Statistica Sinica **28** (2018), 363-398
doi:https://doi.org/10.5705/ss.202015.0267

MULTIPLE HYPOTHESIS TESTS CONTROLLING GENERALIZED ERROR RATES FOR SEQUENTIAL DATA

Jay Bartroff

University of Southern California

Abstract: The γ-FDP and k-FWER multiple testing error metrics, which are tail probabilities of the respective error statistics, have become popular recently as alternatives to the FDR and FWER. We propose general and flexible stepup and stepdown procedures for testing multiple hypotheses about sequential (or streaming) data that simultaneously control both the type I and II versions of γ-FDP, or k-FWER. The error control holds regardless of the dependence between data streams, which may be of arbitrary size and shape. All that is needed is a test statistic for each data stream that controls the 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 procedures can be used with sequential, group sequential, truncated, or other sampling schemes. We give recommendations for the procedures' implementation including closed-form expressions for the needed critical values in some commonly-encountered testing situations. The proposed sequential procedures are compared with each other and with comparable fixed sample size procedures in the context of strongly positively correlated Gaussian data streams. For this setting we conclude that both the stepup and stepdown sequential procedures provide substantial savings over the fixed sample procedures in terms of expected sample size, and the stepup procedure performs slightly but consistently better than the stepdown for γ-FDP control, with the relationship reversed for k-FWER control.

Key words and phrases: False discovery proportion, familywise error, generalized error rate, high-dimensional statistics, multiple comparisons, multiple testing, sequential analysis, sequential hypothesis testing, stepdown procedure, stepup procedure, Wald approximations.

1. Introduction and Summary

Driven in part by modern applications involving high-dimensional models or the need for many comparisons in areas such as high-throughput gene and protein expression data, brain imaging, and astrophysics, there has been much interest and innovation during recent decades in statistical methodology involving multiple testing error rates which are less stringent than the classical *familywise* *error rate (FWER)*, the probability of rejecting at least one true null hypothesis. Hommel and Hoffmann (1988) proposed the k-FWER, the probability of rejecting at least $k \geq 1$ true null hypotheses, and this was independently proposed by Lehmann and Romano (2005). Benjamini and Hochberg (1995) proposed the *false discovery rate (FDR)*, the expectation of the *false discovery proportion (FDP)*, the latter being the proportion of rejected null hypotheses that are true. As a generalization of the FDR, Lehmann and Romano (2005) proposed using the probability that the FDP exceeds a fixed value $\gamma \in [0, 1)$, which has come to be known as the γ-FDP. Recently, Guo, He and Sarkar (2014) proposed a further generalization of the γ-FDP. Each of these works supplied procedures to control the respective generalized error rates under various dependence assumptions on the data, ranging from independence to positive regression dependency on subsets (Benjamini and Yekutieli (2001)) to no assumptions at all. Many other authors have also provided innovative new procedures and theory surrounding these generalized error rates and we do not attempt to summarize this large and growing literature here, but instead refer the reader to Guo, He and Sarkar (2014) and the references therein.

All of the mentioned references take as their starting point a set of valid p-values corresponding to fixed sample size tests for the list of null hypotheses of interest. However, in some areas of application the data in a multiple testing setup does not naturally occur in a fixed sample but rather arrives sequentially (or in groups) in time, referred to as “streaming” data in some applications. An obvious example is in biomedical clinical trials with multiple endpoints or arms (e.g., Jennison and Turnbull (2000, Chap. 15)), but others areas with naturally sequential data abound including certain high-throughput sequencing technologies (Salzman, Jiang and Wong (2011); Jiang and Salzman (2012)), multi-channel changepoint detection (Tartakovsky, Li and Yaralov (2003)), biosurveillance (Mei (2010)), acceptance sampling with multiple criteria (Baillie (1987)), financial data (Lai and Xing (2008)), and some agricultural studies (Clements et al. (2014)). Only recently have general and flexible multiple testing procedures suited for the particular needs of sequential data been proposed in the literature. Bartroff and Lai (2010) proposed a sequential version of Holm’s (1979) FWER-controlling procedure. De and Baron (2012a,b) proposed procedures controlling both the type I and II FWER under the restriction that all data streams be sampled until accept/reject decisions can be reached for all null hypotheses simultaneously, and Bartroff and Song (2014b) proposed a procedure lifting this restriction. Like Holm’s procedure, each of these sequential procedures mentioned so far has guaranteed FWER control under arbitrary dependence of data streams. Bartroff and Song (2014a) proposed an analogous procedure controlling FDR and its type II analog, the false nondiscovery rate, on sequential data.

The purpose of this paper is to provide general and flexible procedures for controlling k-FWER and γ-FDP on sequential data. By "general and flexible" we mean procedures that can test $J \geq 2$ arbitrary null hypotheses $H^{(1)}, \ldots, H^{(J)}$ about J data streams

$$\begin{aligned} &\text{Data stream 1:} \quad X_1^{(1)}, X_2^{(1)}, \ldots, \\ &\text{Data stream 2:} \quad X_1^{(2)}, X_2^{(2)}, \ldots, \\ &\qquad\qquad\vdots \\ &\text{Data stream } J\text{:} \quad X_1^{(J)}, X_2^{(J)}, \ldots, \end{aligned} \tag{1.1}$$

respectively, of arbitrary size, shape, and dependence. In particular, each data point $X_n^{(j)}$ may itself be the vector of observations from the nth group, corresponding to group sequential sampling. This setup is formalized below. In particular we define stepdown and stepup procedures that require only arbitrary sequential test statistics $\Lambda^{(j)}(n) = \Lambda^{(j)}(n)(X_1^{(j)}, \ldots, X_n^{(j)})$ for each stream $j = 1, \ldots, J$ that control the conventional type I and II error probabilities for each individual null hypothesis $H^{(j)}$, and combine them to give a sequential multiple testing procedure as a collection of J sequential stopping and decision rules for each data stream, that controls k-FWER or γ-FDP at a prescribed level under arbitrary dependence structure between data streams. In this regard our procedures can be viewed as extensions to the sequential realm of the procedures of Lehmann and Romano (2005) and Romano and Shaikh (2006a,b) who accomplished this in the fixed sample setup. Indeed, our approach owes much to the work of these authors and, in particular, we utilize the same stepup and stepdown values as they do. Sarkar (2007, 2008) and Guo, He and Sarkar (2014) have furthered the work of these authors by developing stepup and stepdown fixed sample size procedures that utilize the joint null distribution of the p-values and, in some cases, dominate previously proposed procedures while controlling generalized error rates. While we expect that these innovations by Sarkar and his coauthors can be similarly extended to the sequential domain, since our goal here is to propose procedures that do not require knowledge (or modeling) of joint distributions, we have not pursued those extensions here.

An additional aspect of our approach is that our procedures may be able to simultaneously control both the type I and II versions of the generalized error metrics at prescribed values, which is a possibility opened up by the sequential setting considered. As with single hypothesis testing, if the prescribed type II error rate (or equivalently, power) is not well-motivated, then its strict control can be dispensed with or used as a surrogate for other operating characteristics of interest, such as average sample size. For this reason and others discussed in Section 4, there we provide versions of the procedures that control the type I generalized error rate but not necessarily the type II version, and these procedures can be used with arbitrary acceptance rules for the null hypotheses.

Regarding our sequential setup, we remark that in order to sequentially test $J \geq 2$ null hypotheses, one could simply apply J chosen sequential stopping rules to the data streams, calculate the (appropriately adjusted) p-values upon stopping, and then apply a fixed-sample procedure to the p-values. However, this "naive" method will in general be inefficient compared to the procedures proposed herein since the stopping rules do not explicitly take the multiple testing error metric into account. Moreover, the naive method will not in general control both the type I and II multiple testing error rates which, even when this feature is not a priority of the statistician, means that the relationship between the naive method's stopping rule and its power is not well understood, unlike the proposed procedures. Nonetheless, our approach could be applied by taking the test statistics $\Lambda^{(j)}(n)$ to be sequential p-values, making it look more like the fixed sample size procedures. Instead we have chosen to use arbitrary sequential test statistics to maintain generality and to make the resulting procedure more user-friendly, given that other types of test statistics like log-likelihood ratios (or simple functions thereof) are much more commonly used with sequential data than sequential p-values. This is perhaps due to the complexity and non-uniqueness of sequential p-values in all but the simplest cases; see Jennison and Turnbull (2000, Chaps 8.4 and 9).

The remainder of this paper is organized as follows. After introducing the notation and setup in Section 2, in Section 3.1 we define a "generic" sequential stepdown procedure that accepts arbitrary stepdown values (2.4), special cases of which are given in Sections 3.1.2 and 3.1.3, that control type I and II γ-FDR and k-FWER, respectively. An analogous development of stepup procedures in given in Section 3.2. Section 4 gives versions of these procedures with only explicit rejection rules for use when the type II error rate is not well motivated or there is a restriction on maximum sample size. In Section 5 we give recommendations for implementing the procedures by reviewing how to implement sequential single-hypothesis tests in some commonly-encountered situations, and we give closed-form expressions for the needed critical values in Theorem 7. Section 5.2 discusses how to implement group sequential sampling. Section 6 contains the results of a numerical study comparing the proposed stepup, stepdown, and comparable fixed sample size procedures in a setting of strongly positively correlated Gaussian data streams. In Section 7 we summarize our recommendations. All proofs can be found in a supplemental document.

2. Setup

2.1. Data streams, hypotheses and error metrics

Assume that there are $J \geq 2$ data streams (1.1). In general we make no assumptions about the dimension of the sequentially-observed data $X_n^{(j)}$, which may themselves be vectors of varying size, nor about the dependence structure of within-stream data $X_n^{(j)}, X_m^{(j)}$ or between-stream data $X_n^{(j)}, X_m^{(j')}$ $(j \neq j')$. In particular there can be arbitrary "overlap" between data streams, an extreme case being that all the data streams are the same, which is equivalent to testing multiple hypotheses about a single data source. For any positive integer j let $[j] = \{1, \ldots, j\}$. For each data stream, indexed by $j \in [J]$, assume that there is a parameter vector $\theta^{(j)} \in \Theta^{(j)}$ determining that distribution of the stream $X_1^{(j)}, X_2^{(j)}, \ldots$, and it is desired to test a null hypothesis $H^{(j)}$ versus the alternative hypothesis $G^{(j)}$, where $H^{(j)}$ and $G^{(j)}$ are disjoint subsets of the parameter space $\Theta^{(j)}$ containing $\theta^{(j)}$. The null $H^{(j)}$ is considered *true* if $\theta^{(j)} \in H^{(j)}$, and *false* if $\theta^{(j)} \in G^{(j)}$. The global parameter $\theta = (\theta^{(1)}, \ldots, \theta^{(J)})$ is the concatenation of the individual parameters and is contained in the global parameter space $\Theta = \Theta^{(1)} \times \cdots \times \Theta^{(J)}$. Let

$$\mathcal{T}(\theta) = \{j \in [J] : \theta^{(j)} \in H^{(j)}\} \tag{2.1}$$

denote the indices of the true null hypotheses when θ is the true global parameter, and

$$\mathcal{F}(\theta) = \{j \in [J] : \theta^{(j)} \in G^{(j)}\} \tag{2.2}$$

the indices of the false null hypotheses.

It may appear that the notation (1.1) for the data streams restricts us to fully-sequential sampling where the streamwise sample sizes may take any value $1, 2, \ldots$ *ad infinitum*. However, since the observations $X_n^{(j)}$ themselves may be of arbitrary size and shape, group sequential (and even variable-stage size) sampling fits into this framework. To wit, the nth "observation" $X_n^{(j)}$ in the jth stream may actually be the nth group $X_n^{(j)} = (X_{n,1}^{(j)}, \ldots, X_{n,\ell}^{(j)})$ of size ℓ. Moreover, the group size ℓ may vary with n and may even be data-dependent, e.g., determined by some type of adaptive sampling. Similarly, truncated sampling can be implemented for the jth stream by defining $X_n^{(j)} = \emptyset$ for all $n > \overline{N}^{(j)}$ for some stream-specific truncation point $\overline{N}^{(j)}$, or globally for all streams by replacing statements like "for some n" in what follows with "for some $n \leq \overline{N}$," for some global truncation point $\overline{N}$.

The FDP is formally defined as

$$\mathrm{FDP}(\theta) = \begin{cases} \dfrac{\text{the number of } H^{(j)} \text{ rejected, } j \in \mathcal{T}(\theta)}{\text{the number of } H^{(j)} \text{ rejected}}, & \text{if the denominator is positive,} \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

For example, as mentioned above, Benjamini and Hochberg's (1995) FDR is the expectation $E_\theta(\mathrm{FDP}(\theta))$ of the FDP. Since we will consider procedures that simultaneously control both the type I and type II versions of the generalized error rates, we also define the type II analog of FDP, which we call the *false nondiscovery proportion (FNP)*,

$$\mathrm{FNP}(\theta) = \begin{cases} \dfrac{\text{the number of } H^{(j)} \text{ accepted, } j \in \mathcal{F}(\theta)}{\text{the number of } H^{(j)} \text{ accepted}}, & \text{if the denominator is positive,} \\ 0, & \text{otherwise.} \end{cases}$$

With FDP and FNP nailed down, for $\gamma_1, \gamma_2 \in [0, 1)$ we define

$$\gamma_1\text{-FDP}(\theta) = P_\theta(\mathrm{FDP}(\theta) > \gamma_1) \quad \text{and} \quad \gamma_2\text{-FNP}(\theta) = P_\theta(\mathrm{FNP}(\theta) > \gamma_2).$$

Similarly, for k-FWER we will distinguish the type I and II versions by, for $k_1, k_2 \in [J]$, defining

$$k_1\text{-FWER}_1(\theta) = P_\theta(\text{at least } k_1 \text{ null hypotheses } H^{(j)} \text{ rejected, } j \in \mathcal{T}(\theta)),$$
$$k_2\text{-FWER}_2(\theta) = P_\theta(\text{at least } k_2 \text{ null hypotheses } H^{(j)} \text{ accepted, } j \in \mathcal{F}(\theta)).$$

We will omit the argument θ from these quantities in what follows when it causes no confusion.

2.2. Test statistics and critical values

The building blocks of our sequential procedures are J individual sequential test statistics $\{\Lambda^{(j)}(n)\}_{j \in [J],\, n \geq 1}$, where $\Lambda^{(j)}(n)$ is the statistic for testing $H^{(j)}$ vs. $G^{(j)}$ based on the data $X_1^{(j)}, X_2^{(j)}, \ldots, X_n^{(j)}$ available from the jth stream at time n. For example, $\Lambda^{(j)}(n)$ may be a sequential log likelihood ratio statistic for testing $H^{(j)}$ vs. $G^{(j)}$. Our stepup and stepdown procedures are defined in terms of given constants

$$0 \le \alpha_1 \le \ldots \le \alpha_J \le 1 \quad \text{and} \quad 0 \le \beta_1 \le \ldots \le \beta_J \le 1, \tag{2.4}$$

which we refer to as *step values*, the α_j corresponding to type I error control and the β_j to type II. These values are used in a similar way as in fixed sample size stepdown and stepup procedures, which we review now for comparison. Based on p-values $p^{(j_1)} \le \ldots \le p^{(j_J)}$ with $p^{(j)}$ corresponding to $H^{(j)}$, the stepdown procedure based on constants α_j satisfying (2.4) rejects $H^{(j_1)}, \ldots, H^{(j_d)}$ where $d = \max\{i \in [J] : p^{(j_{i'})} \le \alpha_{i'} \text{ for all } i' \le i\}$ (accepting all nulls if the maximum does not exist), whereas the stepup procedure rejects $H^{(j_1)}, \ldots, H^{(j_u)}$ where $u = \max\{i \in [J] : p^{(j_i)} \le \alpha_i\}$ (accepting all nulls if the maximum does not exist). Here $d \le u$ so that the stepup procedure rejects at least as many null hypotheses as the corresponding stepdown procedure using the same step values.

Given step values $\{\alpha_j, \beta_j\}_{j \in [J]}$, for each test statistic $\Lambda^{(j)}(n)$ we assume the existence of critical values $\{A_w^{(j)}, B_w^{(j)}\}_{w \in [J]}$ such that

$$P_{\theta^{(j)}}(\Lambda^{(j)}(n) \ge B_w^{(j)} \text{ some } n, \Lambda^{(j)}(n') > A_1^{(j)} \text{ all } n' < n) \le \alpha_w \text{ for all } \theta^{(j)} \in H^{(j)}, \tag{2.5}$$

$$P_{\theta^{(j)}}(\Lambda^{(j)}(n) \le A_w^{(j)} \text{ some } n, \Lambda^{(j)}(n') < B_1^{(j)} \text{ all } n' < n) \le \beta_w \text{ for all } \theta^{(j)} \in G^{(j)} \tag{2.6}$$

for all $w \in [J]$. The critical values $A_1^{(j)}, B_1^{(j)}$ are simply the critical values for the sequential test that samples until $\Lambda^{(j)}(n) \notin (A_1^{(j)}, B_1^{(j)})$, and this test has type I and II error probabilities bounded above by α_1 and β_1, respectively. The values $B_w^{(j)}$, $w \in [J]$, are then such that the similar sequential test with critical values $A_1^{(j)}$ and $B_w^{(j)}$ has type I error probability α_w, which is just a restatement of (2.5), with an analogous statement holding for critical values $A_w^{(j)}$ and $B_1^{(j)}$, type II error probability β_w, and (2.6). The reason that critical values $A_1^{(j)}$ and $B_w^{(j)}$ are considered in (2.5) for type I error probability control and not, say, $A_w^{(j)}$ and $B_w^{(j)}$ is that the procedures defined below sample during the ith stage using critical values $A_w^{(j)}$ and $B_{w'}^{(j)}$ for some fixed values $w, w' \in [J]$ determined by the data in the previous stages $1, \ldots, i-1$. The probability that, during the ith stage, $\Lambda^{(j)}(n) \ge B_{w'}^{(j)}$ before $\Lambda^{(j)}(n) \le A_w^{(j)}$ is then be bounded above by the corresponding statement with $A_w^{(j)}$ replaced by $A_1^{(j)}$, using the fact that $A_1^{(j)} \le A_w^{(j)}$ by (2.7), and thus this probability related to (2.5) after bounding w'. Analogous statements apply regarding bounding the type II error probability.

In all commonly-encountered testing situations there are standard sequential statistics whose critical values can be chosen that satisfy these error bounds, for any given $\{\alpha_j, \beta_j\}_{j \in [J]}$ (Bartroff and Song (2014b) give examples). Without loss of generality we assume that, for each $j \in [J]$,

$$A_1^{(j)} \le A_2^{(j)} \le \ldots \le A_J^{(j)} \le B_J^{(j)} \le B_{J-1}^{(j)} \le \ldots \le B_1^{(j)}, \tag{2.7}$$

$$A_w^{(j)} = A_{w+1}^{(j)} \text{ if and only if } \beta_w = \beta_{w+1}, \tag{2.8}$$

$$B_w^{(j)} = B_{w+1}^{(j)} \text{ if and only if } \alpha_w = \alpha_{w+1}. \tag{2.9}$$

A simplistic example of how critical values (2.7) are used in our sequential multiple testing procedure is given in the last two paragraphs of Section 3.1.1.

Our sequential multiple testing procedures involve ranking the test statistics associated with different data streams, which may be on completely different scales in general, so for each stream j we introduce a *standardizing function* $\varphi^{(j)}(\cdot)$ which is applied to the statistic $\Lambda^{(j)}(n)$ before ranking. The standardizing functions $\varphi^{(j)}$ can be any increasing functions such that $\varphi^{(j)}(A_w^{(j)})$ and $\varphi^{(j)}(B_w^{(j)})$ do not depend on j, and we let

$$a_w = \varphi^{(j)}(A_w^{(j)}) \quad \text{and} \quad b_w = \varphi^{(j)}(B_w^{(j)}), \quad j, w \in [J], \tag{2.10}$$

denote these common values. Given critical values $\{A_w^{(j)}, B_w^{(j)}\}_{j,w\in[J]}$ satisfying (2.5)-(2.6), one may choose arbitrary values $\{a_w, b_w\}_{w\in[J]}$ satisfying the same monotonicity conditions as the $\{A_w^{(j)}, B_w^{(j)}\}$ according to (2.8)-(2.9) and then define the standardizing functions $\varphi^{(j)}(\cdot)$ to be increasing, piecewise linear functions satisfying (2.10). For example, if all the α_w are distinct and the β_w are distinct then a simple choice for the $\{a_j, b_j\}$ are the integers

$$a_1 = -J, \quad a_2 = -J+1, \quad \ldots, \quad a_J = -1, \quad b_J = 1, \quad b_{J-1} = 2, \quad \ldots, \quad b_1 = J.$$

In any case, the assumptions on the critical values and standardizing functions imply that the a_w must be nondecreasing and the b_w nonincreasing. Finally, we denote $\widetilde{\Lambda}^{(j)}(n) = \varphi^{(j)}(\Lambda^{(j)}(n))$ and then (2.5)-(2.6) can be written as

$$P_{\theta^{(j)}}(\widetilde{\Lambda}^{(j)}(n) \ge b_w \text{ some } n, \widetilde{\Lambda}^{(j)}(n') > a_1 \text{ all } n' < n) \le \alpha_w \quad \text{for all} \quad \theta^{(j)} \in H^{(j)}, \tag{2.11}$$

$$P_{\theta^{(j)}}(\widetilde{\Lambda}^{(j)}(n) \le a_w \text{ some } n, \widetilde{\Lambda}^{(j)}(n') < b_1 \text{ all } n' < n) \le \beta_w \quad \text{for all} \quad \theta^{(j)} \in G^{(j)}, \tag{2.12}$$

for all $j, w \in [J]$.

3. Procedures Controlling Type I and II Generalized Error Rates

3.1. Stepdown procedures

3.1.1. The generic sequential stepdown procedure

Here we define a generic sequential stepdown procedure, special cases of which are used to define the type I and II k-FWER and γ-FDP controlling sequential procedures. We assume that step values $\{\alpha_j, \beta_j\}_{j\in[J]}$ satisfying (2.4) are given and that the test statistics and critical values satisfy the assumptions in Section 2.2 with respect to these values.

We describe the procedure in terms of stages of sampling, between which reject/accept decisions are made. Let $\mathcal{J}_i \subseteq [J]$ $(i = 1, 2, \ldots)$ denote the index set of the *active* data streams, those whose corresponding null hypothesis $H^{(j)}$ has been neither accepted nor rejected yet, at the beginning of the ith stage of sampling, and n_i denote the cumulative sample size of any active test statistic up to and including the ith stage. The total number of null hypotheses that have been rejected (resp. accepted) at the beginning of the ith stage is denoted by r_i (resp. c_i). Accordingly, set $\mathcal{J}_1 = [J]$, $n_0 = 0$, $r_1 = c_1 = 0$. Let $|\cdot|$ denote set cardinality. Then the ith stage of sampling $(i = 1, 2, \ldots)$ of the **Generic Sequential Stepdown Procedure** with step values $\{\alpha_j, \beta_j\}_{j\in[J]}$ proceeds as follows.

1. Sample the active streams $\{X_n^{(j)}\}_{j\in\mathcal{J}_i,\, n>n_{i-1}}$ until n equals
$$n_i = \inf\left\{n > n_{i-1} : \widetilde{\Lambda}^{(j)}(n) \notin (a_{c_i+1}, b_{r_i+1}) \quad \text{for some} \quad j \in \mathcal{J}_i\right\}. \tag{3.1}$$

2. Order the active test statistics
$$\widetilde{\Lambda}^{(j(n_i,1))}(n_i) \le \widetilde{\Lambda}^{(j(n_i,2))}(n_i) \le \ldots \le \widetilde{\Lambda}^{(j(n_i,|\mathcal{J}_i|))}(n_i),$$
where $j(n_i, \ell)$ denotes the index of the ℓth ordered active statistic at the end of stage i.

3. (a) If the upper boundary in (3.1) has been crossed, $\widetilde{\Lambda}^{(j)}(n_i) \ge b_{r_i+1}$ for some $j \in \mathcal{J}_i$, then reject the $m_i \ge 1$ null hypotheses
$$H^{(j(n_i,|\mathcal{J}_i|))}, H^{(j(n_i,|\mathcal{J}_i|-1))}, \ldots, H^{(j(n_i,|\mathcal{J}_i|-m_i+1))}, \tag{3.2}$$
where
$$m_i = \max\big\{m \in [|\mathcal{J}_i|] : \widetilde{\Lambda}^{(j(n_i,\ell))}(n_i) \ge b_{r_i+|\mathcal{J}_i|-\ell+1} \\ \text{for all} \quad \ell = |\mathcal{J}_i| - m + 1, \ldots, |\mathcal{J}_i|\big\},$$
and set $r_{i+1} = r_i + m_i$. Otherwise set $r_{i+1} = r_i$.

   (b) If the lower boundary in (3.1) was crossed, $\widetilde{\Lambda}^{(j)}(n_i) \le a_{c_i+1}$ for some $j \in \mathcal{J}_i$, then accept the $m_i' \ge 1$ null hypotheses
$$H^{(j(n_i,1))}, H^{(j(n_i,2))}, \ldots, H^{(j(n_i,m_i'))},$$
where

$$m_i' = \max\left\{ m \in [|\mathcal{J}_i|] : \widetilde{\Lambda}^{(j(n_i,\ell))}(n_i) \le a_{c_i+\ell} \quad \text{for all} \quad \ell = 1, \ldots, m \right\},$$

and set $c_{i+1} = c_i + m_i'$. Otherwise set $c_{i+1} = c_i$.

4. Stop if there are no remaining active hypotheses, $r_{i+1}+c_{i+1} = J$. Otherwise, let $\mathcal{J}_{i+1}$ be the indices of the remaining active hypotheses and continue on to stage $i+1$.

Thus the procedure samples all active data streams until at least one of the active null hypotheses can be accepted or rejected, indicated by the stopping rule (3.1). At that point, stepdown rejection/acceptance rules are used in steps 3a/3b to reject/accept some active null hypotheses. After updating the list of active hypotheses, the process is repeated until no active hypotheses remain.

Remark 1. (A) The relationships (2.7)-(2.10) ensure that there is never a conflict between the rejections in Step (3a) and the acceptances in Step (3b).

(B) Ties in the order statistics $\widetilde{\Lambda}_n^{(j)}$ in Step 2 can be broken arbitrarily (at random, say) without affecting any of the error control properties in our Theorems 1 and 2.

(C) If common critical values can be used for all data streams, $A_w^{(j)} = A_w^{(j')} = A_w$ and $B_w^{(j)} = B_w^{(j')} = B_w$ for all $j, j', w \in [J]$, then the standardizing functions can be dispensed with and we take $\varphi^{(j)}(x) = x$ giving $a_j = A_j$ and $b_j = B_j$ for all $j \in [J]$.

(D) The critical values $A_w^{(j)}, B_w^{(j)}$ may also depend on the current sample size n of the test statistic $\Lambda^{(j)}(n)$ being compared with them, with only notational changes in the definition of the generic procedure and the properties proved below; for simplicity we omit this from the presentation. Such standard group sequential stopping boundaries like Pocock, O'Brien-Fleming, power family, and any others (see Jennison and Turnbull (2000, Chaps 2 and 4)) can be utilized for the individual test statistics in this way.

(E) The stopping time n_i of the ith stage, given by (3.1), is determined by the numbers c_i and r_i of null hypotheses that have been rejected and accepted, respectively, during prior stages $1, \ldots, i-1$. Therefore this stopping rule is completely determined before the start of the ith stage and, in particular, unambiguously defined.

Example 1. To show the mechanics of the procedure we summarize an example in Bartroff and Song (2014b, p. 104); details of the test statistics and critical values are given there and omitted here. There are $J = 3$ data streams, null/alternative hypothesis pairs $(H^{(j)}, G^{(j)})$, and sequential test statistics $\Lambda^{(j)}(n)$ with common critical values $A_w^{(j)} = A_w^{(j')} = A_w$ and $B_w^{(j)} = B_w^{(j')} = B_w$ for all $j, j', w \in \{1, 2, 3\}$, which are given in the header of Table 1. In particular, per Remark C we take $a_j = A_j$, $b_j = B_j$, and $\widetilde{\Lambda}^{(j)}(n) = \Lambda^{(j)}(n)$ in the definition of the procedure. Table 1 contains three simulated sample paths and the critical values are given in the table's header. We focus on how the critical values determine the procedure's decisions to stop or continue sampling; the values of the stopped test statistics are given in bold in the table.

On sample path 1, sampling proceeds until time $n_1 = 7$ when $H^{(1)}$ and $H^{(2)}$ are rejected because this is the first time any of the three test statistics exceed B_1 or fall below A_1. In particular, $H^{(1)}$ is rejected because $\Lambda^{(1)}(7) = 2.03 \geq B_1 = 1.93$ and $H^{(2)}$ is also rejected at this time because $\Lambda^{(2)}(7) = 2.03 \geq B_2 = 1.53$, while one null hypothesis $H^{(1)}$ has already been rejected; the fact that $\Lambda^{(2)}(7)$ also exceeds B_1 was not necessary for rejecting $H^{(2)}$. Sampling of stream 3 is continued until time $n_2 = 10$ when $H^{(3)}$ is accepted because its test statistic falls below $A_1 = -2.43$. Similarly, on sample path 2, after rejecting $H^{(1)}$ at time $n_1 = 7$, $H^{(2)}$ is then rejected at time $n_2 = 8$ because $\Lambda^{(2)}(8)$ exceeds $B_2 = 1.53$ and $H^{(1)}$ has already been rejected. $H^{(3)}$ is also accepted at time $n_2 = 8$ for the same reason as above. On sample path 3, all three null hypotheses are rejected at time $n_1 = 7$ because $\Lambda^{(1)}(7) = 2.03 \geq B_1$, $\Lambda^{(2)}(7) = 2.03 \geq B_2$ and $H^{(1)}$ has already been rejected, and $\Lambda^{(3)}(7) = 1.22 \geq B_3$ and $H^{(1)}$ and $H^{(2)}$ have already been rejected.

3.1.2. A stepdown procedure controlling γ_1-FDP and γ_2-FNP

The following step values[1] were proposed by Romano and Shaikh (2006a). For $v \in [J]$ and $\gamma \in [0, 1)$, take

$$\bar{j}(t, v, \gamma) = \min\{J, J + t - v, \lceil t/\gamma \rceil - 1\} \quad \text{for} \quad t \in [\lfloor \gamma J \rfloor + 1], \tag{3.3}$$

$$\bar{t}(v, \gamma) = \min\left\{ \lfloor \gamma J \rfloor + 1, v, \left\lfloor \frac{\gamma(J - v)}{1 - \gamma} \right\rfloor + 1 \right\}, \tag{3.4}$$

omitting the third term in the minimum in (3.3) if $\gamma = 0$. Given a nondecreasing sequence $0 \leq \delta_1 \leq \ldots \leq \delta_J \leq 1$, for $v \in [J]$ and $\gamma \in [0, 1)$ define

$$\varepsilon(t, v, \gamma, \{\delta_j\}) = \delta_{\bar{j}(t,v,\gamma)} \quad \text{for} \quad t \in [\lfloor \gamma J \rfloor + 1],$$

[1] See Remark 3.

Table 1. Three sample paths of a stepdown procedure for $J = 3$ hypotheses using critical values $A_1 = -2.34$, $A_2 = -1.94$, $A_3 = -1.27$, $B_1 = 1.93$, $B_2 = 1.53$, $B_3 = 0.86$. The values of the stopped sequential statistics are in bold. From Bartroff and Song (2014b, p. 104).

Data Stream		$n = 1$	2	3	4	5	6	7	8	9	10
Sample Path 1											
1	$X_n^{(1)}$	0	1	1	1	1	1	1			
	$\Lambda^{(1)}(n)$	−0.41	0.00	0.41	0.81	1.22	1.62	**2.03**			
2	$X_n^{(2)}$	1	0	1	1	1	1	1			
	$\Lambda^{(2)}(n)$	0.41	0.00	0.41	0.81	1.22	1.62	**2.03**			
3	$X_n^{(3)}$	0	1	0	0	1	0	0	0	0	0
	$\Lambda^{(3)}(n)$	−0.41	0.00	−0.41	−0.81	−0.41	−0.81	−1.22	−1.62	−2.03	−**2.43**
Sample Path 2											
1		0	1	1	1	1	1	1			
		−0.41	0.00	0.41	0.81	1.22	1.62	**2.03**			
2		1	0	0	1	1	1	1	1		
		0.41	0.00	−0.41	0.00	0.41	0.81	1.22	**1.62**		
3		0	1	0	0	0	0	0	0		
		−0.41	0.00	−0.41	−0.81	−1.22	−1.62	−2.03	−**2.43**		
Sample Path 3											
1		1	0	1	1	1	1	1			
		0.41	0.00	0.41	0.81	1.22	1.62	**2.03**			
2		1	1	1	0	1	1	1			
		0.41	0.81	1.22	0.81	1.22	1.62	**2.03**			
3		0	1	0	1	1	1	1			
		−0.41	0.00	−0.41	0.00	0.41	0.81	**1.22**			

$$S_1(v, \gamma, \{\delta_j\}) = v \sum_{t=1}^{\bar{t}(v,\gamma)} \frac{\varepsilon_t - \varepsilon_{t-1}}{t} \quad \text{where} \quad \varepsilon_t = \varepsilon(t, v, \gamma, \{\delta_j\}) \quad \text{and} \quad \varepsilon_0 = 0,$$

$$D_1(\gamma, \{\delta_j\}) = \max_{0 \le v \le J} S_1(v, \gamma, \{\delta_j\}).$$

These quantities also depend on the total number J of null hypotheses but we have suppressed this in the notation since J is fixed throughout.

Theorem 1. *Fix $\alpha, \beta \in (0, 1)$ and $\gamma_1, \gamma_2 \in [0, 1)$. Given any sequences of constants $0 \le \delta_1 \le \ldots \le \delta_J \le 1$ and $0 \le \eta_1 \le \ldots \le \eta_J \le 1$, take*

$$\alpha_j = \frac{\alpha \delta_j}{D_1(\gamma_1, \{\delta_{j'}\})}, \quad \beta_j = \frac{\beta \eta_j}{D_1(\gamma_2, \{\eta_{j'}\})}, \quad j \in [J]. \tag{3.5}$$

If the test statistics and critical values satisfy the assumptions in Section 2.2 for these $\{\alpha_j, \beta_j\}_{j \in [J]}$, then the sequential stepdown procedure with step values (3.5) *satisfies*

$$\gamma_1\text{-}FDP(\theta) \leq \alpha \quad \textit{and} \quad \gamma_2\text{-}FNP(\theta) \leq \beta \quad \textit{for all} \quad \theta \in \Theta$$

regardless of the dependence between data streams.

Remark 2. A special case of the theorem is

$$\delta_j = \frac{\lfloor \gamma_1 j \rfloor + 1}{J + \lfloor \gamma_1 j \rfloor + 1 - j}, \quad \eta_j = \frac{\lfloor \gamma_2 j \rfloor + 1}{J + \lfloor \gamma_2 j \rfloor + 1 - j}, \quad j \in [J]. \tag{3.6}$$

Of course there are other possibilities, such as $\delta_j = \eta_j = j/J$, which give step values proportional to the ones used in the original FDR-controlling procedure of Benjamini and Hochberg (1995), although Romano and Shaikh (2006a, p. 44) found these to be smaller (and thus less desirable) than the step values (3.5) given by (3.6), for the most part.

Remark 3. The third term in (3.4) is a slight improvement over the corresponding third term in Romano and Shaikh (2006a, Eq. (3.11)), and our proof holds in their fixed sample size setting, giving a slightly improved upper bound for the number of true hypotheses.

3.1.3. A stepdown procedure controlling the k_1-FWER$_1$ and k_2-FWER$_2$

The stepdown procedure in the following theorem utilizes step values proposed by Lehmann and Romano (2005).

Theorem 2. *Fix* $\alpha, \beta \in (0,1)$, $k_1, k_2 \in [J]$, *and take*

$$\alpha_j = \frac{k_1 \alpha}{J - (j - k_1)^+}, \quad \beta_j = \frac{k_2 \beta}{J - (j - k_2)^+}, \quad j \in [J], \tag{3.7}$$

where $x^+ = \max\{x, 0\}$. *If the test statistics and critical values satisfy the assumptions in Section* 2.2 *for these* $\{\alpha_j, \beta_j\}_{j \in [J]}$, *then the sequential stepdown procedure with step values* (3.7) *satisfies*

$$k_1\text{-}FWER_1(\theta) \leq \alpha \quad \textit{and} \quad k_2\text{-}FWER_2(\theta) \leq \beta \quad \textit{for all} \quad \theta \in \Theta \tag{3.8}$$

regardless of the dependence between data streams.

Remark 4. Lehmann and Romano (2005, Thm. 2.3) exhibit a distribution of fixed sample size p-values for which the achieved (type I) FWER is exactly the prescribed value α. By taking $X_1^{(j)}$ in (1.1) to be the fixed sample size data and $X_n^{(j)} = \emptyset$ for $n > 1$, applying their example to both true and false null hypotheses shows that there is a distribution for the data such that the inequalities in (3.8) are equalities. In this sense the bounds (3.8) are sharp.

3.2. Stepup procedures

In this section we develop stepup procedures analogously to what was done for stepdown procedures in Section 3.1.

3.2.1. The generic sequential stepup procedure

Here we define a generic sequential stepup procedure, special cases of which are used to define our type I and II k-FWER and γ-FDP controlling sequential procedures. We assume that step values $\{\alpha_j, \beta_j\}_{j\in[J]}$ satisfying (2.4) are given and that the test statistics and critical values satisfy the assumptions in Section 2.2 with respect to these values.

We describe the stepup procedure in terms of stages of sampling, between which reject/accept decisions are made, and we use the notation $\mathcal{J}_i$, n_i, r_i, and c_i as before, with $\mathcal{J}_1 = [J]$, $n_0 = 0$, and $r_1 = c_1 = 0$. Then the ith stage of sampling ($i = 1, 2, \ldots$) of the **Generic Sequential Stepup Procedure** with step values $\{\alpha_j, \beta_j\}_{j\in[J]}$ proceeds as follows.

1. Sample the active data streams $\{X_n^{(j)}\}_{j\in\mathcal{J}_i,\, n>n_{i-1}}$ until n equals
$$n_i = \inf\{n > n_{i-1} : \widetilde{\Lambda}^{(j(n,\ell))}(n) \notin (a_{c_i+\ell}, b_{r_i+|\mathcal{J}_i|-\ell+1}) \quad \text{for some} \quad \ell \in [|\mathcal{J}_i|]\}, \tag{3.9}$$
where $j(n,\ell)$ denotes the index of the ℓth ordered active standardized statistic at sample size n.

2. (a) If an upper boundary in (3.9) was crossed,
$$\widetilde{\Lambda}^{(j(n_i,\ell))}(n_i) \geq b_{r_i+|\mathcal{J}_i|-\ell+1} \quad \text{for some} \quad \ell \in [|\mathcal{J}_i|],$$
then reject the $m_i \geq 1$ null hypotheses
$$H^{(j(n_i,|\mathcal{J}_i|))}, H^{(j(n_i,|\mathcal{J}_i|-1))}, \ldots, H^{(j(n_i,|\mathcal{J}_i|-m_i+1))},$$
where
$$m_i = \max\left\{m \in [|\mathcal{J}_i|] : \widetilde{\Lambda}^{(j(n_i,|\mathcal{J}_i|-m+1))}(n_i) \geq b_{r_i+m}\right\}, \tag{3.10}$$
and set $r_{i+1} = r_i + m_i$. Otherwise set $r_{i+1} = r_i$.

   (b) If a lower boundary in (3.9) was crossed,
$$\widetilde{\Lambda}^{(j(n_i,\ell))}(n_i) \leq a_{c_i+\ell} \quad \text{for some} \quad \ell \in [|\mathcal{J}_i|],$$
then accept the $m_i' \geq 1$ null hypotheses
$$H^{(j(n_i,m_i'))}, H^{(j(n_i,m_i'-1))}, \ldots, H^{(j(n_i,1))},$$
where

$$m_i' = \max\left\{m \in [|\mathcal{J}_i|] : \widetilde{\Lambda}^{(j(n_i,m))}(n_i) \le a_{c_i+m}\right\},$$

and set $c_{i+1} = c_i + m_i'$. Otherwise set $c_{i+1} = c_i$.

3. Stop if there are no remaining active hypotheses, $r_{i+1}+c_{i+1} = J$. Otherwise, let $\mathcal{J}_{i+1}$ be the indices of the remaining active hypotheses and continue on to stage $i+1$.

Thus the procedure samples all active data streams until at least one of the active null hypotheses can be accepted or rejected, indicated by the stopping rule (3.9). At that point, stepup rejection/acceptance rules are used in steps 2a/2b to reject/accept some active null hypotheses. After updating the list of active hypotheses, the process is repeated until no active hypotheses remain.

Remark 5. Points analogous to those of Remark 1 apply to the generic sequential stepup procedure as well.

3.2.2. A stepup procedure controlling γ_1-FDP and γ_2-FNP

The following step values were proposed by Romano and Shaikh (2006b). Given a nondecreasing sequence $0 \le \delta_1 \le \ldots \le \delta_J \le 1$, for $\gamma \in [0,1)$ and $v \in [J]$ take

$$S_2(v,\gamma,\{\delta_j\}) = v\delta_1 + v \sum_{v-J+1<s\le v,\ v\ge\lfloor\gamma(J-v+s)\rfloor+1} \frac{\delta_{J-v+s} - \delta_{J-v+s-1}}{s \vee (\lfloor\gamma(J-v+s)\rfloor+1)},$$

$$D_2(\gamma,\{\delta_j\}) = \max_{v\in[J]} S_2(v,\gamma,\{\delta_j\}).$$

Here $x \vee y = \max\{x,y\}$. These quantities also depend on J but we have suppressed this in the notation since J is fixed throughout.

Theorem 3. *Fix $\alpha,\beta \in (0,1)$ and $\gamma_1,\gamma_2 \in [0,1)$. Given any sequences of constants $0 \le \delta_1 \le \ldots \le \delta_J \le 1$ and $0 \le \eta_1 \le \ldots \le \eta_J \le 1$, take*

$$\alpha_j = \frac{\alpha\delta_j}{D_2(\gamma_1,\{\delta_{j'}\})}, \quad \beta_j = \frac{\beta\eta_j}{D_2(\gamma_2,\{\eta_{j'}\})}, \quad j \in [J]. \tag{3.11}$$

If the test statistics and critical values satisfy the assumptions in Section 2.2 for these $\{\alpha_j,\beta_j\}_{j\in[J]}$, then the sequential stepup procedure with step values (3.11) satisfies

$$\gamma_1\text{-}FDP(\theta) \le \alpha \quad \text{and} \quad \gamma_2\text{-}FNP(\theta) \le \beta \quad \text{for all} \quad \theta\in\Theta$$

regardless of the dependence between data streams.

Remark 6. A special case of the theorem is given by (3.6). Of course there are other possibilities, such as $\delta_j = \eta_j = j/J$, which give step values proportional to the ones used in the original FDR-controlling procedure of Benjamini and Hochberg (1995), although Romano and Shaikh (2006b, p. 1,865) found these to be smaller (and thus less desirable), for the most part, than the step values (3.11) given by (3.6).

Remark 7. Romano and Shaikh (2006a, Thm. 4.1(ii)) exhibit a joint distribution of p-values under which the procedure using step values (3.12) achieves γ_1-FDP$(\theta) = \alpha$. Since, as mentioned in Remark 4, the fixed-sample setting is a special case of the sequential setting, their example applies here as well, and the same argument gives a joint distribution under which γ_2-FNP $= \beta$. Thus, their result provides a weak optimality property of the sequential stepup procedure.

3.2.3. A Stepup Procedure Controlling k_1-FWER$_1$ and k_2-FWER$_2$

The following step values were proposed by Romano and Shaikh (2006b). Given a nondecreasing sequence $0 \leq \delta_1 \leq \ldots \leq \delta_J \leq 1$, for $k, v \in [J]$ let

$$S_3(v,k,\{\delta_j\}) = \frac{v\delta_{J-v+k}}{k} + v \sum_{k<s\leq v} \frac{\delta_{J-v+s} - \delta_{J-v+s-1}}{s},$$

$$D_3(k,\{\delta_j\}) = \max_{k\leq v\leq J} S_3(v,k,\{\delta_j\}).$$

These quantities also depend on J but we have suppressed this in the notation since J is fixed throughout.

Theorem 4. *Fix $\alpha, \beta \in (0,1)$ and $k_1, k_2 \in [J]$. Given any sequences of constants $0 \leq \delta_1 \leq \ldots \leq \delta_J \leq 1$ and $0 \leq \eta_1 \leq \ldots \leq \eta_J \leq 1$, take*

$$\alpha_j = \frac{\alpha\delta_j}{D_3(k_1,\{\delta_{j'}\})}, \quad \beta_j = \frac{\beta\eta_j}{D_3(k_2,\{\eta_{j'}\})}, \quad j \in [J]. \tag{3.12}$$

If the test statistics and critical values satisfy the assumptions in Section 2.2 for these $\{\alpha_j, \beta_j\}_{j\in[J]}$, then the sequential stepup procedure with step values (3.12) satisfies

$$k_1\text{-}FWER_1(\theta) \leq \alpha \quad \text{and} \quad k_2\text{-}FWER_2(\theta) \leq \beta \quad \text{for all} \quad \theta \in \Theta$$

regardless of the dependence between data streams.

Remark 8. A special case of the theorem is given by the constants

$$\delta_j = \frac{k_1}{J-(j-k_1)^+}, \quad \eta_j = \frac{k_2}{J-(j-k_2)^+}, \quad j \in [J], \tag{3.13}$$

which are proportional to those proposed by Hommel and Hoffmann (1988) and Lehmann and Romano (2005), as well as (3.7) in the proposed stepdown procedure. Other possibilities exist, such as $\delta_j = \eta_j = j/J$, but Romano and Shaikh

(2006b, p. 1,859) computed the resulting step values (3.12) for these choices and found those given by (3.13) to be larger (and hence more desirable) than those given by j/J for large or small values of j, and smaller for moderate values of j, but differing by relatively little in this case.

Remark 9. Romano and Shaikh (2006b, Thm. 3.1(ii)) exhibited a joint distribution of p-values under which the procedure using step values (3.12) achieves k_1-FWER$_1(\theta) = \alpha$. Since the fixed-sample setting is a special case of the sequential setting, their example applies here as well, and the same argument gives a joint distribution under which k_2-FWER$_2(\theta) = \beta$. Thus, their result provides a weak optimality property of the sequential stepup procedure.

4. Versions of the Procedures Controlling only the Type I Generalized Error Rate

In this section we describe versions of our procedures which only stop early to reject (rather than accept) null hypotheses and thus which only explicitly control the corresponding type I generalized error rate, recorded in Theorems 5 and 6. For this reason we refer to them as "rejective" versions of the procedures. The rejective procedures may be preferable in certain situations such as when (a) a null hypothesis being true represents the system being "in control" and therefore continued sampling (rather than stopping) is desirable, (b) there is a maximum sample size imposed on the data streams preventing achievement of the error bounds (2.5)-(2.6), or (c) the type II generalized error rate β is not well-motivated. In any of theses cases, the statistician might prefer to drop the requirement that the type II generalized error rate be strictly controlled at β and use one of the rejective procedures which, roughly speaking, are similar but ignore the lower stopping boundaries $A_w^{(j)}$. Even if β is not well motivated but the statistician prefers early stopping under the null hypotheses, then we encourage the use of our procedures while treating β as a parameter to be chosen to give a procedure with other desirable operating characteristics, such as expected total or streamwise maximum sample size.

The setup for rejective procedures requires a few modifications. Let the data streams $X_n^{(j)}$, test statistics $\Lambda^{(j)}(n)$, and parameters $\theta^{(j)}$ and θ be as in Section 2. Since only the type I error rate, γ_1-FDP or k_1-FWER$_1$, will be explicitly controlled we only require specification of null hypotheses $H^{(j)} \subseteq \Theta^{(j)}$ and not of alternative hypotheses $G^{(j)}$. Accordingly we modify the definition of the false hypotheses (2.2) to be

$$\mathcal{F}(\theta) = \{j \in [J] : \theta^{(j)} \notin H^{(j)}\},$$

and the true hypotheses $\mathcal{T}(\theta)$ are still given by (2.1). We focus on rejective procedures with a streamwise maximum sample size (or "truncation point") $\overline{N}$. With only notational changes, what follows could be formulated without a truncation point or with sample sizes other than $1, \ldots, \overline{N}$.

Given a sequence of step values $0 \leq \alpha_1 \leq \ldots \leq \alpha_J \leq 1$, we assume that the test statistics $\Lambda^{(j)}(n)$ have associated critical values $B_1^{(j)}, \ldots, B_J^{(j)}$ satisfying

$$P_{\theta^{(j)}}\left(\Lambda^{(j)}(n) \geq B_w^{(j)} \text{ for some } n \leq \overline{N}\right) \leq \alpha_w \quad \text{for all} \quad \theta^{(j)} \in H^{(j)}, \tag{4.1}$$

for each $w \in [J]$, as well as (2.7) and (2.9) without loss of generality. We let the standardizing functions $\varphi^{(j)}$ be any increasing functions such that $b_w = \varphi^{(j)}(B_w^{(j)})$ does not depend on j, and $\widetilde{\Lambda}^{(j)}(n) = \varphi^{(j)}(\Lambda^{(j)}(n))$ denote the standardized statistics.

We give the rejective versions of the generic stepdown and stepup procedures in Sections 3.1.1 and 3.2.1, respectively, and state their type I generalized error control properties in Theorems 5 and 6. The proofs are similar to the proofs of the corresponding theorems in Section 3 and are thus omitted.

4.1. Rejective sequential stepdown procedures

With $x \wedge y = \min\{x, y\}$ and with the notation as in Section 3.1.1, the ith stage ($i = 1, 2, \ldots$) of the **Generic Rejective Sequential Stepdown Procedure** with step values $\{\alpha_j\}_{j \in [J]}$ proceeds as follows.

1. Sample the active streams $\{X_n^{(j)}\}_{j \in \mathcal{J}_i,\, n > n_{i-1}}$ until n equals
$$n_i = \overline{N} \wedge \inf\left\{n > n_{i-1} : \widetilde{\Lambda}^{(j)}(n) \geq b_{r_i+1} \quad \text{for some} \quad j \in \mathcal{J}_i\right\}. \tag{4.2}$$

2. If $n_i = \overline{N}$ and no test statistic has crossed the critical value in (4.2), accept all active null hypotheses and terminate the procedure. Otherwise, proceed to Step 3.

3. Order the active test statistics
$$\widetilde{\Lambda}^{(j(n_i,1))}(n_i) \leq \widetilde{\Lambda}^{(j(n_i,2))}(n_i) \leq \ldots \leq \widetilde{\Lambda}^{(j(n_i,|\mathcal{J}_i|))}(n_i)$$
and reject the $m_i \geq 1$ null hypotheses
$$H^{(j(n_i,|\mathcal{J}_i|))}, H^{(j(n_i,|\mathcal{J}_i|-1))}, \ldots, H^{(j(n_i,|\mathcal{J}_i|-m_i+1))},$$
where
$$m_i = \max\{m \in [|\mathcal{J}_i|] : \widetilde{\Lambda}^{(j(n_i,\ell))}(n_i) \geq b_{r_i+|\mathcal{J}_i|-\ell+1}$$

$$\text{for all} \quad \ell = |\mathcal{J}_i| - m + 1, \ldots, |\mathcal{J}_i|\}.$$

4. If $r_i + m_i = J$ or $n_i = \overline{N}$, terminate the procedure. Otherwise, set $r_{i+1} = r_i + m_i$, let $\mathcal{J}_{i+1}$ be the indices of the remaining hypotheses, and continue on to stage $i + 1$.

Remark 10. Points analogous to Remark 1, aside from Point A, apply to the generic rejective sequential stepdown procedure as well.

Theorem 5. *Fix* $\alpha \in (0, 1)$.

1. *Fix* $\gamma_1 \in [0, 1)$. *Given any sequence of constants* $0 \leq \delta_1 \leq \ldots \leq \delta_J \leq 1$ *let* α_j *be given by* (3.5). *If the test statistics and critical values satisfy the assumptions for these* α_j*, then the rejective sequential stepdown procedure with step values* (3.5) *satisfies* γ_1*-FDP*$(\theta) \leq \alpha$ *regardless of the dependence between data streams.*

2. *Fix* $k_1 \in [J]$ *and let* α_j *be given by* (3.7). *If the test statistics and critical values satisfy the assumptions for these* α_j*, then the rejective sequential stepdown procedure with step values* (3.7) *satisfies* k_1*-FWER*$_1(\theta) \leq \alpha$ *regardless of the dependence between data streams.*

Remark 11. As mentioned in Remark 2, the δ_j given in (3.6) may be useful in practice for the procedure in Part 1 of the theorem; the weak optimality mentioned in Remark 4 applies as well to the rejective procedure in Part 2 of the theorem.

4.2. Rejective sequential stepup procedures

With the same notation as in Section 3.2.1, the ith stage ($i = 1, 2, \ldots$) of the **Generic Rejective Sequential Stepup Procedure** with step values $\{\alpha_j\}_{j\in[J]}$ proceeds as follows.

1. Sample the active data streams $\{X_n^{(j)}\}_{j\in\mathcal{J}_i,\, n>n_{i-1}}$ until n equals
$$n_i = \overline{N} \wedge \inf\left\{n > n_{i-1} : \widetilde{\Lambda}^{(j(n,\ell))}(n) \geq b_{r_i+|\mathcal{J}_i|-\ell+1} \quad \text{for some} \quad \ell \in [|\mathcal{J}_i|]\right\}. \tag{4.3}$$

2. If $n_i = \overline{N}$ and no test statistic has crossed its corresponding critical value in (4.3), accept all active null hypotheses and terminate the procedure. Otherwise, proceed to Step 3.

3. Reject the $m_i \geq 1$ null hypotheses

$$H^{(j(n_i,|\mathcal{J}_i|-m_i+1))}, H^{(j(n_i,|\mathcal{J}_i|-m_i+2))}, \ldots H^{(j(n_i,|\mathcal{J}_i|))},$$

where

$$m_i = \max\left\{m \in [|\mathcal{J}_i|] : \widetilde{\Lambda}^{(j(n_i,|\mathcal{J}_i|-m+1))}(n_i) \geq b_{r_i+m}\right\}.$$

4. If $r_i + m_i = J$ or $n_i = \overline{N}$, terminate the procedure. Otherwise, set $r_{i+1} = r_i + m_i$, let $\mathcal{J}_{i+1}$ be the indices of the remaining hypotheses, and continue on to stage $i+1$.

Remark 12. Points analogous to Remark 1, aside from Point A, apply to the generic rejective sequential stepup procedure as well.

Theorem 6. *Fix* $\alpha \in (0,1)$.

1. *Fix* $\gamma_1 \in [0,1)$. *Given any sequence of constants* $0 \leq \delta_1 \leq \ldots \leq \delta_J \leq 1$ *let* α_j *be given by* (3.11). *If the test statistics and critical values satisfy the assumptions for these* α_j, *then the rejective sequential stepup procedure with step values* (3.11) *satisfies* γ_1*-FDP*$(\theta) \leq \alpha$ *regardless of the dependence between data streams.*

2. *Fix* $k_1 \in [J]$. *Given any sequence of constants* $0 \leq \delta_1 \leq \ldots \leq \delta_J \leq 1$ *let* α_j *be given by* (3.12). *If the test statistics and critical values satisfy the assumptions for these* α_j, *then the rejective sequential stepup procedure with step values* (3.12) *satisfies* k_1*-FWER*$_1(\theta) \leq \alpha$ *regardless of the dependence between data streams.*

Remark 13. As mentioned in Remarks 6 and 8, the δ_j given in (3.6) and (3.13) may be useful in practice for the procedures in Parts 1 and 2 of the theorem, respectively; the weak optimality mentioned in Remarks 7 and 9 applies as well to the rejective procedures in Parts 1 and 2 of the theorem, respectively.

5. Implementation

5.1. Simple vs. simple hypotheses

In this section we briefly discuss constructing individual test statistics and critical values satisfying (2.5)-(2.6) (or (4.1) for the rejective versions of the procedures). More complete discussions, including discussion of testing more general composite hypotheses and examples, are given in Bartroff and Song (2014a,b). Here we focus on simple hypotheses and those that can be approximated by simple hypotheses, and in Theorem 7 we give closed-form expressions for the critical values $A_w^{(j)}, B_w^{(j)}$, satisfying (2.5)-(2.6) to a very close approximation, that are based on the closed-form, widely-used Wald approximations for the sequential probability ratio test (SPRT). Sequential test statistics and critical values for other testing situations, including composite hypotheses and nuisance parameter problems, are covered in the texts Bartroff, Lai and Shih (2013) and Siegmund (1985).

Focusing on a stream j for which $H^{(j)}$ and $G^{(j)}$ are simple hypotheses, a natural choice for the test statistic $\Lambda^{(j)}(n)$ is the log-likelihood ratio because of its strong optimality property of the resulting (single hypothesis) test, the SPRT; see Chernoff (1972). In order to express the likelihood ratio test in a simple form, we now make the additional assumption that each data stream $X_1^{(j)}, X_2^{(j)}, \ldots$ constitutes independent and identically distributed data. This independence assumption is limited to *within* each stream so that, for example, elements of $X_1^{(j)}, X_2^{(j)}, \ldots$ may be correlated with (or even identical to) elements of another stream $X_1^{(j')}, X_2^{(j')}, \ldots$. Formally we represent the simple null and alternative hypotheses $H^{(j)}$ and $G^{(j)}$ by the corresponding distinct density functions $h^{(j)}$ (null) and $g^{(j)}$ (alternative) with respect to some common σ-finite measure $\mu^{(j)}$. The parameter space $\Theta^{(j)}$ corresponding to this data stream is the set of all densities f with respect to $\mu^{(j)}$, and $H^{(j)}$ is considered true if the actual density $f^{(j)}$ satisfies $f^{(j)} = h^{(j)}$ $\mu^{(j)}$-a.s., and is false if $f^{(j)} = g^{(j)}$ $\mu^{(j)}$-a.s. The SPRT for testing $H^{(j)} : f^{(j)} = h^{(j)}$ vs. $G^{(j)} : f^{(j)} = g^{(j)}$ with type I and II error probabilities α and β, respectively, utilizes the simple log-likelihood ratio test statistic

$$\Lambda^{(j)}(n) = \sum_{i=1}^{n} \log\left(\frac{g^{(j)}(X_i^{(j)})}{h^{(j)}(X_i^{(j)})}\right) \tag{5.1}$$

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

$$P_{h^{(j)}}(\Lambda^{(j)}(n) \geq B \text{ some } n, \Lambda^{(j)}(n') > A \text{ all } n' < n) \leq \alpha, \tag{5.2}$$

$$P_{g^{(j)}}(\Lambda^{(j)}(n) \leq A \text{ some } n, \Lambda^{(j)}(n') < B \text{ all } n' < n) \leq \beta. \tag{5.3}$$

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 $a + b \leq 1$ and a fixed quantity $\rho \geq 0$. See Hoel, Port and Stone (1971, Sec. 3.3.1) for a derivation of the $\rho = 0$ case and, based on Brownian motion approximations, Siegmund (1985, p. 50 and Chap. X) derives the value $\rho = 0.583$ which has been used to improve the approximation for continuous random variables. 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, Port and Stone (1971) show that the actual type I and II error probabilities can only exceed α or β by a 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 utilize fractions of the actual prescribed error rates.

We use the Wald approximations to construct closed-form critical values $A_w^{(j)}$, $B_w^{(j)}$ satisfying (2.5)-(2.6) up to Wald's approximation. Specifically, given step values $\{\alpha_j, \beta_j\}$, we show that when using (5.5), the left-hand-sides of (2.5)-(2.6) are the same quantities one would get using Wald's approximations with α_j, β_j in place of α, β. This generalizes results of Bartroff and Song (2014a,b) which gave Wald approximations for the specific step values $\{\alpha_j, \beta_j\}$ proposed for the FWER- and FDR-controlling procedures, respectively, given there.

Theorem 7. *Fix $\{\alpha_j, \beta_j\}_{j\in[J]}$ satisfying (2.4) and $\alpha_1 + \beta_1 \le 1$, and $\rho \ge 0$. Suppose that, for a certain data stream j, the associated hypotheses $H^{(j)} : f^{(j)} = h^{(j)}$ and $G^{(j)} : f^{(j)} = g^{(j)}$ are simple. For $a, b \in (0,1)$ such that $a + b \le 1$ let $\alpha_W^{(j)}(a,b)$ and $\beta_W^{(j)}(a,b)$ be the values of the probabilities on the left-hand sides of (5.2) and (5.3), respectively, when $\Lambda^{(j)}(n)$ 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 $w \in [J]$ let*

$$\widetilde{\alpha}_w = \frac{\alpha_1(1-\beta_w)}{1-\beta_1} \quad \text{and} \quad \widetilde{\beta}_w = \frac{\beta_1(1-\alpha_w)}{1-\alpha_1},$$

and let $p_w^{(j)}$ and $q_w^{(j)}$ denote the left-hand-sides of (2.5) and (2.6), respectively, with $A_w^{(j)}$, $B_w^{(j)}$ given by

$$A_w^{(j)} = \log\left(\frac{\beta_w(1-\beta_1)}{1-\beta_1-\alpha_1(1-\beta_w)}\right) + \rho, \quad B_w^{(j)} = \log\left(\frac{1-\alpha_1-\beta_1(1-\alpha_w)}{\alpha_w(1-\alpha_1)}\right) - \rho. \tag{5.5}$$

Then, for all $w \in [J]$,

$$\alpha_w + \widetilde{\beta}_w \le 1, \quad \widetilde{\alpha}_w + \beta_w \le 1, \tag{5.6}$$

$$p_w^{(j)} = \alpha_W^{(j)}(\alpha_w, \widetilde{\beta}_w), \quad \text{and} \quad q_w^{(j)} = \beta_W^{(j)}(\widetilde{\alpha}_w, \beta_w) \tag{5.7}$$

and therefore (2.5)-(2.6) hold, up to Wald's approximation, when using the critical values (5.5).

We remark that the $\rho = 0$ case of Theorem 7 holds without the independence assumption on $X_1^{(j)}, X_2^{(j)}, \ldots$ made in this section, since this original form of Wald's approximations does not require this.

5.2. Group sequential testing

The setup considered here is general enough to admit group sequential sampling as a special case and the popular methods for choosing group sequential stopping boundaries, such as Pocock's (1977) test and O'Brien and Fleming's (1979) test, can be utilized. See also Jennison and Turnbull (2000, Chaps 2.4 and 2.5) for these tests, whose setup we follow. Pocock's and O'Brien and Fleming's tests, in their original forms, utilize a fixed maximum number g of groups and only allow early stopping to reject the corresponding null hypothesis; if the null is not rejected at or before the gth group then it is accepted. This is precisely the form of the rejective procedures defined in Section 4, which we now consider; the last paragraph in this section discusses group sequential tests that allow early rejection or acceptance of the null hypothesis. To utilize Pocock's test of the null hypothesis $H^{(j)} : \theta^{(j)} = 0$ about the average difference $\theta^{(j)}$ in treatment effects with at most g groups all of size m (although groups of unequal sizes can be handled with only minor notational burden), let $X_n^{(j)} = (D_{(n-1)m+1}^{(j)}, D_{(n-1)m+2}^{(j)}, \ldots, D_{nm}^{(j)})$, $n \in [g]$, be the vector of observed differences $D_i^{(j)}$ in the nth group. Pocock's test statistic can be written

$$\Lambda^{(j)}(n) = \left| \frac{1}{\sqrt{nm\sigma^2}} \sum_{i=1}^{nm} D_i^{(j)} \right| \quad \text{for} \quad n \in [g], \tag{5.8}$$

where σ^2 is the known variance of the $D_i^{(j)}$. Given $\alpha \in (0,1)$, the α-level version of the test stops after group $n \in [g]$ and rejects $H^{(j)}$ if $\Lambda^{(j)}(n) \geq C_P(\alpha)$, accepting $H^{(j)}$ if no rejection has occurred by the gth group. Here $C_P(\alpha)$ is a constant (the subscript P for Pocock) calculated to make the type I error probability of this test no greater than α,

$$P_{\theta^{(j)}=0}(\Lambda^{(j)}(n) \geq C_P(\alpha) \text{ for some } n \in [g]) \leq \alpha \quad \text{for any} \quad \alpha \in (0,1). \tag{5.9}$$

Calculation of $C_P(\alpha)$ is well-understood and included in many standard software packages; see Jennison and Turnbull (2000, Chap. 19).

To utilize the Pocock test as the jth component test in a rejective sequential stepup or stepdown procedure defined in Section 4, let $\overline{N} = g$, $\Lambda^{(j)}(n)$ be as in (5.8) for $n \in [g]$, and $B_w^{(j)} = C_P(\alpha_w)$ for $w \in [J]$ where α_w are the given step values. By these definitions and those of the rejective procedures we see that $H^{(j)}$ will be rejected at the first stage $n \in [\overline{N}] = [g]$ where $\Lambda^{(j)}(n)$ crosses a certain boundary $B_w^{(j)}$, and accepted otherwise. All that remains to check is that Theorems 5 and 6 are in force is to verify that (4.1) holds, whose left-hand side is

$$P_{\theta^{(j)}=0}(\Lambda^{(j)}(n) \geq C_P(\alpha_w) \text{ for some } n \in [g])$$

which, by (5.9), is no greater than α_w.

O'Brien and Fleming's test can be applied similarly but with

$$\Lambda^{(j)}(n) = \left| \frac{1}{\sqrt{gm\sigma^2}} \sum_{i=1}^{nm} D_i^{(j)} \right| \quad \text{for} \quad n \in [g], \tag{5.10}$$

which differs from (5.8) by a factor of $\sqrt{g/n}$. This test stops to reject $H^{(j)}$ at the earliest stage $n \in [g]$ such that $\Lambda^{(j)}(n) \geq C_{OF}(\alpha)$, constants satisfying

$$P_{\theta^{(j)}=0}(\Lambda^{(j)}(n) \geq C_{OF}(\alpha) \text{ for some } n \in [g]) \leq \alpha \quad \text{for any} \quad \alpha \in (0,1). \tag{5.11}$$

Using this test as a component test in a rejective procedure is similar to that for Pocock's test but taking $B_w^{(j)} = C_{OF}(\alpha_w)$. As above, (5.11) guarantees that (4.1) holds, and hence Theorems 5 and 6 are in force.

Neither Pocock's nor O'Brien and Fleming's tests stop early to accept the null hypothesis, but other popular group sequential tests do allow this behavior, such as power family tests (see Jennison and Turnbull (2000, Chap. 5)). These tests can be used as component tests in the sequential stepup or stepdown procedures in Section 3 in a similar way for rejective procedures with the minor notational burden of including a maximum sample size $\overline{N}$, equal to the maximum number of groups in this group sequential setting. Of course the choice of $\overline{N}$, as well as the group size (e.g., m in the discussion above) may affect the ability to achieve the needed type I and II error probabilities (2.5) and (2.6), but this issue is not unique to multiple testing considerations and must be considered in group sequential testing of a single null hypothesis as well.

6. Numerical Comparisons

6.1. Introduction and setup

Although a comprehensive comparison of the sequential stepup and stepdown procedures proposed here is beyond the scope of this article, in this section we give a comparison in the particular setting of inference about the means of strongly positively correlated Gaussian data streams; Müller, Parmigiani and Rice (2007) note that this setting is still one of the most widely used in applications involving multiple testing.

If a fixed sample stepup procedure uses the same (or larger) step values $\{\alpha_j\}$ as a stepdown procedure, then the stepup procedure is preferred because it will

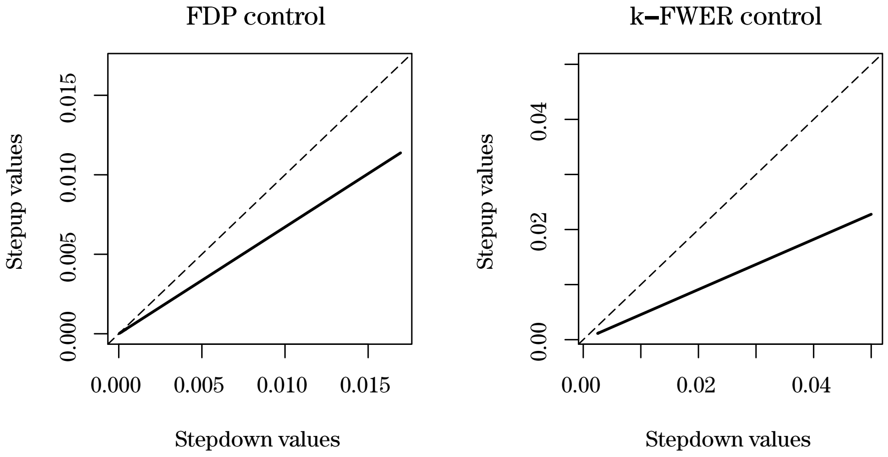

Figure 1. Stepdown versus stepup values (solid lines) for testing $J = 500$ null hypotheses with $\alpha = 0.05$. In the left panel, the stepdown and stepup values α_j are given by (3.5) and (3.11), respectively, both with δ_j given by (3.6) and $\gamma_1 = 0.1$. In the right panel, the stepdown and stepup values α_j are given by (3.7) and (3.12), respectively, with $k_1 = 25$ and δ_j given by (3.13) in the latter case. The identity line is dashed.

reject more null hypotheses and hence be more powerful while not exceeding the prescribed multiple testing error bound α. The same statement holds about the rejective sequential procedures in Section 4, and an analogous statement holds about the sequential procedures in Section 3 which control both type I and II generalized error rates and their step values $\{\alpha_j, \beta_j\}$, in which case "more powerful" means less conservative type I and II error control below the prescribed values α and β. However there is no such simple "dominating" relationship between the values of the stepup and stepdown procedures proposed above. For example, Figure 1 contains plots of the stepdown versus stepup values α_j defined in Sections 3.1 and 3.2, respectively, for $\alpha = 0.05$, $J = 500$ null hypotheses, and $\gamma_1 = 0.1$ for FDP control (left panel) and $k_1 = 25$ for k_1-FWER$_1$ control (right panel). In both panels the solid line is below the dotted identity line indicating that each stepdown value exceeds its corresponding stepup value.

Thus, to investigate the efficiency and overall performance of the proposed sequential stepdown and stepup procedures, simulation studies were performed to estimate their operating characteristics. For this, J streams of Gaussian data were repeatedly simulated in order to consider a battery of tests of the form

$$H^{(j)} : \theta^{(j)} \leq 0 \quad \text{vs.} \quad G^{(j)} : \theta^{(j)} \geq 1 \tag{6.1}$$

about the mean $\theta^{(j)}$ of the jth data stream. The proposed procedures, with their strict error control regardless of dependence, will probably be most useful in settings with strongly positively correlated data streams. For example, about multiple testing problems which arise in genetic association studies by comparing many possible statistical models for genetic data, Zheng et al. (2012, p. 24) remark that typically "all genetic models under consideration are positively correlated." And in randomized multi-arm clinical trials, Freidlin et al. (2008, p. 4,369) note that "individual comparisons are positively correlated due to the use of the same control arm." To create such a setting of strongly positively correlated data streams, the collection $(X_n^{(1)}, \ldots, X_n^{(J)})$ of the nth observations from the J data streams were simulated as a J-dimensional multivariate normal distribution with mean $\theta = (\theta^{(1)}, \ldots, \theta^{(J)})$ and covariance matrix

$$\sigma^2 \begin{bmatrix} 1 & 0.95 & \cdots & 0.95 \\ 0.95 & 1 & \cdots & 0.95 \\ \vdots & \vdots & \ddots & \vdots \\ 0.95 & 0.95 & \cdots & 1 \end{bmatrix}. \tag{6.2}$$

Constant correlation models such as this have recently been popular in the study of genetic correlation structure (Lee et al. (2011); Hardin, Garcia and Golan (2013)), and for us (6.2) provides a convenient way of generating a large number of data streams with strong positive correlation. In the studies that follow we have chosen $\sigma = 2$ to give tests of reasonable length. The collection $(X_n^{(1)}, \ldots, X_n^{(J)})$ of the nth observations was generated using this distribution, with successive observations $(X_n^{(1)}, \ldots, X_n^{(J)})$, $(X_{n+1}^{(1)}, \ldots, X_{n+1}^{(J)})$ generated independently. The test statistics (5.1) were used with $\theta^{(j)} = 0$ vs. $\theta^{(j)} = 1$ as surrogate hypotheses, reducing to

$$\Lambda^{(j)}(n) = \frac{1}{\sigma^2} \left(\sum_{i=1}^{n} X_i^{(j)} - \frac{n}{2} \right)$$

in this case, and the critical values (5.5) were used with $\rho = 0.583$ and $\{\alpha_j, \beta_j\}$ as described below. The results of simulation studies in this setting are reported in Section 6.2 for γ_1-FDP and γ_2-FNP control, and Section 6.3 for k_1-FWER$_1$ and k_2-FWER$_2$ control. Finally, in Section 6.4, the assumption of known variance is dropped and Student's t-tests of composite hypotheses are considered.

6.2. Study of procedures controlling γ_1-FDP and γ_2-FNP

Table 2 contains some operating characteristics under various settings of the sequential stepdown and stepup procedures, denoted Seq$_D$ and Seq$_U$, using step values (3.5) and (3.11) (both with δ_j given by (3.6)), respectively, and which control γ_1-FDP $\leq \alpha = 0.05$ and γ_2-FNP $\leq \beta = 0.2$. The operating characteristics are the *expected streamwise average sample size* $E_\theta N$ which is the average sample size over the J streams, $N = \sum_{j=1}^{J} N_j/J$, where N_j denotes the sample size of the jth stream, its standard error SE, and the achieved generalized error rates γ_1-FDP and γ_2-FNP. Each operating characteristic estimate is the result of 10,000 Monte Carlo simulated ensembles of J data streams. The parameter values $\gamma_1 = \gamma_2 = 0.1$ were used and three states of nature, in terms of the number of true null hypotheses $H^{(j)}$, were considered for both of the $J = 500$ and $J =$ 1,000 scenarios, with the true $H^{(j)}$ are simulated using $\theta^{(j)} = 0$ and the false $H^{(j)}$ with $\theta^{(j)} = 1$, representing the "worst case" with respect to distinguishability of the null and alternative hypotheses.

To provide a point of reference for these sequential procedures, the performance of comparable fixed sample size stepdown and stepup procedures, denoted by Fix_D and Fix_U, were also estimated. These are the procedures defined in Section 2.2 that use the same step values α_j as Seq_D and Seq_U, respectively. Since these values α_j determine the type I generalized error rate γ_1-FDP, to obtain procedures comparable to the sequential ones, the fixed sample sizes for Fix_D and Fix_U were chosen as the values yielding the type II generalized error rate γ_2-FNP most closely matching that of the sequential procedure with the smallest $E_\theta N$, the more efficient of Seq_D and Seq_U, whose row is shaded in each scenario in the table. The fixed sample size procedures are *more* conservative than the sequential procedures since the error probabilities tend to decrease as sample size increases; in this sense this comparison is conservative. Because the sample sizes of Fix_D and Fix_U are fixed, their SE is left blank. The final column of the table shows that savings in $E_\theta N$ of each sequential procedure relative to its fixed sample counterpart.

The sequential procedures in Table 2 show a dramatic savings in average sample size relative to the fixed sample size procedures of at least 50% in all cases, and as high as 65%. The sequential procedures also have less conservative error control than their fixed sample size counterparts, most evident in the type I generalized error rate γ_1-FDP which was not used for "matching" the fixed sample procedures as the type II version was. This less conservative error control is perhaps due to the sequential procedures' smaller average sample size. Nonetheless, all the procedures still have quite conservative error control relative to the prescribed values of $\alpha = 0.05$ and $\beta = 0.2$ even on this highly positively correlated data. Another notable feature of the results in Table 2 is that the sequential stepup procedures are slightly but consistently more efficient than the stepdown

Table 2. Expected streamwise average sample size $E_\theta N$, its standard error SE, achieved error rates γ_1-FDP and γ_2-FNP, and savings in $E_\theta N$ of the sequential (denoted Seq_D and Seq_U) and fixed sample size (denoted Fix_D and Fix_U) procedures described in Section 6.2 for testing J null hypotheses about the means of Gaussian data streams. The parameter values are $\alpha = 0.05$, $\beta = 0.2$, and $\gamma_1 = \gamma_2 = 0.1$ and each estimate is the result of 10,000 simulated ensembles of J data streams. The shaded row in each scenario is the procedure with the smallest $E_\theta N$.

# True $H^{(j)}$	Procedure	$E_\theta N$	SE	γ_1-FDP(θ)	γ_2-FNP(θ)	$E_\theta N$ Savings
$J = 500$						
100	Seq_D	63.63	0.60	0.007	0.015	53%
	Seq_U	54.17	0.67	0.008	0.012	55%
	Fix_D	136		0.002	0.012	
	Fix_U	120		0.001	0.012	
	Fix'_D	129		0.007	0.015	
	Fix'_U	110		0.008	0.012	
250	Seq_D	60.66	0.40	0.004	0.026	55%
	Seq_U	53.39	0.40	0.003	0.016	58%
	Fix_D	135		0.001	0.015	
	Fix_U	128		0.001	0.016	
400	Seq_D	56.98	0.58	0.006	0.039	57%
	Seq_U	45.97	0.57	0.007	0.022	65%
	Fix_D	134		0.001	0.022	
	Fix_U	131		0.001	0.022	
$J = 1000$						
250	Seq_D	67.26	0.52	0.006	0.008	54%
	Seq_U	54.93	0.58	0.003	0.009	56%
	Fix_D	147		0.002	0.009	
	Fix_U	125		0.001	0.009	
500	Seq_D	65.54	0.39	0.009	0.021	50%
	Seq_U	54.06	0.41	0.002	0.027	54%
	Fix_D	130		0.001	0.026	
	Fix_U	118		0.001	0.026	
750	Seq_D	63.16	0.55	0.001	0.026	54%
	Seq_U	49.72	0.53	0.002	0.023	61%
	Fix_D	136		0.001	0.023	
	Fix_U	129		0.001	0.023	

procedures in each scenario, in terms of minimizing $E_\theta N$. In the next section we will see that the reverse is true in a similar study of procedures controlling k_1-FWER_1 and k_2-FWER_2.

Because of the highly conservative error control of all the procedures in Table 2, especially the fixed sample size procedures, another type of comparison that may shed light on how much of the efficiency gained by the sequential procedures is due the sequential sampling itself rather than the differing achieved error rates. Included in the first scenario of Table 2 are two more fixed sample size procedures (denoted Fix'_D and Fix'_U) which match *both* error rates γ_1-FDP and γ_2-FNP of Seq_D and Seq_U, respectively. These were found by exhaustively searching over values of the fixed streamwise sample size N and a grid of values for the nominal γ_1-FDP rate α for Fix'_D and Fix'_U. The procedure Fix'_D uses $\alpha = 0.092$ and $N = 129$ to match the error rates γ_1-FDP $= 0.007$ and γ_2-FNP $= 0.015$ of Seq_D, and Fix'_U uses $\alpha = 0.112$ and $N = 110$ to match γ_1-FDP $= 0.008$ and γ_2-FNP $= 0.012$ of Seq_U. The increase in nominal α required for this matching is roughly a factor of 2, and the decrease in sample size is modest, leaving the sample sizes of Fix'_D and Fix'_U still substantially larger than their sequential counterparts even though they do not have proven error control at the $\alpha = 0.05$ level. This suggests the efficiency gains of the sequential procedures relative to the fixed sample size procedures are due more to the sequential sampling than their less conservative error control.

6.3. Study of procedures controlling k_1-FWER$_1$ and k_2-FWER$_2$

Table 3 contains the results of a study similar to Table 2 but for procedures controlling k_1-FWER$_1$ and k_2-FWER$_2$. In Table 3, Seq_D and Seq_U denote the stepdown and stepup procedures defined in Sections 3.1.2 and 3.2.2 using step values (3.7) and (3.12), respectively, with δ_j given by (3.13) for the latter. The parameters $k_1 = k_2 = 25$ were used for the $J = 500$ scenario and $k_1 = k_2 = 50$ for the $J = 1{,}000$ scenario, and the same prescribed error bounds $\alpha = 0.05$, $\beta = 0.2$ were used. The operating characteristics and simulation settings are otherwise the same as the previous section. As there, the stepdown and stepup fixed sample size procedures Fix_D and Fix_U are those defined in Section 2.2 that use the same step values α_j as Seq_D and Seq_U, respectively; the fixed sample sizes of these procedures was chosen to match their type II generalized error rate k_2-FWER$_2$ as closely as possible to the sequential procedure with the smallest $E_\theta N$, whose row is shaded in the table in each scenario.

The sequential procedures in Table 3 show a substantial savings of roughly 50% to 60% in average sample size relative to the fixed sample size procedures, and less conservative error control than their fixed sample size counterparts, most evident in the type I generalized error rate k_1-FWER$_1$ that was not used for "matching" the fixed sample procedures as the type II version was. All the procedures have quite conservative error control relative to the prescribed values of $\alpha = 0.05$ and $\beta = 0.2$. Unlike Table 2, the sequential stepdown procedures in

Table 3. Expected streamwise average sample size $E_\theta N$, its standard error SE, achieved error rates k_1-FWER$_1$ and k_2-FWER$_2$, and the savings in $E_\theta N$ of the sequential (denoted Seq$_D$ and Seq$_U$) and fixed sample size (denoted Fix$_D$ and Fix$_U$) procedures described in Section 6.3 for testing J null hypotheses about the means of Gaussian data streams. The parameter values are $\alpha = 0.05$ and $\beta = 0.2$ and each estimate is the result of 10,000 simulated ensembles of J data streams. The shaded row in each scenario is the procedure with the smallest $E_\theta N$.

# True $H^{(j)}$	Procedure	$E_\theta N$	SE	k_1-FWER$_1(\theta)$	k_2-FWER$_2(\theta)$	$E_\theta N$ Savings
$J = 500$, $k_1 = k_2 = 25$						
100	Seq$_D$	38.39	0.48	0.020	0.039	49%
	Seq$_U$	44.91	0.59	0.009	0.034	54%
	Fix$_D$	75		0.023	0.039	
	Fix$_U$	97		0.002	0.040	
	Fix$'_D$	77		0.020	0.039	
	Fix$'_U$	95		0.009	0.034	
250	Seq$_D$	36.81	0.32	0.017	0.047	57%
	Seq$_U$	43.32	0.38	0.011	0.041	55%
	Fix$_D$	86		0.005	0.047	
	Fix$_U$	97		0.001	0.046	
400	Seq$_D$	32.12	0.46	0.007	0.067	60%
	Seq$_U$	38.17	0.53	0.009	0.065	57%
	Fix$_D$	80		0.030	0.066	
	Fix$_U$	89		0.001	0.066	
$J = 1{,}000$, $k_1 = k_2 = 50$						
250	Seq$_D$	37.45	0.42	0.015	0.033	58%
	Seq$_U$	44.07	0.51	0.005	0.042	56%
	Fix$_D$	89		0.009	0.034	
	Fix$_U$	100		0.002	0.034	
500	Seq$_D$	36.73	0.31	0.012	0.050	57%
	Seq$_U$	42.46	0.38	0.008	0.044	56%
	Fix$_D$	86		0.005	0.051	
	Fix$_U$	96		0.001	0.049	
750	Seq$_D$	33.27	0.41	0.012	0.065	59%
	Seq$_U$	39.93	0.46	0.006	0.040	57%
	Fix$_D$	82		0.003	0.063	
	Fix$_U$	92		0.001	0.064	

Table 3 were more efficient than the stepup procedures in terms of smaller $E_\theta N$.

Similar to Table 2, the first scenario in Table 3 also includes fixed sample size procedures Fix$'_D$ and Fix$'_U$ whose values of streamwise sample size N and nominal k_1-FWER$_1$ bound α were searched over to find values giving attained k_1-FWER$_1$ and k_2-FWER$_2$ equal to those of the sequential procedures Seq$_D$ and Seq$_U$, respectively. The procedure Fix$'_D$ uses $\alpha = 0.048$ and $N = 77$ to match the error rates k_1-FWER$_1 = 0.020$ and k_2-FWER$_2 = 0.039$ of Seq$_D$, and Fix$'_U$ uses $\alpha = 0.080$ and $N = 95$ to match k_1-FWER$_1 = 0.009$ and k_2-FWER$_2 = 0.034$ of Seq$_U$. Whereas Fix$'_U$ uses a slightly smaller sample size (and larger α) than Fix$_U$ because the latter is more conservative than Seq$_U$ in terms of error rates, Fix$'_D$ uses a slightly larger sample size (and smaller α) than Fix$_D$ because the latter is actually less conservative than Seq$_D$. In any case, the change in sample size of these modified fixed sample procedures is slight and the fixed sample sizes remain substantially larger than their sequential counterparts, indicating that the increased efficiency is due to the sequential sampling rather than differing achieved error rates, as in Table 2.

6.4. Composite hypotheses: Student's t-tests

In this section we consider a setting similar to the Gaussian mean testing problem (6.1) of the previous sections but drop the assumption of known variance σ^2, making both the null and alternative in (6.1) composite hypotheses. First we briefly describe a sequential approach to this Student's t-test problem and then give the results of a simulation study in a similar setting to Section 6.3 for k_1-FWER$_1$ and k_2-FWER$_2$ control.

Suppose that the data $X_1^{(j)}, X_2^{(j)}, \ldots$ from a certain data stream are i.i.d. Gaussian data with mean μ and variance σ^2, both unknown, and it is desired to test the null hypothesis $\mu \le 0$ versus the alternative $\mu \ge \delta$, for some given $\delta > 0$. Formally, this is a special case of the setup in Section 2 by taking $\theta^{(j)} = (\mu, \sigma)^T$, $\Theta^{(j)} = \mathbb{R} \times (0, \infty)$, $H^{(j)} = \{(\mu, \sigma)^T \in \Theta^{(j)} : \mu \le 0\}$, and $G^{(j)} = \{(\mu, \sigma)^T \in \Theta^{(j)} : \mu \ge \delta\}$. Bartroff and Song (2014b, Sec. 3.2) suggest sequential log generalized likelihood ratio (GLR) statistics for a general class of composite hypotheses when the data is from an exponential family, including this t-test setting for which the sequential log GLR statistic is (see Bartroff (2006, p. 106))

$$\Lambda^{(j)}(n) = \begin{cases} +\sqrt{2n\Lambda_H(n)}, & \text{if } \overline{X}_n^{(j)} \ge \delta/2, \\ -\sqrt{2n\Lambda_G(n)}, & \text{otherwise,} \end{cases} \tag{6.3}$$

where $\Lambda_H(n) = \dfrac{n}{2} \log \left[1 + \left(\dfrac{\overline{X}_n^{(j)}}{\widehat{\sigma}_n} \right)^2 \right]$, $\Lambda_G(n) = \dfrac{n}{2} \log \left[1 + \left(\dfrac{\overline{X}_n^{(j)} - \delta}{\widehat{\sigma}_n} \right)^2 \right]$, and $\overline{X}_n^{(j)}$ and $\widehat{\sigma}_n^2$ are the usual MLE estimates of μ and σ^2, respectively, based on $X_1^{(j)}, \ldots, X_n^{(j)}$. Bartroff and Song (2014b, Lem. 3.1) also give formulas for certain upper bounds on the probabilities in (2.5)-(2.6) involving only properties of the standard normal distribution, allowing critical values $\{A_w^{(j)}, B_w^{(j)}\}_{w \in [J]}$ to be computed satisfying (2.5)-(2.6) for given step values $\{\alpha_w, \beta_w\}_{w\in[J]}$ by either recursive numerical integration or Monte Carlo simulation of standard normal variates.

Table 4 contains the results of a study similar to Table 3 but for sequential and fixed sample size t-tests. In Table 4, Seq_D and Seq_U denote the stepdown and stepup procedures defined in Sections 3.1.2 and 3.2.2 using step values (3.7) and (3.12), respectively, with δ_j given by (3.13) for the latter. The sequential procedures use the statistics (6.3) with critical values computed by Monte Carlo as described in the previous paragraph. The stepdown and stepup fixed sample size procedures Fix_D and Fix_U are those defined in the first paragraph of Section 2.2 with p-values for sample size n computed in the standard way as $1-T_{n-1}(\overline{X}_n^{(j)}\sqrt{n-1}/\widehat{\sigma}_n)$, where $T_{n-1}(\cdot)$ denotes the c.d.f. of the Student's t distribution with $n-1$ degrees of freedom, and which use the same step values α_j as Seq_D and Seq_U, respectively. As in Section 6.3, the fixed sample sizes of these procedures was chosen to match their type II generalized error rate k_2-FWER$_2$ as closely as possible to the sequential procedure with the smallest $E_\theta N$, whose row is shaded in the table in each scenario. To give a view of the procedures' performance under a different dependency structure for the Gaussian data streams, they were simulated not as highly correlated but rather nearly independent with correlation coefficient 0.05 replacing 0.95 in (6.2). The data was simulated using the same value $\sigma = 2$ as above, but not assumed to be known.

Comparing Table 4 with the first half of Table 3, one sees that the additional task of estimating the unknown variance in the t-test setting, plus the near-independence of the data streams, only cause a modest increase in sample size of all the procedures. The relationship between the sequential and fixed sample size procedures is otherwise remarkably similar to that in Table 3, with the stepdown procedure Seq_D being slightly more efficient than the stepup procedure Seq_U for FWER control, and both being roughly 50-60% more efficient than the fixed sample size procedures in terms of expected sample size. Also like Table 3, all procedures are very conservative in terms of error control, with the sequential procedures tending to be less so (but not uniformly – see Fix_D in the case of 400 true $H^{(j)}$) because of their smaller average sample size.

7. Conclusions and Discussion

We have proposed general and flexible multiple testing procedures for controlling generalized error rates on sequential data whose error control holds re-

Table 4. Expected streamwise average sample size $E_\theta N$, its standard error SE, achieved error rates k_1-FWER$_1$ and k_2-FWER$_2$, and the savings in $E_\theta N$ of the sequential (denoted Seq$_D$ and Seq$_U$) and fixed sample size (denoted Fix$_D$ and Fix$_U$) procedures described in Section 6.4 for testing J null hypotheses about the means of Gaussian data streams with unknown variances. The parameter values are $\alpha = 0.05$ and $\beta = 0.2$ and each estimate is the result of 10,000 simulated ensembles of J data streams. The shaded row in each scenario is the procedure with the smallest $E_\theta N$.

# True $H^{(j)}$	Procedure	$E_\theta N$	SE	k_1-FWER$_1(\theta)$	k_2-FWER$_2(\theta)$	$E_\theta N$ Savings
			$J = 500$, $k_1 = k_2 = 25$			
100	Seq$_D$	40.22	0.09	0.003	0.053	48%
	Seq$_U$	47.58	0.11	0.006	0.021	54%
	Fix$_D$	77		0.003	0.053	
	Fix$_U$	103		0.002	0.053	
250	Seq$_D$	38.79	0.04	0.018	0.060	56%
	Seq$_U$	44.79	0.05	0.003	0.018	56%
	Fix$_D$	89		0.006	0.059	
	Fix$_U$	101		0.001	0.060	
400	Seq$_D$	33.62	0.09	0.009	0.059	60%
	Seq$_U$	37.10	0.11	0.011	0.057	60%
	Fix$_D$	85		0.037	0.060	
	Fix$_U$	93		0.002	0.059	

gardless of dependence between data streams. We have given both stepdown and stepup procedures for controlling the tail probabilities of FDP and k-FWER, as well as their type II versions. In the numerical studies of their performance in Section 6 in the setting of highly positively correlated Gaussian data streams we found that, in terms of achieving smaller expected sample size, the stepup procedures performed better for controlling FDP, and the stepdown procedures performed better for controlling k-FWER. Although this study was limited to the specific setting of testing hypotheses about the means of Gaussian data streams with covariance matrix (6.2), these are our working recommendations for what to use in practice until further study is possible.

The simulation studies also show the procedures to be highly conservative in the situation considered, in terms of having generalized error rates substantially smaller than the prescribed values α and β. However, it is apparent that this is not related to the sequential nature of the procedures proposed here because the fixed sample versions also have this property and even more so. This is not surprising since the error rates tend to decrease as sample size increases and efficient sequential procedures will have smaller expected sample sizes than their fixed sample counterparts. On the other hand, the results of Lehmann and Romano (2005) and Romano and Shaikh (2006a) (see Remarks 4, 7, and 9) show that the error bounds are indeed "sharp" and cannot be improved without more restrictive assumptions on the joint distribution of the data streams. Less conservative error control (or equivalently, more efficiency in terms of smaller expected sample sizes) may be possible by assumptions about (or direct modeling of) this joint distribution, which was not the focus of this paper but may be a fruitful area of future work.

Our procedures, as well as those in Bartroff and Song (2014a,b) for FDR/FNR and type I/II FWER control, are all special cases of the generic sequential procedures in Sections 3.1.1 and 3.2.1 and all use the same step values as the corresponding fixed sample size procedures: the Bartroff and Song (2014a,b) procedures utilize the same step values as the Benjamini and Hochberg (1995) and Holm (1979) procedures, respectively, and the procedures in this paper utilize the step values of Lehmann and Romano (2005) and Romano and Shaikh (2006a,b). Thus, the theme that emerges from this body of work is that, with the appropriate care, fixed sample size step values can be used with the suitable sequential test.

Supplementary Materials

Proofs and auxiliary results for this paper appear in an online supplement.

Acknowledgment

The author thanks an associate editor and the reviewers for helpful comments. Bartroff's work was partially supported by National Science Foundation grant DMS-1310127 and National Institutes of Health grant R01 GM068968.

References

Baillie, D. (1987). Multivariate acceptance sampling – some applications to defence procurement. *Journal of the Royal Statistical Society D* **36**, 465–478.

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

Bartroff, J. and Lai, T. L. (2010). Multistage tests of multiple hypotheses. *Communications in Statistics – Theory and Methods (Special Issue Honoring M. Akahira, M. Aoshima, ed.)* **39**, 1597–1607.

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

Design and Analysis. Springer, New York.

Bartroff, J. and Song, J. (2014a). Sequential tests of multiple hypotheses controlling false discovery and nondiscovery rates. Under review. `https://arxiv.org/abs/1311.3350`.

Bartroff, J. and Song, J. (2014b). Sequential tests of multiple hypotheses controlling type I and II familywise error rates. *Journal of Statistical Planning and Inference* **153**, 100–114.

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.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29**(4), 1165–1188.

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

Clements, N., Sarkar, S. K., Zhao, Z. and Kim, D.-Y. (2014). Applying multiple testing procedures to detect change in east african vegetation. *The Annals of Applied Statistics* **8**, 286–308.

De, S. and Baron, M. (2012a). Sequential Bonferroni methods for multiple hypothesis testing with strong control of family-wise error rates I and II. *Sequential Analysis* **31**(2), 238–262.

De, S. and Baron, M. (2012b). Step-up and step-down methods for testing multiple hypotheses in sequential experiments. *Journal of Statistical Planning and Inference* **142**, 2059–2070.

Freidlin, B., Korn, E. L., Gray, R. and Martin, A. (2008). Multi-arm clinical trials of new agents: Some design considerations. *Clinical Cancer Research* **14**(14), 4368–4371.

Guo, W., He, L. and Sarkar, S. K. (2014). Further results on controlling the false discovery proportion. *The Annals of Statistics* **42**(3), 1070–1101.

Hardin, J., Garcia, S. R. and Golan, D. (2013). A method for generating realistic correlation matrices. *The Annals of Applied Statistics* **7**(3), 1733–1762.

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

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.

Hommel, G. and Hoffmann, T. (1988). Controlled uncertainty. In Bauer, P., Hommel, G. and Sonnemann, E., editors, *Multiple Hypothesenprüfung/Multiple Hypotheses Testing*, pages 154–161. Springer, Heidelberg.

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

Jiang, H. and Salzman, J. (2012). Statistical properties of an early stopping rule for resampling-based multiple testing. *Biometrika* **99**(4), 973–980.

Lai, T. L. and Xing, H. (2008). *Statistical Models and Methods for Financial Markets*. Springer, New York.

Lee, S. H., Wray, N. R., Goddard, M. E. and Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics* **88**(3), 294–305.

Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. *The Annals of Statistics* **33**(3), 1138–1154.

Mei, Y. (2010). Efficient scalable schemes for monitoring a large number of data streams.

Biometrika **97**(2), 419–433.

Müller, P., Parmigiani, G. and Rice, K. (2007). FDR and Bayesian multiple comparisons rules. In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M. and West, M., editors, *Bayesian Statistics 8: Proceedings of the Eighth Valencia International Meeting, June 2-6, 2006*, pages 349–370. Oxford University Press.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549–556.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**(2), 191–199.

Romano, J. P. and Shaikh, A. M. (2006a). On stepdown control of the false discovery proportion. In Rojo, J., editor, *Optimality: The Second Erich L. Lehmann Symposium*, volume 49 of *IMS Lecture Notes–Monograph Series*, pages 33–50, Beachwood, Ohio, USA. Institute of Mathematical Statistics.

Romano, J. P. and Shaikh, A. M. (2006b). Stepup procedures for control of generalizations of the familywise error rate. *The Annals of Statistics* **34**, 1850–1873.

Salzman, J., Jiang, H. and Wong, W. H. (2011). Statistical modeling of RNA-seq data. *Statistical Science* **26**, 62–83.

Sarkar, S. (2007). Stepup procedures controlling generalized FWER and generalized FDR. *The Annals of Statistics* **35**, 2405–2420.

Sarkar, S. (2008). Generalizing Simes' test and Hochberg's stepup procedure. *The Annals of Statistics* **36**, 337–363.

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

Tartakovsky, A., Li, X. and Yaralov, G. (2003). Sequential detection of targets in multichannel systems. *Information Theory, IEEE Transactions on* **49**(2), 425–445.

Zheng, G., Yang, Y., Zhu, X. and Elston, R. (2012). *Analysis of Genetic Association Studies.* Springer, New York.

Department of Mathematics, University of Southern California, 3620 S Vermont Ave, KAP 104,
Los Angeles, CA 90089, USA.

E-mail: bartroff@usc.edu

(Received August 2015; accepted December 2016)