Statistics in Biosciences
https://doi.org/10.1007/s12561-024-09469-y

ORIGINAL PAPER

# Sequential FDR and pFDR Control Under Arbitrary Dependence, with Application to Pharmacovigilance Database Monitoring

**Michael Hankin**[1] · **Jay Bartroff**[2]

Received: 3 June 2024 / Revised: 9 November 2024 / Accepted: 25 November 2024


**Abstract**
We propose sequential multiple testing procedures which control the false discovery rate (FDR) or the positive-false discovery rate (pFDR) under arbitrary dependence between the data streams. This is accomplished by "optimizing" an upper bound on these error metrics for a class of step-down sequential testing procedures. Both open-ended and truncated versions of these sequential procedures are given, both being able to control both the type I multiple testing metric (FDR or pFDR) at specified levels, and the former being able to control both the type I and type II (e.g., FDR *and* the false nondiscovery rate, FNR). In simulation studies, these procedures provide 45–65% savings in average sample size over their fixed-sample competitors. We illustrate our procedures on drug data from the United Kingdom's Yellow Card Pharmacovigilance Database.

**Keywords** False discovery rate · Knapsack problem · Multiple testing · Optimization · Positive-false discovery rate · Sequential analysis · Step down procedure

## 1 Introduction

The majority of the procedures proposed in the statistics literature for multiple testing for fixed-sample size or sequential data can be bifurcated into either step-up or step-down procedures. Step-up procedures decide whether to accept or reject null hypotheses in order of increasing significance, whereas step-down procedures operate in the reverse order. For testing $J \geq 2$ null hypotheses, the

✉ Jay Bartroff
bartroff@austin.utexas.edu

[1] Mothball Labs, San Francisco, USA

[2] University of Texas at Austin, Austin, USA

Published online: 19 December 2024

Benjamini-Hochberg [1, hereafter BH] procedure is step-up, and is known to control the false discovery rate (FDR) at the nominal level $\alpha$ under independence of $p$ values, and at no more than the inflated level $C\alpha$, where $C = C(J) := \sum_{j=1}^{J} 1/j$, under arbitrary dependence among $p$ values, with $C \approx \log(J)$ for large $J$. Although [2] showed that FDR can be controlled at the uninflated level $\alpha$ under certain types of positive dependence, [3] showed that the FDR bound $C\alpha$ is sharp for the step-up BH procedure in the sense that there exists a joint distribution of $p$ values for which the FDR equals this. Analogous results were established by [4] and [5] about step-up procedures for sequential data.

In this sense, the "worst-case" performance of step-up procedures is essentially completely understood, at least without imposing additional restrictions or assumptions on the joint distribution of $p$ values. So in this paper, we turn out attention to studying and controlling the worst-case performance of step-down procedures, in the setting of sequential data, i.e., $J$ data streams or arbitrary dependence. More specifically, we give a step-down procedure for sequential data and the smallest possible constant $D < C$ whose FDR is bounded above by $D\alpha$ where, again, $\alpha$ is the nominal FDR level.

Hart and Weiss [6] used linear programming to find a sharp upper bound on the probability of rejection for a class of multiple testing procedures in terms of the true, marginal c.d.f.s of the $p$ values. Calculating their bound, therefore, requires knowledge of at least the marginals of the true joint distribution. Guo and Rao [3] also used optimization but to find an upper bound on the false discovery rate (FDR) in terms of a step-down procedure's step values, that holds under arbitrary joint distribution of the $p$ values. By choosing the step values to control the upper bound, they produce a step-down procedure whose FDR control holds under arbitrary joint distributions. The current paper extends this line of research in the following ways. First, by extending to the sequential setting an FDR-controlling procedure without requiring or imposing assumptions on the joint distribution of the $p$ values, the open-ended versions of these sequential procedures allow the possibility of simultaneously controlling both type I and type II error metrics (e.g., FDR *and* the false nondiscovery rate, FNR). Second, by removing an implicit assumption on the $p$ values in Guo and Rao's proof that is stronger than validity of $p$ values and applies to both the sequential and fixed-sample settings; see the Appendix for details. Finally, by extending this methodology to positive-false discovery rate (pFDR) and its type II analog.

The rest of this paper is organized as follows. After giving the notational setup in Sect. 2, Sect. 3 covers untruncated (i.e., open-ended) sequential procedures and records the FDR/FNR- and pFDR/pFNR-controlling procedures in Theorems 3.1 and 3.2, respectively, as special cases of a generic untruncated procedure given in Sect. 3.1. Truncated sequential procedures are handled analogously in Sect. 4 with the corresponding FDR- and pFDR-controlling procedures given in Theorems 4.1 and 4.2, respectively. To aide the reader, Table 1 is a "directory" of these theorems and cases. In Sect. 5, we present simulation studies on Binomial and Poisson data of the proposed sequential procedures and compare with their fixed-sample counterparts, and in Sect. 6, we apply our proposed procedures to

data from the UK's Yellow Card Pharmacovigilance Database. We end with conclusions and discussion in Sect. 7. Proofs are given in the Appendix.

### 1.1 Dedication

This paper is dedicated to the memory of Prof. Tze Leung Lai, who influenced and inspired the authors directly and indirectly. While the current paper is not a direct offshoot of one of Prof. Lai's works, it overlaps with some of Lai's most active research areas including sequential hypothesis testing [7–9], adaptive designs [10–12], multiple comparisons [13, 14], optimization [15], quality control [16], longitudinal data [17], and biomedical applications [18], to name a few references.

## 2 Setup

### 2.1 Data Streams, Hypotheses, and Error Metrics

Our general form of data streams and hypotheses is as follows. Assume that there are $J \geq 2$ data streams

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

In order to implement our proposed procedures, the marginal type I and II error probabilities of component test statistics on each stream will need to be controlled at certain levels; see (4)–(5) below. Beyond that, 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 $j \in [J]$, 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 an alternative hypothesis $G^{(j)}$, where $H^{(j)}$ and $G^{(j)}$ are disjoint subsets of the parameter space $\Theta^{(j)}$ containing $\theta^{(j)}$. It is *not* required that $H^{(j)} \cup G^{(j)} = \Theta^{(j)}$, e.g., one-sided alternatives or separated hypotheses are possible. The null hypothesis $H^{(j)}$ is considered *true*

**Table 1** Directory of this paper's error control theorems

| Error metrics | Untruncated, type I and II controlling | Truncated, rejective |
|---|---|---|
| FDR/FNR | Theorem 3.1 | Theorem 4.1 |
| pFDR/pFNR | Theorem 3.2 | Theorem 4.2 |

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 $\Theta = \Theta^{(1)} \times \cdots \times \Theta^{(J)}$.

The general notation (1) includes fully sequential sampling where the streamwise sample sizes may take any value $1, 2, \ldots$ *ad infinitum*, but other sampling setups fit this as well including group sequential, truncated, and even fixed-sample size testing. For example, the $n$th "observation" $X_n^{(j)}$ in the $j$th stream may actually be the $n$th group $X_n^{(j)} = (X_{n,1}^{(j)}, \ldots, X_{n,\ell}^{(j)})$ of size $\ell$. Moreover, the group size $\ell$ may vary with $n$ and may even be data-dependent, e.g., determined by some type of adaptive sampling. Similarly, truncated sequential (or group sequential) sampling can be implemented for the $j$th 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 $n \le \overline{N}$," for some global truncation point $\overline{N}$. One may represent a fixed-sample size in the $j$th stream in this way by taking $\overline{N}^{(j)} = 1$, or in all streams with $\overline{N} = 1$.

For any multiple testing procedure under consideration, let $V$ denote the number of true null hypotheses it rejects (i.e., the number of false positives), $W$ the number of false null hypotheses it accepts (i.e., the number of false negatives), and $R$ the number of null hypotheses it rejects. The number of null hypotheses accepted is, therefore, $J - R$. Under the true value of the parameter $\theta$, the false discovery and nondiscovery rates [1, FDR,FNR] are

$$\mathrm{FDR} = \mathrm{FDR}(\theta) = E_\theta\left(\frac{V}{R \vee 1}\right), \quad \mathrm{FNR} = \mathrm{FNR}(\theta) = E_\theta\left(\frac{W}{(J-R) \vee 1}\right),$$

where $x \vee y = \max\{x, y\}$. Similarly, the positive-false discovery rate pFDR [19] and its type II analog, the positive-false nondiscovery rate pFNR, are defined as

$$\mathrm{pFDR} = \mathrm{pFDR}(\theta) = E_\theta\left(\frac{V}{R}\middle| R \ge 1\right) \quad \text{and} \quad \mathrm{pFNR} = \mathrm{pFNR}(\theta) = E_\theta\left(\frac{W}{J-R}\middle| J - R \ge 1\right). \tag{2}$$

### 2.2 Test Statistics and Critical Values

The building blocks of the sequential procedures defined below are $J$ individual sequential test statistics $\{\Lambda^{(j)}(n)\}_{j\in[J],\, n\ge 1}$, 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 $j$th stream at time $n$. For example, in parametric settings $\Lambda^{(j)}(n)$ may be a sequential log (generalized) likelihood ratio statistic for testing $H^{(j)}$ vs. $G^{(j)}$. Like the fixed-sample size procedures mentioned above, our sequential procedures will utilize *step values* (or *step value vectors*) which are $J$-long vectors of nondecreasing values in [0, 1], such as

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J) \quad \text{with} \quad 0 < \alpha_1 \le \alpha_2 \le \ldots \le \alpha_J \le 1. \tag{3}$$

Given two step value vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we say that a sequential test statistic $\{\Lambda^{(j)}(n)\}_{n\geq 1}$ is *implemented with step values* $\boldsymbol{\alpha}, \boldsymbol{\beta}$ if there are critical values $\{A_k^{(j)}, B_k^{(j)}\}_{k\in[J]}$ such that

$$P_{\theta^{(j)}}(\Lambda^{(j)}(n) \geq B_k^{(j)} \text{ some } n,\ \Lambda^{(j)}(n') > A_1^{(j)} \text{ all } n' < n) \leq \alpha_k \quad \text{for all} \quad \theta^{(j)} \in H^{(j)}, \tag{4}$$

$$P_{\theta^{(j)}}(\Lambda^{(j)}(n) \leq A_k^{(j)} \text{ some } n,\ \Lambda^{(j)}(n') < B_1^{(j)} \text{ all } n' < n) \leq \beta_k \quad \text{for all} \quad \theta^{(j)} \in G^{(j)}, \tag{5}$$

for all $k \in [J]$. 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 $\alpha_1$ and $\beta_1$, respectively. The values $B_k^{(j)}$, $k \in [J]$, are then such that the similar sequential test with critical values $A_1^{(j)}$ and $B_k^{(j)}$ has type I error probability $\alpha_k$, which is just a restatement of (4), with an analogous statement holding for critical values $A_k^{(j)}$ and $B_1^{(j)}$, type II error probability $\beta_k$, and (5). We say that the test statistics $\{\Lambda^{(j)}(n)\}_{j\in[J],\, n\geq 1}$ for all the streams are *implemented with step values* $\boldsymbol{\alpha}, \boldsymbol{\beta}$ if they are for each stream. In all commonly encountered testing situations there are standard sequential statistics which can be implemented with given step values $\boldsymbol{\alpha}, \boldsymbol{\beta}$. [5, 20] and [4] give examples.

Without loss of generality we assume that, for each $j \in [J]$,

$$A_1^{(j)} \leq A_2^{(j)} \leq \ldots \leq A_J^{(j)} \leq B_J^{(j)} \leq B_{J-1}^{(j)} \leq \ldots \leq B_1^{(j)}, \tag{6}$$

$$A_k^{(j)} = A_{k+1}^{(j)} \text{ if and only if } \beta_k = \beta_{k+1}, \tag{7}$$

$$B_k^{(j)} = B_{k+1}^{(j)} \text{ if and only if } \alpha_k = \alpha_{k+1}. \tag{8}$$

The sequential multiple testing procedures proposed below will 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 will be applied to the statistic $\Lambda^{(j)}(n)$ before ranking. The standardizing functions $\varphi^{(j)}$ can be any increasing functions such that $\varphi^{(j)}(A_k^{(j)})$ and $\varphi^{(j)}(B_k^{(j)})$ do not depend on $j$, and we let

$$a_k = \varphi^{(j)}(A_k^{(j)}) \quad \text{and} \quad b_k = \varphi^{(j)}(B_k^{(j)}), \quad j, k \in [J], \tag{9}$$

denote these common values. Given critical values $\{A_k^{(j)}, B_k^{(j)}\}_{j,k\in[J]}$ satisfying (4)–(5), one may choose arbitrary values $\{a_k, b_k\}_{k\in[J]}$ satisfying the same monotonicity conditions as the $\{A_k^{(j)}, B_k^{(j)}\}$ according to (7)–(8) and then define the standardizing functions $\varphi^{(j)}(\cdot)$ to be increasing, piecewise linear functions satisfying (9). For example, if all the $\alpha_k$ are distinct and the $\beta_k$ 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 $a_k$ must be nondecreasing and the $b_k$ nonincreasing. Finally, we denote $\widetilde{\Lambda}^{(j)}(n) = \varphi^{(j)}(\Lambda^{(j)}(n))$ and then (4)–(5) can be written

$$P_{\theta^{(j)}}(\widetilde{\Lambda}^{(j)}(n) \geq b_k \text{ some } n,\ \widetilde{\Lambda}^{(j)}(n') > a_1 \text{ all } n' < n) \leq \alpha_k \quad \text{for all} \quad \theta^{(j)} \in H^{(j)}, \tag{10}$$

$$P_{\theta^{(j)}}(\widetilde{\Lambda}^{(j)}(n) \leq a_k \text{ some } n,\ \widetilde{\Lambda}^{(j)}(n') < b_1 \text{ all } n' < n) \leq \beta_k \quad \text{for all} \quad \theta^{(j)} \in G^{(j)}, \tag{11}$$

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

The theorems in this paper will give upper bounds on the FDR and pFDR, and their type II versions, for sequential procedures implemented with arbitrary step values $\boldsymbol{\alpha}, \boldsymbol{\beta}$. In practice we recommend using some of the commonly used values such as those recommended by [1] or [21]. For desired value $q_1 \in (0, 1)$ of the type I error metric (FDR or pFDR), the former are

$$\boldsymbol{\alpha}_j = \boldsymbol{\alpha}(q_1)_j := q_1 j/J \quad \text{for} \quad j \in [J], \tag{12}$$

and the latter are

$$\boldsymbol{\alpha}_j = \boldsymbol{\alpha}(q_1)_j = 1 - \left(1 - \left(1 \wedge \frac{q_1 J}{J-j+1}\right)\right)^{1/(J-j+1)} \quad \text{for} \quad j \in [J]. \tag{13}$$

Here, $x \wedge y = \min\{x, y\}$. We recommend using the same expressions for $\boldsymbol{\beta} = \boldsymbol{\beta}(q_2)$ for desired value $q_2 \in (0, 1)$ of the type II error metric (FNR or pFNR). In our examples below we will refer to (12) and (13) as the BH and BL step values, respectively.

### 2.3 Simple Hypotheses and Wald Approximations

In practice, many testing situations can be reduced to, or approximated by, testing simple vs. simple hypotheses. [22, Sect. 1] point out that testing a battery of simple vs. simple hypothesis tests has been the standard setup for use of FDR in the literature. In this section we show how to construct the test statistics $\Lambda^{(j)}(n)$ and critical values $A_s^{(j)}, B_s^{(j)}$ satisfying (4)–(5) for any data stream $j$ such that $H^{(j)}$ and $G^{(j)}$ are both simple hypotheses. In this case the test statistic $\Lambda^{(j)}(n)$ can be taken to be log-likelihood ratio (density under $G^{(j)}$ divided by the density under $H^{(j)}$) yielding the sequential probability ratio test (SPRT; see [23]}. See [5, Sect. 5.1] for a more formal description of hypotheses and parameter space in this setup. In the single null hypothesis testing setup, the SPRT samples until $\Lambda^{(j)}(n) \notin (A, B)$ where the critical values $A$, $B$ are chosen so that

$$P_{H^{(j)}}(\Lambda^{(j)}(n) \geq B \text{ some } n,\ \Lambda^{(j)}(n') > A \text{ all } n' < n) \leq \alpha, \tag{14}$$

$$P_{G^{(j)}}(\Lambda^{(j)}(n) \leq A \text{ some } n,\ \Lambda^{(j)}(n') < B \text{ all } n' < n) \leq \beta \tag{15}$$

for desired type I and II error probabilities $\alpha$ and $\beta$. The most common way of choosing the critical values $A, B$ is to use the closed-form *Wald approximations*

$$A = A_W(\alpha, \beta) := \log\left(\frac{\beta}{1-\alpha}\right) + \rho, \quad B = B_W(\alpha, \beta) := \log\left(\frac{1-\beta}{\alpha}\right) - \rho \tag{16}$$

for which it is assumed that $\alpha + \beta \leq 1$ and $\rho \geq 0$ is a fixed adjustment term to account for the test statistic's excess over the boundary upon stopping. See [24, Sect. 3.3.1] for a derivation of Wald's [25] original $\rho = 0$ case and, based on Brownian motion approximations, [26, p. 50 and Chapter X] derives the value $\rho = 0.583$ which has been used to improve the approximation for continuous random variables. With our multiple testing examples below we recommend using Siegmund's $\rho = 0.583$.

In order to apply the above to the $j$th stream in the multiple testing setup for given step value vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, a slight extension of [5, Theorem 5.1] shows that choosing

$$A_k^{(j)} = A_W(\widetilde{\alpha}_k, \beta_k), \quad B_k^{(j)} = B_W(\alpha_k, \widetilde{\beta}_k) \quad \text{for} \quad k \in [J], \tag{17}$$

where

$$\widetilde{\alpha}_k := \frac{\alpha_1(1-\beta_k)}{1-\beta_1} \quad \text{and} \quad \widetilde{\beta}_k := \frac{\beta_1(1-\alpha_k)}{1-\alpha_1} \quad \text{for} \quad k \in [J], \tag{18}$$

satisfies (4)–(5) up to Wald's approximations, and that $\widetilde{\alpha}_k + \beta_k \leq 1$ and $\alpha_k + \widetilde{\beta}_k \leq 1$ for all $k \in [J]$. Although, strictly speaking, the inequalities (4)–(5) will only be guaranteed to hold approximately using these approximations, [24] show that the actual type I and II error probabilities can only exceed the nominal values by a negligibly small amount in the worst case, and the difference approaches 0 for small nominal values, which is relevant in the present multiple testing situation where we will utilize fractions of $\alpha$ and $\beta$. Alternatives to the Wald approximations in the simple vs. simple setup are Monte Carlo or to replace the terms in (16) by the values $\log\beta$ and $\log(1/\alpha)$, respectively, for which (14)–(15) hold conservatively (see [24]) and proceed similarly.

## 3 Procedures Controlling Type I and II Error Metrics

### 3.1 The Generic Sequential Step-Down Procedure

Next we review the generic sequential step-down procedure, defined in [4]. Versions of this procedure, implemented with certain step values $\boldsymbol{\alpha}, \boldsymbol{\beta}$, will produce our proposed type I and II FDR- and pFDR-controlling sequential procedures. We assume that test statistics are implemented with $\boldsymbol{\alpha}, \boldsymbol{\beta}$, and the critical values referred to are those satisfying (4)–(5).

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 $i$th stage of sampling, and $n_i$ denote the cumulative sample size of any active test statistic up to and including the $i$th stage. The total number of null hypotheses that have been rejected (resp. accepted) at the beginning of the $i$th 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 $i$th stage of sampling ($i = 1, 2, \ldots$) of the *Generic Sequential Step-Down Procedure* implemented with step values $\boldsymbol{\alpha}, \boldsymbol{\beta}$ proceeds as follows.

1. Sample the active streams $\{X_n^{(j)}\}_{j \in \mathcal{J}_i,\, n > n_{i-1}}$ until $n$ equals

$$n_i = \inf\left\{ n > n_{i-1} \,:\, \widetilde{\Lambda}^{(j)}(n) \notin (a_{c_i+1}, b_{r_i+1}) \quad \text{for some} \quad j \in \mathcal{J}_i \right\}. \tag{19}$$

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 $\ell$th ordered active statistic at the end of stage $i$.

3. (a) If the upper boundary in (19) has been crossed, $\widetilde{\Lambda}^{(j)}(n_i) \ge b_{r_i+1}$ for some $j \in \mathcal{J}_i$, then reject the $t_i \ge 1$ null hypotheses

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

where

$$t_i = \max\left\{ t \in [|\mathcal{J}_i|] : \widetilde{\Lambda}^{(j(n_i,\ell))}(n_i) \ge b_{r_i+|\mathcal{J}_i|-\ell+1} \quad \text{for all} \quad \ell = |\mathcal{J}_i| - t + 1, \ldots, |\mathcal{J}_i| \right\},$$

and set $r_{i+1} = r_i + t_i$. Otherwise set $r_{i+1} = r_i$.

(b) If the lower boundary in (19) was crossed, $\widetilde{\Lambda}^{(j)}(n_i) \le a_{c_i+1}$ for some $j \in \mathcal{J}_i$, then accept the $t_i' \ge 1$ null hypotheses

$$H^{(j(n_i,1))}, H^{(j(n_i,2))}, \ldots, H^{(j(n_i,t_i'))},$$

where

$$t_i' = \max\left\{ t \in [|\mathcal{J}_i|] \,:\, \widetilde{\Lambda}^{(j(n_i,\ell))}(n_i) \le a_{c_i+\ell} \quad \text{for all} \quad \ell = 1, \ldots, t \right\},$$

and set $c_{i+1} = c_i + t_i'$. 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 (19). At that point, step-down 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.

### 3.2 Untruncated Sequential Procedures Controlling FDR/FNR or pFDR/pFNR

For $m \in \{0, 1, \ldots, J\}$ and step values $\boldsymbol{\alpha}$, define

$$D(\boldsymbol{\alpha}, m) = m\left(\sum_{j=1}^{J-m+1} \frac{\alpha_j - \alpha_{j-1}}{j} + (J-m)\sum_{j=J-m+2}^{J} \frac{\alpha_j - \alpha_{j-1}}{j(j-1)}\right), \tag{20}$$

$$D(\boldsymbol{\alpha}) = \max_{0 \le m \le J} D(\boldsymbol{\alpha}, m). \tag{21}$$

Our main results in this section, Theorems 3.1 and 3.2, show that the generic sequential step-down procedure applied with step values $\boldsymbol{\alpha}$ have error metrics (e.g., FDR) bounded above by $D(\boldsymbol{\alpha}, m_0)$ where $m_0$ is the number of true null hypotheses, which is in turn bounded above by $D(\boldsymbol{\alpha})$. Since $m_0$ is usually unknown in practice, this latter bound is typically more useful in practice, and the former is primarily of interest for theoretical understanding of the procedure, although there are situations wherein the number of true nulls is assumed to be known (e.g., [27, 28]). It follows from $D(\boldsymbol{\alpha}, m)$ being linear in $\boldsymbol{\alpha}$ that

$$D(c\boldsymbol{\alpha}) = cD(\boldsymbol{\alpha}) \quad \text{for any c>0.} \tag{22}$$

Hence, an error metric of a procedure utilizing step values $\boldsymbol{\alpha}$ being bounded above by $D(\boldsymbol{\alpha})$ is equivalent to saying that, for a desired value $q \in (0, 1)$ of the error metric, the procedure utilizing step values $\widetilde{\boldsymbol{\alpha}} = q\boldsymbol{\alpha}/D(\boldsymbol{\alpha})$ has the error metric bounded above by $q$.

The proofs of Theorems 3.1 and 3.2 are delayed until the Appendix.

**Theorem 3.1** *The generic sequential step-down procedure implemented with step values* $\boldsymbol{\alpha}, \boldsymbol{\beta}$ *satisfies*

$$\mathrm{FDR}(\theta) \le D(\boldsymbol{\alpha}, m_0) \le D(\boldsymbol{\alpha}) \quad \text{and} \quad \mathrm{FNR}(\theta) \le D(\boldsymbol{\beta}, m_1) \le D(\boldsymbol{\beta}) \quad \text{for all} \quad \theta \in \Theta \tag{23}$$

*regardless of the dependence between data streams, where D is as defined in* (20)–(21) *and* $m_0$ *and* $m_1$ *are the numbers of true and false null hypotheses, respectively. In particular, given* $q_1, q_2 \in (0, 1)$, *the generic sequential step-down procedure implemented with step values* $\widetilde{\boldsymbol{\alpha}} = q_1\boldsymbol{\alpha}/D(\boldsymbol{\alpha})$, $\widetilde{\boldsymbol{\beta}} = q_2\boldsymbol{\beta}/D(\boldsymbol{\beta})$ *satisfies*

$$\mathrm{FDR}(\theta) \le q_1 \quad \text{and} \quad \mathrm{FNR}(\theta) \le q_2 \quad \text{for all} \quad \theta \in \Theta \tag{24}$$

*regardless of the dependence between data streams.*

For the next theorem, recall that $R$ denotes the number of null hypotheses rejected by the procedure in question.

**Theorem 3.2** *Let $D$ be as defined in (20)–(21) and $m_0$ and $m_1$ denote the numbers of true and false null hypotheses, respectively. Fix arbitrary $\theta \in \Theta$.*

1. *If $\gamma_1, \gamma_2 > 0$ are values such that the generic sequential step-down procedure implemented with step values $\boldsymbol{\alpha}, \boldsymbol{\beta}$ satisfies*

$$P_\theta(R > 0) \geq \gamma_1 \quad \text{and} \quad P_\theta(R < J) \geq \gamma_2, \tag{25}$$

   then this procedure satisfies

$$\text{pFDR}(\theta) \leq \frac{D(\boldsymbol{\alpha}, m_0)}{P_\theta(R > 0)} \leq \frac{D(\boldsymbol{\alpha})}{\gamma_1} \quad \text{and} \quad \text{pFNR}(\theta) \leq \frac{D(\boldsymbol{\beta}, m_1)}{P_\theta(R < J)} \leq \frac{D(\boldsymbol{\beta})}{\gamma_2} \tag{26}$$

   regardless of the dependence between data streams.
2. *In particular, given $q_1, q_2 \in (0, 1)$, the generic sequential step-down procedure implemented with step values $\widetilde{\boldsymbol{\alpha}} = q_1\gamma_1\boldsymbol{\alpha}/D(\boldsymbol{\alpha})$, $\widetilde{\boldsymbol{\beta}} = q_2\gamma_2\boldsymbol{\beta}/D(\boldsymbol{\beta})$ satisfies*

$$\text{pFDR}(\theta) \leq q_1 \quad \text{and} \quad \text{pFNR}(\theta) \leq q_2 \tag{27}$$

   regardless of the dependence between data streams, if $\gamma_1, \gamma_2 > 0$ satisfy (25) for this procedure.
3. *For $j \in [J]$ let*

$$\begin{aligned} \gamma_{1j} &= P_{\theta^{(j)}}(\widetilde{\Lambda}^{(j)}(n) \geq b_1 \text{ some } n,\ \widetilde{\Lambda}^{(j)}(n') > a_J \text{ all } n' < n), \\ \gamma_{2j} &= P_{\theta^{(j)}}(\widetilde{\Lambda}^{(j)}(n) \leq a_1 \text{ some } n,\ \widetilde{\Lambda}^{(j)}(n') < b_J \text{ all } n' < n). \end{aligned} \tag{28}$$

   If $\max_{j\in[J]} \gamma_{1j} > 0$ and $\max_{j\in[J]} \gamma_{2j} > 0$, then (25)–(26) hold with $\gamma_i = \max_{j\in[J]} \gamma_{ij}$, $i = 1, 2$.

## 4 Truncated, Rejective Procedures

In this section we describe versions of our procedures which only stop early to reject (rather than accept) null hypotheses and, thus, only explicitly control the corresponding type I multiple testing error rate (FDR or pFDR), recorded in Theorems 4.1 and 4.2. This setting naturally corresponds with having a streamwise maximum sample size (or "truncation point") $\overline{N}$ which we assume throughout this section. For this reason we refer to them as "truncated, rejective" versions of the procedures. These 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 possible preventing simultaneous achievement of the nominal error bounds (4)–(5), or (c) the type II multiple testing error rate (e.g, FNR) $q_2$ is not well motivated. In any of theses cases, one may prefer to drop the requirement that the type II multiple testing error rate be strictly controlled at an arbitrary level $q_2$ and use one of the rejective procedures which, roughly speaking,

are similar but ignore the lower stopping boundaries $A_k^{(j)}$. On the other hand, if $q_2$ is not well motivated but the statistician prefers early stopping under the null hypotheses, we encourage the use of one of our procedures in Sect. 3 with both early stopping to reject and accept null hypotheses, and treat $q_2$ as a parameter to be chosen to give a procedure with other desirable operating characteristics, such as expected total or streamwise maximum sample size.

### 4.1 Setup and Critical Values

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 $\theta$ be as in Sect. 2. Since only the type I multiple testing error rate, FDR or pFDR, will be explicitly controlled we only require specification of null hypotheses $H^{(j)} \subseteq \Theta^{(j)}$ and not of alternative hypotheses $G^{(j)}$. We assume a streamwise maximum sample size $\overline{N}$ for each stream, but with only notational changes, what follows could be formulated by stream-specific truncation points $\{\overline{N}^{(j)}\}_{j\in[J]}$ or with sample sizes other than $1, \ldots, \overline{N}$.

Given step values $\boldsymbol{\alpha}$, we adapt our definition from Sect. 2.2 of the test statistics $\{\Lambda^{(j)}(n)\}_{j,n}$ being *implemented with step values* $\boldsymbol{\alpha}$ to this truncated, rejective setting if, for all $j, k \in [J]$, the critical values $B_1^{(j)}, \ldots, B_J^{(j)}$ satisfy

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

as well as (6) and (8) without loss of generality. We let the standardizing functions $\varphi^{(j)}$ be any increasing functions such that $b_k = \varphi^{(j)}(B_k^{(j)})$ does not depend on $j$, and $\widetilde{\Lambda}^{(j)}(n) = \varphi^{(j)}(\Lambda^{(j)}(n))$ denote the standardized statistics.

In the next section we give the truncated, rejective version of the generic step-down procedure, and then in Theorems 4.1 and 4.2 state their FDR- and pFDR-controlling properties. The proofs are similar to the proofs of Theorems 3.1 and 3.2 and are sketched in the Appendix.

### 4.2 The Generic Rejective Sequential Step-Down Procedure

With the notation of Sect. 3.1, the $i$th stage ($i = 1, 2, \ldots$) of the *Generic Rejective Sequential Step-Down Procedure* with step values $\boldsymbol{\alpha}$ proceeds as follows.

1. Sample the active streams $\{X_n^{(j)}\}_{j\in\mathcal{J}_i,\, n>n_{i-1}}$ until $n$ equals

$$n_i = \overline{N} \wedge \inf\left\{n > n_{i-1} : \widetilde{\Lambda}^{(j)}(n) \geq b_{r_i+1} \quad \text{for some} \quad j \in \mathcal{J}_i\right\}. \tag{29}$$

2. If $n_i = \overline{N}$ and no test statistic has crossed the critical value in (29), accept all active null hypotheses and terminate the procedure. Otherwise, proceed to Step 3.
3. Order the active test statistics

$$\widetilde{\Lambda}^{(j(n_i,1))}(n_i) \le \widetilde{\Lambda}^{(j(n_i,2))}(n_i) \le \ldots \le \widetilde{\Lambda}^{(j(n_i,|\mathcal{J}_i|))}(n_i)$$

and reject the $t_i \ge 1$ null hypotheses

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

where

$$t_i = \max\left\{t \in [|\mathcal{J}_i|]: \widetilde{\Lambda}^{(j(n_i,\ell))}(n_i) \ge b_{r_i+|\mathcal{J}_i|-\ell+1} \quad \text{for all} \quad \ell = |\mathcal{J}_i| - t + 1, \ldots, |\mathcal{J}_i|\right\}.$$

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

### 4.3 Truncated, Rejective Procedures Controlling FDR or pFDR

**Theorem 4.1** *The generic rejective sequential step-down procedure implemented with step values* $\boldsymbol{\alpha}$ *satisfies*

$$\text{FDR}(\theta) \le D(\boldsymbol{\alpha}, m_0) \le D(\boldsymbol{\alpha}) \quad \text{for all} \quad \theta \in \Theta \tag{30}$$

*regardless of the dependence between data streams, where D is as defined in* (20)–(21) *and* $m_0$ *is the number of true null hypotheses. In particular, given* $q_1 \in (0, 1)$, *the generic rejective sequential step-down procedure implemented with step values* $\widetilde{\boldsymbol{\alpha}} = q_1\boldsymbol{\alpha}/D(\boldsymbol{\alpha})$ *satisfies*

$$\text{FDR}(\theta) \le q_1 \quad \text{for all} \quad \theta \in \Theta \tag{31}$$

*regardless of the dependence between data streams.*

**Theorem 4.2** *Let D be as defined in* (20)–(21) *and* $m_0$ *denote the numbers of true null hypotheses. Fix arbitrary* $\theta \in \Theta$.

1. *If* $\gamma_1 > 0$ *is such that the generic rejective sequential step-down procedure implemented with step values* $\boldsymbol{\alpha}$ *satisfies*

$$P_\theta(R > 0) \ge \gamma_1, \tag{32}$$

then this procedure satisfies

$$\text{pFDR}(\theta) \le \frac{D(\boldsymbol{\alpha}, m_0)}{P_\theta(R > 0)} \le \frac{D(\boldsymbol{\alpha})}{\gamma_1} \tag{33}$$

regardless of the dependence between data streams.

2. *In particular, given* $q_1 \in (0, 1)$, *the generic rejective sequential step-down procedure implemented with step values* $\widetilde{\boldsymbol{\alpha}} = q_1\gamma_1\boldsymbol{\alpha}/D(\boldsymbol{\alpha})$ *satisfies*

$$\text{pFDR}(\theta) \le q_1 \tag{34}$$

regardless of the dependence between data streams, if $\gamma_1 > 0$ satisfies (25) for this procedure.

3. *For* $j \in [J]$ *let*

$$\gamma_{1j} = P_{\theta^{(j)}}(\widetilde{\Lambda}^{(j)}(n) \geq b_1 \text{ some } n \leq \overline{N}). \tag{35}$$

If $\max_{j\in[J]} \gamma_{1j} > 0$, then (32)–(33) hold with $\gamma_1 = \max_{j\in[J]} \gamma_{1j}$.

## 5 Simulation Studies

In this section we present the results of simulation studies of the sequential step-down procedures described above in Sects. 3 and 4 in order to evaluate their operating characteristics and compare them to analogous fixed-sample procedures. All computations in this section and the analysis of the UK Yellow Card pharmacovigilance database in Sect. 6 were performed using our Python package available from github.com/bartroff792/mult-seq-dependence. First we describe the distributional setups for our simulation studies, and in then discuss the results in Tables 2 and 3 and conclude by comparing our proposed sequential procedures with appropriate fixed-sample competitors.

In order to examine the performance of the proposed sequential procedures on correlated Poisson and Binomial data we (i) generated correlated normally distributed data from a Gaussian copula [29], (ii) applied in the inverse Gaussian c.d.f. to this data to obtain quantiles, and (iii) applied Poisson and Binomial (respectively) c.d.f.s to the quantiles to obtain correlated data of those marginal distributions. That is, for each time point $i$ we first generated a $J$-dimensional multivariate normal vector $Y_i \sim N_J(0, \Sigma)$, using the Toeplitz covariance matrix $\Sigma_{jj'} = \rho^{|j-j'|}$ for all $j,j' \in [J]$ and where $\rho$ is a chosen value. Then the $J$-dimensional vector of quantiles $Q_i = \Phi^{-1}(Y_i)$ was computed where $\Phi$ is the standard normal c.d.f. Finally the data values $X_i^{(j)}$, $j \in [J]$, were set to be

**Table 2** Expected sample size $EN_{\text{Seq}}$ and achieved FDR and FNR of sequential step-down procedures controlling FDR and FNR, evaluated on negatively correlated Binomial data generated for different numbers $m_0$ of true null hypotheses; standard errors (SE) are given following each estimated quantity

| $m_0$ | $N_{\text{FSS}}$ | $EN_{\text{Seq}}$ | SE | FDR | SE | FNR | SE |
|---|---|---|---|---|---|---|---|
| 0 | 101 | 36.0 | 0.34 | 0.000 | 0.000 | 0.111 | 0.010 |
| 1 | 105 | 39.8 | 0.34 | 0.009 | 0.001 | 0.079 | 0.006 |
| 3 | 101 | 45.9 | 0.33 | 0.027 | 0.002 | 0.049 | 0.004 |
| 5 | 97 | 50.5 | 0.32 | 0.047 | 0.003 | 0.031 | 0.002 |
| 7 | 103 | 53.9 | 0.33 | 0.069 | 0.004 | 0.020 | 0.002 |
| 9 | 113 | 55.1 | 0.32 | 0.109 | 0.007 | 0.007 | 0.001 |
| 10 | – | 55.2 | 0.32 | 0.168 | 0.012 | 0.000 | 0.000 |

$N_{\text{FSS}}$ is the sample size of the corresponding fixed-sample test using the same nominal FDR rate $q_1 = 0.25$ and which matches the achieved FNR rate of the sequential procedure

**Table 3** Expected sample size $EN_{\text{Seq}}$ and achieved FDR and FNR of sequential step-down procedures controlling FDR and FNR, evaluated on negatively correlated Poisson data generated for different numbers $m_0$ of true null hypotheses; standard errors (SE) are given following each estimated quantity

| $m_0$ | $N_{\text{FSS}}$ | $EN_{\text{Seq}}$ | SE | FDR | SE | FNR | SE |
|---|---|---|---|---|---|---|---|
| 0 | 83 | 31.6 | 0.22 | 0.000 | 0.000 | 0.107 | 0.010 |
| 1 | 79 | 34.3 | 0.23 | 0.009 | 0.001 | 0.077 | 0.006 |
| 3 | 77 | 38.1 | 0.24 | 0.029 | 0.002 | 0.052 | 0.004 |
| 5 | 73 | 40.4 | 0.25 | 0.050 | 0.003 | 0.038 | 0.003 |
| 7 | 79 | 41.6 | 0.26 | 0.077 | 0.004 | 0.023 | 0.002 |
| 9 | 99 | 41.0 | 0.26 | 0.119 | 0.008 | 0.007 | 0.001 |
| 10 | – | 40.1 | 0.23 | 0.172 | 0.012 | 0.000 | 0.000 |

$N_{\text{FSS}}$ is the sample size of the corresponding fixed-sample test using the same nominal FDR rate $q_1 = 0.25$ and which matches the achieved FNR rate of the sequential procedure

$$X_i^{(j)} = \begin{cases} \max\left\{n : F_{\lambda_j}(n) \le Q_{ij}\right\}, & \text{for the Poisson case,} \\ \mathbf{1}\{Q_{ij} \le p_j\}, & \text{for the Binomial case.} \end{cases}$$

Here $F_\lambda$ is the c.d.f. of the Poisson distribution with mean $\lambda$, and the $\lambda_j$ and $p_j$ are the specified means of the data in the $j$th stream in the Poisson and Binomial cases, respectively. This process was repeated independently at each needed time point $i$. The specified value $\rho$ above equals correlation of "adjacent" elements of $Y_i$, corresponding to adjacent data streams, and we note that choosing $\rho < 0$, as we do in some of our simulations below, leads to negative dependence among data streams and hence is outside the PRDS condition under which step-up and down procedures are already well known (e.g., [30]) to control FDR.

For the Binomial data on which the procedures in Table 2 were evaluated, individual Bernoulli observations $X_i^{(j)}$ were generated as described above with mean $p_H = 0.05$ under the null and $p_G = 0.15$ under the alternative. For Table 3, the observations $X_i^{(j)}$ were generated as Poisson with mean $\lambda_H = 1.5$ under the null, $\lambda_G = 2.0$ under the alternative. For both the Binomial and Poisson data generated to evaluate the procedures in Tables 2 and 3, respectively, $J = 10$ data streams were utilized scenarios with $m_0 = 0, 1, 3, 5, 7, 9, 10$ true nulls and $\rho = -.6$ was used in the generating copula. Tables 2 and 3 contain the operating characteristics of the untruncated, FDR and FNR controlling procedures described in Theorem 3.1. In both tables, the sequential procedures were implemented with the nominal FDR and FNR error bounds $q_1 = 0.25$ and $q_2 = 0.15$ and using the Wald approximations (17)–(18) with the null/alternative pairs $\lambda_H, \lambda_G$ and $p_H, p_G$ for the Poisson and Binomial scenarios, respectively, with the BH step values (12). The truncated sequential procedures were implemented using Monte Carlo under the known, null parameter values $\lambda_H, p_H$ to find the critical values, as described in Sect. 4.1.

The achieved FDR and FNR of the sequential procedures in Tables 2 and 3 are substantially smaller than their nominal values $q_1 = 0.25$ and $q_2 = 0.15$, although FNR is less so. To provide a meaningful comparison of these sequential procedures' expected sample size $EN_{\text{Seq}}$ with a comparable fixed-sample size alternative procedure, the sample size $N_{\text{FSS}}$ needed for a fixed-sample BH procedure with the same

nominal FDR control level $q_1 = 0.25$ to match corresponding achieved FNR values are given in the second columns of the tables. Thus, the sequential procedures provide a large savings in term of sample sizes, at least roughly 45% savings in all cases and more than 60% savings in the small $m_0$ cases in both tables, where the adaptivity of the sequential procedures is pronounced. And although the sequential procedures are conservative in the sense of having achieved FDR and FNR values smaller than their nominal values, the corresponding fixed-sample procedures are even more conservative in terms of achieved FDR value due to the larger sample size, while their FNR values were matched for comparison.

## 6 Application: The UK's Yellow Card Scheme Pharmacovigilance Database

### 6.1 Introduction

The United Kingdom's Yellow Card pharmacovigilance Database (yellowcard.mhra.gov.uk/information), run by the Medicines and Healthcare Products Regulatory Agency (MHRA), collects voluntary reports on safety and side effects from the public and healthcare professionals in the UK on a host of healthcare treatments including medicines, vaccines, medical devices, and even e-cigarettes; in what follows, for simplicity we refer to the treatments in Yellow Card generically as "drugs." The Yellow Card data collection scheme, which now includes a mobile phone app, began in 1964, spurred by the thalidomide crisis. Its use has grown steadily since then, now receiving more than 20,000 reports of possible side effects each year, and totaling more than half a million reports in the scheme's first 40 years.

We created a Python script to download and analyze "Interactive Drug Analysis Profiles" PDF reports from Yellow Card on all of the roughly 2800 different drugs in the database through February 2016. Figure 1 is a heatmap showing the number of drugs (indicated by color) in these reports as a function of their total number of reported adverse action reactions over this period ($y$ axis, on the $\log_{10}$ scale) and number of years ($x$ axis) the drug has been in the database, calculated from the starting date for collection of reports on the drug until the closure date of the drug's most recent summary report. The figure shows that some of the drugs have many thousands of reaction reports, collected over decades. However, the majority of drug entries have fewer reports, collected over less than one decade, emphasizing the need for nimble, sequential monitoring of this data.

### 6.2 Illustrating the Sequential Procedures' Performance on the Database

Some Yellow Card data has been considered in the statistics literature on multiple testing (e.g., [31]), but the inherently streaming nature of the data has not been taken into account, with only fixed-sample multiple testing methods being applied to it so far, to our knowledge. In order to demonstrate the behavior of the

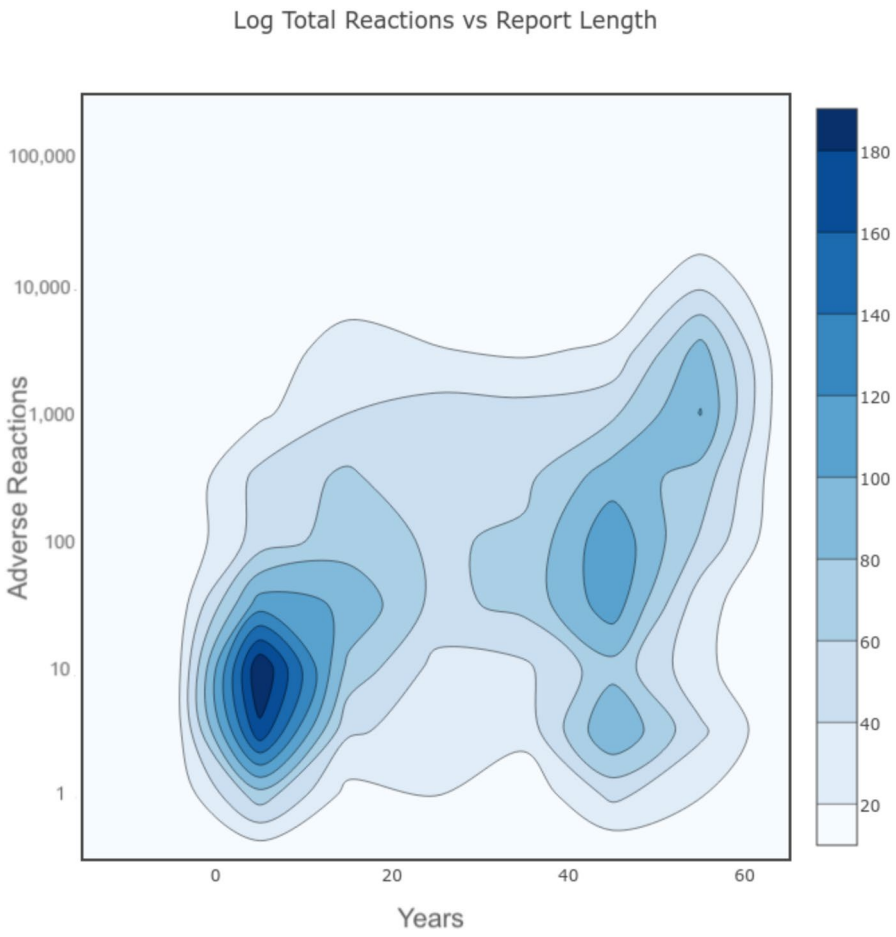


**Fig. 1** Number of drugs (indicated by heatmap color) in the Yellow Card database in February 2016 as a function of total adverse reaction reports ($y$ axis, on the $\log_{10}$ scale) and the number of years ($x$ axis) the drug has been in the database

sequential procedures proposed above on data like that in Yellow Card, we focus on a particular type of side effect report, amnesia, in the Yellow Card reports.

The MHRA informed us at the time of our data download that the exact time series of the reports to Yellow Card were not being made available, so it is not possible to apply our sequential procedures to the data precisely as it was being received by Yellow Card. But, for each of the drug reports downloaded as described above and illustrated in Fig. 1, we were able to obtain the average number of amnesia and other side effect reports per year, the starting date for collection of reports for each drug, and the date of the closure of the drug's most recent summary report. So here we apply our sequential procedures to data streams simulated using the actual yearly average rates obtained from Yellow Card. This

approach, similar to the parametric Bootstrap [32], makes our simulation as close as possible to real Yellow Card data, given the limitations in data availability.

The Yellow Card data was used to simulate future drug reports as follows. At each time step a Poisson number of amnesia and non-amnesia reports for each drug was generated as correlated Poisson processes with rates

$$\lambda_{amn}^{(j)} = (|\{\text{amnesia reports from drug } j\}| + 1)/T_j, \tag{36}$$

$$\begin{aligned} \lambda_{-amn}^{(j)} &= (|\{\text{non-amnesia reports from drug } j\}| + 1)/T_j, \quad \text{where} \\ T_j &= (\text{drug } j \text{ most recent report}) - (\text{drug } j \text{ start date}). \end{aligned} \tag{37}$$

Note that we have added 1 amnesia and non-amnesia report to each drug in (36)–(37) to account for issues with rare or exotic drugs like inherently higher variance in use and reporting. These rates are visualized in Fig. 2, where the blue, green, and red points are the drugs with "low," "medium," and "high" amnesia report rates according to (38) and (39).

In order to generate a medically realistic correlation structure for the drugs, we utilized BioGPT [33], a generative language model trained on biomedical research articles. For each drug under consideration, we calculated the element-wise mean of the BioGPT's *token embedding* (its numerical representation

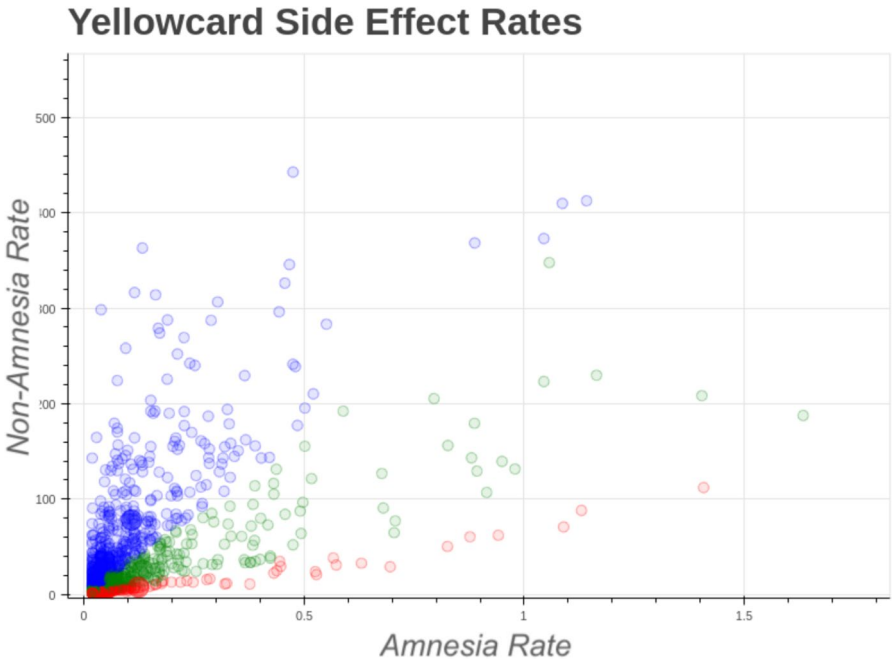


**Fig. 2** Drug side effect rates (36)–(37) from the Yellow Card database data. Blue points are those drugs $j$ with $p^{(j)} \leq p_H$, green those with $p_H < p^{(j)} < p_G$, and red those with $p^{(j)} \geq p_G$, as defined by (38) and (39) (color figure online)

of phrases, words, or characters) for "[DrugName] as a drug" and then applied $K$-means clustering to obtain $K = 15$ clusters of the drugs. Within each cluster, the copula method described in Sect. 5 was then used to generate the Poisson data using the Toeplitz covariance matrix given there, but with the correlation parameter $\rho$ generated as a Beta(4, 2) random variable, shifted and scaled to have support $[-1, 1]$. In this way, drugs that were medically related according to BioGPT were more likely to receive non-trivial correlation.

In order to account accurately for the differing number of reports that drugs with different prescription and use rates in the database have, we set up the hypothesis tests not in terms of the Poisson rate parameters (36)–(37) but rather a scaled parameter:

$$p^{(j)} = \frac{\lambda_{amn}^{(j)}}{\lambda_{amn}^{(j)} + \lambda_{-amn}^{(j)}} \approx \frac{\#\text{amnesia reports for drug } j}{\text{total } \#\text{ reports for drug } j}. \tag{38}$$

The data generation scheme above is equivalent to a Binomial number of amnesia reports being generated at each time step, with success probability (38) and number of trials equal to the Poisson random total number of reports.

To set up our hypothesis tests, we utilize the Yellow Card database to calculate

$$p_H = 50\text{th percentile of } \{p^{(j)}\}, \quad \text{and} \quad p_G = 90\text{th percentile of } \{p^{(j)}\}, \tag{39}$$

which we use as surrogates for "typical" and "extreme" Binomial rates of amnesia reports, respectively. After calculating $p_H$ and $p_G$ on the entire Yellow Card database, we then applied our sequential procedure to the $J = 1800$ drugs with the highest total side effect reports in the database. This reduction still maintains a large number of data streams but also allows us to filter out drugs that have only recently been made available with too small a number of reports, but also the opposite extreme of filtering out obscure and antiquated drugs that exist in the database (e.g., "Mammalian Blood" and "Wool Fat").

For each of these $J = 1800$ drugs we tested the hypotheses

$$H^{(j)} : p^{(j)} \leq p_H \quad \text{vs.} \quad G^{(j)} : p^{(j)} \geq p_G > p_H$$

representing that the drug has a "typical" versus abnormally high rate of amnesia reports. For these we implemented the open-ended sequential test described in Theorem 3.1 for Binomial data using FDR controlled at $q_1 = 0.05$ and FNR at $q_2 = 0.15$, with the Wald approximations (17)–(18) the BH step values (12). Running this data generation scheme with 1800 data streams a single time is computationally intensive even on a Google Cloud Computing instance, so rather than report Monte Carlo statistics, in Table 4 we give the output of a typical run and there list the first 15 drugs whose streams terminated to accept, and the first 15 to reject, the null hypothesis $H^{(j)}$. Since the simulation is setup where each "observation" is the number of reports in a year, in the table the Termination Step corresponds to the number of years required for termination of each stream. The Termination Level is a measure of the strength of the terminal action, where smaller

**Table 4** The earliest 15 Yellow Card database drugs to terminate to reject, or accept, the null hypothesis $H^{(j)}$ in a single run of the sequential step-down procedure controlling FDR at $q_1 = 0.05$ and FNR at $q_2 = 0.15$

| Drug | Terminal action | Termination step | Termination level |
|---|---|---|---|
| Bupropion | Accept $H^{(j)}$ | 1 | 2 |
| Clozapine | Accept $H^{(j)}$ | 1 | 2 |
| Etanercept | Accept $H^{(j)}$ | 2 | 8 |
| Venlafaxine | Accept $H^{(j)}$ | 2 | 8 |
| Varenicline | Accept $H^{(j)}$ | 2 | 8 |
| Paroxetine | Accept $H^{(j)}$ | 2 | 8 |
| Clostridium Tetani | Accept $H^{(j)}$ | 2 | 8 |
| Telaprevir | Accept $H^{(j)}$ | 2 | 8 |
| Adalimumab | Accept $H^{(j)}$ | 3 | 10 |
| Rofecoxib | Accept $H^{(j)}$ | 3 | 10 |
| Ranibizumab | Accept $H^{(j)}$ | 4 | 21 |
| Amoxycillin | Accept $H^{(j)}$ | 4 | 21 |
| Oseltamivir | Accept $H^{(j)}$ | 4 | 21 |
| Trimethoprim | Accept $H^{(j)}$ | 4 | 21 |
| Ciprofloxacin | Accept $H^{(j)}$ | 4 | 21 |
| Zopiclone | Reject $H^{(j)}$ | 7 | 1 |
| Simvastatin | Reject $H^{(j)}$ | 9 | 2 |
| Zolpidem | Reject $H^{(j)}$ | 13 | 3 |
| Bifonazole | Reject $H^{(j)}$ | 15 | 5 |
| Rimonabant | Reject $H^{(j)}$ | 15 | 5 |
| Enzalutamide | Reject $H^{(j)}$ | 25 | 7 |
| Gabapentin | Reject $H^{(j)}$ | 25 | 7 |
| Atorvastatin | Reject $H^{(j)}$ | 31 | 8 |
| cAC10-vcMMAE | Reject $H^{(j)}$ | 32 | 9 |
| Dutasteride | Reject $H^{(j)}$ | 33 | 10 |
| Mitotane | Reject $H^{(j)}$ | 36 | 11 |
| Boceprevir | Reject $H^{(j)}$ | 39 | 12 |
| Lurasidone | Reject $H^{(j)}$ | 42 | 14 |
| Asenapine | Reject $H^{(j)}$ | 42 | 14 |
| Retigabine | Reject $H^{(j)}$ | 44 | 16 |

The Termination Step corresponds to the number of years, and termination level is the cumulative value of the level $\ell$ in step 3a or 3b of the procedure as defined in Sect. 3.1. cAC10-vcMMAE is an abbreviation for *Chimeric Human Murine Monoclonal Antibody Cac10 Anti Cd30 Linked To Cytotoxic Molecule Sgd 1006*

values indicate a stronger acceptance/rejection of the null, and equals the *cumulative* value of the level $\ell$ in step 3a or 3b of the procedure as defined in Sect. 3.1.

### 6.3 Discussion

While we draw no medical conclusions from our simulations, we note that some of the drugs identified by our sequential procedure by stopping to reject their null hypotheses in favor of an abnormally high level of amnesia reports, have been associated with amnesia in the biomedical literature (e.g., [34]). Monitoring pharmacovigilance databases for amnesia and other cognitive impairment adverse events is a real-world priority of statisticians and regulators around the world; see [35] and [36] for examples of such analyses in the U.S.'s Food and Drug Administration Adverse Event Reporting System (FAERS; see www.fda.gov/media/97567/download), and [37] for an analysis of the Korean Adverse Event Reporting System Database. The example presented here provides a possible new statistical tool for this important task that takes into account the real-time, streaming nature of the data as well as the inherent multiple comparisons issues with such data. Although we are aware of proposed approaches that separately take into account the time-sequential nature of the data (e.g., [38]) and the multiple comparison aspect, this is the first proposal to incorporate both that we are aware of.

The dominant statistical methods in this existing literature are "disproportionality analyses" (see [39, 40, 41, 42]), utilizing odds (or log odds) ratios of the probability of reporting an adverse event, which is equivalent to our use of the probabilities of such reports in (38)–(39). There we use the probabilities of such reports across the entire database as the levels of "typical" and "extreme" in (39) to set our hypotheses. This approach, known as "signal of disproportional reporting," is common in the literature (e.g., [42–44]), although different approaches exist on how to choose these cutoffs and some of these authors use individual-level covariates, such as age in [38], in setting these cutoffs, which was not available in our example. However, signal of disproportional reporting is not a requirement of our method proposed here and those levels could be set by other methods such as from the medical literature, regulatory concerns, etc.

The approach proposed here could be used with other types of drugs and side effects, although domain-specific details the particular application will likely affect the choices made in implementation. For example, in monitoring adverse side effects of statins in the Korean pharmacovigilance database, [45] grouped together statin-specific symptoms and measured severity on the WHO-specified scale, and the life-threatening nature of adverse immune-mediated reactions associated with certain immunotherapies (e.g., [46]) would be taken into account in setting the early stopping properties of the procedure.

## 7 Conclusion

We have proposed both open-ended and truncated sequential multiple testing procedures which can control FDR or pFDR (and their type II analogs, FNR or pFNR in the open-ended case) under arbitrary dependence between the data streams. These procedures have shown large savings in average sample size compared to their fixed-sample counterparts in our simulation studies, and in Sect. 6, we demonstrate how these procedures may be used to monitor data streams like those coming from a pharmacovigilance database, like the UK's Yellow Card database.

## Appendix

### Proof of Theorem 3.1

We first prove the inequalities for FDR in (23) for the sequential step-down procedure with step values $\boldsymbol{\alpha}, \boldsymbol{\beta}$, and then discuss those for FNR. The inequalities in (24) then follow, using (22). Our proof largely follows the arguments of [3], with a few important differences to account for the sequential nature of our procedures.

Fix arbitrary $\theta$ and omit it from the notation, and below let $P(\cdot)$ and $E(\cdot)$ denote the probability and expectation under $\theta$ and an arbitrary joint distribution between the data streams. Without loss of generality assume $H^{(1)}, \ldots, H^{(m_0)}$ are the true hypotheses for some $m_0 \in [J]$, the $m_0 = 0$ case being trivial since FDR $= 0$ in this case. For $i \in [J]$ let $\tau_i$ denote the time at which the $i$th stream $\widetilde{\Lambda}^{(i)}$ terminates, and for $i, j \in [J]$ let

$$F_{ij} = \{\widetilde{\Lambda}^{(i)}(\tau_i) \in [b_j, b_{j-1})\}, \tag{40}$$

setting $b_0 = \infty$ to handle the $j = 1$ case. The event $F_{ij}$ is similar (in fact, contained in) the event in (10), but it specifies which interval $[b_j, b_{j-1})$ the $i$th statistic $\widetilde{\Lambda}^{(i)}$ is in when stopping to reject its corresponding null $H^{(i)}$; we will refer to this as $H^{(i)}$ *being rejected at the jth level*. Recalling that $R$ and $V$ denote the number of null and true null hypotheses, respectively, rejected by the procedure, define

$$p_{ijk} = P(F_{ij} \cap \{R = k\}) \quad \text{for} \quad i, j, k \in [J], \tag{41}$$

with which we can write

$$\text{FDR} = \sum_{i=1}^{m_0} \sum_{j=1}^{J} \sum_{k=j}^{J} \frac{p_{ijk}}{k}, \tag{42}$$

which is analogous to similar expressions obtained for FDR in the fixed-sample setting obtained by [2] and [47]. To see why (42) holds here for the sequential procedure, write

$$\text{FDR} = E\Big(\frac{V}{R \vee 1}\Big) = \sum_{k=1}^{J} \frac{1}{k} E(V\mathbf{1}\{R=k\}) = \sum_{k=1}^{J}\sum_{i=1}^{m_0} \frac{1}{k} P(\{H^{(i)} \text{ rejected}\} \cap \{R=k\})$$
$$= \sum_{k=1}^{J}\sum_{i=1}^{m_0}\sum_{j=1}^{k} \frac{1}{k} P(F_{ij} \cap \{R=k\}) = \sum_{i=1}^{m_0}\sum_{j=1}^{J}\sum_{k=j}^{J} \frac{1}{k} p_{ijk},$$

where in the last equality, we have reordered the sums and used the definition of $p_{ijk}$.

We further extend the notation to expand (42). Given $k \in [J]$, $\ell \in [k \wedge m_0]$, and a pair of vectors $(\boldsymbol{i},\boldsymbol{j})$ in

$$\Omega_{\ell k} = \{(\boldsymbol{i},\boldsymbol{j}) \in [m_0]^{\ell} \times [k]^{\ell} \,:\, 1 \le i_1 < i_2 < \ldots < i_{\ell} \le m_0\},$$

define

$$p_{\boldsymbol{ij}k} = P\left(\bigcap_{d=1}^{\ell} F_{i_d j_d} \cap \{R=k\} \cap \{V=\ell\}\right), \tag{43}$$

which is the probability of $k$ rejections, of which the false rejections are $H^{(i_d)}$ being rejected at the $j_d$th level, $d \in [\ell]$. Let

$$\Omega_{\ell k}(i,j) = \{(\boldsymbol{i},\boldsymbol{j}) \in \Omega_{\ell k} \,:\, (i,j) = (i_d,j_d) \text{ for some } d \in [\ell]\}$$

be the set of those vector pairs that include $H^{(i)}$ being rejected at the $j$th level. With these definitions we can further expand $p_{ijk}$ in the form:

$$p_{ijk} = \sum_{\ell=1}^{k\wedge m_0} \sum_{(\boldsymbol{i},\boldsymbol{j}) \in \Omega_{\ell k}(i,j)} p_{\boldsymbol{ij}k}. \tag{44}$$

This equality holds because, for distinct pairs $(\boldsymbol{i},\boldsymbol{j}) \in \Omega_{\ell k}(i,j)$, the events in (43) correspond to different nulls being rejected, or at different levels, and are hence disjoint, thus,

$$\sum_{(\boldsymbol{i},\boldsymbol{j}) \in \Omega_{\ell k}(i,j)} p_{\boldsymbol{ij}k} = P\big(F_{ij} \cap \{R=k\} \cap \{V=\ell\}\big).$$

For distinct values of $\ell$, the events in this last are clearly disjoint, hence summing over $\ell$ yields

$$\sum_{\ell=1}^{k\wedge m_0} \sum_{(\boldsymbol{i},\boldsymbol{j}) \in \Omega_{\ell k}(i,j)} p_{\boldsymbol{ij}k} = \sum_{\ell=1}^{k\wedge m_0} P\big(F_{ij} \cap \{R=k\} \cap \{V=\ell\}\big) = P(F_{ij} \cap \{R=k\}) = p_{ijk}.$$

With (44) established, we can substitute it into (42) to obtain

$$\begin{aligned}\text{FDR} = \sum_{i=1}^{m_0}\sum_{j=1}^{J}\sum_{k=j}^{J}\sum_{\ell=1}^{k\wedge m_0}\sum_{(\boldsymbol{i},\boldsymbol{j})\in\Omega_{\ell k}(i,j)}\frac{p_{\boldsymbol{ijk}}}{k} &= \sum_{\ell=1}^{m_0}\sum_{k=\ell}^{J}\sum_{i=1}^{m_0}\sum_{j=1}^{k}\sum_{(\boldsymbol{i},\boldsymbol{j})\in\Omega_{\ell k}(i,j)}\frac{p_{\boldsymbol{ijk}}}{k} \\ &= \sum_{\ell=1}^{m_0}\sum_{k=\ell}^{J}\sum_{(\boldsymbol{i},\boldsymbol{j})\in\Omega_{\ell k}}\frac{\ell\, p_{\boldsymbol{ijk}}}{k}.\end{aligned} \tag{45}$$

The second equality in (45) is obtained by reordering the summations, and the last equality uses

$$\sum_{i=1}^{m_0}\sum_{j=1}^{k}\sum_{(\boldsymbol{i},\boldsymbol{j})\in\Omega_{\ell k}(i,j)}\frac{p_{\boldsymbol{ijk}}}{k} = \sum_{(\boldsymbol{i},\boldsymbol{j})\in\Omega_{\ell k}}\frac{\ell\, p_{\boldsymbol{ijk}}}{k},$$

since each pair $(\boldsymbol{i},\boldsymbol{j}) \in \Omega_{\ell k}$ appears in the $\ell$ sets $\Omega_{\ell k}(i_1,j_1), \ldots, \Omega_{\ell k}(i_\ell,j_\ell)$, and only those sets.

With the expression (45) for FDR established, we now consider constraints that the $p_{\boldsymbol{ijk}}$ must satisfy. For the first constraint, take arbitrary $k \in [J]$, $\ell \in [k \wedge m_0]$, $(\boldsymbol{i},\boldsymbol{j}) \in \Omega_{\ell k}$, and let $j_{(1)} \le \ldots \le j_{(\ell)}$ denote an ordering of the values in $\boldsymbol{j} = (j_1, \ldots, j_\ell)$. For an arbitrary fixed-sample step-down procedure with $p_{\boldsymbol{ijk}}$ defined analogously, [3, Lemma 3.3] show that

$$p_{\boldsymbol{ijk}} = 0 \quad \text{if} \quad k - \ell > m - m_0 \quad \text{or} \quad j_{(d)} > k - \ell + d \text{ for some } d \in [\ell]. \tag{46}$$

Their proof depends only on the step-down structure and the values that $R$, $V$ can take with positive probability, and thus, hold for our sequential step-down procedure as well, so we do not repeat the proof here.

The other constraint relates the $p_{\boldsymbol{ijk}}$ to the error probabilities in (10). First observe that, for any $i \in [m_0]$, the events $F_{i1}, F_{i2}, \ldots$ are disjoint, as are $\{R = 1\}, \{R = 2\}, \ldots$. Then, for any $s \in [J]$, we write (10) as follows

$$\begin{aligned}\alpha_s \ge P\left(\bigcup_{j=1}^{s} F_{ij}\right) = \sum_{j=1}^{s} P(F_{ij}) = \sum_{j=1}^{s} P\left(F_{ij} \cap \bigcup_{k=1}^{J}\{R=k\}\right) &= \sum_{j=1}^{s}\sum_{k=1}^{J} P\big(F_{ij} \cap \{R=k\}\big) \\ &= \sum_{j=1}^{s}\sum_{k=1}^{J} p_{ijk} = \sum_{j=1}^{s}\sum_{k=1}^{J}\sum_{\ell=1}^{k\wedge m_0}\sum_{(\boldsymbol{i},\boldsymbol{j})\in\Omega_{\ell k}(i,j)} p_{\boldsymbol{ijk}},\end{aligned} \tag{47}$$

using (44) for this last equality.

Combining (45), (46), and (47), the goal of finding the worst-case joint distribution that maximizes FDR can be stated as finding the $\{p_{\boldsymbol{ijk}}\}$ that

$$\text{maximize FDR} = \sum_{\ell=1}^{m_0}\sum_{k=\ell}^{J}\sum_{(\boldsymbol{i},\boldsymbol{j})\in\Omega_{\ell k}}\frac{\ell\, p_{\boldsymbol{ijk}}}{k} \quad \text{subject to (46) and (47).} \tag{48}$$

At this point the $\{p_{\boldsymbol{ijk}}\}$ can be completely divorced from their original meaning about a multiple testing procedure and treated as arbitrary variables in the constrained optimization problem (48). Guo and Rao [3] solve a similar problem, identical to

(48) except for the second constraint; their second constraint is sufficient but not necessary for our second constraint (47); about this, see Remark A.1. Guo and Rao proceed to solve this problem by producing an upper bound on FDR that coincides with FDR at its maximum $D(\boldsymbol{\alpha}, m_0)$, and then show that the maximizer satisfies both constraints. Since their second constraint is sufficient for ours, their proof applies here, and we do not repeat more of the details here. This establishes the inequalities for FDR in (23).

The inequalities for FNR in (23) are established in a completely similar way after reversing the roles of the "type I" and "type II" objects, e.g., substituting FNR for FDR, $\boldsymbol{\beta}$ for $\boldsymbol{\alpha}$, the bound (11) for (10), etc. The details are straightforward and, therefore, omitted here. □

***Remark A.1*** Guo and Rao's [3, expression (19)] second constraint involves bounding from above the probability that a $p$ value for a true null falls in the interval $[\alpha_{j-1}, \alpha_j)$ by $\alpha_j - \alpha_{j-1}$. Validity of the $p$ value is not sufficient for this to hold and, for example, it may fail for valid but discrete $p$ values. Therefore, their proof actually requires a condition stronger than validity on the $p$ values, for example having the exact uniform (0, 1) null distribution. We avoid the need for such a stronger condition by summing over $j$ in (47), and thus, only need (10) which is the sequential analog of validity for $p$ values in the fixed-sample setup.

## Sketch of Proof of Theorem 3.2

The proof of the inequalities for pFDR in (26) is similar to the proof of Theorem 3.1 after replacing $p_{ijk}$ in (41) by

$$p_{ijk} = P(F_{ij} \cap \{R = k\} | R > 0)$$

and using (25) to bound $P(R > 0)$ from below, leading to the factor of $1/\gamma_1$ in the bound. The proof of the inequalities for pFNR in (26) proceeds similarly, conditioning on $R < J$ and using $\gamma_2$. Part 2 of Theorem 3.2 then follows using (22). In Part 3, the event in (28) implies rejection of $H^{(j)}$, hence $\max_j \gamma_{1j} \leq P(R > 0)$, with analogous statements applying to the type II version.

## Sketch of Proofs of Theorems 4.1 and 4.2

The proof of Theorem 4.1 follows the proof of FDR control in Theorem 3.1 with $p_{ijk}$ defined in (41) but with

$$F_{ij} = \{\widetilde{\Lambda}^{(i)}(\tau_i) \in [b_j, b_{j-1}), \quad \tau_i < \overline{N}\} \tag{49}$$

rather than (40).

For Theorem 4.2, the proof proceeds the similarly but with $p_{ijk}$ replaced by

$$p_{ijk} = P(F_{ij} \cap \{R = k\} | R > 0),$$

again with $F_{ij}$ given by (49). Then (32) is used to bound $P(R > 0)$ from below, giving the factor of $1/\gamma_1$ in the bound. Part 2 follows from (22), and the event in (35) of Part 3 implies rejection of $H^{(j)}$, hence $\max_j \gamma_{1j} \leq P(R > 0)$.

## Declarations

**Competing interest** The authors declare no competing interests.

## References


1. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 57:289–300
2. Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29(4):1165–1188
3. Guo W, Rao MB (2008) On control of the false discovery rate under no assumption of dependency. J Stat Plan Inference 138(10):3176–3188
4. Bartroff J (2018) Multiple hypothesis tests controlling generalized error rates for sequential data. Stat Sin 28:363–398
5. Bartroff J, Song J (2020) Sequential tests of multiple hypotheses controlling false discovery and nondiscovery rates. Sequ Anal 39:65–91
6. Hart S, Weiss B (1997) Significance levels for multiple tests. Stat Probab Lett 35(1):43–48
7. Lai TL (1981) Asymptotic optimality of invariant sequential probability ratio tests. Ann Stat 9:318–333
8. Lai TL (1988) Nearly optimal sequential tests of composite hypotheses. Ann Stat 16:856–886
9. Lai TL (2001) Sequential analysis: some classical problems and new challenges (with discussion). Stat Sin 11:303–408
10. Lai TL, Robbins H (1979) Adaptive design and stochastic approximation. Ann Stat 7:1196–1221
11. Lai TL (1987) Adaptive treatment allocation and the multi-armed bandit problem. Ann Stat 15:1091–1114
12. Bartroff J, Lai TL (2008) Efficient adaptive designs with mid-course sample size adjustment in clinical trials. Stat Med 27:1593–1611
13. Lai TL (2000) Sequential multiple hypothesis testing and efficient fault detection-isolation in stochastic systems. IEEE Trans Inf Theory 46:595–608
14. Bartroff J, Lai TL (2010) Multistage tests of multiple hypotheses. Commun Stat—Theory Methods 39:1597–1607
15. Han J, Lai TL, Spivakovsky V (2006) Approximate policy optimization and adaptive control in regression models. Comput Econ 27:433–452
16. Lai TL (1995) Sequential changepoint detection in quality control and dynamical systems. J R Stat Soc B 57(4):613–644
17. Lai T, Shih M, Wong S (2006) Flexible modeling via a hybrid estimation scheme in generalized mixed models for longitudinal data. Biometrics 62(1):159–167
18. Bartroff J, Lai TL, Shih M (2013) Sequential experimentation in clinical trials: design and analysis. Springer, New York
19. Storey JD (2002) A direct approach to false discovery rates. J R Stat Soc Ser B 64(3):479–498
20. Bartroff J, Song J (2014) Sequential tests of multiple hypotheses controlling type I and II familywise error rates. J Stat Plan Inference 153:100–114
21. Benjamini Y, Liu W (1999) A step down multiple hypotheses testing procedure that controls the false discovery rate under independence. J Stat Plan Inference 82(1):163–170
22. Müller P, Parmigiani G, Rice K (2007) FDR and Bayesian multiple comparisons rules. In: Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, West M (eds) Bayesian statistics 8: proceedings of the eighth Valencia international meeting, June 2–6, 2006, Oxford University Press, pp 349–370
23. Chernoff H (1972) Sequential analysis and optimal design. Society for Industrial and Applied Mathematics, Philadelphia

24. Hoel PG, Port SC, Stone CJ (1971) Introduction to statistical theory. Houghton Mifflin Co., Boston
25. Wald A (1947) Sequential analysis (Reprinted by Dover). Wiley, New York, p 1973
26. Siegmund D (1985) Sequential analysis: tests and confidence intervals. Springer, New York
27. Song Y, Fellouris G (2017) Asymptotically optimal, sequential, multiple testing procedures with prior information on the number of signals. Electron J Stat 11(1):338–363
28. He X, Bartroff J (2021) Asymptotically optimal sequential FDR and pFDR control with (or without) prior information on the number of signals. J Stat Plan Inference 210:87–99
29. Trivedi PK, Zimmer DM (2007) Copula modeling: an introduction for practitioners. Found Trends Econometr 1:1–111
30. Finner H, Dickhaus T, Roters M (2009) On the false discovery rate and an asymptotically optimal rejection curve. Ann Stat 37:596–618
31. Döhler S (2018) A discrete modification of the Benjamini-Yekutieli procedure. Econometr Stat 5:137–147
32. Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. CRC Press, Boca Raton
33. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, Liu T-Y (2022) BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 23(6):bbac409
34. Chavant F, Favrelière S, Lafay-Chebassier C, Plazanet C, Pérault-Pochat M-C (2011) Memory disorders associated with consumption of drugs: updating through a case/noncase study in the French PharmacoVigilance Database. Br J Clin Pharmacol 72(6):898–904
35. Borchert JS, Wang B, Ramzanali M, Stein AB, Malaiyandi LM, Dineley KE (2019) Adverse events due to insomnia drugs reported in a regulatory database and online patient reviews: comparative study. J Med Internet Res 21(11):e13371
36. Sato K, Mano T, Iwata A, Toda T (2020) Neuropsychiatric adverse events of chloroquine: a real-world pharmacovigilance study using the FDA Adverse Event Reporting System (FAERS) database. Biosci Trends 14(2):139–143
37. Sunwoo Y, Eom SH, Yun JS, Kim Y, Lee J, Lee SH, Shin S, Choi YJ (2024) Real-world data-derived pharmacovigilance on drug-induced cognitive impairment utilizing a nationwide spontaneous adverse reporting system. Medicina 60(7):1028
38. Hocine MN, Musonda P, Andrews NJ, Paddy Farrington C (2009) Sequential case series analysis for pharmacovigilance. J R Stat Soc Ser A 172(1):213–236
39. Bihan K, Lebrun-Vignes B, Funck-Brentano C, Salem J-E (2020) Uses of pharmacovigilance databases: an overview. Therapies 75(6):591–598
40. Lavertu A, Vora B, Giacomini KM, Altman R, Rensi S (2021) A new era in pharmacovigilance: toward real-world data and digital monitoring. Clin Pharmacol Ther 109(5):1197–1202
41. Zhang W, Chen M, Cai X, Zhang M, Hu M, Hu Y, Yang Y, Zhu J, Du Y, Yang C (2024) Detection and analysis of signals of adverse events of memantine based on the US Food and Drug Administration Adverse Event Reporting System. Expert Opin Drug Saf 23(5):617–625
42. Zorych I, Madigan D, Ryan P, Bate A (2013) Disproportionality methods for pharmacovigilance in longitudinal observational databases. Stat Methods Med Res 22(1):39–56
43. Hauben M, Madigan D, Gerrits CM, Walsh L, Van Puijenbroek EP (2005) The role of data mining in pharmacovigilance. Expert Opin Drug Saf 4(5):929–948
44. Hauben M, Reich L (2005) Potential utility of data-mining algorithms for early detection of potentially fatal/disabling adverse drug reactions: a retrospective evaluation. J Clin Pharmacol 45(4):378–384
45. Kim H, Kim N, Lee DH, Kim H-S (2017) Analysis of national pharmacovigilance data associated with statin use in Korea. Basic Clin Pharmacol Toxicol 121(5):409–413
46. Ali AK, Watson DE (2017) Pharmacovigilance assessment of immune-mediated reactions reported for checkpoint inhibitor cancer immunotherapies. Pharmacotherapy: J Hum Pharmacol Drug Ther 37(11):1383–1390
47. Sarkar SK (2002) Some results on false discovery rate in stepwise multiple testing procedures. Ann Stat 30(1):239–257