SEQUENTIAL ANALYSIS
2020, VOL. 39, NO. 1, 65–91
https://doi.org/10.1080/07474946.2020.1726686

Sequential tests of multiple hypotheses controlling false discovery and nondiscovery rates

Jay Bartroff[a] and Jinlin Song[b]

[a]Department of Mathematics, University of Southern California, Los Angeles, California, USA; [b]Analysis Group, Inc., Los Angeles, California, USA

ABSTRACT
We propose a general and flexible procedure for testing multiple hypotheses about sequential (or streaming) data that simultaneously controls both the false discovery rate (FDR) and false nondiscovery rate (FNR) under minimal assumptions about the data streams, which may differ in distribution and dimension and be dependent. All that is needed is a test statistic for each data stream that controls its conventional type I and II error probabilities, and no information or assumptions are required about the joint distribution of the statistics or data streams. The procedure can be used with sequential, group sequential, truncated, or other sampling schemes. The procedure is a natural extension of Benjamini and Hochberg's (1995) widely used fixed sample size procedure to the domain of sequential data, with the added benefit of simultaneous FDR and FNR control that sequential sampling affords. We prove the procedure's error control and give some tips for implementation in commonly encountered testing situations.

ARTICLE HISTORY
Received 21 September 2019
Revised 28 December 2019
Accepted 8 January 2020

KEYWORDS
Generalized likelihood ratio; multiple comparisons; multiple endpoint clinical trials; multiple testing; sequential analysis; sequential hypothesis testing; Wald approximations

SUBJECT CLASSIFICATIONS
62J15; 62L10

1. Introduction

Multiple testing error metrics based on the false discovery proportion, such as its expectation of the false discovery rate (FDR), are widely used in applications involving large or high-dimensional data sets or when many comparisons are needed. These areas include high-throughput gene and protein expression data, brain imaging, and astrophysics; Müller et al. (2007, section 1) give a variety of examples. Since Benjamini and Hochberg's (1995) seminal paper introducing FDR and proving that Simes' (1986) earlier step-up procedure controls FDR, the topic and related problems such as empirical Bayes have been active areas of research; see Efron et al. (2001), Efron and Tibshirani (2002), Genovese and Wasserman (2002), Storey (2002), Newton et al. (2004), Storey et al. (2004), and Cohen and Sackrowitz (2005).

One characteristic of the data in some of the application areas mentioned above is that they arrive sequentially in time, or as data streams. One such area is in certain types of clinical trials, in particular the setting discussed by Berry and Berry (2004) in which treatments are compared on the basis of a long list of adverse events affecting

CONTACT Jay Bartroff bartroff@usc.edu Department of Mathematics, University of Southern California, KAP-104, Los Angeles, CA 90089-2532, USA.
Recommended by B. Chattopadhyay

the patients during the trial; multiple-endpoint clinical trials such as this are discussed in more detail in Sections 2 and 5.3. Other areas of application involving multiple data streams include data from pharmacovigilance drug side effect databases (e.g., Avery et al., 2011), testing for disease clusters over a spatial area (e.g., Sonesson, 2007) and closely related problems of industrial quality control (see Woodall, 2006), and testing for a signal in a noisy image (e.g., Siegmund and Yakir, 2008).

However, the particular needs of sequential data have largely been neglected in the FDR literature, and most papers adopt Benjamini and Hochberg's (1995) starting point, a set of p-values arising from fixed sample size hypothesis tests. Our goal is to introduce an FDR-controlling procedure with as much flexibility as the Benjamini-Hochberg (BH) procedure but tailored for sequential data, by allowing for accept/reject decisions in between sequential sampling of data streams. In Section 3 we introduce such a procedure, which we call the sequential BH procedure that controls FDR as well as its type II analog, the false nondiscovery rate (FNR, both defined below), under independence of data streams and under arbitrary dependence with a small logarithmic inflation of the prescribed values; these results mirror the conditions under which Benjamini and Hochberg (1995) proved FDR control in their original paper. We make minimal assumptions about the data streams, which may differ in distribution and dimension. The only thing the procedure needs is a test statistic for each data stream that controls the conventional type I and II error probabilities; that is, only marginal information about the individual test statistics is needed, and no information or assumptions are required about the joint distribution of the data streams or statistics. Likewise, there are no restrictions on the hypotheses that can be tested (i.e., any combination of simple/composite null and alternative hypotheses) provided that the conventional type I and II error probabilities can be controlled. The procedure can be used with sequential, group sequential, truncated, or other sampling schemes. The simultaneous control of FDR and FNR is a feature of the sequential setting we consider, but if there is a restriction on the maximum sample size of a given stream, it may not be possible to achieve simultaneous FDR and FNR control because the needed error bounds (3.3)–(3.4) may not both be satisfied. For this situation or one in which FNR control is simply not a priority of the statistician, in Section 4 we give a "rejective" version of the procedure that only stops early to reject null hypotheses and explicitly controls FDR but not necessarily FNR. To aid with implementation of either of these procedures, in Section 5 we review how to construct the component sequential tests and give closed-form expressions for the needed critical values in some commonly encountered testing situations. In Section 6 we discuss a simulation study comparing the proposed procedure to its fixed sample size analog, and the article concludes with a discussion of extensions and more suggestions for implementation.

On the one hand, our approach to deriving sequential procedures that control FDR is inspired by recent advances by Sarkar (1998), Benjamini and Yekutieli (2001), and Storey et al. (2004) that broaden the conditions under which the BH procedure controls FDR and that give a better understanding of FDR in general. On the other hand, this work also springs from recent advances (Bartroff and Lai, 2010; Bartroff and Song, 2014, 2015) occurring for sequential procedures controlling familywise error rate (FWER) and other error rates (Bartroff, 2018). Although the sequential BH procedure may appear similar to the sequential procedures of these authors controlling the FWER, the underlying principles of FDR and FWER are fundamentally much different and hence any similarity is only superficial.

A distinct but related sequential multiple testing setup has been considered in recent work by Chen and Arias-Castro (2017) and Javanmard and Montanari (2018). In these papers, a sequence of null hypotheses and their p-values are observed over time and accept/reject decisions are made in a manner that controls FDR. Whereas this setup may be thought of as a single stream of experiments, each one already terminated and a terminal p-value computed, it differs from the current article, which considers multiple streams of data and proposes procedures that decide when to terminate each stream individually.

2. Motivating example: Multiple-endpoint clinical trials

There are economic, administrative, and ethical reasons why data from clinical trials are often analyzed sequentially. Sequential (or group sequential) clinical trials with multiple endpoints, or clinical outcomes of interest represented as hypotheses, are a special case of the general setup that we will address in Section 3.1. Suppose a trial concerns $K \geq 2$ endpoints, each represented by a null hypothesis $H^{(k)}, k \in [K]$, which denotes as $\{1, 2, \ldots, K\}$ throughout. Suppose that patients are evaluated sequentially; the group sequential setting requires only minor modifications mentioned at the end of this paragraph. Measurements or data are taken on the nth patient and we let $X_n^{(k)}$ denote the vector of data on the nth patient concerning the kth endpoint. Certain data points may be relevant for more than one endpoint, so there may be substantial (if not complete) overlap between $X_n^{(k)}$ and $X_n^{(k')}$, say. But because the focus here is on procedures that stop early to accept or reject certain endpoints, we do not assume that $X_n^{(k)}$ and $X_n^{(k')}$ are necessarily identical. Thus, as the trial proceeds, if no endpoints are dropped we observe

$$\begin{array}{lll} X_1^{(1)}, & X_2^{(1)}, & X_3^{(1)}, \\ X_1^{(2)}, & X_2^{(2)}, & X_3^{(2)}, \\ X_1^{(3)}, \text{ then} & X_2^{(3)}, \text{ then} & X_3^{(3)}, \ \ldots \ \text{and soon.} \\ \vdots & \vdots & \vdots \\ X_1^{(K)}, & X_2^{(K)}, & X_3^{(K)}, \end{array} \tag{2.1}$$

If the patients are evaluated in groups, the only notational modification needed is that $X_n^{(k)}$ now denotes the vector of data on the nth group concerning the kth endpoint. The total number of endpoints K may be large, especially with recent advances in genetic testing that allow biomarkers to be included as endpoints. Examples are the adverse event trials discussed by Berry and Berry (2004), mentioned in Section 1.

Another example is the randomized trial described by O'Brien (1984) to compare two diabetes therapies, experimental and conventional, in the form of improvements in nerve function of patients as measured through 34 different electromyographic (EMG) endpoints. In this case, $H^{(k)}$ $(k = 1, \ldots, K = 34)$ is the null hypothesis of no difference between treatment and control in the change (baseline to evaluation) in the kth EMG variable, and $X_n^{(k)}$ is vector of differences in the kth EMG variable between patients in the nth accrued group, which includes patients randomized to both the treatments. Although the trial described by O'Brien (1984) was a fixed sample and hence a single group, a sequential version would result in data of the form (2.1).

In the previous example $X_n^{(1)}, X_n^{(2)}, \ldots, X_n^{(K)}$ are vectors of the same length (i.e., the number of patients in the nth group, or perhaps summary statistics thereof), but this is not a requirement of the procedures proposed below in which these can be of arbitrary size and shape and, further, may be dependent. This feature may be particularly useful in multiple-endpoint clinical trials wherein data associated with different endpoints may be of different size but also are likely to be correlated because they are measurements on the same patient. For example, in clinical trials for AIDS treatments, it is common (e.g., Fischl et al., 1987) to have multiple endpoints of both the continuous and categorical types, like CD4 (T-cell) level, which is commonly modeled as a normal random variable, and the binary indicator of opportunistic infectious disease like pneumonia, modeled as a Bernoulli random variable. For a CD4 endpoint, $X_n^{(k)}$ may be the difference, after minus before, in CD4 count for the nth patient, modeled as normally distributed with unknown mean and variance μ and σ^2, with associated endpoint $H^{(k)} : \mu \leq 0$ of no positive treatment effect on CD4 count, versus the alternative $\mu \geq \delta$, where $\delta > 0$ is some minimal meaningful treatment effect. For an opportunistic infectious disease endpoint, $X_n^{(k)}$ may be the indicator of the disease in the nth patient, modeled as a Bernoulli random variable taking the value 1 with unknown probability p, with associated endpoint $H^{(k)} : p \leq p_0$ where p_0 is some baseline rate of disease occurrence in healthy patients, versus the alternative $p \geq p_1$, where $p_1 > p_0$ is an elevated occurrence rate of interest. We will show how to implement the test statistics and critical values for both of these examples in Section 5.3.

A feature of the sequential BH procedure defined in Section 3.2 is that it allows data streams to be "dropped" (i.e., sampling terminated) when no more information is needed to reach an accept/reject decision about the corresponding hypotheses. The need to drop certain endpoints in a multiple endpoint clinical trial while continuing others occurs frequently in practice because certain measurements are costly or invasive. An example is the well-known Women's Health Initiative (see Anderson et al., 2004; Rossouw et al., 2002), one of the largest multiple-endpoint randomized prevention studies of its kind. The Women's Health Initiative dropped the endpoints designed to investigate the effect of hormone replacement therapy on cardiovascular and cancer outcomes in 2002 and 2005, respectively, but continued to follow-up participants for dementia and other cognition-related endpoints. This portion of the study with the continued endpoints is known as the Women's Health Initiative Memory Study (see Espeland et al., 2004; Shumaker et al., 1998).

3. Control of FDR and FNR

3.1. General notation and setup

The methodology introduced below is to handle a general situation in which there are K sequentially observable data streams:

$$\begin{array}{llll}\text{Data stream } 1 & X_1^{(1)}, X_2^{(1)}, \ldots & \text{from} & \text{Experiment } 1 \\ \text{Data stream } 2 & X_1^{(2)}, X_2^{(2)}, \ldots & \text{from} & \text{Experiment } 2 \\ \vdots & & & \\ \text{Data stream } K & X_1^{(K)}, X_2^{(K)}, \ldots & \text{from} & \text{Experiment } K. \end{array} \tag{3.1}$$

In general we make no assumptions about the dimension of the sequentially observed data $X_n^{(k)}$, which may themselves be vectors of varying size, nor about the dependence structure of within-stream data $X_n^{(k)}, X_{n'}^{(k)}$ or between-stream data $X_n^{(k)}, X_{n'}^{(k')}$ $(k \neq k')$. Assume that for each data stream $k \in [K]$ there is a parameter vector $\theta^{(k)} \in \Theta^{(k)}$ governing that stream $X_1^{(k)}, X_2^{(k)}, \ldots,$ and it is desired to test a null hypothesis $H^{(k)} \subseteq \Theta^{(k)}$ about $\theta^{(k)}$, versus an alternative hypothesis $G^{(k)} \subseteq \Theta^{(k)}$, which is disjoint from $H^{(k)}$. A null hypothesis $H^{(k)}$ is considered *true* if $\theta^{(k)} \in H^{(k)}$, and $H^{(k)}$ is *false* if $\theta^{(k)} \in G^{(k)}$. The global parameter $\theta = \left(\theta^{(1)}, \ldots, \theta^{(K)}\right)$ is the concatenation of the individual parameters and is contained in the global parameter space $\Theta = \Theta^{(1)} \times \cdots \times \Theta^{(K)}$. The FDR and FNR are defined as

$$\mathrm{FDR} = \mathrm{FDR}(\theta) = E_\theta\left(\frac{V}{R \vee 1}\right) \text{ and } \mathrm{FNR} = \mathrm{FNR}(\theta) = E_\theta\left(\frac{U}{S \vee 1}\right),$$

where V is the number of true null hypotheses rejected, R is the number of null hypotheses rejected, U is the number of false null hypotheses accepted, S is the number of null hypotheses accepted, and $x \vee y = \max\{x, y\}$.

For simplicity of presentation, we adopt the fully sequential setting so that n takes the values $1, 2, \ldots$; however, other sampling schemes are possible with only minor changes to what follows and without changing our main result. For example, the method presented here includes group sequential sampling, by either taking each $X_n^{(k)}$ in (3.1) to be a group (i.e., vector) of data or, alternatively, letting n take values in a sample size set $\mathcal{N}$ for which group sequential sampling with at most g groups of size m corresponds to $\mathcal{N} = \{m, 2m, \ldots, gm\}$. For convenience we shall refer to the index n of the data $X_n^{(k)}$ and test statistics as the "sample size" or "time." However, because different streams' data $X_n^{(k)}$ and $X_n^{(k')}$ may be vectors of different sizes, this value n may not refer to the actual sample size in any given stream. For the same reason, n may represent different "information times" across different streams, and does not necessarily have to coincide with the calendar time of a particular analysis.

The BH procedure requires only a valid p-value $p^{(k)}$ for each null hypothesis $H^{(k)}$ and, letting $p^{(k_1)} \leq p^{(k_2)} \leq \ldots \leq p^{(k_K)}$, rejects $H^{(k_1)}, \ldots, H^{(k_u)}$ where $u = \max\{s \in [K] : p^{(s)} \leq s\alpha/K\}$ for a given desired FDR bound α, accepting all $H^{(k)}$ if the maximum does not exist. Playing a role analogous to the p-values $p^{(k)}$, in our sequential setting we utilize a sequential test statistic $\Lambda_n^{(k)} = \Lambda_n^{(k)}\left(X_1^{(k)}, \ldots, X_n^{(k)}\right)$ associated with each data stream $k \in [K]$. What follows could have been formulated completely in terms of sequential p-values, making it look more like the BH procedure; however, we have chosen to use arbitrary sequential test statistics $\Lambda_n^{(k)}$ instead to maintain generality and to make the resulting procedure more user-friendly, given the complexity and non-uniqueness of sequential p-values in all but the simplest cases; see (Jennison and Turnbull, 2000, ch. 8.4 and 9). Nonetheless, sequential p-values can be used for the test statistics $\Lambda_n^{(k)}$.

For each data stream $k \in [K]$, the test statistic $\Lambda_n^{(k)}$ must satisfy certain error probabilities that only depend on its associated data stream $X_1^{(k)}, X_2^{(k)}, \ldots,$ and not on any other data streams in any multivariate way. Specifically, given prescribed bounds $\alpha, \beta \in (0,1)$ on the FDR and FNR, we assume that for each test statistic $\Lambda_n^{(k)}$, $k \in [K]$, there exist critical values

$$A_1^{(k)} \leq A_2^{(k)} \ldots \leq A_K^{(k)} \leq B_K^{(k)} \leq B_{K-1}^{(k)} \leq \ldots \leq B_1^{(k)} \tag{3.2}$$

such that

$$P_{\theta^{(k)}}\left(\Lambda_n^{(k)} \geq B_s^{(k)} \text{ some } n,\ \Lambda_{n'}^{(k)} > A_1^{(k)} \text{ all } n' < n\right) \leq \left(\frac{s}{K}\right)\alpha \text{ for all } \theta^{(k)} \in H^{(k)}, \tag{3.3}$$

$$P_{\theta^{(k)}}\left(\Lambda_n^{(k)} \leq A_s^{(k)} \text{ some } n,\ \Lambda_{n'}^{(k)} < B_1^{(k)} \text{ all } n' < n\right) \leq \left(\frac{s}{K}\right)\beta \text{ for all } \theta^{(k)} \in G^{(k)}, \tag{3.4}$$

for each $s \in [K]$. These error bounds simply guarantee that $\Lambda_n^{(k)}$ has critical values allowing it to achieve conventional (i.e., not in any multiple testing sense) type I and II error probabilities given by certain fractions of α and β, respectively. In particular, (3.3) says that the sequential test on the kth data stream $X_1^{(k)}, X_2^{(k)}, \ldots,$ that samples until $\Lambda_n^{(k)} \notin \left(A_1^{(k)}, B_s^{(k)}\right)$, rejecting (respectively accepting) $H^{(k)}$ if $\Lambda_n^{(k)}$ crosses $B_s^{(k)}$ (respectively $A_1^{(k)}$) first, has type I error probability no greater than $(s/K)\alpha$. Similarly, (3.4) says that the test that samples until $\Lambda_n^{(k)} \notin \left(A_s^{(k)}, B_1^{(k)}\right)$ has type II error probability no greater than $(s/K)\beta$, for any $s \in [K]$. Below we will show that in many cases there are standard sequential statistics that satisfy these error bounds, and there are standard software packages that allow computation of the critical values, as well as even closed-form formulas in some cases. Given critical values satisfying (3.3)–(3.4), the ordering (3.2) holds without loss of generality because otherwise $A_s^{(k)}$ could be replaced by $\tilde{A}_s^{(k)} = \max\{A_1^{(k)}, \ldots, A_s^{(k)}\}$ for which (3.4) would still hold, and similarly for $B_s^{(k)}$. In addition, the critical values $A_s^{(k)}, B_s^{(k)}$ may depend on the sample size n of the test statistic $\Lambda_n^{(k)}$ being compared with them; however, we omit this from the notation in order to avoid making it too cumbersome. This is because, although different test statistics will be ranked and compared with the critical values at different stages, the test statistics being compared always have the same current sample size n at the time of comparison (see the definition of the $\tilde{\Lambda}_n^{(k)}$ in step 1 of the procedure's definition, below) making this dependence possible. That is, $\tilde{\Lambda}_n^{(1)}, \tilde{\Lambda}_n^{(2)}, \tilde{\Lambda}_n^{(3)}, \ldots,$ will be ranked and compared, but never with any $\tilde{\Lambda}_{n'}^{(k)}$ for $n' \neq n$.

The individual sequential test statistics $\Lambda_n^{(k)}$ form the building blocks of our sequential BH procedure, which we define in the next section. Like the original BH procedure, which compares p-values, the sequential procedure involves ranking test statistics. At the current level of generality, the sequential test statistics may be on completely different scales, so we must introduce standardizing functions $\varphi^{(k)}$, $k \in [K]$, which are applied to the test statistics $\Lambda_n^{(k)}$ before ranking them, and the sequential BH procedure in the next section is defined in terms of standardized test statistics $\tilde{\Lambda}_n^{(k)} = \varphi^{(k)}\left(\Lambda_n^{(k)}\right)$. The only required property of the standardizing functions is that they are increasing functions such that $\varphi^{(k)}\left(A_s^{(k)}\right)$ and $\varphi^{(k)}\left(B_s^{(k)}\right)$ do not depend on k. For simplicity, here we take the $\varphi^{(k)}$ to be piecewise linear functions such that

$$\varphi^{(k)}\left(A_s^{(k)}\right) = -(K - s + 1) \text{ and } \varphi^{(k)}\left(B_s^{(k)}\right) = K - s + 1 \text{ for all } k, s \in [K]. \tag{3.5}$$

That is, for $k \in [K]$ define

$$\varphi^{(k)}(x) = \begin{cases} x - A_1^{(k)} - K, & \text{for } x \leq A_1^{(k)} \\ \dfrac{x - A_s^{(k)}}{A_{s+1}^{(k)} - A_s^{(k)}} - (K - s + 1), & \text{for } A_s^{(k)} \leq x \leq A_{s+1}^{(k)} \text{ if } A_{s+1}^{(k)} > A_s^{(k)},\ 1 \leq s < K \\ \dfrac{2\left(x - A_K^{(k)}\right)}{B_K^{(k)} - A_K^{(k)}} - 1, & \text{for } A_K^{(k)} \leq x \leq B_K^{(k)} \\ \dfrac{x - B_s^{(k)}}{B_{s-1}^{(k)} - B_s^{(k)}} + K - s + 1, & \text{for } B_s^{(k)} \leq x \leq B_{s-1}^{(k)} \text{ if } B_{s-1}^{(k)} > B_s^{(k)},\ 1 < s \leq K \\ x - B_1^{(k)} + K, & \text{for } x \geq B_1^{(k)}. \end{cases}$$

3.2. The sequential BH procedure controlling FDR and FNR

The sequential BH procedure is defined iteratively by defining its jth stage of sampling ($j = 1, 2, \ldots$), between which null hypotheses are accepted or rejected. Let $\mathcal{I}_j$ denote the indices of the active null hypotheses (i.e., the null hypotheses that have not been accepted or rejected yet) at the beginning of the jth stage of sampling, let a_j (respectively r_j) be the number of null hypotheses that have been accepted (respectively rejected) at the beginning of the jth stage of sampling, and let n_j denote the cumulative sample size of the active data streams at the end of the jth stage of sampling. Accordingly, set $\mathcal{I}_1 = [K]$, $a_1 = r_1 = 0$, and $n_0 = 0$. The jth stage of sampling ($j = 1, 2, \ldots$) proceeds as follows.

1. Sample the active data streams $\left\{X_n^{(k)}\right\}_{k \in \mathcal{I}_j,\ n > n_{j-1}}$ until n equals

   $$n_j = \inf\left\{n > n_{j-1} : \tilde{\Lambda}_n^{(i(n,\ell))} \notin \left(-\left(K - a_j - \ell + 1\right), a_j + \ell\right),\ \text{some } \ell \in \left[|\mathcal{I}_j|\right]\right\}, \tag{3.6}$$

   where $\tilde{\Lambda}_n^{(k)} = \varphi^{(k)}\left(\Lambda_n^{(k)}\right)$ and $i(n, \ell) \in [K]$ denotes the index of the ℓth ordered active standardized statistic at sample size n.

2. (a) If a lower boundary in (3.6) has been crossed—that is, if

$$\tilde{\Lambda}_{n_j}^{\left(i\left(n_j,\ell\right)\right)} \le -\left(K-a_j-\ell+1\right) \text{ for some } \ell\in\left[|\mathcal{I}_j|\right], \tag{3.7}$$

then accept the $m_j \ge 1$ null hypotheses

$$H^{(i(n_j,1))}, H^{(i(n_j,2))}, \ldots, H^{\left(i(n_j,m_j)\right)},$$

where

$$m_j = \max\left\{m \le |\mathcal{I}_j| : \tilde{\Lambda}_{n_j}^{\left(i(n_j,m)\right)} \le -(K-a_j-m+1)\right\}, \tag{3.8}$$

and set $a_{j+1}=a_j+m_j$. Otherwise, set $a_{j+1}=a_j$.

(b) If an upper boundary in (3.6) has been crossed—that is, if

$$\tilde{\Lambda}_{n_j}^{\left(i\left(n_j,\ell\right)\right)} \ge a_j+\ell \text{ for some } \ell\in\left[|\mathcal{I}_j|\right],$$

then reject the $m'_j \ge 1$ null hypotheses

$$H^{\left(i\left(n_j,|\mathcal{I}_j|-m'_j+1\right)\right)}, H^{\left(i\left(n_j,|\mathcal{I}_j|-m'_j+2\right)\right)}, \ldots, H^{\left(i\left(n_j,|\mathcal{I}_j|\right)\right)}, \tag{3.9}$$

where

$$m'_j = \max\left\{m \le |\mathcal{I}_j| : \tilde{\Lambda}_{n_j}^{\left(i\left(n_j,|\mathcal{I}_j|-m+1\right)\right)} \ge K-r_j-m+1\right\}, \tag{3.10}$$

and set $r_{j+1}=r_j+m'_j$. Otherwise, set $r_{j+1}=r_j$.

3. Stop if there are no remaining active hypotheses; that is, if $a_{j+1}+r_{j+1}=K$. Otherwise, let $\mathcal{I}_{j+1}$ be the indices of the remaining active hypotheses and continue on to stage $j+1$.

In other words, the procedure samples all active data streams until at least one of the active null hypotheses will be accepted or rejected, indicated by the stopping rule (3.6). At this point, "step-up" acceptance and rejection rules (3.8) and (3.9), related to the BH procedure's rule, are used to accept or reject some active hypotheses in steps 2a and 2b, respectively. After updating the list of active hypotheses, the process is repeated until no active hypotheses remain.

Before stating our main result in Theorem 3.1 that this procedure controls both FDR and FNR, we make some remarks about its definition.

(A) There will never be a conflict between the acceptances in Step 2a and the rejections in Step 2b. Suppose (toward contradiction) that at some stage j the rule in Step 2a said to accept $H^{(k)}$, whereas the rule in Step 2b said to reject $H^{(k)}$. Then $k=i\left(n_j,\ell\right)$ for some $\ell \le m_j$ and $\ell \ge |\mathcal{I}_j|-m'_j+1$. The former implies

$$\tilde{\Lambda}_{n_j}^{(k)} = \tilde{\Lambda}_{n_j}^{\left(i(n_j,m)\right)} \le \tilde{\Lambda}_{n_j}^{\left(i(n_j,m_j)\right)} \le -(K-a_j-m_j+1) < 0,$$

whereas the latter implies

$$\tilde{\Lambda}_{n_j}^{(k)} = \tilde{\Lambda}_{n_j}^{\left(i(n_j,m)\right)} \ge \tilde{\Lambda}_{n_j}^{\left(i\left(n_j,|\mathcal{I}_j|-m'_j+1\right)\right)} \ge K-r_j-m'_j+1 > 0,$$

a contradiction.

(B) Ties in the order statistics $\tilde{\Lambda}_n^{(k)}$ can be broken arbitrarily (at random, say) without affecting the error control proved in Theorem 3.1.

(C) As mentioned above, the critical values $A_s^{(k)}, B_s^{(k)}$ can also depend on the current sample size n of the test statistic $\Lambda_n^{(k)}$ being compared to them, with only notational changes in the definition of the procedure and the properties proved below; to avoid overly cumbersome notation, we have omitted this from the presentation. Standard group sequential stopping boundaries, such as Pocock, O'Brien-Fleming, power family, and any others (see Jennison and Turnbull, 2000, ch. 2 and 4), can be utilized for the individual test statistics in this way.

Our main result, given in Theorem 3.1, is that this procedure controls both FDR and FNR at the prescribed levels α and β when the test statistics are independent, and controls them at slightly inflated values of α and β under arbitrary dependence of data streams, with the inflation factor given by $\sum_{k=1}^K 1/k$, which is asymptotically equivalent to $\log K$ for large K. This result generalizes the original result of Benjamini and Hochberg (1995) for their fixed-sample-size procedure by building on the arguments of Benjamini and Yekutieli (2001), and is proved in the Appendix.

Theorem 3.1. *Fix $\alpha, \beta \in (0,1)$ and suppose that (3.3)–(3.4) hold. Let K_0 and K_1 denote the number of true and false null hypotheses $H^{(k)}$, respectively, and let $\Delta = \sum_{k=1}^K 1/k$. Then, regardless of the dependence between the data streams, the sequential BH procedure defined above satisfies*

$$\mathrm{FDR}(\theta) \le \Delta\left(\frac{K_0}{K}\right)\alpha \le \Delta\alpha \text{ and} \tag{3.11}$$

$$\mathrm{FNR}(\theta) \le \Delta\left(\frac{K_1}{K}\right)\beta \le \Delta\beta \text{ for all } \theta \in \Theta. \tag{3.12}$$

Further, if the K_0 data streams corresponding to the true null hypotheses are independent, then the sequential BH procedure satisfies

$$\mathrm{FDR}(\theta) \le \left(\frac{K_0}{K}\right)\alpha \le \alpha \text{ for all } \theta \in \Theta. \tag{3.13}$$

If the K_1 data streams corresponding to the false null hypotheses are independent, then the sequential BH procedure satisfies

$$\mathrm{FNR}(\theta) \le \left(\frac{K_1}{K}\right)\beta \le \beta \text{ for all } \theta \in \Theta. \tag{3.14}$$

4. A rejective sequential BH procedure controlling FDR

In some applications where sequential sampling is called for, the statistician is primarily concerned with stopping and rejecting a null hypothesis $H^{(k)}$ if it appears to be false, but is content to continue sampling for a very long time if $H^{(k)}$ appears to be true. Such tests have been called "power one tests" (see Mukhopadhyay and De Silva [2009, ch. 5] or Siegmund [1985, ch. IV]). Some examples of this scenario are sequential monitoring of a process (such as manufacturing) where the null hypothesis represents the process being "in control" or monitoring a drug being used in a population and the null hypothesis represents the drug being safe. In this section we present a version of the sequential BH procedure with this property, which is obtained from the sequential BH procedure in the previous section, by, roughly speaking, ignoring the lower boundaries $A_s^{(k)}$ for the test statistics, plus a few other minor modifications.

In addition to the scenarios above, this version may be useful in applications where there is a restriction on the maximum sample size. When this occurs, it may not be possible to achieve the bounds (3.3) and (3.4), and one alternative available to the statistician is to drop the requirement of guaranteed FNR control while still achieving guaranteed FDR control, which the procedure introduced below provides by only specifying rejections (and not acceptances) of null hypotheses. For this reason we call it the rejective sequential BH procedure. Even in the presence of a maximum sample size or truncation point, it may still be possible to achieve (3.4), which is simply a (marginal) power condition on the kth component test statistic. The statistician can verify that (3.4) is possible by checking whether the most stringent case, the $s=1$ case, of (3.4) holds for any values $A_1^{(k)}, B_1^{(k)}$. See the Discussion for an alternative approach of using β as a parameter for obtaining multiple testing procedures with desirable properties.

Let the data streams $X_n^{(k)}$, test statistics $\Lambda_n^{(k)}$, and parameters $\theta^{(k)}$ and θ be as in Section 3.1. Because only FDR will be explicitly controlled, we only require specification of null hypotheses $H^{(k)} \subset \Theta^{(k)}$ and not alternative hypotheses $G^{(k)}$, and $H^{(k)}$ is *true* if $\theta^{(k)} \in H^{(k)}$ and *false* otherwise. As mentioned above, we also modify the fully sequential sampling setup of Section 3.1 to incorporate a maximum streamwise sample size (or "truncation point") $\overline{N}$ in (4.1) below because this is most natural in the scenarios mentioned above, although what follows could be formulated without a truncation point or with sample sizes other than $1, \ldots, \overline{N}$ by replacing statements like $n < \overline{N}$ in what follows by $n \in \mathcal{N}$ for an arbitrary sample size set $\mathcal{N}$, with only notational changes. Without the need for lower stopping boundaries, given a desired FDR bound $\alpha \in (0,1)$, for each test statistic $\Lambda_n^{(k)}$, $k \in [K]$, we only require the existence of "upper" critical values $B_K^{(k)} \le B_{K-1}^{(k)} \le \ldots \le B_1^{(k)}$ satisfying

$$P_{\theta^{(k)}}\left(\Lambda_n^{(k)} \ge B_s^{(k)} \text{ some } n < \overline{N}\right) \le \left(\frac{s}{K}\right)\alpha \text{ for all } s \in [K], \quad \theta^{(k)} \in H^{(k)}. \tag{4.1}$$

Similar to (3.3), this is just a bound on the type I error probability of the sequential test that stops and rejects $H^{(k)}$ at time $n < \overline{N}$ if $\Lambda_n^{(k)} \ge B_s^{(k)}$ and accepts $H^{(k)}$ otherwise. The standardizing functions $\varphi^{(k)}$ can be any increasing functions such that $\varphi^{(k)}\left(B_s^{(k)}\right)$ does not depend on k. Here we take

$$\varphi^{(k)}(x) = \begin{cases} x - B_K^{(k)} + 1, & \text{for } x \le B_K^{(k)} \\ \dfrac{x - B_s^{(k)}}{B_{s-1}^{(k)} - B_s^{(k)}} + K - s + 1, & \text{for } B_s^{(k)} \le x \le B_{s-1}^{(k)} \text{ if } B_{s-1}^{(k)} > B_s^{(k)}, \quad 1 < s \le K \\ x - B_1^{(k)} + K, & \text{for } x \ge B_1^{(k)}, \end{cases}$$

for all $k, s \in [K]$, giving $\varphi^{(k)}\left(B_s^{(k)}\right) = K - s + 1$.

Letting $\mathcal{I}_j$, n_j be as in Section 3.2, the jth stage ($j = 0, 1, \ldots$) of the rejective sequential BH procedure is defined as follows.

1. Sample the active data streams $\left\{X_n^{(k)}\right\}_{k \in \mathcal{I}_j,\ \ n > n_{j-1}}$ until n equals
$$n_j = \overline{N} \wedge \inf\left\{ n > n_{j-1} :\ \ \tilde{\Lambda}_n^{(i(n,\ell))} \geq \ell,\ \text{ some } \ell \in \left[|\mathcal{I}_j|\right] \right\}, \tag{4.2}$$
where $\tilde{\Lambda}_n^{(k)} = \varphi^{(k)}\left(\Lambda_n^{(k)}\right)$ and $i(n, \ell)$ denotes the index of the ℓth-ordered active standardized statistic at sample size n.
2. If $n_j < \overline{N}$, then reject the null hypotheses
$$H^{\left(i\left(n_j, \ell_j\right)\right)}, H^{\left(i\left(n_j, \ell_j+1\right)\right)}, \ldots, H^{\left(i\left(n_j, |\mathcal{I}_j|\right)\right)},$$
where
$$\ell_j = \min\left\{ \ell \in \left[|\mathcal{I}_j|\right] : \Lambda_{n_j}^{i\left(n_j, \ell\right)} \geq \ell \right\}. \tag{4.3}$$
Set $\mathcal{I}_{j+1}$ to be the indices of the remaining hypotheses and proceed to stage $j + 1$.
3. Otherwise, $n = \overline{N}$, so accept all active hypotheses $H^{(k)}$, $k \in \mathcal{I}_j$, and stop.

Like the sequential BH procedure, this procedure samples all active test statistics until at least one of them will be rejected, indicated by the stopping rule (4.3), which is similar to the BH rejection rule. Then a step-up rejection rule is used in Step 2 to reject certain hypotheses before the next stage of sampling begins. When the truncation point $\overline{N}$ is reached, all remaining active hypotheses are accepted. The next theorem shows that, similar to the sequential BH procedure, the rejective procedure has guaranteed FDR control under independence of true hypotheses, and FDR control with a slight inflation factor under arbitrary dependence.

Theorem 4.1. *Fix $\alpha \in (0, 1)$. In the above setup, suppose that there are K_0 true null hypotheses $H^{(k)}$ and that* (4.1) *holds. Then the rejective sequential BH procedure defined above satisfies* (3.13) *if the K_0 data streams corresponding to the true null hypotheses are independent, and it satisfies* (3.11) *under arbitrary dependence between data streams.*

The proof of Theorem 4.1 is similar to that of Theorem 3.1 and thus is omitted.

5. Implementation

In this section we discuss constructing sequential test statistics and critical values satisfying (3.3)–(3.4) (or 4.1 for the rejective version of the procedure) for individual data streams, and give some examples. Unlike many fixed-sample-size settings, critical values for sequential (or group sequential) test statistics can rarely can be written down as exact, closed-form expressions. However, critical values for sequential test statistics are routinely computed to sufficient accuracy using software packages, Monte Carlo, or some form of distributional approximation, asymptotic or otherwise. In Section 5.1 we give closed-form expressions for the critical values $A_s^{(k)}, B_s^{(k)}$ satisfying (3.3)–(3.4) to a very close approximation and that are based on the simple and widely used Wald approximations for the critical values of the sequential probability ratio test (SPRT) for testing simple-vs.-simple hypotheses, which are routinely used as surrogates for more complicated testing situations by monotone likelihood ratio, least favorable distributions, and other similar considerations; see Lehmann and Romano (2005, ch. 3.4 and 3.8) and Siegmund (1985, ch. II.3). For more complicated testing situations, sequential generalized likelihood ratio statistics and their signed-root normal approximations are discussed in Section 5.2. Though these approaches will address many of the commonly encountered testing situations, they do not cover every possible testing situation, so we stress that the multiple testing procedure's FDR and FNR control will hold no matter the form of the hypotheses and test statistic provided that (3.3)–(3.4) are satisfied; hence, the critical values obtained in ways other than those discussed here may be used.

5.1. Simple hypotheses and their use as surrogates for certain composite hypotheses

In this section we show how to construct the test statistics $\Lambda_n^{(k)}$ and critical values $A_s^{(k)}, B_s^{(k)}$ satisfying (3.3)–(3.4) for any data stream k such that $H^{(k)}$ and $G^{(k)}$ are both simple hypotheses. This setting is of interest in practice because many more complicated composite hypotheses can be reduced to simple hypotheses. Indeed, Müller et al. (2007, section 1) pointed out that testing a battery of simple-vs.-simple hypothesis tests is the standard setup in most discussions of FDR in the literature. In this case, the test statistics $\Lambda_n^{(k)}$ will be taken to be log-likelihood ratios because of their strong optimality properties of the resulting test, the SPRT; see Chernoff (1972). In order to express the likelihood ratio tests in simple form, we now make the additional assumption that each data stream $X_1^{(k)}, X_2^{(k)}, \ldots,$ constitutes independent and identically distributed (i.i.d.) data. However, we stress that this independence assumption is limited to *within* each stream so that, for example, elements of $X_1^{(k)}, X_2^{(k)}, \ldots,$ may be correlated with (or even identical to) elements of another stream $X_1^{(k')}, X_2^{(k')}, \ldots.$

Formally we represent the simple null and alternative hypotheses $H^{(k)}$ and $G^{(k)}$ by the corresponding distinct density functions $h^{(k)}$ (null) and $g^{(k)}$ (alternative) with respect to some common σ-finite measure $\mu^{(k)}$. The parameter space $\Theta^{(k)}$ corresponding to this data stream is the set of all densities f with respect to $\mu^{(k)}$, and $H^{(k)}$ is considered *true* if the actual density $f^{(k)}$ satisfies $f^{(k)} = h^{(k)}$ $\mu^{(k)}$ a.s. and is *false* if $f^{(k)} = g^{(k)}$ $\mu^{(k)}$ a.s. The SPRT for testing $H^{(k)} : f^{(k)} = h^{(k)}$ vs. $G^{(k)} : f^{(k)} = g^{(k)}$ with type I and II error probabilities α and β, respectively, utilizes the simple log-likelihood ratio test statistic

$$\Lambda_n^{(k)} = \sum_{j=1}^{n} \log\left(\frac{g^{(k)}\left(X_j^{(k)}\right)}{h^{(k)}\left(X_j^{(k)}\right)}\right) \tag{5.1}$$

and samples sequentially until $\Lambda_n^{(k)} \notin (A, B)$, where the critical values A, B satisfy

$$P_{h^{(k)}}\left(\Lambda_n^{(k)} \geq B \text{ some } n, \ \Lambda_{n'}^{(k)} > A \text{ all } n' < n\right) \leq \alpha \tag{5.2}$$

$$P_{g^{(k)}}\left(\Lambda_n^{(k)} \le A \text{ some } n,\ \Lambda_{n'}^{(k)} < B \text{ all } n' < n\right) \le \beta. \tag{5.3}$$

The most simple and widely used method for finding A and B is to use the closed-form Wald approximations $A = A_W(\alpha, \beta)$ and $B = B_W(\alpha, \beta)$, where

$$A_W(a, b) = \log\left(\frac{b}{1-a}\right) + \rho, \quad B_W(a, b) = \log\left(\frac{1-b}{a}\right) - \rho \tag{5.4}$$

for $a, b \in (0, 1)$ such that $a + b \le 1$ and a fixed $\rho \ge 0$. The quantity ρ is an adjustment to the boundaries to account for continuous test statistics whose excess over the boundary upon stopping may be smaller than discrete statistics. See Hoel et al. (1971, section 3.3.1) for a derivation of Wald's (1947) original $\rho = 0$ case and, based on Brownian motion approximations, Siegmund (1985, p. 50 and ch. X) derives the value $\rho = 0.583$, which has been used to improve the approximation for continuous random variables. With our multiple testing procedure we recommend using Siegmund's $\rho = 0.583$ for continuous test statistics and $\rho = 0$ for discrete statistics.

Although, in general, the inequalities in (5.2)–(5.3) only hold approximately when using the Wald approximations $A = A_W(\alpha, \beta)$ and $B = B_W(\alpha, \beta)$, Hoel et al. (1971) show that the actual type I and II error probabilities can only exceed α or β by a negligibly small amount in the worst case, and the difference approaches 0 for small α and β, which is relevant in the present multiple testing situation where we will utilize fractions of α and β. Next we use the Wald approximations to construct closed-form critical values $A_s^{(k)}$, $B_s^{(k)}$ satisfying (3.3)–(3.4). The simulations performed in Section 6 show that this approximation does not lead to any exceedances of the desired FDR and FNR bounds even in the case of highly correlated data streams. Alternative approaches would be to use a software package or Monte Carlo or to replace (5.4) by $\log b$ and $-\log a$, respectively, for which (5.2)–(5.3) always hold (see Hoel et al., 1971) and proceed similarly. The next theorem, proved in the Appendix, gives simple, closed-form critical values (5.5) that can be used in lieu of these other methods to calculate the $2K$ critical values $\left\{A_s^{(k)}, B_s^{(k)}\right\}_{s \in [K]}$ for a given data stream with simple hypotheses $H^{(k)}, G^{(k)}$ in the sequential BH procedure. Specifically, we show that when using (5.5), the left-hand sides of (3.3)–(3.4) equal the same quantities one would get using Wald's approximations with $s\alpha/K$ and $s\beta/K$ in place of α and β; hence, the inequalities in (3.3)–(3.4) hold up to Wald's approximation.

Theorem 5.1. *Fix $\alpha, \beta \in (0, 1)$ such that $\alpha + \beta \le 1$. Suppose that, for a certain data stream k, the associated hypotheses $H^{(k)} : f^{(k)} = h^{(k)}$ and $G^{(k)} : f^{(k)} = g^{(k)}$ are simple. For $a, b \in (0, 1)$ such that $a + b \le 1$, let $\alpha_W^{(k)}(a, b)$ and $\beta_W^{(k)}(a, b)$ be the values of the probabilities on the left-hand sides of (5.2) and (5.3), respectively, when $\Lambda_n^{(k)}$ is given by (5.1) and $A = A_W(a, b)$ and $B = B_W(a, b)$ are given by the Wald approximations (5.4). For $s \in [K]$ define*

$$\alpha_s = \frac{\alpha(K - s\beta)}{K(K - \beta)}, \quad \beta_s = \frac{\beta(K - s\alpha)}{K(K - \alpha)}.$$

Finally, for $k \in [K]$, let $\alpha_{BH,s}^{(k)}$ and $\beta_{BH,s}^{(k)}$ denote the left-hand sides of (3.3) and (3.4), respectively, with $A_s^{(k)}$, $B_s^{(k)}$ given by

$$A_s^{(k)} = \log\left(\frac{s\beta}{(1-\alpha_s)K}\right) + \rho, \quad B_s^{(k)} = \log\left(\frac{(1-\beta_s)K}{s\alpha}\right) - \rho. \tag{5.5}$$

Then, for all $s \in [K]$,

$$s\alpha/K + \beta_s \le 1, \quad \alpha_s + s\beta/K \le 1, \tag{5.6}$$

$$\alpha_{BH,s}^{(k)} = \alpha_W^{(k)}(s\alpha/K, \beta_s), \text{ and } \beta_{BH,s}^{(k)} = \beta_W^{(k)}(\alpha_s, s\beta/K), \tag{5.7}$$

and therefore (3.3)–(3.4) hold, up to Wald's approximation when using the critical values (5.5).

5.1.1. Example: Exponential families

Suppose that a certain data stream k is composed of i.i.d. d-dimensional random vectors $X_1^{(k)}, X_2^{(k)}, \ldots$, from a multiparameter exponential family of densities

$$X_n^{(k)} \sim f_{\theta^{(k)}}(x) = \exp\left[\theta^{(i)T}x - \psi^{(k)}\left(\theta^{(k)}\right)\right], \quad n = 1, 2, \ldots, \tag{5.8}$$

where $\theta^{(k)}$ and x are d-vectors, $(\cdot)^T$ denotes transpose, $\psi : \mathbb{R}^d \to \mathbb{R}$ is the cumulant generating function, and it is desired to test

$$H^{(k)} : \theta^{(k)} = \eta \text{ vs.} \mathrm{G}^{(\mathrm{k})} : \theta^{(\mathrm{k})} = \gamma \tag{5.9}$$

for given $\eta, \gamma \in \mathbb{R}^d$. Letting $S_n^{(k)} = \sum_{j=1}^n X_j^{(k)}$, the log-likelihood ratio (5.1) in this case is

$$\Lambda_n^{(k)} = (\gamma - \eta)^T S_n^{(k)} - n\left[\psi^{(k)}(\gamma) - \psi^{(k)}(\eta)\right] \tag{5.10}$$

and, by Theorem 5.1, the critical values (5.5) can be used and satisfy (3.3)–(3.4) up to Wald's approximation.

As mentioned above, many more complicated testing situations reduce to this setting. For example, to test the hypotheses $p^{(k)} \le p_0$ vs. $p^{(k)} \ge p_1$ about the the success probability $p^{(k)}$ of Bernoulli trials and given values $p_0 < p_1$, one may wish to instead test $H^{(k)} : p^{(k)} = p_0$ vs. $G^{(k)} : p^{(k)} = p_1$ by considering the worst-case error probabilities of the original hypotheses; such simplifications are, of course, routine in practice. For this case, the exponential family (5.8) and hypotheses (5.9) are given by

$$\theta^{(k)} = \log\left[p^{(k)}\Big/\left(1 - p^{(k)}\right)\right] \tag{5.11}$$

$$\psi^{(k)}\left(\theta^{(k)}\right) = \log\left[1 + \exp\left(\theta^{(k)}\right)\right] \tag{5.12}$$

$$\eta = \log\left[p_0/(1 - p_0)\right] \tag{5.13}$$

$$\gamma = \log\left[p_1/(1 - p_1)\right]. \tag{5.14}$$

A simulation study of the proposed procedure's performance in this setting is presented in Section 6.1.

5.2. Other composite hypotheses

Though many composite hypotheses can be reduced to the simple-vs.-simple situation in Section 5.1, the generality of Theorem 3.1 (and Theorem 4.1 for the rejective version) does not require this and allows any type of hypothesis to be tested as long as the corresponding sequential statistics satisfy (3.3)–(3.4) (or 4.1 in the rejective case). In this section we discuss the more general case of how to proceed to apply Theorem 3.1 when a certain data stream k is described by a multiparameter exponential family (5.8) but simple hypotheses are not appropriate; Theorem 4.1 and the rejective setting are discussed further below.

Letting ∇ denote the gradient, let

$$I\left(\theta^{(k)},\lambda^{(k)}\right)=\left(\theta^{(k)}-\lambda^{(k)}\right)^{T}\nabla\psi^{(k)}\left(\theta^{(k)}\right)-\left[\psi^{(k)}\left(\theta^{(k)}\right)-\psi^{(k)}\left(\lambda^{(k)}\right)\right]$$

denote the Kullback-Leibler information number for the distribution (5.8), and suppose it is desired to test

$$H^{(k)}:u\left(\theta^{(k)}\right)\le u_0\text{ vs. }G^{(k)}:u\left(\theta^{(k)}\right)\ge u_1,\tag{5.15}$$

where $u(\cdot)$ is some continuously differentiable real-valued function such that

$$\text{for all fixed }\theta^{(k)},I\left(\theta^{(k)},\lambda^{(k)}\right)\text{ is}\begin{pmatrix}\text{decreasing}\\ \text{increasing}\end{pmatrix}\text{in }u\left(\lambda^{(k)}\right)\begin{pmatrix}<\\ >\end{pmatrix}u\left(\theta^{(k)}\right),\tag{5.16}$$

and $u_0<u_1$ are chosen real numbers. In other words, (5.16) says that for any $\lambda^{(k)},\tilde{\lambda}^{(k)}$ such that $u\left(\theta^{(k)}\right)<u\left(\lambda^{(k)}\right)\le u\left(\tilde{\lambda}^{(k)}\right)$ we have $I\left(\theta^{(k)},\lambda^{(k)}\right)\le I\left(\theta^{(k)},\tilde{\lambda}^{(k)}\right)$ and a similar statement with all inequalities reversed. The family of models (5.8) and general form (5.15) of the hypotheses contain a large number of situations frequently encountered in practice, including various two-population (or more) comparison tests and testing problems with nuisance parameters. For example, the sequential Student's t-test problem mentioned in Section 2 is a special case of this setup; details are given in the next section.

The hypotheses (5.15) can be tested with the flexible and powerful sequential generalized likelihood ratio (GLR) statistics. Letting

$$\hat{\theta}_n^{(k)}=\left(\nabla\psi^{(k)}\right)^{-1}\left(\frac{1}{n}\sum_{j=1}^{n}X_j^{(k)}\right)$$

denote the maximum likelihood estimate of θ based on the data from the first n observations, define

$$\Lambda_{H,n}^{(k)}=n\left[\inf_{\lambda:u(\lambda)=u_0}I\left(\hat{\theta}_n^{(k)},\lambda\right)\right],\tag{5.17}$$

$$\Lambda_{G,n}^{(k)}=n\left[\inf_{\lambda:u(\lambda)=u_1}I\left(\hat{\theta}_n^{(k)},\lambda\right)\right],\tag{5.18}$$

$$\Lambda_n^{(k)}=\begin{cases}+\sqrt{2n\Lambda_{H,n}^{(k)}}, & \text{if }u\left(\hat{\theta}_n^{(k)}\right)>u_0\text{ and }\Lambda_{H,n}^{(k)}\ge\Lambda_{G,n}^{(k)}\\ -\sqrt{2n\Lambda_{G,n}^{(k)}}, & \text{otherwise.}\end{cases}\tag{5.19}$$

The statistics (5.17) and (5.18) are the log-GLR statistics for testing against $H^{(k)}$ and against $G^{(k)}$, respectively. To find the critical values that satisfy (3.3) and (3.4), Monte Carlo simulation or software packages for sequential (or group sequential) sampling of Gaussian data utilizing the large-n limiting distribution of the signed roots of (5.17)–(5.18) in (5.19) under $u\left(\theta^{(k)}\right) = u_0$ and u_1, respectively, can be used; see Jennison and Turnbull (1997, theorem 2).

Another commonly encountered testing situation is testing the simple null hypotheses versus the composite alternative

$$H^{(k)} : \theta^{(k)} = \theta_0^{(k)} \text{ vs. } G^{(k)} : \theta^{(k)} \neq \theta_0^{(k)} \tag{5.20}$$

for a give value $\theta_0^{(k)} \in \mathbb{R}^d$. However, by considering true values of $\theta^{(k)}$ arbitrarily close to $\theta_0^{(k)}$, it is clear that no test of (5.20) can control the type II error probability for all $\theta^{(k)} \in G^{(k)}$ in general, hence it may not be possible to find a test satisfying (3.4). If "early stopping" under the null hypothesis is not a priority, then the rejective sequential BH procedure in Section 4 can be used with the GLR statistic (5.17), as discussed above. On the other hand, in order to use the sequential BH version that allows early stopping under the null as well, one may need to restrict $G^{(k)}$ in some way for that to be possible; for example, by modifying $G^{(k)}$ to be only the $\theta^{(k)}$ such that $||\theta^{(k)} - \theta_0^{(k)}|| \geq \delta$ for some smooth norm $||\cdot||$ (such as l^2 norm) and value $\delta > 0$. This restricted form is a special case of the framework (5.15) by choosing $u\left(\theta^{(k)}\right) = ||\theta^{(k)} - \theta_0^{(k)}||$, $u_0 = 0$, and $u_1 = \delta$.

5.3. Example revisited: Multiple-endpoint clinical trials

In this section we discuss how to implement the two component hypothesis tests given as examples of endpoints in Section 2 for multiple-endpoint clinical trials.

The example of testing $H^{(k)} : p \leq p_0$ vs. $G^{(k)} : p \geq p_1 > p_0$ about the success probability p of i.i.d, Bernoulli data $X_1^{(k)}, X_2^{(k)}, \ldots,$ was discussed in Section 5.1.1 where the monotone likelihood ratio allowed reduction to simple hypotheses and thus the closed-form expressions (5.5) can be used for the critical values. A simulation study of the sequential BH procedure's performance on data streams of this type is presented in Section 6.1.

For testing

$$H^{(k)} : \mu \leq 0 \text{ vs. } G^{(k)} : \mu \geq \delta > 0$$

about the mean μ of i.i.d. normal data $X_1^{(k)}, X_2^{(k)}, \ldots,$ with unknown variance σ^2, the same immediate reduction to simple hypotheses is not possible because of the nuisance parameter σ^2; however, the sequential GLR statistics in Section 5.2 can handle this situation. The statistics (5.17)–(5.19) are

$$\Lambda^{(k)}_{H,n} = (n/2)\log\left[1+\left(\frac{\overline{X}^{(k)}_n}{\hat{\sigma}_n}\right)^2\right], \quad \Lambda^{(k)}_{G,n} = (n/2)\log\left[1+\left(\frac{\overline{X}^{(k)}_n-\delta}{\hat{\sigma}_n}\right)^2\right],$$
$$\text{and} \quad \Lambda^{(k)}_n = \begin{cases} +\sqrt{2n\Lambda^{(k)}_{H,n}}, & \text{if } \overline{X}^{(k)}_n \geq \delta/2 \\ -\sqrt{2n\Lambda^{(k)}_{G,n}}, & \text{otherwise,} \end{cases} \tag{5.21}$$

where $\overline{X}^{(k)}_n$ and $\hat{\sigma}^2_n$ are the usual maximum likelihood estimates of μ and σ^2, respectively, based on $X^{(k)}_1, \ldots, X^{(k)}_n$ (see Bartroff, 2006, p. 106). In order to compute the critical values $\left\{A^{(k)}_s, B^{(k)}_s\right\}_{s\in[K]}$ satisfying (3.3)–(3.4) for (5.21), Bartroff and Song (2014, lemma 3.1) showed that, in this case, the left-hand sides of (3.3) and (3.4) are bounded above by

$$P\left(t_n \geq b_{n,s} \text{ some } n,\ t_{n'} > a_{n',1} \text{ all } n' < n\right) \text{ and } P\left(t_n \leq a_{n,s} \text{ some } n,\ t_{n'} < b_{n',1} \text{ all } n' < n\right), \tag{5.22}$$

respectively, where

$$a_{n,s} = -\left\{(n-1)\left(\exp\left\{\left(A^{(k)}_s/n\right)^2\right\}-1\right)\right\}^{1/2}, \quad b_{n,s} = \left\{(n-1)\left(\exp\left\{\left(B^{(k)}_s/n\right)^2\right\}-1\right)\right\}^{1/2},$$
$$\text{and } t_n = \overline{Z}_n \Bigg/ \left\{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(Z_i-\overline{Z}_n\right)^2\right\}^{1/2} \tag{5.23}$$

in which $Z_1, \ldots, Z_n$ are i.i.d. standard normal random variables. Thus, t_n has the Student's t distribution with n – 1 degrees of freedom. Using (5.22), critical values satisfying (3.3) and (3.4) can be computed using recursive numerical integration, and this is the standard method used in this setting by the many sequential and group sequential software packages that exist; see Jennison and Turnbull (2000, ch. 19) and Bartroff et al. (2013, ch. 4.3). Alternatively, Monte Carlo can be used in which all that is needed is the generation of the i.i.d. standard normal random variables Z_i in (5.23).

In some applications it may be desired to test a composite null hypothesis versus a simple alternative hypothesis, and thus control the FNR at a particular value of the unknown parameter. For example, in the current sequential Student's t-test setting, suppose it is desired to test the composite null $H^{(k)} : \mu \leq 0$ versus the simple alternative $G^{(k)} : (\mu, \sigma) = (\delta, \sigma_1)$, for given values $\delta > 0$, $\sigma_1 > 0$. By arguments similar to those in the proof of Bartroff and Song (2014, lemma 3.1), it can be shown that for the test statistic (5.21), the error probabilities (3.3) and (3.4) are bounded above by the terms in (5.22), but with the $a_{n,s}$ and $b_{n',1}$ in the second term in (5.22) replaced by

$$\tilde{a}_{n,s} = \min\left\{a_{n,s}, -\frac{\delta\sqrt{n-1}}{2\sigma_1}\right\} \text{ and } \tilde{b}_{n',1} = b_{n',1} - \frac{\delta\sqrt{n'-1}}{2\sigma_1},$$

respectively. Using these formulas, the critical values can be computed using only the standard normal distribution by either recursive numerical integration or Monte Carlo, as described in the previous paragraph.

6. Simulation studies

In this section we present simulation studies comparing the sequential BH procedure (denoted SBH throughout this section) to the fixed sample BH procedure (denoted FBH) defined in Section 3.1. We note that there are no existing sequential competitors of SBH with which to compare. Although the traditional BH procedure could be applied to K arbitrary level-α sequential tests performed independent of each other, the resulting procedure would not control the FNR, nor would it likely be very efficient because the stopping rules of the individual tests would not take the other data streams into account.

In Section 6.1 we compare SBH with FBH in the context of Bernoulli data streams, the setup discussed at the end of Section 5.1.1, and in Section 6.2 we consider normal data streams and embed the streams in a multivariate normal distribution in order to simulate various between-stream correlation structures. Both studies use the values $\alpha = 0.05$ and $\beta = 0.2$ as the prescribed FDR and FNR bounds, respectively. This same value of α is used for the FBH procedure and, because the resulting procedure does not guarantee FNR control at a prescribed level, in order to compare "apples with apples" we have varied its fixed sample size in order to make its achieved value of FNR approximately match that of the SBH procedure. For each scenario considered below, we estimate FDR; FNR; their upper bounds $K_0\alpha/K$ and $K_1\beta/K$ under independence in (3.13) and (3.14), respectively; the expected total sample size $EN = E\left(\sum_{k=1}^K N^{(k)}\right)$ over all of the data streams where $N^{(k)}$ is the total sample size of the kth stream; and relative savings in sample size of SBH relative to FBH using 100,000 Monte Carlo simulated batteries of K sequential tests. Finally, we note that these two simulation studies have the property that each data stream and corresponding hypothesis test has the same structure; we emphasize that this is only for the sake of getting a clear picture of the procedures' performance and this property is not required of the SBH that allows arbitrary "mixing" of data stream distributions and types of hypotheses.

6.1. Independent Bernoulli data streams

Table 1 contains the operating characteristics of SBH and FBH for testing K hypotheses of the form

$$H^{(k)} : p^{(k)} \leq .4 \text{ vs. } G^{(k)} : p^{(k)} \geq .6, \quad k = 1, \ldots, K, \tag{6.1}$$

about the probability $p^{(k)}$ of success in the kth stream of i.i.d. Bernoulli data, which, for the sake of illustration, were generated independent of each other; a situation with between-stream dependence is considered in the next section. Standard errors (denoted SE) are given in parentheses. For the SBH procedure, the sequential log likelihood ratio test statistic (5.10)–(5.14) was used for each stream with the Wald approximation critical values (5.5) with $\rho = 0$. For FBH, whose sample size $N^{(k)}$ for each stream k is fixed,

Table 1. Operating characteristics of sequential (SBH) and fixed-sample (FBH) BH procedures for testing the hypotheses (6.1) about the success probabilities of i.i.d. Bernoulli data streams.

K	K_0	Procedure	FDR (SE)	$K_0\alpha/K$	FNR (SE)	$K_1\beta/K$	EN (SE)	Savings (%)
2	2	SBH	0.0314 (0.0063)	0.050	0 (0)	0	50.8 (1.9)	
		FBH	0.0315 (0.0064)		0 (0)		105	51.62
	1	SBH	0.0157 (0.0030)	0.025	0.0772 (0.0059)	0.100	61.9 (1.0)	
		FBH	0.0212 (0.0031)		0.0860 (0.0065)		120	48.42
	0*	SBH	0 (0)	0	0 (0)	0	273.1 (1.5)	
5	5	SBH	0.0264 (0.0035)	0.050	0 (0)	0	166.5 (2.5)	
		FBH	0.0238 (0.0034)		0 (0)		360	53.75
	3	SBH	0.0170 (0.0023)	0.030	0.0412 (0.0027)	0.080	193.7 (1.9)	
		FBH	0.0198 (0.0025)		0.0430 (0.0030)		370	47.65
	2	SBH	0.0115 (0.0017)	0.020	0.0628 (0.0044)	0.120	207.2 (1.8)	
		FBH	0.0188 (0.0020)		0.0629 (0.0044)		375	44.75
	0*	SBH	0 (0)	0	0 (0)	0	767.1 (1.5)	
10	10	SBH	0.0252 (0.0032)	0.050	0 (0)	0	338.0 (3.1)	
		FBH	0.0285 (0.0042)		0 (0)		765	55.82
	8	SBH	0.0195 (0.0026)	0.040	0.0201 (0.0015)	0.040	364.5 (3.3)	
		FBH	0.0291 (0.0034)		0.0280 (0.0018)		760	52.04
	5	SBH	0.0114 (0.0014)	0.025	0.0512 (0.0028)	0.100	430.3 (3.1)	
		FBH	0.0191 (0.0016)		0.0533 (0.0030)		770	44.12
	2	SBH	0.0048 (0.0007)	0.010	0.1015 (0.0046)	0.160	462.1 (3.2)	
		FBH	0.0085 (0.0009)		0.1037 (0.0057)		770	39.99
	0*	SBH	0 (0)	0	0 (0)	0	1,541.7 (3.2)	
20	20	SBH	0.0228 (0.0023)	0.050	0 (0)	0	703.3 (4.4)	
		FBH	0.0206 (0.0021)		0 (0)		1650	57.38
	16	SBH	0.0183 (0.0019)	0.040	0.0191 (0.0010)	0.040	763.5 (4.2)	
		FBH	0.0274 (0.0027)		0.0204 (0.0010)		1760	56.62
	10	SBH	0.0114 (0.0010)	0.025	0.0493 (0.0021)	0.100	891.9 (5.0)	
		FBH	0.0208 (0.0013)		0.0544 (0.0021)		1640	45.62
	4	SBH	0.0047 (0.0005)	0.010	0.0854 (0.0039)	0.160	964.7 (4.5)	
		FBH	0.0074 (0.0007)		0.0945 (0.0040)		1700	43.25
	0*	SBH	0 (0)	0	0 (0)	0	2,141.2 (4.6)	

p-values were computed in the standard way as $1 - F_{N^{(k)},.4}\left(S^{(k)}_{N^{(k)}} - 1\right)$, where $F_{n,p}(\cdot)$ is the cumulative distribution function of the binomial distribution with n trials and probability p of success, and $S^{(k)}_{N^{(k)}}$ is the sum of the $N^{(k)}$ observations from stream k. The data were generated for each data stream with $p^{(k)} = 0.4$ or 0.6, and the second column of Table 1 gives the number K_0 of true null hypotheses—that is, those for which $p^{(k)} = 0.4$—and $K_1 = K - K_0$ is the number of false null hypotheses. The final column, labeled "Savings," gives the percentage decrease in expected total sample size EN of SBH relative to FBH. Note that no standard error is given for the expected sample size of FBH because it is fixed.

The SBH procedure gives a sizable reduction in expected sample size relative to FBH procedure, at least roughly 40% in all scenarios and more than 50% savings in some. Turning our attention to FDR and FNR, note that both procedures routinely have achieved values of FDR and FNR not only less than the prescribed levels $\alpha = 0.05$ and $\beta = 0.2$ but also well below the bounds $K_0\alpha/K$ and $K_1\beta/K$, respectively. The sample size savings of SBH seems to grow with both the number K of hypotheses and the number K_0 of true null hypotheses.

An important consideration when choosing any statistical test, whether sequential or fixed sample, of hypotheses like (6.1) is its performance when $p^{(k)}$ lies in the "indifference region" between p_0 and p_1. Although FDR and FNR are not defined in

Table 2. Operating characteristics of sequential (SBH) and fixed-sample (FBH) BH procedures for testing (6.2) about the means of correlated normal data streams.

Covariance	True θ	Procedure	FDR (SE)	$K_0\alpha/K$	$\Delta\alpha$	FNR (SE)	$K_1\beta/K$	EN	Savings (%)
M_1	(1, 0)	SBH	0.0249 (0.0035)	0.025	0.075	0.0983 (0.0065)	0.100	9.6 (0.1)	
		FBH	0.0248 (0.0033)			0.0970 (0.0075)		16	35.63
M_1	(1, 0)	SBH	0.0228 (0.0047)	0.025	0.075	0.0676 (0.0062)	0.100	10.5 (0.2)	
		FBH	0.0293 (0.0043)			0.0626 (0.0053)		20	47.50
M_3	$(1,0,1,0)$	SBH	0.0212 (0.0030)	0.025	0.104	0.0767 (0.0045)	0.100	24.0 (0.2)	
		FBH	0.0264 (0.0034)			0.0800 (0.0051)		40	40.00
	$(1,1,0,0)$	SBH	0.0163 (0.0036)	0.025	0.104	0.0524 (0.0053)	0.100	24.1 (0.4)	
		FBH	0.0249 (0.0042)			0.0578 (0.0053)		44	45.23
M_4	$(1,0,0,0,0,0)$	SBH	0.0302 (0.0047)	0.042	0.123	0.0213 (0.0016)	0.033	31.3 (0.3)	
		FBH	0.0379 (0.0043)			0.0236 (0.0017)		72	56.53
	$(1,0,0,1,0,0)$	SBH	0.0251 (0.0034)	0.033	0.123	0.0476 (0.0027)	0.067	34.9 (0.3)	
		FBH	0.0324 (0.0037)			0.0483 (0.0029)		66	47.12
	$(1,1,0,0,0,0)$	SBH	0.0225 (0.0038)	0.033	0.123	0.0378 (0.0034)	0.067	35.1 (0.5)	
		FBH	0.0319 (0.0039)			0.0370 (0.0036)		72	51.25
	$(1,1,1,0,0,0)$	SBH	0.0142 (0.0032)	0.025	0.123	0.0478 (0.0044)	0.100	38.3 (0.6)	
		FBH	0.0250 (0.0038)			0.0490 (0.0048)		72	46.81
	$(1,1,0,1,1,0)$	SBH	0.0137 (0.0019)	0.017	0.123	0.0952 (0.0061)	0.133	39.8 (0.4)	
		FBH	0.0181 (0.0021)			0.0879 (0.0061)		66	39.70
	$(1,1,1,1,0,0)$	SBH	0.0113 (0.0025)	0.017	0.123	0.0826 (0.0057)	0.133	40.2 (0.5)	
		FBH	0.0175 (0.0027)			0.0884 (0.0052)		66	39.09
	$(1,1,1,1,1,0)$	SBH	0.0069 (0.0014)	0.008	0.123	0.1174 (0.0091)	0.167	41.1 (0.4)	
		FBH	0.0095 (0.0016)			0.1226 (0.0081)		66	37.73

this case, we recommend still considering other operating characteristics such as expected sample size for these values of the parameters. For example, in the setting of Table 1, if $p^{(k)} = 0.5$ for all $K = 10$ streams, then the expected sample size EN of the SBH procedure is 640.9, with a standard error of 2.3, based on 100,000 Monte Carlo replications. Because FDR and FNR are not defined in this case, there is no natural way to match error rates to compare with a fixed-sample-size procedure. However, this value of SBH's EN is substantially smaller than FBH's EN even in the more favorable scenarios (i.e., with all $p^{(k)}$ *outside* the indifference region) for the latter in the $K = 10$ cases of Table 1, in which EN = 760, 770, and 770, respectively.

6.2. Correlated normal data streams

Table 2 contains the operating characteristics of SBH and FBH for testing the hypotheses

$$H^{(k)} : \theta^{(k)} \le 0 \text{ vs. } G^{(k)} : \theta^{(k)} \ge \delta, \quad k = 1, ..., K, \tag{6.2}$$

about the mean $\theta^{(k)}$ of the kth stream of normal observations with variance 1 and where $\delta = 1$. As discussed in Section 5.2, this alternative hypothesis $G^{(k)}$ can be thought of as a surrogate for the alternative hypothesis $\theta^{(k)} > 0$. In order to generate K normal data streams under various correlation structures, the K streams were generated as components of a K-dimensional multivariate normal distribution with mean $\theta = \left(\theta^{(1)}, ..., \theta^{(K)}\right)$, given in the second column of Table 2, and various non-identity covariance matrices M_i, given in the Appendix. These covariance matrices provide a variety of different scenarios with positively and/or negatively correlated data streams. The Wald approximation critical values (5.5) were used with the continuity correction $\rho = 0.583$ suggested by Siegmund (1985, p. 50 and ch. X). The other columns have the same meaning as in Table 1. The p-values for FBH were computed in the standard way as $1 - \Phi\left(S^{(k)}_{N^{(k)}}/\sqrt{N^{(k)}}\right)$, where Φ is the cumulative distribution function of the standard normal distribution and $S^{(k)}_{N^{(k)}}$ is the sum of the $N^{(k)}$ observations from stream k.

Despite the correlations present between different data streams, the interaction of these various combinations of correlations with various true or false null hypotheses all show behavior somewhat similar to the case of independent data streams in the previous section in that SBH has sizably smaller expected sample size than FBH in all cases, roughly a 40% reduction in most cases. Though the independent case of Theorem 3.1 no longer applies because of the dependence, we note that in each scenario the achieved FDR and FNR rates are all less than $K_0\alpha/K$ and $K_1\beta/K$ in (3.13) and (3.14), respectively.

7. Discussion

We have proposed a flexible procedure to combine basic sequential hypothesis tests into an FDR-and FNR-controlling multiple testing procedure tailored to sequential data. The error control in Theorems 3.1 and 4.1 is proved under arbitrary dependence with a small logarithmic inflation Δ of the prescribed levels α and β that may be dispensed with under independence. These were the same conditions under which Benjamini and Hochberg (1995) proved FDR control in their original paper and, as mentioned in the Introduction, recent work by Benjamini and Yekutieli (2001) has broadened this from independence to positive regression dependence. We fully expect to be able to similarly extend the conditions under which (uninflated) FDR control holds in the sequential domain too, but the distributional complications introduced by sequential sampling present additional challenges. Our conjecture is supported by the simulation studies in Section 6.2 and other simulation studies we have performed under strong positive dependence, in which not a single instance of achieved FDR or FNR has exceeded the uninflated levels $K_0\alpha/K$ and $K_1\alpha/K$ of the independent case. Moreover, the setting of Section 6.2 in which dependence exists between data streams but it may be impossible for the statistician to know or model a priori is a prime example of where the proposed procedure may be useful since FDR and FNR can still be controlled by only knowing something about the marginal distributions of the test statistics through (3.3)–(3.4). The results of this section are encouraging that a sequential analog of Storey and Tibshirani's (2003) argument that the BH procedure controls FDR asymptotically as $K \to \infty$ under arbitrary dependence will hold as well.

The simultaneous control of FDR and FNR achievable by the sequential BH procedure is a by-product of the sequential setting and is analogous to the situation in classical single hypothesis testing where there exist sequential tests simultaneously controlling both type I and II error probabilities at arbitrary levels (Stein, 1945), a feat that is impossible in general for fixed-sample-size tests (Dantzig, 1940). Also analogous to the classical setting, it may be that the statistician has a well-motivated value of the FDR bound α in mind but not necessarily a value of the FNR bound β (or the value u_1

in the composite alternative 5.15). In this case, the rejective version of the sequential BH procedure in Section 4 may be used, which only stops early to reject null hypotheses; that is, when the data indicates that an alternative hypothesis is true. If substantial early stopping is also desired when null hypotheses are true, then we encourage the statistician to utilize the sequential BH procedure and to treat β as a parameter that may be chosen to give a procedure with other desirable operating characteristics, such as expected total or streamwise maximum sample size.

In addition to the widely available software packages for computing group sequential critical values and the formulas (5.5) that can both be used to compute the $2K^2$ critical values $\left\{A_s^{(k)}, B_s^{(k)}\right\}_{s,k\in[K]}$ of the individual sequential tests satisfying (3.3)–(3.4), we mentioned Monte Carlo as an alternative. Although $2K^2$ critical values are needed in general, raising the specter of $2K^2$ different simulations studies, there are features of the problem making the actual number much smaller and indicating that it is somewhat immune to the curse of dimensionality in many cases, which afflicts many problems in high-dimensional statistics. For simplicity let us focus on the rejective version of the sequential BH procedure in Section 4; however, similar statements apply to the general version. In the rejective version, the K^2 critical values $\left\{B_s^{(k)}\right\}_{s,k\in[K]}$ satisfying (4.1) are needed in general. However, in settings like the simulation studies in Section 6 where multiple data streams utilize test statistics of the same form, the actual number may be much smaller; for examle, K if all tests are of the same form. Moreover, because of the nested nature of the error probabilities (4.1), these K values can be simulated in a *single* Monte Carlo study by letting B_s be the upper $(s\alpha/K)$ quantile of the simulated empirical distribution of the statistic $\max_{1\le n\le \overline{N}}\Lambda_n$.

Appendix: Proofs and details of simulation studies

Proof of Theorem 3.1. Fix $\theta\in\Theta$ and omit it from the notation. First we prove (3.11) and (3.13). Without loss of generality, let $H^{(1)},\ldots,H^{(K_0)}$ denote the true null hypotheses, some $1\le K_0\le K$. For $k\in[K_0]$ and $s\in[K]$ define the events

$$W_{k,s}=\left\{\tilde{\Lambda}_n^{(k)}\ge K-s+1 \text{ some } n,\ \tilde{\Lambda}_{n'}^{(k)}>-K \text{ all } n'<n\right\}$$

and, by (3.5) and (3.3), we have

$$P(W_{k,s})=P\left(\Lambda_n^{(k)}\ge B_s^{(k)} \text{ some } n,\ \Lambda_{n'}^{(k)}>A_1^{(k)} \text{ all } n'<n\right)\le s\alpha/K. \tag{A.1}$$

For $v\in[K_0]$ and $t\in\{0,\ldots,K-K_0\}$, let

$$\begin{aligned}\Omega_v&=\{\omega\subseteq[K_0]:|\omega|=v\},\\ V_{v,t}^{\omega}&=\left\{H^{(k)},k\in\omega,\text{ and } t \text{ false hypotheses rejected}\right\}\text{ for }\omega\in\Omega_v,\text{ and}\\ V_{v,t}&=\bigcup_{\omega\in\Omega_v}V_{v,t}^{\omega}=\{v\text{ true and }t\text{ false hypotheses rejected}\},\end{aligned}$$

and note that this union is disjoint.

We begin by showing that

$$P\left(W_{k,v+t}\cap V_{v,t}^{\omega}\right)\ge 1\{k\in\omega\}P\left(V_{v,t}^{\omega}\right), \tag{A.2}$$

which is trivial for $k \notin \omega$. To show that (A.2) holds for $k \in \omega$, we will show that $V^{\omega}_{v,t} \subseteq W_{k,v+t}$ in this case. Consider any outcome in $V^{\omega}_{v,t}$. Because $H^{(k)}$ is rejected on this outcome, let j denote the stage at which $H^{(k)}$ is rejected. By the definition of step 2b of the procedure, $k = i(n_j, \ell)$ for some $\ell \geq |\mathcal{I}_j| - m'_j + 1$, so

$$\tilde{\Lambda}^{(k)}_{n_j} = \tilde{\Lambda}^{(i(n_j,\ell))}_{n_j} \geq \tilde{\Lambda}^{(i(n_j,|\mathcal{I}_j|-m'_j+1))}_{n_j} \geq K - \left(r_j + m'_j\right) + 1, \tag{A.3}$$

this last inequality by (3.10). Because $r_j + m'_j = r_{j+1}$ and this value is no greater than the total number $v + t$ of null hypotheses rejected on $V^{\omega}_{v,t}$, (A.3) gives

$$\tilde{\Lambda}^{(k)}_{n_j} \geq K - (v + t) + 1. \tag{A.4}$$

Now suppose toward contradiction that $\tilde{\Lambda}^{(k)}_{n'} \leq -K$ for some $n' < n_j$. Then, by (3.7), $H^{(k)}$ would have been accepted at some stage j' prior to j because

$$\tilde{\Lambda}^{(k)}_{n'} \leq -K \leq -\left(K - a_{j'} - \ell + 1\right)$$

for any possible value of $a_{j'} \geq 0$ and any $\ell \geq 1$, contradicting the assumption that $H^{(k)}$ is rejected at stage j. Thus, it must be that $\tilde{\Lambda}^{(k)}_{n'} > -K$ for all $n' < n_j$, and combining this with (A.4) shows that this outcome is in $W_{k,v+t}$, finishing the proof of (A.2).

With (A.2) established, we now follow the argument of Benjamini and Yekutieli (2001, section 4) more directly with a few modifications.

$$\begin{aligned}\sum_{k=1}^{K_0} P(W_{k,v+t} \cap V_{v,t}) &= \sum_{k=1}^{K_0} \sum_{\omega \in \Omega_v} P\left(W_{k,v+t} \cap V^{\omega}_{v,t}\right) \geq \sum_{k=1}^{K_0} \sum_{\omega \in \Omega_v} 1\{k \in \omega\} P\left(V^{\omega}_{v,t}\right) \\ &= \sum_{\omega \in \Omega_v} \sum_{k=1}^{K_0} 1\{k \in \omega\} P\left(V^{\omega}_{v,t}\right) = \sum_{\omega \in \Omega_v} |\omega| P\left(V^{\omega}_{v,t}\right) = vP(V_{v,t}).\end{aligned} \tag{A.5}$$

Using this and the definition of FDR,

$$\begin{aligned}\mathrm{FDR} &= \sum_{t=0}^{K-K_0} \sum_{v=1}^{K_0} \frac{v}{v+t} P(V_{v,t}) \leq \sum_{t=0}^{K-K_0} \sum_{v=1}^{K_0} \frac{v}{v+t} \left(\frac{1}{v} \sum_{k=1}^{K_0} P(W_{k,v+t} \cap V_{v,t}) \right) \\ &= \sum_{t=0}^{K-K_0} \sum_{v=1}^{K_0} \frac{1}{v+t} \sum_{k=1}^{K_0} P(W_{k,v+t} \cap V_{v,t}).\end{aligned} \tag{A.6}$$

Define $U_{v,t,k}$ as the event in which, if $H^{(k)}$ is rejected, then v – 1 other true and t false null hypotheses are also rejected, so that $W_{k,v+t} \cap V_{v,t} = W_{k,v+t} \cap U_{v,t,k}$. Let $U_{s,k} = \cup_{v+t=s} U_{v,t,k}$ and note that, for any k, $U_{1,k}, \ldots, U_{K,k}$ partition the sample space. Then, starting at (A.6),

$$\mathrm{FDR} \leq \sum_{t=0}^{K-K_0} \sum_{v=1}^{K_0} \frac{1}{v+t} \sum_{k=1}^{K_0} P(W_{k,v+t} \cap U_{v,t,k}) = \sum_{k=1}^{K_0} \sum_{s=1}^{K} \frac{1}{s} P(W_{k,s} \cap U_{s,k}). \tag{A.7}$$

With the convention $W_{k,0} = \emptyset$, define

$$p_{k,\ell,s} = P((W_{k,\ell} \setminus W_{k,\ell-1}) \cap U_{s,k}) \text{ for } k \in [K_0], \ell \in [s], s \in [K].$$

Note that $W_{k,\ell-1} \subseteq W_{k,\ell}$, so $W_{k,s} = \cup_{\ell=1}^{s} (W_{k,\ell} \setminus W_{k,\ell-1})$ and this union is disjoint. Writing $W_{k,s}$ in this way in (A.7), we have

$$\begin{aligned}
\mathrm{FDR} &\le \sum_{k=1}^{K_0}\sum_{s=1}^{K}\frac{1}{s}\sum_{\ell=1}^{s}p_{k,\ell,s} \le \sum_{k=1}^{K_0}\sum_{s=1}^{K}\sum_{\ell=1}^{s}\frac{p_{k,\ell,s}}{\ell} \le \sum_{k=1}^{K_0}\sum_{s=1}^{K}\sum_{\ell=1}^{K}\frac{p_{k,\ell,s}}{\ell} = \sum_{k=1}^{K_0}\sum_{\ell=1}^{K}\frac{1}{\ell}\sum_{s=1}^{K}p_{k,\ell,s}\\
&= \sum_{k=1}^{K_0}\sum_{\ell=1}^{K}\frac{1}{\ell}P(W_{k,\ell}\setminus W_{k,\ell-1}) = \sum_{k=1}^{K_0}\sum_{\ell=1}^{K}\frac{1}{\ell}[P(W_{k,\ell})-P(W_{k,\ell-1})]\\
&= \sum_{k=1}^{K_0}\left[\sum_{\ell=1}^{K}\frac{P(W_{k,\ell})}{\ell}-\sum_{\ell=0}^{K-1}\frac{P(W_{k,\ell})}{\ell+1}\right]\\
&= \sum_{k=1}^{K_0}\left[\sum_{\ell=1}^{K-1}\frac{P(W_{k,\ell})}{\ell(\ell+1)}+\frac{P(W_{k,K})}{K}-P(W_{k,0})\right] \le \sum_{k=1}^{K_0}\left[\sum_{\ell=1}^{K-1}\frac{\alpha}{K(\ell+1)}+\frac{\alpha}{K}\right] \qquad (\text{by}(\text{A.1}))\\
&= \sum_{k=1}^{K_0}\sum_{\ell=1}^{K}\frac{\alpha}{K\ell} = \Delta\left(\frac{K_0}{K}\right)\alpha.
\end{aligned}$$

If data streams $k\in[K_0]$ are independent, returning to (A.7) we have

$$\begin{aligned}
\mathrm{FDR} &\le \sum_{k=1}^{K_0}\sum_{s=1}^{K}\frac{1}{s}P(W_{k,s})P(U_{s,k}) \le \sum_{k=1}^{K_0}\sum_{s=1}^{K}\frac{1}{s}\left(\frac{s\alpha}{K}\right)P(U_{s,k}) = \frac{\alpha}{K}\sum_{k=1}^{K_0}\sum_{s=1}^{K}P(U_{s,k}) = \frac{\alpha}{K}\sum_{k=1}^{K_0}1\\
&= \left(\frac{K_0}{K}\right)\alpha,
\end{aligned}$$

where the second equality holds because $U_{1,k},\ldots,U_{K,k}$ partition the sample space.

The proof of FNR control is entirely symmetric and so is omitted here. □

Proof of Theorem 5.1. We verify the first parts of (5.6) and (5.7); the second parts are verified similarly. Using that $\beta\le 1-\alpha$ and some calculus, we have

$$\frac{s\alpha}{K}+\beta_s = \frac{s\alpha}{K}+\frac{\beta(K-s\alpha)}{K(K-\alpha)} \le \frac{s\alpha}{K}+\frac{(1-\alpha)(K-s\alpha)}{K(K-\alpha)} = \frac{s}{K}-\int_\alpha^1\frac{(K-1)(s-1)}{(K-a)^2}da \le \frac{s}{K}\le 1.$$

The forms of $A_s^{(k)}$ and $B_s^{(k)}$ in (5.5) can equivalently be written as $A_W(\alpha_s, s\beta/K)$ and $B_W(s\alpha/K,\beta_s)$, respectively, and it is simple algebra to check that $A_W(s\alpha/K,\beta_s)=A_1^{(k)}$ for all $s\in[K]$. Then

$$\begin{aligned}
\alpha_{BH,s}^{(k)} &= P_{h^{(k)}}\left(\Lambda_n^{(k)}\ge B_s^{(k)}\text{ some } n, \Lambda_{n'}^{(k)} > A_1^{(k)} \text{ all } n'<n\right)\\
&= P_{h^{(k)}}\left(\Lambda_n^{(k)}\ge B_W(s\alpha/K,\beta_s)\text{ some } n,\ \Lambda_{n'}^{(k)} > A_W(s\alpha/K,\beta_s) \text{ all } n'<n\right)\\
&= \alpha_W^{(k)}(s\alpha/K,\beta_s),
\end{aligned}$$

by definition of $\alpha_W^{(k)}$. □

Details of Simulation Studies

The four covariance matrices used in the simulations in Section 6.2 are as follows:

$$M_1 = \begin{pmatrix}1 & 0.8\\ 0.8 & 1\end{pmatrix}$$

$$M_2 = \begin{pmatrix}1 & -0.8\\ -0.8 & 1\end{pmatrix}$$

$$M_3 = \begin{pmatrix} 1 & 0.8 & -0.6 & -0.8 \\ 0.8 & 1 & -0.6 & -0.8 \\ -0.6 & -0.6 & 1 & 0.8 \\ -0.8 & -0.8 & 0.8 & 1 \end{pmatrix}$$

$$M_4 = \begin{pmatrix} 1 & 0.8 & 0.6 & -0.4 & -0.6 & -0.8 \\ 0.8 & 1 & 0.8 & -0.4 & -0.6 & -0.8 \\ 0.6 & 0.8 & 1 & -0.4 & -0.6 & -0.8 \\ -0.4 & -0.4 & -0.4 & 1 & 0.8 & 0.6 \\ -0.6 & -0.6 & -0.6 & 0.8 & 1 & 0.8 \\ -0.8 & -0.8 & -0.8 & 0.6 & 0.8 & 1 \end{pmatrix}.$$

Funding

Bartroff's work was partially supported by grant DMS-1310127 from the National Science Foundation, grants R01-AA026368-01 and R01-GM068968 from the National Institutes of Health, and grant 1U01FD006549 from the U.S. Food and Drug Administration. The majority of Song's work was completed while a Ph.D. student in the Department of Mathematics at the University of Southern California.

References

Anderson, G. L., Limacher, M. C., Assaf, A. R., Bassford, T., Beresford, S. A., Black, H. R., Bonds, D. E., Brunner, R. L., Brzyski, R. G., Caan, B., et al. (2004). Effects of Conjugated Equine Estrogen in Postmenopausal Women with Hysterectomy: The Women's Health Initiative Randomized Controlled Trial, *Journal of American Medical Association* 291(14): 1701–1712.

Avery, A. J., Anderson, C., Bond, C., Fortnum, H., Gifford, A., Hannaford, P. C., Hazell, L., Krska, J., Lee, A., Mclernon, D. J., et al. (2011). Evaluation of Patient Reporting of Adverse Drug Reactions to the UK "Yellow Card Scheme": Literature Review, Descriptive and Qualitative Analyses, and Questionnaire Surveys, *Health Technology Assessment* 15: iii–227. doi:10.3310/hta15200

Bartroff, J. (2006). Efficient three-stage *t*-tests, in *Recent Developments in Nonparametric Inference and Probability: Festschrift for Michael Woodroofe*, vol. 50 of IMS Lecture Notes Monograph Series, pp. 105–111, Hayward: Institute of Mathematical Statistics.

Bartroff, J. (2018). Multiple Hypothesis Tests Controlling Generalized Error Rates for Sequential Data, *Statistica Sinica* 28: 363–398. doi:10.5705/ss.202015.0267

Bartroff, J. and Lai, T. L. (2010). Multistage Tests of Multiple Hypotheses, *Communications in Statistics – Theory and Methods* 39: 1597–1607. doi:10.1080/03610920802592852

Bartroff, J., Lai, T. L., and Shih, M. (2013). *Sequential Experimentation in Clinical Trials: Design and Analysis*, New York: Springer.

Bartroff, J. and Song, J. (2014). Sequential Tests of Multiple Hypotheses Controlling Type I and II Familywise Error Rates, *Journal of Statistical Planning and Inference* 153: 100–114. doi:10.1016/j.jspi.2014.05.010

Bartroff, J. and Song, J. (2015). A Rejection Principle for Sequential Tests of Multiple Hypotheses Controlling Familywise Error Rates, *Scandinavian Journal of Statistics* 43: 3–19. doi:10.1111/sjos.12161

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. doi:10.1111/j.2517-6161.1995.tb02031.x

Benjamini, Y. and Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency, *Annals of Statistics* 29(4): 1165–1188.

Berry, S. M. and Berry, D. A. (2004). Accounting for Multiplicities in Assessing Drug Safety: A Three-Level Hierarchical Mixture Model, *Biometrics* 60(2): 418–426. doi:10.1111/j.0006-341X.2004.00186.x

Chen, S. and Arias-Castro, E. (2017). Sequential Multiple Testing, ArXiV preprint, http://arxiv.org/abs/1705.10190.

Chernoff, H. (1972). *Sequential Analysis and Optimal Design*, Philadelphia: Society for Industrial and Applied Mathematics.

Cohen, A. and Sackrowitz, H. B. (2005). Decision Theory Results for One-Sided Multiple Comparison Procedures, *Annals of Statistics* 33: 126–144. doi:10.1214/009053604000000968

Dantzig, G. B. (1940). On the Non-existence of Tests of Student's Hypothesis Having Power Functions Independent of σ, *Annals of Mathematical Statistics* 11: 186–192. doi:10.1214/aoms/1177731912

Efron, B. and Tibshirani, R. (2002). Empirical Bayes Methods and False Discovery Rates for Microarrays, *Genetic Epidemiology* 23(1): 70–86. doi:10.1002/gepi.1124

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes Analysis of a Microarray Experiment, *Journal of the American Statistical Association* 96(456): 1151–1160. doi:10.1198/016214501753382129

Espeland, M. A., Rapp, S. R., Shumaker, S. A., Brunner, R., Manson, J. E., Sherwin, B. B., Hsia, J., Margolis, K. L., Hogan, P. E., Wallace, R., et al. (2004). Conjugated Equine Estrogens and Global Cognitive Function in Postmenopausal Women: Women's Health Initiative Memory Study, *Journal of American Medical Association* 291(24): 2959–2968.

Fischl, M. A., et al. (1987). The Efficiency of Azidothymidine (AZT) in the Treatment of Patients with AIDS and AIDS-Related Complex, *New England Journal of Medicine* 317: 185–191. doi:10.1056/NEJM198707233170401

Genovese, C. and Wasserman, L. (2002). Operating Characteristics and Extensions of the False Discovery Rate Procedure, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(3): 499–517. doi:10.1111/1467-9868.00347

Hoel, P. G., Port, S. C., and Stone, C. J. (1971). *Introduction to Statistical Theory*, Boston: Houghton Mifflin Co.

Javanmard, A. and Montanari, A. (2018). Online Rules for Control of False Discovery Rate and False Discovery Exceedance, *Annals of Statistics* 46(2): 526–554. doi:10.1214/17-AOS1559

Jennison, C. and Turnbull, B. W. (1997). Group Sequential Analysis Incorporating Covariate Information, *Journal of American Statistical Association* 92: 1330–1341. doi:10.1080/01621459.1997.10473654

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*, New York: Chapman & Hall/CRC.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*, third edition, New York: Springer.

Mukhopadhyay, N. and De Silva, B. (2009). *Sequential Methods and Their Applications*, New York: Chapman & Hall/CRC.

Müller, P., Parmigiani, G., and Rice, K. (2007). FDR and Bayesian Multiple Comparisons Rules, in *Bayesian Statistics 8:* Proceedings of the Eighth Valencia International Meeting, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and W. West, eds., pp. 349–370, Oxford, UK: Oxford University Press.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting Differential Gene Expression with a Semiparametric Hierarchical Mixture Method, *Biostatistics* 5(2): 155–176. doi:10.1093/biostatistics/5.2.155

O'Brien, P. C. (1984). Procedures for Comparing Samples with Multiple Endpoints, *Biometrics* 40: 1079–1087.

Rossouw, J. E., Anderson, G. L., Prentice, R. L., LaCroix, A. Z., Kooperberg, C., Stefanick, M. L., Jackson, R. D., et al. (2002). Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal Women: Principal Results from the Women's Health Initiative Randomized Controlled Trial, *Journal of the American Medical Association* 288(3): 321–333.

Sarkar, S. K. (1998). Some Probability Inequalities for Ordered MTP_2 Random Variables: A Proof of the Simes' Conjecture, *Annals of Statistics* 26(2): 494–504. doi:10.1214/aos/1028144846

Shumaker, S. A., Reboussin, B. A., Espeland, M. A., Rapp, S. R., McBee, W. L., Dailey, M., Bowen, D., Terrell, T., and Jones, B. N. (1998). The Women's Health Initiative Memory Study (WHIMS): A Trial of the Effect of Estrogen Therapy in Preventing and Slowing the Progression of Dementia, *Controlled Clinical Trials* 19(6): 604–621. doi:10.1016/S0197-2456(98)00038-5

Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*, New York: Springer-Verlag.

Siegmund, D. and Yakir, B. (2008). Detecting the Emergence of a Signal in a Noisy Image, *Statistics and Its Inference* 1: 3–12. doi:10.4310/SII.2008.v1.n1.a1

Simes, R. J. (1986). An Improved Bonferroni Procedure for Multiple Tests of Significance, *Biometrika* 73(3): 751–754. doi:10.1093/biomet/73.3.751

Sonesson, C. (2007). A CUSUM Framework for Detection of Space–Time Disease Clusters Using Scan Statistics, *Statistics in Medicine* 26: 4770–4789. doi:10.1002/sim.2898

Stein, C. (1945). A Two-Sample Test for a Linear Hypothesis Whose Power Is Independent of the Variance, *Annals of Mathematical Statistics* 16: 243–258. doi:10.1214/aoms/1177731088

Storey, J. D. (2002). A Direct Approach to False Discovery Rates, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(3): 479–498. doi:10.1111/1467-9868.00346

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66(1): 187–205. doi:10.1111/j.1467-9868.2004.00439.x

Storey, J. D. and Tibshirani, R. (2003). Statistical Significance for Genomewide Studies, *Proceedings of National Academy of Sciences* 100(16): 9440–9445. doi:10.1073/pnas.1530509100

Wald, A. (1947). *Sequential Analysis*, New York: Wiley. Repr., Dover, 1973.

Woodall, W. H. (2006). The Use of Control Charts in Health-Care and Public-Health Surveillance, *Journal of Quality Technology* 38(2): 89–104. doi:10.1080/00224065.2006.11918593