retain a form of control on the FCP even if δ is chosen data dependently. Define $\mathrm{FCP}(\delta')$ to be the FCP of a post-selection inference procedure at level δ'.

Corollary 1. *The following is always true for the e-BY procedure (regardless of selection rule or dependence in* $\mathbf{X}$*):*

$$\mathbb{E}\left[\sup_{\delta' \in (0,1]} \frac{\mathrm{FCP}(\delta')}{\delta'}\right] \leq 1. \tag{3}$$

Here, $\mathrm{FCP}(\delta')$ *is FCP of the e-BY procedure at level* δ'*, i.e.,* $\alpha_i = \delta'|\mathcal{S}|/K$ *for all* $i \in \mathcal{S}$*.*

Satisfying (3) is a stronger condition than FCR control, since (3) implies FCR control for any fixed $\delta \in [0,1]$. This error guarantee follows from an argument similar to the proof of Theorem 2. When testing K hypothesis to control the false discovery rate, we note that a similar result for the false discovery proportion (FDP) of the e-BH procedure is implied by this corollary:

$$\mathbb{E}\left[\sup_{\delta' \in (0,1]} \frac{\mathrm{FDP}(\delta')}{\delta'}\right] \leq 1.$$

This can also be observed directly from the proof of e-BH and we include the details in Appendix B.2 for completeness. When specialized to a single hypothesis test, the latter observation amounts to saying that for any e-value E, we have

$$\mathbb{E}\left[\sup_{\delta' \in (0,1]} \frac{\mathbf{I}\{E \geq 1/\delta'\}}{\delta'}\right] \leq 1,$$

which is captured in [32, Lemma 1]. One can view the post-hoc type-I safety discussed by Grünwald [11] and post-hoc validity of Koning [18] as a rephrasing of this latter result.

3.1. Comparison with the BY procedure

The FCR guarantee for e-BY in Theorem 2 solely relies on the definition of an e-CI. Since the expectation property of e-values holds regardless of the type of dependence, proving FCR control of e-BY does not require fine grained analysis of how the distribution of an e-value changes when conditioned on other e-values. On the other hand, the BY procedure does require more detailed analysis because its proof only uses the marginal coverage property of each CI. Hence, the BY algorithm requires different levels of corrections based on the dependence structure and selection rule.

Definition 7. The *BY procedure* at level $\delta \in [0,1]$ sets $\alpha_i = \delta R_i^{\min}/K$ if $X_1, \dots X_K$ are independent (or PRDS [4]), $R_i^{\min}$, as formulated in (4), is known, and $R_i^{\min} > |\mathcal{S}|/\ell_K$. Otherwise, it sets $\alpha_i = \delta|\mathcal{S}|/(K\ell_K)$.

Fact 1 (Theorems 1 and 4 from BY [4])**.** *The BY procedure ensures that* FCR $\leq \delta$.

A key difference in the tightness of CIs between the e-BY and BY procedures is how the selection rule affects the confidence levels chosen for the CIs. In the e-BY procedure, the confidence levels are proportional to $|\mathcal{S}|$, the number of selected parameters. On the other hand, the BY procedure requires knowledge of $R_i^{\min}$, which is formulated as follows for each $i \in \mathcal{S}$:

$$R_i^{\min} := \min \{|\mathfrak{S}(\mathbf{X}_{i\to x_i})| : x_i \in \mathcal{X}_i,\ i \in \mathfrak{S}(\mathbf{X}_{i\to x_i})\}, \tag{4}$$

where $\mathbf{X}_{i\to x_i} \in \mathcal{X}$ is the data $\mathbf{X}$ with the ith component set to x_i.

Computing $R_i^{\min}$ requires knowledge and analysis of the selection rule Clearly, $R_i^{\min}$ is a quantity that has to be calculated from the selection rule, $\mathfrak{S}$. In practice, however, $R_i^{\min}$ may be impossible to calculate. The selection rule may not be fully specified or the counterfactual behavior of the selection rule may be difficult to analyze. Post-selection inference is often used for *exploratory data analysis* in the sense described by Goeman and Solari [10]. Hence, the selection rule is often *not* known by the scientist ahead of time — she is developing her selection plan on the fly as she explores the data! Another common scenario is that scientist carrying out the experiment, performing analysis, and ultimately producing the CIs may not be the one deciding upon the selection to report e.g. the scientist's manager makes the decision. In either case, the selection rule is not known. As an aside, the e-BY procedure can also be used by the manager in the prior example to produce CIs herself, since she does not need to know the experimental design and dependencies in the data to construct CIs using the e-BY procedure. Further, even if the selection rule was known beforehand, it still might be difficult to analyze $R_i^{\min}$. For example, the selection procedure could be iterative: at each iteration, the scientist considers a selection set, observes some property of the CIs produced by the e-BY procedure for the candidate selection set and then selects a new candidate selection set based on her observations. Eventually, the scientist may finalize the selection set based

on some sort of convergence criteria. In this selection scheme, the full procedure is not specified, and quite difficult to specify since it requires knowing what the scientist would do at every iteration if experimental data were different. Thus, it is often impractical to assume $R_i^{\min}$ is known.

The other main advantage of the e-BY procedure can be seen when the dependence between CIs is arbitrary — there is a clear gap between the confidence levels for the BY procedure and the e-BY procedure. The BY procedure requires an extra correction of $\ell_K \approx \log K$ on the error probability under arbitrary dependence, while the e-BY procedure behaves the same way under any form of dependence. Thus, if the CIs are e-CIs, the e-BY procedure yields tighter CIs than the BY procedure when the data is arbitrarily dependent or the selection rule is unknown or has $R_i^{\min} \leq |\mathcal{S}|/\ell_K$.

3.2. BY is special case of e-BY through e-CI calibration

By using a specific calibrator with the method for calibrating CIs into e-CIs discussed in Section 2.3, we can show that the BY procedure (with no assumptions on dependence or selection rule) is a special case of the e-BY procedure. Define the following calibrator when there are K hypotheses and the desired FCR control is δ:

$$f^{\mathrm{BY}(\alpha,K)}(x) \coloneqq \frac{K}{\alpha}\left(\left\lceil \frac{x}{\alpha/(K\ell_K)} \right\rceil \vee 1\right)^{-1} \mathbf{I}\left\{x \leq \alpha/\ell_K\right\}.$$

Proposition 3. *$f^{\mathrm{BY}(\delta,K)}$ is an upper semicontinuous calibrator for all $\delta \in (0,1)$ and $K \in \mathbb{N}$.*

Proof. Clearly, $f^{\mathrm{BY}(\delta,K)}$ is nonincreasing, and we note that:

$$\int_0^1 f^{\mathrm{BY}(\delta,K)}(x)\, dx = \sum_{j=1}^{K} \frac{\delta}{K\ell_K} \cdot \frac{K}{j\delta} = \frac{1}{\ell_K}\sum_{j=1}^{K}\frac{1}{j} = 1.$$

Hence, we proved that $f^{\mathrm{BY}(\delta,K)}$ satisfies the properties to be a calibrator. □

By our result in Theorem 1, we can now show that the e-BY procedure with $f^{\mathrm{BY}(\delta,K)}$ calibrated e-CIs is identical to the BY procedure under arbitrary dependence.

Corollary 2. *For any arbitrary CI constructors, denoted as $C_i : [0,1] \mapsto 2^{\Theta_i}$ for each $i \in [K]$, that are continuous from below. The BY procedure produces the same CIs as the e-BY procedure with e-CI constructors C_i^{cal} that were calibrated through $f^{\mathrm{BY}(\delta,K)}$ for each $i \in [K]$.*

Note that one take the limit from below at every $\alpha \in [0,1]$ of a nonincreasing CI constructor C to create a CI constructor C' that is continuous from below. C' has valid coverage and is no wider than the original C for every α, so the continuity from below is only a technical requirement.

The above corollary follows from the fact that Theorem 1 and the choice of $f^{\mathrm{BY}(\delta,K)}$ as calibrator implies the following:

$$C_i\left(\frac{\delta j}{K\ell_K}\right) = C_i^{\mathrm{cal}}\left(f^{\mathrm{BY}(\delta,K)^{-1}}\left(\frac{K\ell_K}{\delta j}\right)\right) = C_i^{\mathrm{cal}}\left(\frac{\delta j}{K}\right) \text{ for all } j \in [K].$$

Thus, the output of the BY procedure is equivalent to the output of the e-BY procedure with calibrated e-CIs for all possible sizes of the selection set $\mathcal{S}$. As a result, in a situation where there may be a mix of e-CIs and regular CIs that are arbitrarily dependent and the regular CIs, one can always calibrate the regular CIs to e-CIs using $f^{\mathrm{BY}(\delta,K)}$ and then apply the e-BY procedure to produce CIs that are as tight as directly applying the BY procedure.

Remark 2. We have not compared other calibrators, mostly because there is no uniformly best choice of calibrator to use: different ones will perform better in different settings. We recommend using the BY calibrator by default. The BY calibrator is a natural choice since it allows the resulting output of the e-BY procedure on calibrated CI to be no worse than the BY procedure (under dependence or unknown selection rule). It is also exact, in the sense for the resulting e-CI, the e-value corresponding to the true value of the parameter, $E(\theta^*)$, will have an expectation exactly equal to 1 if the original CI has exact coverage.

We have not defined admissibility for CI calibrators, but we conjecture that for some suitable definition, the BY calibrator we introduce is an admissible CI calibrator of CIs into e-CIs (since the corresponding p-to-e value calibrator is admissible). Furthermore, it is well suited to use with the e-BY procedure, since e-BY only produces CIs with error levels at multiples of δ/K — we also conjecture that it is admissible among CI calibrators w.r.t. to the tightness of CIs produced by e-BY. We think it is likely that the BY calibrator is the only calibrator that allows one to produce CIs that were never worse than BY, making it a good default choice.

4. FCR control of e-BY is sharp

To complement our upper bound on the FCR, we demonstrate the FCR bound is indeed sharp for the e-BY procedure by formulating a setting where the FCR is arbitrarily close to δ. First, let $\mathbb{R}^-$ and $\mathbb{R}_0^+$ denote the negative and nonnegative reals, respectively. Now, let

$$Z_i^t(\theta) = 1 + W_i^t - \theta t \text{ for each } \theta \in \mathbb{R}^-,\ t \in \mathbb{R}_0^+, \text{ and } i \in [K],$$

where $(W_i^t)_{t\in\mathbb{R}_0^+}$ is a Brownian motion with drift θ_i^* (unknown to the scientist) and volatility 1. Define $(E_i^t(\theta))_{t\in\mathbb{R}_0^+}$ as the process obtained by stopping $(Z_i^t)_t$ when it hits 0:

$$E_i^t(\theta) = Z_i^t(\theta)\ \mathbf{I}\left\{\inf_{s\in[0,t]} Z_i^s(\theta) > 0\right\}.$$

Assume that (W_i^t) are independent across $i \in [K]$. Clearly, $(E_i^t(\theta_i^*))$ is a nonnegative martingale (and an e-process) for θ_i^*, since a Brownian motion is

a martingale, and stopped martingales are martingales as well. We imagine the scientist to be interested in estimating parameters whose true value θ_i^* is positive, and she uses the following experimentation setup and selection rule to do so. Fix $\gamma \in [1, 1/\delta)$ as the threshold for the processes $(E_i^t(0))$ to be seen as "interesting" (and chosen for further experimentation) but not yet conclusive.

We consider a two-step procedure. In the first step, we stop the i-th experiment at time

$$\tau_i := \inf\{t \geq 0 : W_i^t \geq \gamma - 1 \text{ or } W_i^t \leq -1\},$$

and we select S such that $i \in S$ if and only if $W_i^{\tau_i}(0) \geq \gamma - 1$ (i.e., $E_i^{\tau_i}(0) \geq \gamma$). That is, we select i such that $W_i^t + 1 = E_i^t(0)$ reaches γ (interesting) before it reaches 0 (discarded). In other words, the set $\mathcal{S}$ contains all interesting experiments in the first screening. If $\gamma = 1$ then we simply select all experiments. Let $\beta = K/(\delta|\mathcal{S}|) > \gamma$. Note that β is a function of $\mathcal{S}$. (Here the index t of W_i^t does not necessarily represent time, and it does not have to synchronize across $i \in [K]$. The scientist can finish all step 1 experiments before moving on with the selected ones for step 2.)

In the second step, define another stopping time for the i-th experiment:

$$\eta_i = \inf\{t \geq 0 : W_i^t \geq \beta - 1 \text{ or } W_i^t \leq -1\}.$$

That is, we will stop the experiment if either $E_i^t(0)$ reaches a very high level β or it is discarded. Note that $\eta_i > \tau_i$ for $i \in \mathcal{S}$ and $\eta_i = \tau_i$ for $i \notin \mathcal{S}$, meaning that we only continue those experiments that were deemed as interesting in the first step. As a result, $\{E_i^{\eta_i}(\theta)\}_{\theta \in \Theta}$ forms a family of e-values for each $i \in \mathcal{S}$, and we can define the following one-sided e-CIs:

$$\begin{aligned} C_i(\alpha) &:= \left\{\theta \in \mathbb{R} : E_i^{\eta_i}(\theta) < \frac{1}{\alpha}\right\} \\ &= \left\{\theta \in \mathbb{R} : 1 + W_i^{\eta_i} - \theta\eta_i < \frac{1}{\alpha} \text{ or } \inf_{s \in [0,\eta_i]} 1 + W_i^s - \theta s \leq 0\right\} \\ &= \left(\min\left(\frac{1 + W_i^{\eta_i} - \frac{1}{\alpha}}{\eta_i}, \inf_{s \in [0,\eta_i]} \frac{1 + W_i^s}{s}\right), \infty\right). \end{aligned} \tag{5}$$

Theorem 3. *Let e-CIs be specified by (5). Then,* $\lim_{\boldsymbol{\theta}^* \uparrow \mathbf{0}} \mathrm{FCR} = \delta$, *i.e., limit of FCR for the e-BY procedure approaches* δ *as* $\boldsymbol{\theta}^*$ *approaches* $\mathbf{0}$ *(all zero vector) from below.*

A full proof of this theorem can be found in Appendix A.2.

Remark 3. The other source of looseness of the FCR bound in Theorem 2 is in the relationship between selection event of each parameter and the e-value of the true parameter, i.e., $E_i(\theta_i^*)$. In this example, we can see that the selection rule and the true value of θ^* are conflicting in some sense. The scientist wants to select parameters with positive values, so when the parameters are actually all negative values, the CI for a selected parameter will probably not cover the corresponding true value θ_i^*. We discuss this further in Appendix D and observe

some small improvements we may make on the e-BY procedure under stronger assumptions about the e-CIs.

Remark 4. While the Brownian motion data generating process in our setup is stylized, it does reflect some realistic aspects of sequential data collection. The early stopping rule that halts as soon as an e-CI shows significance, and the selection of promising experiments for further testing, are both natural strategies for sequential experimentation (e.g., two stage designs, follow up studies). Hence, depending on the relationship between selection rule and the true value of the parameters (as illustrated in our example), it is plausible that the true FCR could be close to the FCR bound.

On the other hand, the design of the parameters in this example is highly adversarial: $\boldsymbol{\theta}^*$ needs to be very close to 0 from below to approach this upper FCR bound. Therefore, in practice, we do not always expect the FCR of e-BY to be close to its upper bound, but the bound is still unimprovable in general.

5. Admissibility of the e-BY procedure

Not only does the e-BY procedure provide sharp error guarantees, it is also admissible w.r.t. a general class of e-CI post-selection inference methods. We will formally define the universe of *CI reporting procedures* that e-BY belongs to, i.e., procedures which report a CI for the true value of each selected parameter. We will show that, among CI reporting procedures that control the FCR when given e-CIs, there exists no other procedure that produces strictly tighter CIs on selected parameters than e-BY. To prove this, we define a notion of dominance for FCR controlling CI reporting procedures, and prove that e-BY is indeed admissible among this set of procedures.

Our notion of admissibility essentially requires that there is no FCR controlling procedure that produces narrower CIs on all the selected parameters uniformly over all possible inputs of e-CIs (by narrower, we mean at least as narrow on all instances, but strictly narrower on at least one instance). To develop our formalisms, we will first describe a notion of an e-CI constructor, i.e., a function that maps an error level α to its corresponding $(1-\alpha)$-e-CI.

5.1. Key properties of all e-CIs

Recall that the $(1-\alpha)$-e-CI *associated* with a family of e-values $\{E(\theta)\}_{\theta\in\Theta}$ is given by

$$C(\alpha) = \left\{\theta \in \Theta : E(\theta) < \frac{1}{\alpha}\right\}, \quad \text{for each } \alpha \in [0,1], \tag{6}$$

where the convention $1/0 = \infty$ is used. Note that it is important to formulate the e-CI by using the strict inequality $E(\theta) < 1/\alpha$ instead of $E(\theta) \leq 1/\alpha$. Otherwise all admissibility arguments fail, since one could always improve from $\{\theta \in \Theta : E(\theta) \leq 1/\alpha\}$ to $\{\theta \in \Theta : E(\theta) < 1/\alpha\}$, as smaller CIs are statistically stronger. We define an e-CI constructor as the collection of e-CIs for each error level $\alpha \in [0,1]$.

Definition 8. Let $\mathrm{E} \coloneqq \{E(\theta)\}_{\theta\in\Theta}$ be a family of e-values for a parameter that takes values in Θ. The corresponding *e-CI constructor* is the function $C^{\mathrm{E}} : [0,1] \to 2^{\Theta}$ which is defined as:

$$C^{\mathrm{E}}(\alpha) = \left\{\theta \in \Theta : E(\theta) < \frac{1}{\alpha}\right\}, \quad \text{for each } \alpha \in [0,1].$$

It is important to note that C^{E} we define above is a *random function*. We have not separated designations for random variables or functions and their realizations in this paper before, but we emphasize their distinction in this section. This distinction allows us to define the domain of realizations of e-CI, which in turn, is used to the notion of a CI reporting procedure.

Denote by the set $\mathrm{ECI}(\Theta)$ the set of possible realizations of e-CI constructors for the parameter space Θ, that is, $C \in \mathrm{ECI}(\Theta)$ iff C is a possible realization of an e-CI constructor C^{E} for some family of e-values $\mathrm{E} = \{E(\theta)\}_\theta$ via (6) under some probability $\mathsf{P} \in \mathcal{P}$. We abuse vocabulary in this section and refer to the e-CI constructor, C, as an "e-CI" as well. $\mathrm{ECI}(\Theta)$ captures the domain of allowed input of realized e-CIs, similar to $[0,1]^K$ for input p-values and $[0,\infty]^K$ for input e-values in multiple testing. It is important to formally describe $\mathrm{ECI}(\Theta)$ as it will be used as part of the domain of CI reporting procedures that we discuss later.

Lemma 1. *For a set function $C : [0,1] \mapsto 2^{\Theta}$, $C \in \mathrm{ECI}(\Theta)$ if and only if C is nonincreasing and continuous from below.*

We provide the proof in Appendix A.3. We see from Lemma 1 that $\mathrm{ECI}(\Theta)$ is precisely the set functions from $[0,1]$ to 2^{Θ} which are decreasing and continuous from below; note that these properties are common for any CI. In other words, for a given CI which may be computed with other methods than e-values, we cannot exclude the possibility that it is an e-CI just by looking at the CI.

Lemma 2. *For a given $C \in \mathrm{ECI}(\Theta)$, it is realized by the family of e-values $\{E(\theta)\}_\theta$ if and only if $E(\theta)$ takes the value $E(\theta) = t(\theta)$ whenever $E(\theta) \geq 1$ for all $\theta \in \Theta$, where t is given in (10).*

This is a direct consequence of (6) and (11). Due to Lemma 2, for any family of e-values for Θ and $\theta \in \Theta$, the probability of realizing $C \in \mathrm{ECI}(\Theta)$ is at most $1/t(\theta)$ for t given in (10).

5.2. CI reporting procedures

Next, fix K parameter spaces $\boldsymbol{\Theta} \coloneqq (\Theta_1, \ldots, \Theta_K)$ of interest. ECI_K denotes the set of all K-tuples of realizations of e-CIs, $\mathbf{C} = (C_i)_{i\in[K]}$, for the true values of these parameters (we omit the parameter spaces in ECI_K since they are fixed from now on). Define $\mathbf{2}^{\Theta} \coloneqq \prod_{i=1}^{K} 2^{\Theta_i}$ as the product space of the power set of each parameter's range. Using Lemma 1, ECI_K is the collection of $\mathbf{C} : [0,1] \to \mathbf{2}^{\Theta}$ whose components are nonincreasing and continuous from below. Now, we define a notion of a CI reporting procedure.

Definition 9. A *CI reporting procedure* $\Phi : [K] \times ([0,1] \to \mathbf{2}^{\Theta}) \to \mathbf{2}^{\Theta}$ takes a selected set of indices $S \subseteq [K]$ and the collection of e-CIs, $\mathbf{C} = (C_1, \ldots, C_K) \in \mathrm{ECI}_K$, as input and it outputs a vector of intervals, $\Phi(S, \mathbf{C}) \in \mathbf{2}^{\Theta}$, satisfying

(a) $\Phi(S, \mathbf{C})_i = \emptyset$ for all $i \notin S$.
(b) $\Phi(S, \mathbf{C}) = \Phi(S, \mathbf{C}')$ if $\mathbf{C}_S = \mathbf{C}'_S$ where $\mathbf{C}_S = (C_i)_{i \in S}$; that is, for a selected set of indices S, if two input vectors of CIs are identical for indices in S, then Φ does not distinguish them.

Restriction (a) is a simplifying assumption since the purpose of CI reporting procedures is to provide CIs for parameters in the selected set S. Hence, we just report the empty set for unselected parameters. The above restriction (b) does not seem to be dispensable from the proof of admissibility which we provide later (see Remark 6 in Appendix A.4). It is a reasonable assumption: all the CIs that are discarded or uninteresting should not affect how we report the selected CIs. Both the e-BY and the weighted version of the e-BY procedure (which we introduce and describe in Appendix C) satisfy this requirement.

In a statistical experiment, let $\mathrm{E}_i \coloneqq \{E_i(\theta)\}_{\theta \in \Theta_i}$ denote the family of e-values used to construct e-CIs for θ_i^*. $\mathbf{C}$ are realizations of the e-CIs corresponding to the families of e-values $\mathbf{E} \coloneqq (\mathrm{E}_i)_{i \in [K]}$ via

$$C_i^{\mathbf{E}}(\alpha) \coloneqq C^{\mathrm{E}_i}(\alpha) = \left\{ \theta \in \Theta_i : E_i(\theta) < \frac{1}{\alpha} \right\} \text{ for each } \alpha \in [0,1],\ i \in [K].$$

Let $\mathbf{C}^{\mathbf{E}} \coloneqq (C_1^{\mathbf{E}}, \ldots, C_K^{\mathbf{E}})$ denote the random vector of e-CIs that arise from $\mathbf{E}$. Recall that a CI reporting procedure only has access to the realized selected set and the e-CIs, i.e., S and $\mathbf{C}$, respectively but not the random e-CIs $\mathbf{C}^{\mathbf{E}}$ themselves. The FCR of the CI reporting procedure Φ for the selection $\mathcal{S}$, vector of e-value families $\mathbf{E}$, and true parameters $\theta_1^* \in \Theta_1, \ldots, \theta_K^* \in \Theta_K$ is given by

$$\mathrm{FCR}(\Phi) = \mathbb{E}\left[\frac{\sum_{i \in \mathcal{S}} \mathbf{I}\left\{ \theta_i^* \notin \Phi(\mathcal{S}, \mathbf{C}^{\mathbf{E}})_i \right\}}{|\mathcal{S}| \vee 1} \right],$$

where the expectation is taken under the true distribution P^*.

A CI reporting procedure Φ has FCR level $\delta \in [0,1]$ if $\mathrm{FCR}(\Phi) \leq \delta$ for any selection $\mathcal{S}$, e-value families $\mathbf{E}$, and true values of parameters $\boldsymbol{\theta}^*$.

Definition 10 (Dominance)**.** We say that the CI reporting procedure Φ *dominates* another one Φ' if $\Phi(S, \mathbf{C})_i \subseteq \Phi'(S, \mathbf{C})_i$ for each $i \in S$, i.e., Φ produces narrower CIs for all parameters in S than the CIs produced by Φ', for all $S \subseteq [K]$ and $\mathbf{C} \in \mathrm{ECI}_K$, and *strictly dominates* if further there exists $i \in S$ s.t. $\Phi(S, \mathbf{C})_i \subsetneq \Phi'(S, \mathbf{C})_i$ for some $S \subseteq [K]$ and $\mathbf{C} \in \mathrm{ECI}_K^*$ where ECI_K^* is the set of all $\mathbf{C} \in \mathrm{ECI}_K$ with strictly decreasing components.

Now, that we have a notion of dominance between e-CI reporting procedures, we can define admissibility.

Definition 11 (Admissible CI reporting procedure)**.** A CI reporting procedure with FCR level $\delta \in (0,1)$ is *admissible* if it is not strictly dominated by any CI reporting procedure with FCR level δ.

Theorem 4. *The e-BY procedure at level $\delta \in (0,1)$ is an admissible CI reporting procedure with FCR level δ.*

We defer the proof of this theorem to Appendix A.4. Further, using the same arguments as in the proof of Theorem 4, we can also show that a weighted version of the e-BY procedure, defined in Appendix C, is also admissible.

6. Numerical simulations

To understand the practical difference in the precision of the BY procedure and the e-BY procedure, we conducted simulations in two different data settings. The first is a nonparametric setting, where we only make the assumption that the data is bounded. In the second setting, we simulate the sharp FCR scenario for e-BY described in Section 4, where the data are stopped Brownian motions.

6.1. Bounded setting

We wish to estimate K different means in this setting. For each $i \in [K]$, let $X_i \in [-1,1]$ be a bounded random variable of interest, and let $\mathbf{X}_i = (X_i^1, \ldots, X_i^n)$ represent n i.i.d. draws from the distribution of X_i. Thus, the ith parameter of interest is $\theta_i^* = \mathbb{E}[X_i]$. The distribution of X_i is standard normal distribution truncated to $[-\sigma, \sigma]$ that is normalized to be supported on $[-1,1]$. We set $\sigma = 100$.

Let our desired level of FCR control be $\delta = 0.1$. The selection rule selects all parameters with p-values less than δ. Let $\widehat{\theta}_i$ denote the sample mean of $X_i^1, \ldots, X_i^n$ for each $i \in [K]$. Our p-value of choice for the ith parameter is $P_i = (2\exp(-\widehat{\theta}_i^2/(2n))) \wedge 1$. Such P_i is a two-sided p-value for bounded random variables derived from Hoeffding's inequality that tests the null hypothesis $H_i : \theta_i^* = 0$ for each $i \in [K]$.

The CI for the BY procedure, and the e-CI for the e-BY procedure as defined as follows:

$$C_i(\alpha) := \left(\widehat{\theta}_i \pm \sqrt{\frac{2\log(2/\alpha)}{n}}\right),$$

$$C_i(\alpha) := \left(\widehat{\theta}_i \pm \sqrt{\frac{2\log(2/\alpha)}{n}} \cdot \frac{\log(2/\alpha) + \log(4/\delta)}{2\sqrt{\log(2/\alpha)\log(4/\delta)}}\right).$$

These two CIs are equivalent when $\alpha = \delta/2$ — both are derived from Hoeffding's inequality. We compare three different post-selection inference methods: the e-BY procedure, the BY procedure with both an independence and $R_i^{\min} = |\mathcal{S}|$ assumption (which is satisfied in this setting), and the BY procedure without any assumptions. We refer to the BY procedure with the assumptions as "BY-wA", and the BY procedure without assumptions as "BY-woA".

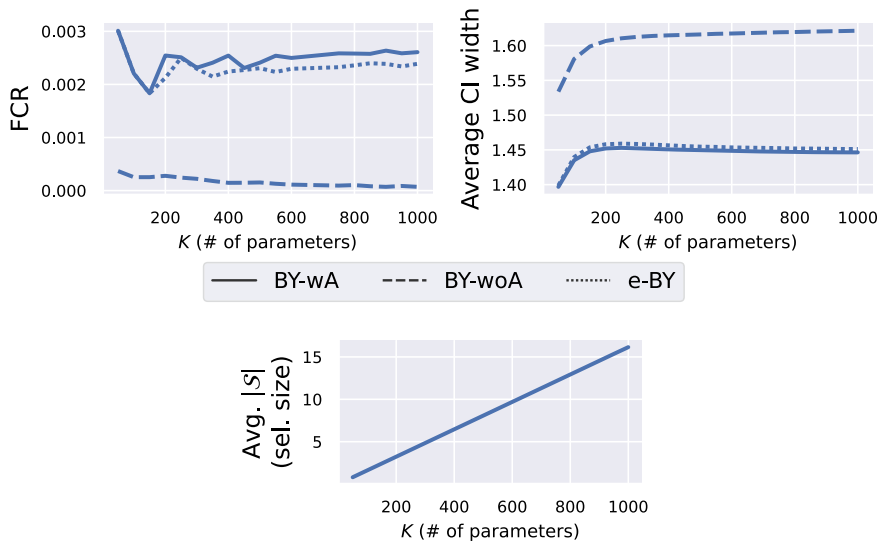

FIG 3. *Comparison of the FCR and average CI width for the BY vs. the e-BY procedure in the bounded setting. We also plot the average number of parameters selected (i.e., $|\mathcal{S}|$) for each K. "BY-wA" refers to the BY procedure under the assumptions that the CIs are independent (or PRDS) and $R_i^{\min} = |\mathcal{S}|$, while "BY-woA" is with no assumptions. The average CI width of the e-BY procedure is nearly the same as the BY procedure with assumptions, and both are vastly tighter than BY with no assumptions.*

Results In Figure 3, we see that e-BY and BY-wA have nearly the same expected width. On the other hand, BY-woA has a much larger expected width since it is uniformly dominated by BY-wA. This is a setting where one should use the BY-wA procedure for post-selection inference, since there is no dependence. However, if the selection rule or dependence is unknown beforehand, one can safely use the e-BY procedure, and attain CIs that are nearly as tight as the ones obtained through BY-wA.

6.2. Stopped Brownian motion setting

We also simulate the sharp FCR setting from Section 4 for different true drift parameters, $\theta_i^* = \theta^*$ for all $i \in [K]$, and number of parameters, K to compare the e-BY and BY procedures under a setting where the dependence structure and selection rule is more complex. In this setting, the assumptions of the BY procedure are not satisfied — the X_i are neither independent nor PRDS across $i \in [K]$. Hence, BY-wA does not have guaranteed FCR control in this setting. We compare the procedures on choices of $K \in \{10, 30, 100, 300, 1000\}$ and $\theta^* \in \{10^{-1}, 10^{-2}, 10^{-3}\}$. We set $\delta = 0.05$ and $\gamma = 2$. The Brownian motion processes, (W_i^t) for each $i \in [K]$, are discretized into time steps of size 10^{-5}. We use the e-CI formulated in (5) for the e-BY procedure and its running intersection and for the BY procedure.

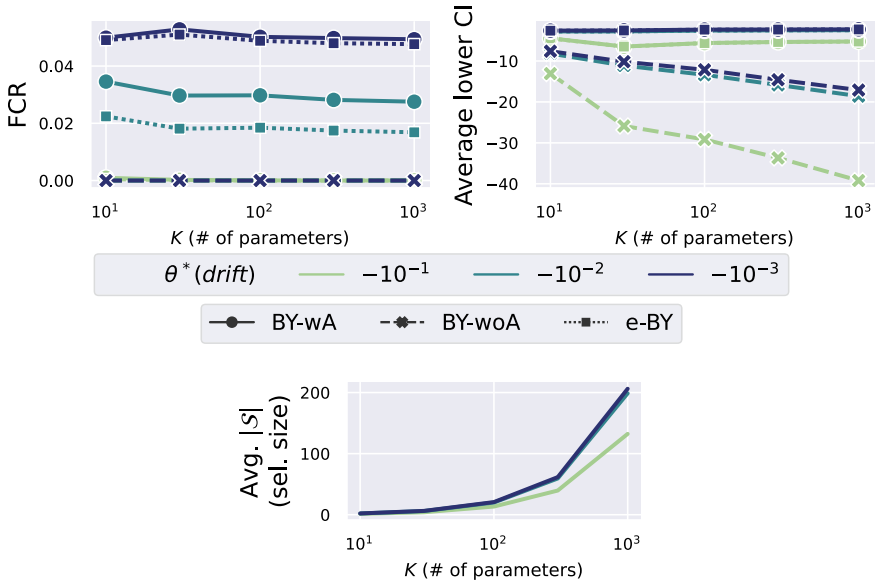

FIG 4. *Comparison of the e-BY procedure vs. the BY procedure for the stopped Brownian motion setting. In this setting, all CIs produced are one-sided, and* $\theta^* < 0$*, so we desire the average lower boundary to be as large as possible. The e-BY procedure and the BY-wA procedure have tight FCR as* $\theta^* \uparrow 0$ *and vastly higher average lower boundaries than the the BY-woA procedure. BY-woA has extremely conservative (nearly 0) FCR control.*

Results In Figure 4, we plot the empirical FCR and average CI lower bound (since the CIs are all one-sided) of each method for different values of the true parameter θ^*. Although it has no theoretical guarantees, BY-wA has an empirical FCR that is under δ for all experiment parameters. The e-BY procedure, which does have an FCR guarantee, also has FCR control, and has an average lower CI endpoint that is nearly as large as BY-wA. On the other hand, BY-woA has an extremely conservative empirical FCR and produces much smaller average lower CI endpoint than the other two methods. This reflects the best-of-both worlds behavior of the e-BY procedure in practice — e-BY has provable FCR control and CI widths comparable to the tightest possible widths produced by BY.

Thus, the e-BY procedure provides FCR control and CIs competitive with the BY procedure while requiring minimal assumptions. We will next show that this is also the case when using the e-BY procedure on real world data from A/B testing.

7. Application: Decision making in A/B testing

One natural application of post-selection inference is in A/B testing — A/B testing is the use of randomized control trials to decide which features (among

many) have a positive impact on users and should be shipped with the product. This method is typically used by information technology companies to determine whether a new version of a software product improves over the current version on certain user metrics. Each A/B test is run in a sequential fashion, where users are continually being added to the experiment over time. The goal of an A/B test is to stop as quickly as possible and allow the data scientist to make a confident decision about which version of software to ship based on estimates of the population mean of the user metrics. The e-BY procedure fits the A/B testing setup well for two key reasons: (1) the measurements of different metrics have a complex dependence structure since multiple measurements are made of a single user, and (2) the setup of the A/B testing is already sequential, so e-CIs, specifically stopped CSes, should already be the default for inference on each metric.

To simulate the behavior of data scientists choosing metrics to justify shipping decisions, we derive a selection method from the criteria for justifying shipping decisions of a single team within Twitter, a large information technology company. The shipping criteria provides guidelines for which combinations of metrics for which the difference between the treatment and control versions need to be statistically significant in a favorable direction, and which metrics that should not be significant in an unfavorable direction. We implemented a simplified version of this shipping criteria to be our selection method. Hence, our selection rule is the same across both procedures, and we are simply testing how the choice of post-selection inference procedure affects when the shipping criteria is considered satisfied.

A/B testing dataset from Twitter We compare the e-BY procedure vs. the BY procedure on a a dataset of actual A/B testing experiments from Twitter in a 1.5 year period and ran for at least two weeks. There are a total of 263 experiments in this dataset. Each of these experiments kept track of all 15 metrics ($K = 30$ since we treating the control and treatment versions of each metric as separate parameters) that were necessary for the shipping criteria at the daily level.

Since the user data is collected sequentially in these experiments, we use the following CS from Waudby-Smith et al. [36] for the e-BY procedure:

$$C_t^i(\alpha) = \left(\widehat{\mu}_i^t \ \pm 1.7 \sqrt{\widehat{\sigma}_{i,t}^2 \cdot \frac{\log\log(2t) + 0.72\log(5.2/\alpha)}{t}} \right),$$

where $\widehat{\mu}_i^t$ and $\widehat{\sigma}_{i,t}^2$ are the empirical mean and variance, respectively, of ith metric for the first t users. This is an asymptotic CS, i.e., the boundaries approach that of a true CS as the sample size increases under similar conditions as the central limit theorem for a traditional fixed-time asymptotic CI [36]. Since the sample size in these experiments are on the order of 10^6 users even on the first day, this asymptotic CS is very close to the true CS boundary. We use $\bar{C}_t^i = \cap_{i=1}^t C_t^i$, i.e., the intersection of all intervals so far for the BY procedure.

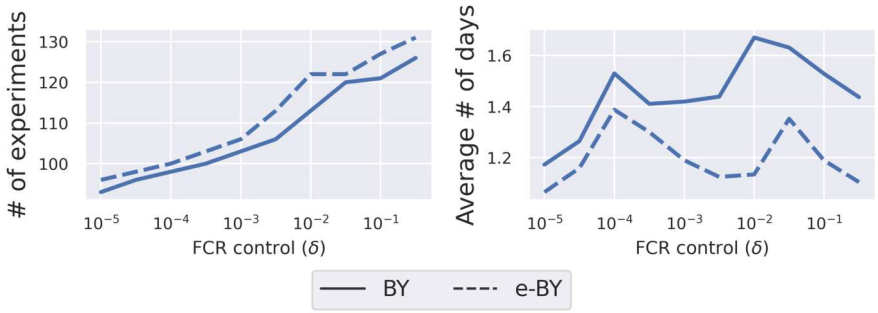

FIG 5. *Results of the e-BY procedure vs. the BY procedure on the real Twitter A/B testing data across different levels of* $\delta \in [0, 1]$ *for each procedure. The left figure depicts the total number of experiments where each procedure satisfied shipping criterion. The right figure depicts the average number of days an experiment ran before the procedure satisfied the shipping criteria. The set of experiments where the e-BY procedure satisfies shipping criteria is a strict superset of the set where the BY procedure satisfies shipping criteria (since e-BY dominates BY under dependence). Hence, the average number of days to satisfying the criteria is calculated over only the experiments for which the BY (and e-BY) procedure satisfied the shipping criteria.*

Results The results of our analysis in Figure 5 show that the e-BY procedure is better than the BY procedure in both number of experiments where a shipping decision can be justified, and the average time within each experiment to have sufficient evidence to satisfy the shipping criteria. For example, for a reasonable level of FCR control of $\delta = 0.1$, the e-BY procedure justified shipping decisions for 127 experiments — 5 more than 122 for the BY procedure. Note that the set of experiments where shipping was justified by the e-BY procedure is a superset of the experiments that the BY procedure justified shipping for. Consequently, we can also compare the average number of days before a shipping decision is justified for the e-BY and BY procedures, by taking the average over the experiments that the BY procedure justified shipping decisions on. We see that the the e-BY procedure took 1.2 days on average to satisfy the shipping criteria — 0.3 days less than the BY procedure which took on average 1.5 days. Clearly, the tightness of the e-BY procedure makes a practical impact on the efficiency of A/B testing.

8. Conclusion

We have shown that the e-BY procedure is a versatile method for controlling FCR in post-selection inference and improves inference in a broad class of settings. While e-BY is restricted use only on e-CIs, universal inference and stopped confidence sequences [34, 15, 16] are already e-CIs. Further, the e-BY procedure maintains FCR control with no assumptions on the dependence structure or selection rule, and produces CIs with widths that are never larger (and usually smaller) than those of the BY procedure. In addition, we demonstrated that the e-BY procedure is optimal for e-CIs in some sense, as the FCR bound it

guarantees is sharp, and that it is admissible in the domain of e-CI reporting procedures.

With respect to the utility of e-CIs, we observe that most of the tightest CIs we can construct in nonparametric and sequential settings are already e-CIs. In the sequential setting, we showed that the e-BY procedure produces practically useful CIs for justifying shipping decisions on data from real A/B tests. In addition, we discuss how to construct nontrivial e-CIs in any settings where CIs exist in Section 2.3: we extend the p-to-e calibration methods introduced by Vovk [29] to calibrating arbitrary CIs into e-CIs.

In this vein, an important direction for future study is to understand in what settings are the effectively tightest CIs also e-CIs. In many settings (e.g. Gaussian) where the exact (often asymptotic) distribution of a statistic under each parameter is known and identical across parameters, the tightest CIs (e.g. CI based on Gaussian CDF) are not e-CIs. On the other hand, e-CIs are often the only kind of CI we can construct when we are testing complicated composite nulls. Particularly in sequential settings where we desire an anytime-valid e-CI, Ramdas et al. [21] have shown that e-CIs are nearly admissible. Hence, further study of e-CIs would improve the utility of the e-BY procedure.

Appendix

The appendix is organized as follows. We provide all the proofs we had omitted in the body of the main paper in Appendix A, in order of their mention in the paper. Appendix B describes explicit connections between the post-selection inference procedures described in this paper for FCR control with FDR controlling procedures for multiple testing. We then present a weighted version of the e-BY procedure in Appendix C. To complement our previous discussion of admissibility, we also describe an improvement we can make to the e-BY procedure in Appendix D if we discard some of the restrictions made on CI-reporting procedures. We highlight the difference between conventional Chernoff based CIs and Chernoff e-CIs in Appendix E by introducing a notion of generalized e-CIs — this notion was applied in our simulations in the bounded setting in Section 6.1. Lastly, we discuss the advantages of FCR and how FCR relates to other error metrics that could be considered for the post-selection inference problem in Appendix F.

Appendix A: Proofs

We produce the proofs that we had omitted in the main body of the paper in this section.

A.1. Proof of Theorem 1

We first note that Definition 5 implies that $f(P)$ is an e-value for any calibrator f and p-value P — even if P is uniformly distributed.

To connect this idea with CIs, observe that the "dual" form of a CI is a p-value — a CI represents a collection of hypothesis tests for a fixed type I error (i.e., confidence level) across a set of hypotheses. In essence, it is the set of hypotheses that would not be rejected at some fixed level α with the current realization of the sampled data. In contrast, a p-value represents a collection of hypothesis tests for a fixed hypothesis, but across different rejection levels α. Here, the p-value itself is the type I error of the test with the smallest rejection level that rejects the fixed null hypothesis.

Definition 12. A *p-value* P for a set of distributions $\mathcal{Q}$ is a random variable supported on $[0,1]$ that satisfies the following:

$$\sup_{\mathtt{P}\in\mathcal{Q}} \mathbb{P}_{\mathtt{P}}(P \leq \alpha) \leq \alpha \text{ for all } \alpha \in [0,1].$$

Thus, we can calibrate the implicit p-value associated with every CI constructing procedure to an e-value, and then produce the e-CI associated with the e-value to calibrate from a CI to an e-CI.

Let $C(\alpha)$ produce a $(1-\alpha)$ CI for any $\alpha \in [0,1]$. Consequently, for a fixed parameter $\theta \in \Theta$, the following is a p-value for the set $\mathcal{Q}_\theta \coloneqq \{\mathtt{P} \in \mathcal{P} : T(\mathtt{P}) = \theta\}$:

$$P^{\text{dual}}(\theta) \coloneqq \inf\left\{\alpha \in [0,1] : \theta \notin C(\alpha)\right\}.$$

Consequently,

$$E^{\text{cal}}(\theta) \coloneqq f(P^{\text{dual}}(\theta))$$

forms a family of e-values. Hence,

$$C^{\text{cal}}(\alpha) \coloneqq \left\{\theta \in \Theta : E^{\text{cal}}(\theta) < \frac{1}{\alpha}\right\}$$

is a $(1-\alpha)$-e-CI. To show the equality between C^{cal} and C, we make the following observations.

$$\begin{aligned} C^{\text{cal}}(\alpha) &= \left\{\theta \in \Theta : E^{\text{cal}}(\theta) < \frac{1}{\alpha}\right\} = \left\{\theta \in \Theta : f(P^{\text{dual}}(\theta)) < \frac{1}{\alpha}\right\} \\ &\overset{\text{(i)}}{=} \left\{\theta : P^{\text{dual}}(\theta) > \max\left\{p : f(p) \geq \frac{1}{\alpha}\right\}\right\} \\ &= \left\{\theta : P^{\text{dual}}(\theta) > f^{-1}\left(\frac{1}{\alpha}\right)\right\} \\ &\overset{\text{(ii)}}{=} C\left(f^{-1}\left(\frac{1}{\alpha}\right)\right). \end{aligned}$$

The equality at (i) is a result of f being nonincreasing and upper semicontinuous at $1/\alpha$. Hence the supremum is achieved and the equality holds. Similarly, the equality at (ii) is because C is nonincreasing and continuous from below at $f^{-1}(1/\alpha)$. If $P^{\text{dual}}(\theta) = f^{-1}(1/\alpha)$, then $\theta \notin C_i(P^{\text{dual}}(\theta))$ by C being continuous from below.

A.2. Proof of Theorem 3

To prove the sharpness of the above situation, we require the following fact about Brownian motions.

Fact 2 (Borodin and Salminen 9, p.223, equation 2.1.2)**.** *Let $(W_t)_{t\in\mathbb{R}_0^+}$ be a Brownian motion with drift μ and initial value x. Define stopping times $\tau^a := \inf\{t \geq 0 : W_t \geq x + a\}$, and $\tau^b := \inf\{t \geq 0 : W_t \leq x - b\}$, where $a, b > 0$. Then,*

$$\mathbb{P}\left(\tau^a < \tau^b\right) = f(\mu, a, b) := 1 - e^{-\mu b}\frac{\sinh(|\mu|a)}{\sinh(|\mu|(a+b))}.$$

We note that $f(\mu, a, b)$ is increasing in μ since the larger the drift term, the more likely that the Brownian motion hits the upper boundary $x+a$. Moreover,

$$\lim_{\mu\uparrow 0} f(\mu, a, b) = 1 - \lim_{\mu\uparrow 0} e^{-\mu b}\frac{e^{|\mu|a} - e^{-|\mu|a}}{e^{|\mu|(a+b)} - e^{-|\mu|(a+b)}} = 1 - \frac{a}{a+b} = \frac{b}{a+b}, \quad (7)$$

which recovers the simple case in which $\mu = 0$ (i.e., the case that the Brownian motion is a martingale). Only the above properties of f (not its precise formula) will be used in the result below.

Using Fact 2, for $i \in [K]$, the probability that W_i hits $\gamma - 1$ before hitting -1 is given by

$$\mathbb{P}\left(W_i^{\tau_i} = \gamma - 1\right) = f(\theta_i^*, \gamma - 1, 1).$$

We can compute the mean of $|\mathcal{S}|$ as

$$\begin{aligned}\mathbb{E}[|\mathcal{S}|] &= \mathbb{E}\left[\sum_{i\in[K]} \mathbf{I}\left\{W_i^{\tau_i} = \gamma - 1\right\}\right] \\ &= \sum_{i\in[K]} \mathbb{P}(W_i^{\tau_i} = \gamma - 1) = \sum_{i\in[K]} f(\theta_i^*, \gamma - 1, 1). \end{aligned} \quad (8)$$

For $i \in \mathcal{S}$, using Fact 2 again and noting that $W_i^{\tau_i} = \gamma - 1$, we have

$$\mathbb{P}\left(W_i^{\tau_i} = \beta - 1 \mid \mathcal{S}\right) = f(\theta_i^*, \beta - \gamma, \gamma). \quad (9)$$

Let us suppose that $\theta_i^* < 0$ for each $i \in [K]$ (but this is not known to the scientist); that is, $X^i(0)$ is a strict supermartingale. In this case, $E_i(0) = \beta$ only if $E_i(\theta_i^*) > \beta$.

Recall that $C_i(\alpha) = \{\theta \in \Theta : E_i(\theta) < 1/\alpha\}$. We can now compute FCR, using (9), as

$$\begin{aligned}\mathrm{FCR} &= \mathbb{E}\left[\frac{\sum_{i\in\mathcal{S}} \mathbf{I}\left\{\theta_i^* \notin C_i(1/\beta)\right\}}{|\mathcal{S}| \vee 1}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{i\in\mathcal{S}} \mathbf{I}\left\{E_i(\theta_i^*) \geq \beta\right\}}{|\mathcal{S}| \vee 1} \mid \mathcal{S}\right]\right]\end{aligned}$$

$$= \mathbb{E}\left[\sum_{i\in\mathcal{S}} \frac{\mathbb{P}\left(E_i(\theta_i^*) \geq \beta \mid \mathcal{S}\right)}{|\mathcal{S}| \vee 1}\right]$$
$$\geq \mathbb{E}\left[\sum_{i\in\mathcal{S}} \frac{\mathbb{P}\left(E_i(0) = \beta \mid \mathcal{S}\right)}{|\mathcal{S}| \vee 1}\right] = \mathbb{E}\left[\frac{\sum_{i\in\mathcal{S}} f(\theta_i^*, \beta-\gamma, \gamma)}{|\mathcal{S}| \vee 1}\right].$$

Obviously, FCR is a function of $\boldsymbol{\theta}^* := (\theta_1, \ldots, \theta_K)$. We consider the situation where $\theta_i^* \uparrow 0$ (i.e., θ_i^* approaches 0) for each $i \in [K]$, denoted by $\boldsymbol{\theta}^* \uparrow \mathbf{0}$. By applying (7), we get

$$f(\theta_i^*, \beta-\gamma, \gamma) \uparrow \frac{\gamma}{\beta} = \frac{\delta|\mathcal{S}|\gamma}{K},$$

as $\theta_i^* \uparrow 0$. Moreover, when $\boldsymbol{\theta}^* \uparrow \mathbf{0}$, (7) and (8) together yield

$$\mathbb{E}_{\boldsymbol{\theta}^*}[|\mathcal{S}|] \uparrow \frac{K}{\gamma},$$

where we emphasize that the expectation is taken with respect to $\boldsymbol{\theta}^*$ (which varies). Using the above results and the monotone convergence theorem, we get

$$\lim_{\boldsymbol{\theta}^*\uparrow\mathbf{0}} \mathbb{E}_{\boldsymbol{\theta}^*}\left[\frac{\sum_{i\in\mathcal{S}} f(\theta_i^*, \beta-\gamma, \gamma)}{|\mathcal{S}| \vee 1}\right] = \lim_{\boldsymbol{\theta}^*\uparrow\mathbf{0}} \mathbb{E}_{\boldsymbol{\theta}^*}\left[\frac{\delta|\mathcal{S}|\gamma}{K}\mathbf{I}\{|\mathcal{S}| > 0\}\right]$$
$$= \lim_{\boldsymbol{\theta}^*\uparrow\mathbf{0}} \mathbb{E}_{\boldsymbol{\theta}^*}\left[\frac{\delta|\mathcal{S}|\gamma}{K}\right] = \delta.$$

Since the FCR is upper bounded by δ as a result of using the e-BY procedure (Theorem 2), we have shown our desired statement.

Remark 5. From the analysis above, we can see that the value δ of FCR of e-BY is sharp when $\boldsymbol{\theta}^*$ approaches $\mathbf{0}$ from below. Since both convergences in the above computation are monotone in $\boldsymbol{\theta}^*$, we have, in general, that FCR $\leq \delta$ for $\boldsymbol{\theta}^* \leq \mathbf{0}$. On the other hand, if some $\theta_i^* > 0$, then $E_i(\theta_i^*) < E_i(0) \leq \beta$, and hence $\mathbb{P}\left(E_i(\theta_i^*) > \beta\right) = 0$, leading also to a smaller FCR for the above procedure. (If one chooses a sightly larger threshold $\beta > K/(\delta|\mathcal{S}|)$, then the scenario $\boldsymbol{\theta}^* \approx \mathbf{0}$ with $\boldsymbol{\theta}^* \geq 0$ also has an FCR close to δ.)

The independence of the e-processes is only used in (9) in the second step. This condition can be weakened to the independence between the e-processes in step 1 and their increment processes in step 2. The independence assumption does not reduce the FCR, at least not in an obvious way.

A.3. Proof of Lemma 1

We first show the "only if" statement. Suppose that $C \in \mathrm{ECI}(\Theta)$ with associated e-value E. It is clear that C is nonincreasing. To show its continuity from above, note that $E(\theta) < 1/\alpha$ if and only if $E(\theta) < 1/\beta$ for some $\beta > \alpha$. Therefore, we have $\bigcup_{\beta>\alpha} C(\beta) = C(\alpha)$, and thus C is continuous from below.

We next show the "if" statement. Let the function $t : \Theta \to [0, \infty]$ be given by

$$t(\theta) = \sup \left\{ \frac{1}{\alpha} : \alpha \in [0,1),\ \theta \notin C(\alpha) \right\}, \tag{10}$$

where the convention is $\sup \emptyset = 1$ (because we can always set $C(1) = \emptyset$ so that $\theta \notin C(1)$ In the two extreme cases, if $C(\alpha)$ is Θ on $[0,1)$, then $t(\theta) = 1$ for each θ, and if $C(\alpha)$ is $\emptyset$ on $[0,1)$, then $t(\theta) = \infty$ for each θ.

Since C is nonincreasing and continuous from below, (10) implies $\theta \notin C(t(\theta))$. Thus, we get the following equivalence:

$$t(\theta) < \frac{1}{\alpha} \iff \theta \in C(\alpha). \tag{11}$$

Let U be a standard uniform random variable under each of the parameter $\theta \in \Theta$, and define

$$E(\theta) = t(\theta)\mathbf{I}\left\{ U < \frac{1}{t(\theta)} \right\}, \quad \theta \in \Theta.$$

It is clear that $E(\theta) \geq 0$ and $\mathbb{E}_\theta[E(\theta)] = 1$, and (6) holds if $E(\theta) = t(\theta)$ for each $\theta \in \Theta$, which is a possible realization of E.

A.4. Proof of Theorem 4

We fix a universal probability space $\mathfrak{P}$ and parameter ranges $\Theta_1, \ldots, \Theta_K$.

We will proceed to prove Theorem 4 by contradiction. Suppose that there exists a CI reporting procedure Φ which strictly dominates the e-BY procedure. That means for all $S \subseteq [K]$ and $\mathbf{C} \in \mathrm{ECI}_K$, $\Phi(S, \mathbf{C})_i \subseteq C_i(\delta|S|/K)$, and

$$\begin{gathered}\text{there exists some } S \subseteq [K], \mathbf{C} \in \mathrm{ECI}_K \text{ and } i^* \in S \\ \text{s.t. } \Phi(S, \mathbf{C})_{i^*} \subsetneq C_{i^*}(\delta|S|/K).\end{gathered}$$

From now on, fix the above $(S, \mathbf{C}, i^*)$ — these values will be key components for our construction, which will result in a contradiction. Clearly, S is not empty.

We assume that there exists random variables $B_{\mathcal{S}}$ and U in each $\mathsf{P} \in \mathcal{P}$ (the existence is guaranteed by enlarging the probability space if needed) that are distributed as follows:

$$B_{\mathcal{S}} \sim \mathrm{Bern}(|S|/K) \text{ and } U \sim \mathrm{Uniform}[0,1],$$

and we let $B_{\mathcal{S}}$ and U be independent of each other.

We will show that $\mathrm{FCR}(\Phi) > \delta$ for some random selection $\mathcal{S}$, vector of families of e-values $\mathbf{E} = (\{E_1(\theta)\}_{\theta \in \Theta_1}, \ldots, \{E_K(\theta)\}_{\theta \in \Theta_K})$ and true values of parameters $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_K^*)$. The main work is to construct such a setting.

First, we define the distribution of our selection rule $\mathcal{S}$. Let $T := [K] \setminus S$ be the complement of S. We specify $\mathcal{S} = S$ if $B_{\mathcal{S}} = 1$ and $\mathcal{S} = T$ otherwise. Hence, under all vectors of possible parameter values in $\Theta_1 \times \cdots \times \Theta_K$, $\mathcal{S}$ has the following distribution:

$$\mathbb{P}(\mathcal{S} = S) = \mathbb{P}(B_{\mathcal{S}} = 1) = |S|/K, \text{ and } \mathbb{P}(\mathcal{S} = T) = \mathbb{P}(B_{\mathcal{S}} = 0) = |T|/K,$$

In the following, statements for $i \in T$ can be ignored if T is empty.

We briefly explain the main idea behind our construction. We first note that any construction of $\mathbf{E}$ and $\boldsymbol{\theta}^*$, if the i-th component of Φ is equal to that of e-BY for all $i \neq i^*$. Then, for $i \in \mathcal{S} \setminus \{i^*\}$,

$$\begin{aligned}\mathbb{P}\left(\theta_i \notin \Phi(\mathcal{S}, \mathbf{C}^{\mathbf{E}})_i,\ \mathcal{S} = S\right) &= \mathbb{P}\left(\theta_i \notin C_i^{\mathbf{E}}\left(\frac{\delta|\mathcal{S}|}{K}\right),\ \mathcal{S} = S\right) \\ &\leq \mathbb{P}\left(E_i(\theta_i) \geq \frac{K}{\delta|\mathcal{S}|}\right) \leq \frac{\delta|\mathcal{S}|}{K},\end{aligned}$$

and for $i \in T$,

$$\begin{aligned}\mathbb{P}\left(\theta_i \notin \Phi(T, \mathbf{C}^{\mathbf{E}})_i,\ \mathcal{S} = T\right) &= \mathbb{P}\left(\theta_i \notin C_i^{\mathbf{E}}\left(\frac{\delta|T|}{K}\right),\ \mathcal{S} = T\right) \\ &\leq \mathbb{P}\left(E_i(\theta_i) \geq \frac{K}{\delta|T|}\right) \leq \frac{\delta|T|}{K}.\end{aligned}$$

We will need to construct suitable $\mathbf{E}$ and $\boldsymbol{\theta}^*$ such that all above inequalities hold as equalities approximately by taking a limit, and for the index i^* we have

$$\mathbb{P}\left(\theta_{i^*} \notin \Phi(\mathcal{S}, \mathbf{C}^{\mathbf{E}})_{i^*},\ \mathcal{S} = S\right) > \frac{\delta|\mathcal{S}|}{K} + c.$$

where $c > 0$ is a constant that does not shrink to 0 when we take a limit. If we are able to achieve the above, then the FCR of Φ will be close to $\delta + c/|\mathcal{S}|$.

Let us specify our choice of $\boldsymbol{\theta}^*$. First, take a small number $\epsilon \in (0, 1)$ which will later shrink to 0.

(a) We take $\theta^*_{i^*} \in C_{i^*}\left(\delta|\mathcal{S}|/K\right) \setminus \Phi(\mathcal{S}, \mathbf{C})_{i^*}$ which is possible since $\Phi(\mathcal{S}, \mathbf{C})_{i^*} \subsetneq C_{i^*}\left(\delta|\mathcal{S}|/K\right)$.
(b) For $i \in \mathcal{S} \setminus \{i^*\}$, we take $\theta^*_i \in C_i\left((1-\epsilon)\delta|\mathcal{S}|/K\right) \setminus C_i\left(\delta|\mathcal{S}|/K\right)$, which is possible since $\mathbf{C}$ is strictly decreasing.
(c) For $i \in T$, we take $\theta^*_i \in C_i\left((1-\epsilon)\delta|T|/K\right) \setminus C_i\left(\delta|T|/K\right)$.

Note that by this construction, we have $\theta^*_i \notin \Phi(\mathcal{S}, \boldsymbol{C})_i$ for each $i \in [K]$. Further, we can arbitrarily select the true values for each parameter, $\boldsymbol{\theta}^*$, with the following construction of our universe of probabilities $\mathcal{P}$. We can let our universe of distributions contain joint distributions over $((U, B_{\mathcal{S}}, Y_1), \ldots (U, B_{\mathcal{S}}, Y_K))$ where Y_i is some random variable whose distribution determines θ^*_i for each $i \in [K]$. Let $\mathtt{P}_{Y_i}$ denote the marginal distribution of Y_i in $\mathtt{P}$. We define $\vartheta_i(\mathtt{P}) = \vartheta_{Y_i}(\mathtt{P}_{Y_i})$, where $\vartheta_{Y_i} : \mathcal{P}_{Y_i} \to \Theta_i$ is a functional that maps from the universe marginal distributions of Y_i, $\mathcal{P}_{Y_i}$, to the parameter value space Θ_i. Since the distribution of $B_{\mathcal{S}}$ and U are identical across any $\mathtt{P} \in \mathcal{P}$, we can simply select some $\boldsymbol{\theta}$ and let $\mathtt{P}^* = \mathtt{P}_{\boldsymbol{\theta}}$, where $\mathtt{P}_{\boldsymbol{\theta}}$ is some distribution in $\mathcal{P}$ with true values $\boldsymbol{\theta}$ for its parameters.

For $i \in [K]$, similarly to (10)-(11), define

$$t_i(\theta) = \sup\left\{\frac{1}{\alpha} : \alpha \in [0, 1),\ \theta \notin C_i(\alpha)\right\};$$

$$\text{this gives } \; t_i(\theta) < \frac{1}{\alpha} \iff \theta \in C_i(\alpha). \tag{12}$$

Next, we define some quantities concerning the true parameter $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_K^*)$.

$$r_S^* = \frac{K}{\delta|\mathcal{S}|}, \quad t_S := \max_{i \in \mathcal{S}} t_i(\theta_i^*) \in \left[r_S^*, \frac{r_S^*}{1-\epsilon} \right),$$
$$t_S^* := t_{i^*}(\theta_{i^*}^*) < r_S^*, \quad \ell_S := \frac{t_S - t_S^*}{r_S^*}.$$

The variables marked with an asterisk $(*)$ are not dependent on ϵ — only t_S and ℓ_S depend on ϵ.

Let U be a random variable that is standard uniform and independent of $\mathcal{S}$ under all $\mathtt{P}_{\boldsymbol{\theta}'}$, and the existence of such may be achieved by enlarging the probability space. Define the random variables U_S and U_T indexed by $\boldsymbol{\theta}'$ as follows.

$$U_S(\boldsymbol{\theta}') := \frac{|\mathcal{S}|}{K} U \mathbf{I}\{\mathcal{S} = S\} \left(\mathbf{I}\left\{ U \geq \frac{K}{|\mathcal{S}| t_S} \right\} \vee \mathbf{I}\left\{ \boldsymbol{\theta}' \neq \boldsymbol{\theta}^* \right\} \right)$$
$$+ \left(\frac{|T|}{K} U + \frac{|\mathcal{S}|}{K} \right) \mathbf{I}\{\mathcal{S} = T\}$$
$$U_T := \frac{|T|}{K} U \mathbf{I}\{\mathcal{S} = T\} + \left(\frac{|\mathcal{S}|}{K} U + \frac{|T|}{K} \right) \mathbf{I}\{\mathcal{S} = S\}.$$

Note that $U_S(\boldsymbol{\theta}')$ is 0 with probability $1/t_S$ and uniformly distributed between $[1/t_S, 1]$ under $\mathtt{P}_{\boldsymbol{\theta}'}$ otherwise if and only if $\boldsymbol{\theta}' = \boldsymbol{\theta}^*$. If $\boldsymbol{\theta}' \neq \boldsymbol{\theta}^*$, then under $\mathtt{P}_{\boldsymbol{\theta}'}$, $U_S(\boldsymbol{\theta}')$ and U_T are both uniform random variables over $[0, 1]$.

We are ready to define our e-values. We first define E_{i^*}, which is the most sophisticated object:

$$E_{i^*}(\theta) = \; t_{i^*}(\theta) \mathbf{I}\left\{ U_S < \frac{1}{t_{i^*}(\theta) \vee t_S} \right\}$$
$$+ r_S^* \mathbf{I}\left\{ \frac{1}{t_S} \leq U_S < \frac{1 + \ell(\theta)}{t_S} \right\} \mathbf{I}\{t_{i^*}(\theta) < t_S\},$$

for all each $\theta \in \Theta_{i^*}$, where

$$\ell(\theta) := \frac{t_S - t_{i^*}(\theta)}{r_S^*}.$$

Clearly $E_{i^*} \geq 0$. It remains to show that, for each $\theta \in \Theta_{i^*}$, the expectation is bounded as $\mathbb{E}_\theta[E_{i^*}(\theta)] \leq 1$. We proceed casewise on $t_{i^*}(\theta)$. If $t_{i^*}(\theta) \geq t_S$, then

$$\mathbb{E}_\theta[E_{i^*}(\theta)] \leq \mathbb{E}_\theta \left[t_{i^*}(\theta) \mathbf{I}\left\{ U_S < \frac{1}{t_{i^*}(\theta)} \right\} \right] = 1.$$

The inequality is simply by construction of U_S — U_S is uniform if θ is the true parameter and $t_{i^*}(\theta) > t_S$. When $t_{i^*}(\theta) = t_S$, the following is still true:

$\mathbb{P}(U_S < 1/t_S) = 1/t_S$. If $t_{i^*}(\theta) < t_S$, then

$$\mathbb{E}_\theta\left[E_{i^*}(\theta)\right] = \frac{t_{i^*}(\theta)}{t_S} + \frac{r_S^* \ell(\theta)}{t_S} = \frac{t_{i^*}(\theta)}{t_S} + \frac{t_S - t_{i^*}(\theta)}{t_S} = 1.$$

Note that the first equality is true again because U_S always satisfies

$$\mathbb{P}(U_S < 1/t_S) = 1/t_S$$

and is otherwise uniformly distributed among values greater than $1/t_S$. Therefore, E_{i^*} is an e-value for Θ_{i^*}. Note that in our choice of true parameters $\boldsymbol{\theta}^*$, we are always in the $t_{i^*}(\theta^*_{i^*}) = t^*_S < t_S$ case.

The other e-values are defined as

$$E_i(\theta) = t_i(\theta)\mathbf{I}\left\{U_S < \frac{1}{t_i(\theta)}\right\}, \quad \theta \in \Theta_i, \quad i \in \mathcal{S} \setminus \{i^*\},$$

and

$$E_i(\theta) = t_i(\theta)\mathbf{I}\left\{U_T < \frac{1}{t_i(\theta)}\right\}, \quad \theta \in \Theta_i, \quad i \in T.$$

It is straightforward to check that each $E_i(\theta)$ above is an e-value for $\theta \in \Theta_i$.

Now that we have defined our e-values, we will show there exists a distribution where the FCR under CI-reporting procedure Φ will be strictly larger than δ.

Note that by construction of U_S and U_T, we know the following:

$$\left\{U_S \le \frac{|\mathcal{S}|}{K}\right\} = \{\mathcal{S} = S\} \text{ and } \left\{U_T \le \frac{|T|}{K}\right\} = \{\mathcal{S} = T\}. \tag{13}$$

By (12) and the construction of $(\theta_i^*)_{i\in[K]}$, we have $t_{i^*}(\theta^*_{i^*}) < K/(\delta|\mathcal{S}|)$ as well as

$$\begin{aligned} &\frac{K}{\delta|\mathcal{S}|} \le t_i(\theta_i^*) < \frac{K}{(1-\epsilon)\delta|\mathcal{S}|} && \text{for } i \in \mathcal{S} \setminus \{i^*\} \\ \text{and} \quad &\frac{K}{\delta|T|} \le t_i(\theta_i^*) < \frac{K}{(1-\epsilon)\delta|T|} && \text{for } i \in T. \end{aligned} \tag{14}$$

Putting the above ranges of $t_i(\theta_i^*)$ together, using (13), we get

$$\begin{aligned} &\left\{U_S < \frac{1}{t_i(\theta_i^*)}\right\} \subseteq \{\mathcal{S} = S\} && \text{for all } i \in \mathcal{S} \setminus \{i^*\} \\ \text{and } &\left\{U_T < \frac{1}{t_i(\theta_i^*)}\right\} \subseteq \{\mathcal{S} = T\} && \text{for all } i \in T. \end{aligned} \tag{15}$$

Moreover, As $\epsilon \downarrow 0$, we have $t_S \downarrow r_S^*$, and

$$\ell_S \to \ell_S^* := 1 - \frac{t_S^*}{r_S^*} > 0.$$

The above construction of e-values leads to the important condition under true parameters $\boldsymbol{\theta}^*$, via (12),

$$U_S < \frac{1}{t_S} \implies E_i(\theta_i^*) = t_i(\theta_i^*) \text{ for all } i \in \mathcal{S} \implies \mathbf{C}_S^{\mathbf{E}} = \mathbf{C}_S. \tag{16}$$

Recall that by construction, we have $\theta_i^* \notin \Phi(\mathcal{S}, \mathbf{C})_i$ for each $i \in \mathcal{S}$. Therefore, by using (15) and (16), for $i \in S \setminus \{i^*\}$,

$$\begin{aligned}\mathbb{P}\left(\theta_i^* \notin \Phi(\mathcal{S}, \mathbf{C}^{\mathbf{E}})_i,\ \mathcal{S} = S\right) &\geq \mathbb{P}\left(\mathbf{C}_S^{\mathbf{E}} = \mathbf{C}_S,\ \mathcal{S} = S\right) \\ &\geq \mathbb{P}\left(U_S < \frac{1}{t_S},\ \mathcal{S} = S\right) = \frac{1}{t_S}. \end{aligned} \tag{17}$$

Further, by construction, we know that $t_S^* < t_S$, so we have

$$\begin{aligned}&\mathbb{P}\left(\theta_{i^*}^* \notin \Phi(\mathcal{S}, \mathbf{C}^{\mathbf{E}})_{i^*},\ \mathcal{S} = S\right) \\ &\geq \mathbb{P}\left(\mathbf{C}_S^{\mathbf{E}} = \mathbf{C}_S,\ \mathcal{S} = S\right) + \mathbb{P}\left(E_{i^*}(\theta) \geq r_S^*,\ \mathbf{C}_S^{\mathbf{E}} \neq \mathbf{C}_S,\ \mathcal{S} = S\right) \\ &\geq \frac{1}{t_S} + \mathbb{P}\left(\frac{1}{t_S} \leq U_S < \frac{1+\ell_S}{t_S}\right) = \frac{1+\ell_S}{t_S}. \end{aligned} \tag{18}$$

Write

$$r_T^* := \frac{K}{|T|\delta}, \quad \text{and} \quad t_T := \max_{i \in T} t_i(\theta_i^*) \in \left[r_T^*, \frac{1}{1-\epsilon} r_T^*\right).$$

In case $|T| = 0$ the above quantities are set to ∞.

For $i \in T$, using $\Phi(\mathcal{S}, \mathbf{C}^{\mathbf{E}})_i \subseteq C_i^{\mathbf{E}}(\frac{\delta|\mathcal{S}|}{K})$ and (14)-(15), we have

$$\begin{aligned}\mathbb{P}\left(\theta_i^* \notin \Phi(T, \mathbf{C}^{\mathbf{E}})_i,\ \mathcal{S} = T\right) &\geq \mathbb{P}\left(\theta_i^* \notin C_i^{\mathbf{E}}\left(\frac{\delta|T|}{K}\right),\ \mathcal{S} = T\right) \\ &= \mathbb{P}\left(E_i(\theta_i^*) \geq r_T^*,\ \mathcal{S} = T\right) \\ &\geq \mathbb{P}\left(E_i(\theta_i^*) \geq r_T^*,\ U_T < \frac{1}{t_T},\ \mathcal{S} = T\right) \\ &= \mathbb{P}\left(U_T < \frac{1}{t_T}\right) = \frac{1}{t_T}. \end{aligned} \tag{19}$$

Putting (17), (18) and (19), together, we obtain

$$\begin{aligned}&\mathbb{E}\left[\frac{\sum_{i \in \mathcal{S}} \mathbf{I}\left\{\theta_i^* \notin \Phi(\mathcal{S}, \mathbf{C}^{\mathbf{E}})_i\right\}}{|\mathcal{S}| \vee 1}\right] \\ &= \mathbb{E}\left[\frac{\sum_{i \in S} \mathbf{I}\left\{\theta_i^* \notin \Phi(\mathcal{S}, \mathbf{C}^{\mathbf{E}})_i\right\}}{|S| \vee 1} \mathbf{I}\{\mathcal{S} = S\}\right] \\ &\quad + \mathbb{E}\left[\frac{\sum_{i \in T} \mathbf{I}\left\{\theta_i^* \notin \Phi(T, \mathbf{C}^{\mathbf{E}})_i\right\}}{|T| \vee 1} \mathbf{I}\{\mathcal{S} = T\}\right] \\ &\geq \frac{1}{|\mathcal{S}|}\left(\frac{1+\ell_S}{t_S}\right) + \frac{1}{|\mathcal{S}|} \sum_{i \in S \setminus \{i^*\}} \frac{1}{t_S} + \frac{1}{|T| \vee 1} \sum_{i \in T} \frac{1}{t_T} \end{aligned}$$

$$= \frac{\ell_S}{|\mathcal{S}|t_S} + \frac{1}{t_S} + \frac{1}{t_T}.$$

Note that as $\epsilon \downarrow 0$, $t_S \to r_S^*$, $t_T \to r_T^*$, and $\ell_S \to \ell_S^* > 0$ by $t_S^* < r_S^*$. Therefore,

$$\lim_{\epsilon \downarrow 0} \frac{\ell_S}{|\mathcal{S}|t_S} + \frac{1}{t_S} + \frac{1}{t_T} = \frac{\ell_S^*}{|\mathcal{S}|r_S^*} + \frac{1}{r_S^*} + \frac{1}{r_T^*} > \frac{\delta|\mathcal{S}|}{K} + \frac{|T|\delta}{K} = \delta.$$

Thus, the FCR of the procedure Φ for our constructed setting is large than δ. This shows that the e-BY procedure is admissible.

Remark 6. From the proof of Theorem 4, we can see that the restriction (b) that $\Phi(\mathcal{S}, \mathbf{C}) = \Phi(\mathcal{S}, \mathbf{C}')$ if $\mathbf{C}_S = \mathbf{C}'_S$ is important, because on the event $\mathcal{S} = S$, we only require $\mathbf{C}_S^{\mathbf{E}} = \mathbf{C}_S$, but not $\mathbf{C}^{\mathbf{E}} = \mathbf{C}$. In our construction, we actually have $E_i(\theta_i^*) = 0$ for $i \in T$ on the event $\{\mathcal{S} = S\}$. Recall that we need many inequalities in the FCR formula to be approximately equalities. The only way to produce precisely the event $\{\mathbf{C}^{\mathbf{E}} = \mathbf{C},\ \mathcal{S} = S\}$ is through the event $\{E_i = t_i \text{ for all } i \in [K] \text{ and } \mathcal{S} = S\}$, but for the approximation we need $\mathbb{E}[E_i \mathbf{I}\{\mathcal{S} = T\}]$ for $t \in T$ to be arbitrarily close to 1, so it does not seem to be possible if we "waste" E_i, $i \in T$ to take positive values on the event $\mathcal{S} = S$.

Remark 7. The proof of Theorem 4 also justifies that e-BY has sharp FCR. To see this, for Φ being e-BY and the setting constructed in the proof of Theorem 4 (omitting the special treatment for i^*), using (17) and (19), we get

$$\mathbb{P}\left(\theta_i^* \notin \Phi(\mathcal{S}, \mathbf{C}^{\mathbf{E}})_i,\ \mathcal{S} = S\right) \geq 1/t_S \text{ for } i \in \mathcal{S},$$
$$\mathbb{P}\left(\theta_j^* \notin \Phi(\mathcal{S}, \mathbf{C}^{\mathbf{E}})_j,\ \mathcal{S} = T\right) \geq 1/t_T \text{ for } j \in T.$$

The rest is taking a limit, and we see that the e-BY procedure has an FCR arbitrarily close to δ for this setting.

Appendix B: Multiple testing procedures based on e-BY

We can derive procedures for the multiple testing problem directly from post-selection inference procedures. We will discuss how that can be accomplished, and provide explicit examples of such derivations. In the multiple testing problem, we wish to identify as many parameters that lie outside of a fixed null hypothesis as possible. Formally, the goal is to determine whether $\theta_i^* \in \Theta_i^0$ where $\Theta_i^0 \subseteq \Theta_i$ is the null hypothesis (as opposed to directly estimating θ_i^*). Thus, for each $i \in [K]$, we must output a decision of whether we reject the null hypothesis $H_i^0 : \theta_i^* \in \Theta_i^0$ or not. We denote set of indices where the null hypothesis is rejected as $\mathcal{R}$. The null hypotheses in the rejection set ($i \in \mathcal{R}$ and H_i^0 is true) are called *false discoveries*. Analogous to the FCR, we wish to control the *false discovery rate* (FDR), which is the expectation of the *false discovery proportion* (FDP):

$$\text{FDP} := \frac{\sum_{i \in \mathcal{R}} \mathbf{I}\left\{\theta_i^* \in \Theta_i^0\right\}}{|\mathcal{R}|}, \qquad \text{FDR} := \mathbb{E}[\text{FDP}].$$

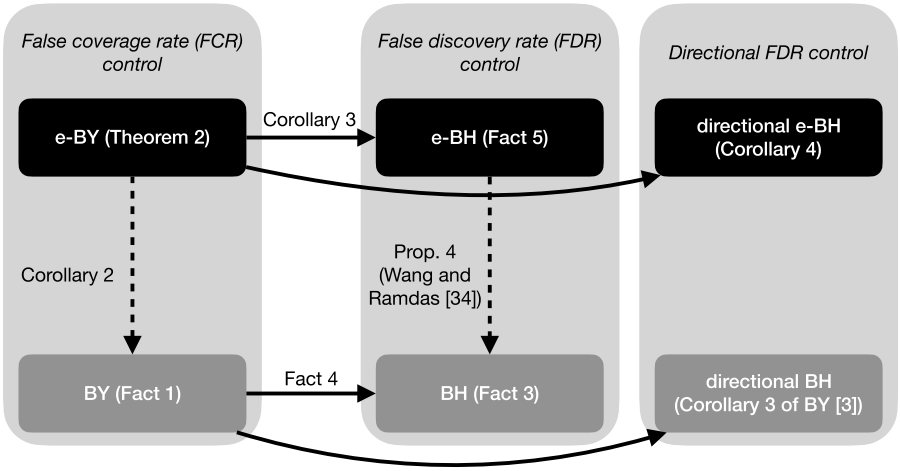

FIG 6. *A diagram depicting the relationships between different procedures and their guarantees for FCR control in post-selection inference, and FDR control in multiple testing. The solid arrows signify that the target result is a special case of the source result. The dashed line signifies that the part of the target result that assumes arbitrary dependence between data is implied by the source result. This diagram illustrates how multiple testing procedures with FDR guarantees are downstream consequence of procedures for the post-selection inference problem with FCR guarantees.*

In Figure 6, we depict how the e-BY and BY procedures subsume results about FDR controlling multiple testing procedures as a special case. The multiple testing analog of the BY procedure is the Benjamini-Hochberg (BH) procedure [2, 3], and BY showed that FCR control of the BY procedure implies FDR control of the BH procedure. The same implication between the e-BY procedure and the e-BH procedure [32]. Further, the e-BY procedure is not a naive application of the e-BH procedure to the post-selection inference problem. In fact, the e-BY procedure is a more powerful result than the e-BH procedure in the sense that Theorem 2 implies FDR control for the e-BH procedure while the converse is not true.

B.1. Benjamini-Hochberg (BH) is a special case of BY

BY showed that their procedure can be used to demonstrate FDR validity (i.e. provably controls FDR under some fixed level δ) of the BH procedure. However, the BH procedure does not directly operate on CIs — instead it takes as input *p-values* i.e. P_i denotes the p-value for the ith null hypothesis for each $i \in [K]$. P_i the superuniformity property under the null: $\mathbb{P}(P_i \leq \alpha) \leq \alpha$ for all $\alpha \in (0,1)$ if $\theta^*_{i^*} \in \Theta^0_i$.

The BH algorithm proceeds as follows. Let $P_1, \ldots, P_K$ be the set of p-values computed from $X_1, \ldots, X_K$ for each hypothesis, and let $P_{(1)}, P_{(2)}, \ldots, P_{(K)}$ denote these p-values ordered from smallest to largest. Define a quantity $D(K) :=$

K if the p-values corresponding to true null hypotheses are independent, or satisfy a positive dependence condition (see [3]) and $D(K) \coloneqq K\ell_K$ otherwise.

$$k_{\mathrm{BH}} \coloneqq \max\left\{k \in [K] : P_{(k)} \leq \delta k / D(K)\right\} \cup \{0\},$$
$$\mathcal{R}_{\mathrm{BH}} \coloneqq \left\{i \in [K] : P_i \leq \delta k_{\mathrm{BH}} / D(K)\right\}.$$

Fact 3 (BH procedure controls FDR [2, 3])**.** *The BH procedure, which rejects the hypotheses indexed by* $\mathcal{R}_{\mathrm{BH}}$*, ensures* $\mathrm{FDR} \leq \delta$*.*

BY observed that the BH procedure can actually be formulated completely in terms of CIs and the BY procedure in the following fashion.

Fact 4 (Corollary 2 of BY)**.** *Let* $\mathcal{S} = \mathcal{R}_{\mathrm{BH}}$*, and define the CIs as follows:*

$$C_i(\alpha) = \begin{cases} \Theta_i \setminus \Theta_i^0 & \text{if } P_i \leq \alpha \\ \Theta_i & \text{if } P_i > \alpha \end{cases}.$$

Recall that, by definition of $\mathcal{R}_{\mathrm{BH}}$*,* $P_i \leq \delta k_{\mathrm{BH}}/D(K)$ *for all* $i \in \mathcal{R}_{\mathrm{BH}}$*. When we apply the BY procedure, we construct exactly* $C_i(\delta k_{\mathrm{BH}}/D(K)) = \Theta_i \setminus \Theta_i^0$ *for each* $i \in \mathcal{R}_{\mathrm{BH}}$*. As a result, a false discovery is made for the* i*th null hypothesis* $(\theta_{i^*}^* \in \Theta_0)$ *if and only if* $\theta_{i^*}^* \notin C_i(\delta k_{\mathrm{BH}}/K)$*. Hence* $\mathrm{FDP} = \mathrm{FCP}$ *and* $\mathrm{FDR} = \mathrm{FCR}$*. Consequently, the FCR guarantee of the BY procedure in Fact 1 implies the BH guarantee of* $\mathrm{FDR} \leq \delta$ *in Fact 3.*

Clearly, the BY procedure is a more powerful and general technique than the BH procedure, which is essentially the BY procedure for a specific choice of CI and selection rule. Hence, a guarantee on FDR by the BH procedure does not imply any guarantee on the FCR the BY procedure, while the reverse implication *does* hold. We will see this paradigm also play out between the e-BY procedure and its multiple testing analog, the e-BH procedure.

B.2. e-BH is a special case of e-BY

We will show this implication by reducing the multiple testing problem to an instance of the post-selection inference problem. The e-BH procedure takes e-values $\mathbf{E} \coloneqq (E_1, \ldots, E_K)$ as input, where E_i is an e-value w.r.t. Θ_i^0 for each $i \in [K]$. Note that both e-BH and e-BY impose no requirements about the dependence between the e-values and e-CIs, respectively. Hence, $E_1, \ldots, E_K$ may be arbitrarily dependent.

Wang and Ramdas [32] show that FDR is controlled not only for e-BH, but for an entire class of "self-consistent" procedures. For a multiple testing procedure $\mathcal{D} : [0, \infty)^K \mapsto 2^{[K]}$ that takes e-values and outputs a rejection set $\mathcal{R}$, $\mathcal{D}$ is *self-consistent* [6] if and only if it satisfies the following property:

$$\text{For any realization of e-values } \mathbf{E} \in [0,\infty)^K, E_i \geq \frac{K}{\delta|\mathcal{D}(\mathbf{E})|} \text{ for every } i \in \mathcal{D}(\mathbf{E}). \tag{20}$$

The e-BH procedure outputs the largest self-consistent set — consequently, any other self-consistent set will be a subset of the rejection set output by the e-BH procedure.

Fact 5 (FDR control of self-consistent procedures [32])**.** *Any self-consistent procedure $\mathcal{D}$ that operates on e-values, i.e., that satisfies* (20)*, ensures* $\mathrm{FDR} \leq \delta$.

Corollary 3 (FCR control for e-BY implies FDR control for any self-consistent procedure)**.** *Let $\mathcal{D}$ be a self-consistent multiple testing procedure, and define the selected parameters to correspond to the rejection set: $\mathcal{S} = \mathcal{R} = \mathcal{D}(\mathbf{E})$. Define a family of e-values $E_i^{\mathrm{ind}}(\theta) := E_i \cdot \mathbf{I}\left\{\theta \in \Theta_i^0\right\}$ and let the e-CIs corresponding to this family be:*

$$C_i^{\mathrm{ind}}(\alpha) := \left\{\theta \in \Theta_i : E_i^{\mathrm{ind}}(\theta) \leq \frac{1}{\alpha}\right\} = \begin{cases} \Theta_i \setminus \Theta_i^0 & \text{if } E_i > 1/\alpha \\ \Theta_i & \text{if } E_i \leq 1/\alpha \end{cases}.$$

The e-BY procedure outputs $C_i^{\mathrm{ind}}(\delta|\mathcal{D}(\mathbf{E})|/K) = \Theta_i \setminus \Theta_i^0$ for each $i \in \mathcal{D}(\mathbf{E})$. Thus, the e-BH procedure makes a false discovery if and only if the e-BY procedure does not cover a CI in $\mathcal{S}$. Hence, we have FDR control of the e-BH procedure through Theorem 2.

Evidently, the e-BY procedure is more general than the e-BH procedure. We show in Appendix B.3 that we can use the e-BY procedure to also get control on a directional form of FDR that requires the procedure to output a sign or direction along with each rejection.

B.3. Controlling the directional false discovery rate dFDR

We can also use the the e-BY procedure to provide guarantees beyond FDR for e-BH. Consider a specialized case of the multiple testing problem, where a scientist is performing only two-sided hypothesis tests e.g. discovering how the presence of certain genotypes have a significant effect on the baseline level of a certain hormone. The scientist may not have an a priori idea of which direction (positive or negative) the effect on a hormone a genotype could have. Thus, when rejecting the null hypothesis and making a discovery for some genotype-hormone interaction, it would also be invaluable to produce an estimate of the *direction* of the effect. This problem of determining the direction along with significance was initiated by Bohrer [7], Bohrer and Schervish [8], and Hochberg [13]. Building on this, BY showed that the BH procedure could ensure control of a directional version of the FDR when the operating on p-values for two-sided hypothesis tests through their FCR result for the BY procedure.

In this section, we describe how we can use e-BY to develop a direction determining version of e-BH, and also ensure control of a directional version of the FDR. For a notion of direction to be well-defined, we consider only parameter spaces where $\Theta_i \subseteq \mathbb{R}$, and for each parameter θ_i^*, we are testing the null hypothesis $H_i^0 : \theta_i^* = 0$. In addition to outputting a rejection set, we also require

the multiple testing procedure to also assign a direction $D_i \in \{\pm 1\}$ for each $i \in \mathcal{R}$. Thus, we can define the directional FDR (dFDR):

$$\text{dFDR} \coloneqq \mathbb{E}\left[\frac{\sum_{i\in\mathcal{R}} \mathbf{I}\{D_i = 1, \theta_i^* \leq 0\} + \mathbf{I}\{D_i = -1, \theta_i^* \geq 0\}}{|\mathcal{R}|}\right].$$

This definition of dFDR coincides with the definition of mixed directional FDR defined in BY. The dFDR is also nearly the same as the *false sign rate* (FSR) that is analyzed in Stephens [27], Weinstein and Ramdas [38] and the "pure directional FDR" discussed in BY. However, we will only consider the dFDR since it is always larger than the other two directional error metrics. Consequently, these all of these metrics are equivalent when no θ_i^* is exactly 0 — Tukey [28] argues happens this occurs in most realistic scenarios as many parameters can be off from 0 by an imperceptible amount so the distinction between these metrics may not be practically critical. We will also require that the input e-values are "two-sided" e-values. Specifically, let $E_i \coloneqq (E_i^+ \vee E_i^-)/2$, where E_i^+ is an e-value w.r.t. $[0, \infty)$ and E_i^- is an e-value w.r.t. $(-\infty, 0]$, and two constituent e-values are inverses of each other $E_i^+ = (E_i^-)^{-1}$.

Corollary 4 (e-BH with two-sided e-values controls dFDR)**.** *Let the selection set still be the output of the e-BH procedure, i.e.,* $\mathcal{S} = \mathcal{R}_{\text{eBH}}$*. In addition, define the following direction decisions for each* $i \in \mathcal{S}$*:*

$$D_i \coloneqq \begin{cases} -1 & \text{if } E_i^+ \geq 2K/(\delta k_{\text{eBH}}) \\ 1 & \text{else.} \end{cases}$$

Define a family of e-values as so: $E^{\text{dir}}(\theta) \coloneqq E_i^+ \mathbf{I}\{\theta \geq 0\} + E_i^- \mathbf{I}\{\theta \leq 0\}$*. We derive the following e-CI procedure from this family:*

$$C_i^{\text{dir}}(\alpha) \coloneqq \begin{cases} (-\infty, 0) & \text{if } E_i^+ \geq 2/\alpha \\ (0, \infty) & \text{if } E_i^- \geq 2/\alpha \\ \mathbb{R} & \text{else.} \end{cases}$$

Consequently, miscoverage occurs if and only if the direction of θ_i^* *for each* $i \in \mathcal{S}$*. Thus, the FCR guarantee from Theorem 2 ensures that* $\text{dFDR} \leq \delta$*.*

As a result, we have shown that e-BY procedure can be used to provide results for more general problems than just e-BH in the multiple testing scenario — e-BY also provides dFDR control in this directional variant of the multiple testing problem. Thus, the FCR control provided by e-BY allows it to be useful, general tool for providing error control guarantees in a variety of problems.

Appendix C: Weighted e-BY: Weights for individual control of the size of each e-CI

The e-BY procedure in Definition 6 assigns a equal confidence level to the CI constructed for each θ_i^*, but there may be cases where the user would desire

Algorithm 1: The weighted e-BY procedure for constructing post-selection CIs with FCR control. Let $w_1, \ldots, w_K$ be nonnegative weights that sum to K. The resulting constructed CIs have FCR $\leq \delta$, where $\delta \in (0,1)$ is a predetermined level of error.

1 Sample data $\mathbf{X} := (X_1, \ldots, X_K)$ for each parameter being estimated — these may be dependent.
2 Let C_i produce *marginal e-CIs* at confidence level $(1-\alpha)$ for any $\alpha \in (0,1)$ for each $i \in [K]$.
3 Select a subset of parameters based on the data for CI construction: $\mathcal{S} := \mathfrak{S}(\mathbf{X})$.
4 Set $\alpha_i := w_i \delta |\mathcal{S}|/K$ for each $i \in \mathcal{S}$.
5 Construct the CI $C_i(\alpha_i)$ for θ_i^* for each $i \in \mathcal{S}$.

tighter CIs for some parameters, and looser CIs for others. For example, the parameters we are estimating may be the severity of different side effects caused by a drug candidate. In that situation, we want to estimate more precisely the side effects that are life threatening or severely injurious to the recipient of the drug, and be willing to have larger CIs for estimating side effects that only cause minor ailments. We present a simple version of the weighted e-BY procedure where the weights are fixed beforehand. However, Ignatiadis, Wang and Ramdas [17] study the weighted e-BY procedure much more deeply and show that, surprisingly, the weights do not need to be normalized to sum to K — one may require the weights to be e-values themselves and still achieve valid FCR control. Further, they show that this approach of using e-values as weights is applicable to a weighted form of the BY procedure as well.

Let $w_1, \ldots, w_K$ be nonnnegative weights assigned to each of the K parameters and let their sum be bounded by K. We can now prove a FCR guarantee about a weighted version of e-BY where the confidence level corresponding to the CI constructed for each θ_i^* is $1 - w_i \delta |\mathcal{S}|/K$.

Theorem 5 (Weighted e-BY controls FCR)**.** *Let $w_1, \ldots, w_K$ be a set of nonnegative weights that sum to at most K. The weighted e-BY procedure formulated in Algorithm 1 ensures* FCR $\leq \delta$.

Proof. The proof follows similarly to that of Theorem 2.

$$\begin{aligned}
\mathrm{FCR} &= \mathbb{E}\left[\frac{\sum_{i\in\mathcal{S}} \mathbf{I}\left\{\theta_i^* \notin C_i\left(w_i\delta|\mathcal{S}|/K\right)\right\}}{|\mathcal{S}| \vee 1}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i\in[K]} \mathbf{I}\left\{E_i(\theta_i^*) w_i \delta |\mathcal{S}|/K > 1\right\} \cdot \mathbf{I}\left\{i \in S\right\}}{|\mathcal{S}| \vee 1}\right] \\
&\leq \sum_{i\in[K]} \mathbb{E}\left[\frac{E_i(\theta_i^*) w_i \delta |\mathcal{S}|}{K(|\mathcal{S}| \vee 1)}\right] = \sum_{i\in[K]} \frac{w_i\delta}{K} \mathbb{E}\left[E_i(\theta_i^*) \cdot \frac{|\mathcal{S}|}{|\mathcal{S}| \vee 1}\right] \leq \delta.
\end{aligned}$$

Here, the last inequality is a consequence of $\sum_{i=1}^{K} w_i \leq K$. □

Appendix D: Improving e-BY when a CI-reporting condition is relaxed

We briefly explain the reason requiring strictly decreasing components of $\mathbf{C} \in \mathrm{ECI}_K^*$ for strict domination in Section 5.2. Suppose that $K = 2$, $S = \{1, 2\}$ and $C_1(0) = C_1(1)$ (i.e., we are certain about θ_1 in a subset of Θ_1, but no further information), and C_2 is strictly decreasing. The e-BY procedure reports $(C_1(\delta), C_2(\delta))$. Consider another procedure Φ which reports $(C_1(0), C_2(\delta+\epsilon)) \subsetneq (C_1(\delta), C_2(\delta))$ for this specific (C_1, C_2) and behaves as e-BY for all other CIs. Clearly, this procedure strictly dominates e-BY. Intuitively, it should be possible to choose $\epsilon > 0$ small enough so that the FCR of Φ is less than or equal to δ (but we did not find a proof). Generally, having a constant $C_i(\alpha)$ for α in an interval leads to the above situation. This situation is not interesting for practice as often an e-CI is strictly decreasing. but it becomes relevant in case the parameter θ_i^* takes finitely many values. We discuss how we could improve on the e-BY procedure if the CIs are not strictly decreasing (and restriction (b) is lifted) in Appendix D.

In this section we explore a minor improvement of e-BY by relaxing the restriction (b) in Section 5.2 in the formulation of a CI reporting procedure, as well as the strict decreasing condition of the e-CIs.

For $(\mathcal{S}, \mathbf{C}) \in 2^{[K]} \times \mathrm{ECI}_K$ where $S \neq \emptyset, [K]$, we define

$$\begin{aligned}
\gamma(\mathcal{S}, \mathbf{C}) &:= \sum_{i \in \mathcal{S}} \sup_{\theta \in \Theta_i} t_i(\theta), \\
\lambda(\mathcal{S}, \mathbf{C}) &:= \sum_{i \in [K] \setminus S} \inf_{\theta \in \Theta_i} \left(t_i(\theta) \mathbf{I}\left\{ t_i(\theta) > 1 \right\} \right), \\
v(\mathcal{S}, \mathbf{C}) &:= 1 + \frac{\lambda(\mathcal{S}, \mathbf{C})}{\gamma(\mathcal{S}, \mathbf{C})},
\end{aligned}$$

where t_i is given in (12), and we set $\infty/\infty = 0$. Note that $\gamma(\mathcal{S}, \mathbf{C}) < \infty$ if and only if for each $i \in S$ there exists $\alpha_i > 0$ such that $C(\alpha_i) = \Theta_i$; similarly, $\lambda(\mathcal{S}, \mathbf{C}) > 0$ if and only if there exists $i \in [K] \setminus S$ such that $C_i(\alpha_i) = \emptyset$ for some $\alpha_i \in (0, 1)$. The condition $v(\mathcal{S}, \mathbf{C}) > 1$ is equivalent to $\gamma(\mathcal{S}, \mathbf{C}) < \infty$ and $\lambda(\mathcal{S}, \mathbf{C}) > 0$. (Admittedly, this is an uncommon situation.)

For a level $w \in (1, \infty)$, we denote by

$$\Sigma(w) := \{(\mathcal{S}, \mathbf{C}) \in 2^{[K]} \times \mathrm{ECI}_K : v(\mathcal{S}, \mathbf{C}) \geq w\}.$$

One can check that the set $\Sigma(w)$ is not empty. In particular, if $S = \{1\}$ and $\mathbf{C}$ is the constant CI of $\Theta_1 \times \emptyset^{K-1}$ (i.e., no information for the first one, and sure false coverage for all others), then $t_1(\theta) = 1$ for all $\theta \in \Theta$, and $t_i(\theta) = \infty$ for all $\theta \in \Theta_i$ and $i \neq 1$. In this case, $\gamma(\mathcal{S}, \mathbf{C}) = 1$, $\lambda(\mathcal{S}, \mathbf{C}) = \infty$, and $v(\mathcal{S}, \mathbf{C}) = \infty$.

For a given level $w \in (1, \infty)$, define a CI reporting procedure Φ by

1. if the input $(\mathcal{S}, \mathbf{C})$ is in $\Sigma(w)$, then report e-BY at level $w\delta$;
2. otherwise, report e-BY at level δ.

Clearly, the above procedure dominates e-BY at level δ, and the domination is strict noting that $\Sigma(w)$ is not empty. If we allow $w = 1$, then it is precisely the e-BY procedure.

Proposition 4. *The above procedure Φ has FCR at most δ.*

Proof. Let $\mathcal{S}$ be a random selection, $\mathbf{E}$ be the vector of e-values, and $\boldsymbol{\theta}^*$ be the true parameter. Denote by $A = \{(\mathcal{S}, \mathbf{C}^{\mathbf{E}}) \in \Sigma(w)\}$, that is, the event that $v(\mathcal{S}, \mathbf{C}^{\mathbf{E}}) \geq w$. Note that by Lemma 2, if $\mathbf{C}^{\mathbf{E}} = \mathbf{C}$ occurs, then $E_i(\theta) = t_i(\theta)$ for all $\theta \in \Theta_i$ and $i \in [K]$ whenever $E_i(\theta) > 1$, where t_i is given by (12), that is,

$$t_i(\theta) = \sup\left\{\frac{1}{\alpha} : \alpha \in [0,1),\ \theta \notin C_i(\alpha)\right\}.$$

In case $(\mathcal{S}, \mathbf{C}^{\mathbf{E}}) = (\mathcal{S}, \mathbf{C}) \in \Sigma(w)$, we have

$$\begin{aligned}
&\sum_{i\in S}(w-1)E_i(\theta_i) - \sum_{i\in[K]\setminus S} E_i(\theta_i) \\
&\leq \sum_{i\in S}(w-1)\sup_{\theta_i\in\Theta_i} t_i(\theta_i) - \sum_{i\in[K]\setminus S}\inf_{\theta_i\in\Theta_i} t_i(\theta_i)\mathbf{I}\{t_i(\theta_i) > 1\} \leq 0.
\end{aligned}$$

Therefore, we have

$$\left(\sum_{i\in\mathcal{S}}(w-1)E_i(\theta_i) - \sum_{i\in[K]\setminus\mathcal{S}} E_i(\theta_i)\right)\mathbf{I}\{A\} \leq 0. \tag{21}$$

Using the above facts, we can compute the following upper bound on the FCR:

$$\begin{aligned}
&\mathrm{FCR}(\Phi) \\
&= \mathbb{E}\left[\frac{\sum_{i\in\mathcal{S}}\mathbf{I}\{\theta_i \notin \Phi(\mathcal{S}, \mathbf{C}^{\mathbf{E}})\}}{|\mathcal{S}| \vee 1}\right] \\
&= \mathbb{E}\left[\frac{\sum_{i\in\mathcal{S}}\mathbf{I}\{E_i(\theta_i) \geq K/(w\delta|\mathcal{S}|)\}}{|\mathcal{S}| \vee 1}\mathbf{I}\{A\}\right] \\
&\quad + \mathbb{E}\left[\frac{\sum_{i\in\mathcal{S}}\mathbf{I}\{E_i(\theta_i) \geq K/(\delta|\mathcal{S}|)\}}{|\mathcal{S}| \vee 1}(1 - \mathbf{I}\{A\})\right] \\
&\leq \mathbb{E}\left[\frac{1}{|\mathcal{S}| \vee 1}\sum_{i\in\mathcal{S}}\frac{w\delta|\mathcal{S}|E_i(\theta_i)}{K}\mathbf{I}\{A\}\right] + \mathbb{E}\left[\frac{1}{|\mathcal{S}| \vee 1}\sum_{i\in\mathcal{S}}\frac{\delta|\mathcal{S}|E_i(\theta_i)}{K}(1 - \mathbf{I}\{A\})\right] \\
&\leq \mathbb{E}\left[\frac{1}{K}\sum_{i\in\mathcal{S}} w\delta E_i(\theta_i)\mathbf{I}\{A\}\right] + \frac{\delta}{K}\mathbb{E}\left[\sum_{i\in[K]} E_i(\theta_i)(1 - \mathbf{I}\{A\})\right].
\end{aligned}$$

Now, we can group the terms that involve $\mathbf{I}\{A\}$ together to simplify our bound on the FCR, and piece together the term we see in (21).

$$\mathrm{FCR}(\Phi) \leq \frac{\delta}{K}\sum_{i\in[K]}\mathbb{E}\left[E_i(\theta_i)\right] + \frac{\delta}{K}\mathbb{E}\left[\left(\sum_{i\in\mathcal{S}} wE_i(\theta_i) - \sum_{i\in[K]} E_i(\theta_i)\right)\mathbf{I}\{A\}\right]$$

$$\leq \delta + \frac{\delta}{K}\mathbb{E}\left[\left(\sum_{i\in\mathcal{S}}(w-1)E_i(\theta_i) - \sum_{i\in[K]\setminus\mathcal{S}} E_i(\theta_i)\right)\mathbf{I}\{A\}\right] \leq \delta,$$

where in the last inequality we used (21). Therefore, $\mathrm{FCR}(\Phi) \leq \delta$. □

We note that this improvement is not very useful as $v(\mathcal{S}, \mathbf{C}) > w$ is a strong requirement and it is only satisfied by limited choices of the input.

Moreover, we clearly see how this procedure uses the information of $\mathbf{C}_i$ for $i \notin \mathcal{S}$, which is used to compute $\lambda(\mathcal{S}, \mathbf{C})$, thus violating the requirement (b) in Section 5.2. Moreover, all input CIs in $\Sigma(w)$ are not strictly decreasing since we need $C_i(\alpha_i) = \Theta_i$ for some $\alpha_i > 0$.

Appendix E: Generalized e-CIs and deriving Chernoff CIs

Extending the notion of e-CI, we can also define a generalized e-CI from a family of e-values that are not only parameterized by a parameter, θ, but also $\alpha' \in (0,1)$ indicating the desired level at which the e-CI is most tight at. Thus, for a family of e-values $E(\theta, \alpha')$ for $\theta \in \Theta$ and $\alpha' \in (0,1)$, we can define the following generalized e-CI:

$$C^{\alpha'}(\alpha) \coloneqq \left\{\theta \in \Theta : E(\theta, \alpha') < \frac{1}{\alpha}\right\}.$$

Note that α' is fixed for the e-CI across all choices of $\alpha \in (0,1)$. Generalized e-CIs are particularly useful in situations where we have a collection of e-CIs for the same parameter space Θ, but may be tighter or looser at different values of α. Hence, it is a way of grouping such e-CIs together in one object, and also clarifying that α' is a fixed parameter for each e-CI in the generalized e-CI and *does not change* with the confidence level $1-\alpha$.

Such a situation arises in the case where our e-values are derived from Chernoff bounds. Consider the case where we draw n i.i.d. samples $X_1, \ldots, X_n$ from a distribution that is bounded in $[0,1]$. Let $\mu \coloneqq \mathbb{E}[X_i]$ — we can derive the following inequality using Hoeffding's lemma:

$$\mathbb{E}[\exp(\lambda(X_i - \mu)] \leq \exp\left(\frac{\lambda^2}{8}\right).$$

Let $\widehat{\mu}_n$ denote the sample mean. Consequently, we can construct the following generalized e-CI from this bound:

$$C^{\alpha'\text{-Hoef}}(\alpha) \coloneqq \left(\widehat{\mu}_n \pm \sqrt{\frac{\log(2/\alpha)}{2n}} \cdot \frac{\log(2/\alpha) + \log(2/\alpha')}{2\sqrt{\log(2/\alpha')\log(2/\alpha)}}\right).$$

In contrast, the typical CI constructed from inverting Hoeffding's inequality is as follows:

$$C^{\mathrm{Hoef}}(\alpha) \coloneqq \left(\widehat{\mu}_n \pm \sqrt{\frac{\log(2/\alpha)}{2n}}\right) = C^{\alpha\text{-Hoef}}(\alpha).$$

Consider the use of $C^{\alpha'\text{-Hoef}}$ and C^{Hoef} in the e-BY and BY procedures, respectively. As a heuristic, we set $\alpha' = \delta$. For the selection set $\mathcal{S}$, and the same desired level of FCR control δ, we get the following intervals:

$$C^{\delta\text{-Hoef}}\left(\frac{\delta|\mathcal{S}|}{K}\right) = \left(\widehat{\mu}_n \pm \sqrt{\frac{\log\left(\frac{2K}{\delta|\mathcal{S}|}\right)}{2n}} \cdot \frac{\log\left(\frac{2}{\delta}\right) + \log\left(\frac{2K}{\delta|\mathcal{S}|}\right)}{2\sqrt{\log\left(\frac{2}{\delta}\right)\log\left(\frac{2K}{\delta|\mathcal{S}|}\right)}}\right),$$

$$C^{\text{Hoef}}\left(\frac{\delta|\mathcal{S}|}{K\ell_K}\right) = \left(\widehat{\mu}_n \pm \sqrt{\frac{\log\left(\frac{2K\ell_K}{\delta|\mathcal{S}|}\right)}{2n}}\right).$$

This results in the following result about $\mathcal{S}$.

Proposition 5. *$C^{\delta\text{-Hoef}}(\delta|\mathcal{S}|/K)$ is as tight as $C^{\text{Hoef}}(\delta|\mathcal{S}|/(K\ell_K))$ when*

$$|\mathcal{S}| \geq \frac{K}{\exp\left(2\sqrt{\log\left(\frac{2}{\delta}\right)\log\ell_K}\right)}.$$

Proof. We note that $C^{\delta\text{-Hoef}}\left(\frac{\delta|\mathcal{S}|}{K}\right)$ is tighter than $C^{\text{Hoef}}\left(\frac{\delta|\mathcal{S}|}{K\ell_K}\right)$ when:

$$\log\left(\frac{2}{\delta}\right) + \log\left(\frac{2K}{\delta|\mathcal{S}|}\right) \leq 2\sqrt{\log\left(\frac{2}{\delta}\right)\log\left(\frac{2K\ell_K}{\delta|\mathcal{S}|}\right)}.$$

This is an inequality on a quadratic expression w.r.t. $\log|\mathcal{S}|$. Hence, we can rearrange and solve by quadratic formula. This gets us the following interval for $\log|\mathcal{S}|$ where the above inequality holds:

$$\log|\mathcal{S}| \in \left[\log K \pm 2\sqrt{\log\left(\frac{2}{\delta}\right)\log\ell_K}\right].$$

Since we know $|\mathcal{S}| \leq K$, we are only interested in the lower bound on this interval, and we get our desired result from the lower bound. □

In Proposition 5, we see that the lower bound for $|\mathcal{S}|$, as a proportion of K, decreases as K increases. Thus, $C^{\delta\text{-Hoef}}(\delta|\mathcal{S}|/K)$ is tighter than $C^{\text{Hoef}}(\delta|\mathcal{S}|/(K\ell_K))$ for an increasing proportion of possible values of $|\mathcal{S}|$ as K grows. We see this advantage of the e-BY procedure over the BY procedure reflected in our simulation results in Section 6 — the CIs output by the e-BY procedure are tighter than the CIs of the BY procedure (under no assumptions on selection rule or dependence) as K grows.

Appendix F: FCR and other measures of statistical validity

We will elaborate on the points made in Section 1 about the other error metrics one may consider applicable to the post-selection inference problem.

Conditional coverage requires knowing about selection rule a priori Before discussing FCR control, we can first consider if a coverage guarantee is possible when we condition on a parameter being selected. In the majority of this paper, we have assumed that each C_i is a *marginal* CI, i.e., for each $i \in [K]$, $\mathbb{P}(\theta_i^* \in C_i(\alpha)) \geq 1-\alpha$ for every $\alpha \in (0,1)$. For conditional coverage, we want a confidence level $(1-\alpha_i)$, for each $i \in [K]$, such that the resulting CI, $C_i(\alpha_i)$, ensures $\mathbb{P}(\theta_i^* \in C_i(\alpha_i) \mid i \in \mathcal{S}) \geq 1-\delta$. However, BY noted that no such correction can be made without knowing the selection rule. To illustrate this, consider the following toy example. Assume that the parameters being estimated are means, and the data distribution for each parameter is Gaussian with variance 1, i.e., $X_i \sim \mathcal{N}(\theta_i^*, 1)$. Then, the CI for the ith parameter is $C_i(\alpha) = [X_i - z_{\alpha_i/2}, X_i + z_{\alpha_i/2}]$ where z_α is the $(1-\alpha)$-quantile of the standard normal distribution. Now, select the ith parameter only if $C_i(\alpha_i)$, the corrected CI, covers solely positive values. If $\theta_i^* \leq 0$, the conditional coverage is 0. Thus, no guarantees can be made about conditional coverage without knowing or constraining the selection rule.

Note that if we consider a constraint on the joint probability of a parameter being selected and miscovered instead, i.e., $\mathbb{P}(\theta_i^* \notin C_i(\alpha_i), i \in \mathcal{S}) \leq \alpha$, then the CI guarantee itself ensures that this constraint is satisfied, i.e., $\mathbb{P}(\theta_i^* \notin C_i(\alpha_i)) \leq \alpha$. The fact that we may bound this joint probability is what motivates the possibility of the use of FCR as a error metric we can control.

In addition, the refutation of conditional coverage guarantees is intended for the general setting, where there is no limitation of the choice of the selection rule. Thus, methods that do provide conditional coverage guarantees must depend on the specific selection rule being used. Zhong and Prentice [40] and Weinstein, Fithian and Benjamini [37] follow this line of inquiry and derive explicit CIs that have the conditional coverage guarantee under specific selection rules and assumptions on the families of parameters. The difficulty in this approach is that the results provided are narrowly applicable to the predetermined selection rule and data distribution families — it requires the user to derive new conditional CIs for different combinations of selection rules and family of distributions. Weinstein and Ramdas [38] come to a similar conclusion about the difficulty of using conditional CIs in an online version of the post-selection inference problem, and also recommend the marginal CI approach used in both BY and e-BY. Further, they note that an additional deficit of the conditional coverage approach is that the conditionally valid CIs are not necessarily "consistent" with the selection procedure. To refer to the toy example, the conditional CIs for a selection rule that aims to only select parameters that have CIs that are completely positive must also include negatives (otherwise the toy example still holds). In this sense, the produced conditional CIs are inconsistent with the goal of the selection rule. Hence, FCR control circumvents these challenges, at the cost of providing a less powerful error guarantee.

FCR control results in smaller CIs than simultaneous coverage Another type of error control to consider is a simultaneous guarantee over the CIs of all selected parameters, i.e., $\mathbb{P}(\forall i \in \mathcal{S} : \theta_i^* \in C_i(\alpha_i)) \geq 1-\delta$, for some choices

of $\alpha_i \in [0,1]$ for each $i \in \mathcal{S}$. This error is called the *simultaneous over selected (SoS)* criterion in Benjamini, Hechtlinger and Stark [1] — they provide methods for controlling the SoS for a few, specific selection rules. Note that control of the SoS criterion implies control of the FCR, but not necessarily the other way around. Hence, SoS control is strictly stronger than FCR control. The only known way to achieve this simultaneous coverage for arbitrary selection rules, however, is to make a Bonferroni correction and ensure simultaneous coverage of all the CIs, i.e., for both selected and non-selected parameters. If all confidence levels are equivalent, this implies $\alpha_i = \delta/K$. This is a much more conservative level of correction than the BY and e-BY procedures.

Acknowledgments

We would like to thank Umashanthi Pavalanathan and Luke Sonnet for insightful discussions about the application of post-selection inference to A/B testing and their help in gathering and releasing the data from Twitter used in our experiments.

Supplementary Material

Code for simulations and figures

(doi: 10.1214/24-EJS2253SUPP; .zip). The ZIP file in Xu et al. [39] contains all the code for the methods in this paper, as well as the code to run the simulations and generate the plots.

References

[1] Benjamini, Y., Hechtlinger, Y. and Stark, P. B. (2019). Confidence Intervals for Selected Parameters. arXiv:1906.00505.

[2] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57** 289–300. MR1325392

[3] Benjamini, Y. and Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics* **29** 1165–1188. MR1869245

[4] Benjamini, Y. and Yekutieli, D. (2005). False Discovery Rate–Adjusted Multiple Confidence Intervals for Selected Parameters. *Journal of the American Statistical Association* **100** 71–81. MR2156820

[5] Bernstein, S. (1927). *Theory of Probability.* Gastehizdat Publishing House, Moscow.

[6] Blanchard, G. and Roquain, E. (2008). Two Simple Sufficient Conditions for FDR Control. *Electronic Journal of Statistics* **2** 963–992. MR2448601

[7] Bohrer, R. (1979). Multiple Three-Decision Rules for Parametric Signs. *Journal of the American Statistical Association* **74** 432–437. MR0548037
[8] Bohrer, R. and Schervish, M. J. (1980). An Optimal Multiple Decision Rule for Signs of Parameters. *Proceedings of the National Academy of Sciences* **77** 52–56. MR0554145
[9] Borodin, A. N. and Salminen, P. (1996). *Handbook of Brownian Motion — Facts and Formulae*. Birkhäuser, Basel. MR1477407
[10] Goeman, J. J. and Solari, A. (2011). Multiple Testing for Exploratory Research. *Statistical Science* **26** 584–597. MR2951390
[11] Grünwald, P. (2023). Beyond Neyman-Pearson. arXiv:2205.00901.
[12] Grünwald, P., de Heide, R. and Koolen, W. (2024 (forthcoming)). Safe Testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.
[13] Hochberg, Y. (1986). Multiple Classification Rules for Signs of Parameters. *Journal of Statistical Planning and Inference* **15** 177–188. MR0873007
[14] Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association* **58** 13–30. MR0144363
[15] Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2020). Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys* **17** 257–317. MR4100718
[16] Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics* **49** 1055–1080. MR4255119
[17] Ignatiadis, N., Wang, R. and Ramdas, A. (2023). E-Values as Unnormalized Weights in Multiple Testing. *Biometrika*. MR4745574
[18] Koning, N. W. (2024). Post-Hoc p-Values. arXiv:2312.08040.
[19] Maurer, A. and Pontil, M. (2009). Empirical Bernstein Bounds and Sample Variance Penalization. In *Conference on Learning Theory*.
[20] Orabona, F. and Jun, K.-S. (2024). Tight Concentrations and Confidence Sequences from the Regret of Universal Portfolio. *IEEE Transactions on Information Theory* **70** 436–455. MR4692683
[21] Ramdas, A., Ruf, J., Larsson, M. and Koolen, W. (2020). Admissible Anytime-Valid Sequential Inference Must Rely on Nonnegative Martingales. arXiv:2009.03167.
[22] Robbins, H. (1970). Statistical Methods Related to the Law of the Iterated Logarithm. *The Annals of Mathematical Statistics* **41** 1397–1409. MR0277063
[23] Sellke, T., Bayarri, M. J. and Berger, J. O. (2001). Calibration of p Values for Testing Precise Null Hypotheses. *The American Statistician* **55** 62–71. MR1818723
[24] Shafer, G. (2021). Testing by Betting: A Strategy for Statistical and Scientific Communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **184** 407–431. MR4255905
[25] Shafer, G. and Vovk, V. (2019). *Game-Theoretic Foundations for Probability and Finance*, First ed. John Wiley & Sons, Ltd. MR1852450

[26] Shafer, G., Shen, A., Vereshchagin, N. and Vovk, V. (2011). Test Martingales, Bayes Factors and p-Values. *Statistical Science* **26** 84–101. MR2849911

[27] Stephens, M. (2017). False Discovery Rates: A New Deal. *Biostatistics (Oxford, England)* **18** 275–294. MR3824755

[28] Tukey, J. W. (1991). The Philosophy of Multiple Comparisons. *Statistical Science* **6** 100–116.

[29] Vovk, V. G. (1993). A Logic of Probability, with Application to the Foundations of Statistics. *Journal of the Royal Statistical Society. Series B (Methodological)* **55** 317–351. MR1224399

[30] Vovk, V. and Wang, R. (2021). E-Values: Calibration, Combination, and Applications. *Annals of Statistics* **49** 1736–1754. MR4298879

[31] Vovk, V. and Wang, R. (2023). Confidence and discoveries with e-values. *Statistical Science*. MR4596762

[32] 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. MR4460577

[33] Wang, H. and Ramdas, A. (2023). Catoni-Style Confidence Sequences for Heavy-Tailed Mean Estimation. *Stochastic Processes and their Applications* **163** 168–202. MR4610125

[34] Wasserman, L., Ramdas, A. and Balakrishnan, S. (2020). Universal Inference. *Proceedings of the National Academy of Sciences* **117** 16880–16890. MR4242731

[35] Waudby-Smith, I. and Ramdas, A. (2023). Estimating Means of Bounded Random Variables by Betting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. MR4716192

[36] Waudby-Smith, I., Arbour, D., Sinha, R., Kennedy, E. H. and Ramdas, A. (2023). Time-uniform central limit theory and asymptotic confidence sequences. arXiv:2103.06476.

[37] Weinstein, A., Fithian, W. and Benjamini, Y. (2013). Selection Adjusted Confidence Intervals With More Power to Determine the Sign. *Journal of the American Statistical Association* **108** 165–176. MR3174610

[38] Weinstein, A. and Ramdas, A. (2020). Online Control of the False Coverage Rate and False Sign Rate. In *International Conference on Machine Learning*.

[39] Xu, Z., Wang, R. and Ramdas, R. (2024). Supplement to "Post-selection inference for e-value based confidence intervals". https://doi.org/10.1214/24-EJS2253

[40] Prentice, R. L. (2008). Bias-Reduced Estimators and Confidence Intervals for Odds Ratios in Genome-Wide Association Studies. *Biostatistics* **9** 621–634. MR2712234