Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics, Vol. 43: 3–19, 2016
doi: 10.1111/sjos.12161

A Rejection Principle for Sequential Tests of Multiple Hypotheses Controlling Familywise Error Rates

JAY BARTROFF
Department of Mathematics, University of Southern California

JINLIN SONG
Analysis Group, Inc.

ABSTRACT. We present a unifying approach to multiple testing procedures for sequential (or streaming) data by giving sufficient conditions for a sequential multiple testing procedure to control the familywise error rate (FWER). Together, we call these conditions a 'rejection principle for sequential tests', which we then apply to some existing sequential multiple testing procedures to give simplified understanding of their FWER control. Next, the principle is applied to derive two new sequential multiple testing procedures with provable FWER control, one for testing hypotheses in order and another for closed testing. Examples of these new procedures are given by applying them to a chromosome aberration data set and finding the maximum safe dose of a treatment.

Key words: closed testing, multiple comparisons, multiple testing, sequential analysis, sequential hypothesis testing, streaming data, testing in order

1. Introduction and background

The need for multiple-comparison-type corrections due to testing more than one null hypothesis occurs in nearly all areas of scientific inquiry in which statistical hypothesis testing is employed. In a number of these areas, the data are inherently sequential, or 'streaming', such as in multiple endpoint (or multi-arm) clinical trials (Jennison & Turnbull, 2000, Chapter 15), multi-channel changepoint detection (Tartakovsky *et al.*, 2003) and its applications to biosurveillance (Mei, 2010), genetics and genomics (e.g., Salzman *et al.,* 2011) and acceptance sampling with multiple criteria (Baillie, 1987). We note that, although the term 'sequential' is often used to describe multiple testing procedures that analyse fixed sample size data in a stepwise fashion (e.g. Goeman and Solari, 2010), here, we primarily use the term to describe data that are inherently sequential in its collection or analysis, as in Siegmund (1985).

Adopting the familywise error rate (FWER) metric, this paper takes a unifying approach to sequential multiple testing procedures that control the FWER. Specifically, we give sufficient conditions for a sequential multiple testing procedure to control the FWER, which turn out to be much simpler and easier to verify in many cases than a comprehensive analysis of the procedure. We call these two sufficient conditions, given in Theorem 1, a *rejection principle for sequential tests*, following and extending the seminal work of Goeman & Solari (2010) who accomplished this for fixed sample size procedures and in turn extended and unified the work of Romano & Wolf (2005), Hommel *et al.* (2007) and Marcus *et al.* (1976). Two overlapping aspects of the problem that we must deal with in the sequential setting that were absent from the fixed sample size setting are how to allow for acceptances of hypotheses as well as rejections and the interplay of sequential sampling with the accept/reject decisions. In the fixed sample size setting, rejecting a hypothesis is equivalent to not accepting it; however, in the sequential setting, these are not necessarily equivalent because there is the third possibility of performing additional sampling. These aspects are dealt with by expanding the notion of a procedure's rejection function, introduced in Section 2, to incorporate not just the already rejected hypotheses as in Goeman and Solari's (2010) fixed sample size setting but also the already accepted hypotheses as well as the current sample size of those data streams that are still being sampled.

The rest of the paper is organized as follows. After briefly reviewing the relevant background in the next paragraph, our rejection principle is introduced in Section 2, and its sufficiency for FWER control is established in Theorem 1. In Section 3, we apply our rejection principle to derive sequential multiple testing procedures, first deriving two general procedures that do not assume a special structure among the hypotheses that control the FWER and both type I and II FWERs (defined in the following), respectively. Then our rejection principle is applied to derive sequential procedures for two settings wherein the special structure of the hypothesis is known: testing hypotheses in order (Rosenbaum, 2008) and closed testing (Marcus *et al.*, 1976). In Section 4, we give examples of these derived procedures, first applying the sequential procedure for testing hypotheses in order to real data from a study (Masjedi *et al.*, 2000) of chromosome aberration effects of an anti-tuberculosis drug, and then applying the closed testing procedure to finding the maximum safe dose of a treatment, wherein the sequential procedure is evaluated in a simulation study. Section 5 provides a summary and discussion.

Separately, multiple testing and sequential testing are both quite mature fields, the former dating back to classical multiple comparison procedures of Fisher (1932), Scheffé (1953), Tukey, and others (see Seber and Lee, 2003) for testing hypotheses about parameter vectors in linear models. Work on sequential hypothesis testing dates back to Wald's (1947) invention of sequential analysis following World War II; see Siegmund (1985) for a summary of the major developments. However, the intersection of these two areas is less well developed in a general setting. One area that has been considered is the adaptation of some classical fixed sample size tests about vector parameters, such as those mentioned earlier, to the sequential sampling setting, including O'Brien and Fleming's (1979) sequential version of Pearson's χ^2 test, and Tang *et al.*'s (1989; 1993) group sequential extensions of O'Brien's (1984) generalized least squares statistic. For bivariate normal populations, Jennison & Turnbull (1993) proposed a sequential test of two one-sided hypotheses about the bivariate mean vector and Cook & Farewell (1994) proposed a sequential test in a similar setting, but where one of the hypotheses is two-sided. A procedure for comparing three treatments was proposed by Siegmund (1993), related to Paulson's (1964) earlier procedure for selecting the largest mean of k normal distributions, which Bartroff & Lai (2010) showed to be a special case of their more general sequential step-down method; this procedure is presented in Section 3.1, where we give a simplified proof of its FWER control using our rejection principle. Recently, Ye *et al.* (2013) proposed a group sequential Holm procedure that is also a special case of the procedure of Bartroff & Lai (2010), which allows arbitrary sampling schemes in addition to group sequential. The first sequential procedures to simultaneously control both the type I and II FWERs were introduced by De and Baron (2012a, 2012b). Bartroff & Song (2014) propose a different approach to sequential control of both type I and II FWERs, and their procedure will also be discussed in Section 3.1 and shown to satisfy this rejection principle.

2. A rejection principle for sequential tests

We present a general framework for testing multiple hypotheses with sequential data, that is, with data streams. Assume that there are $k \geq 2$ data streams

$$X_1^{(j)}, X_2^{(j)}, \ldots, \quad \text{for} \quad j = 1, \ldots, k. \tag{1}$$

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 $X_n^{(j)}, X_{n'}^{(j)}$ or between-stream data $X_n^{(j)}, X_{n'}^{(j')}$ $(j \neq j')$. Assume that for each data stream $j = 1, \ldots, k$, there is a parameter vector $\theta^{(j)} \in \Theta^{(j)}$ governing that stream $X_1^{(j)}, X_2^{(j)}, \ldots$, and it is desired to test a hypothesis $H^{(j)} \subseteq \Theta^{(j)}$ about $\theta^{(j)}$, with $H^{(j)}$ considered true if $\theta^{(j)} \in H^{(j)}$, and false if otherwise. The global parameter $\theta = (\theta^{(1)}, \ldots, \theta^{(k)})$ is the concatenation of the individual parameters and is contained in the global parameter space $\Theta = \Theta^{(1)} \times \cdots \times \Theta^{(k)}$. Each $\theta \in \Theta$ indexes a probability measure P_θ. With $\mathcal{H} = \{H^{(1)}, \ldots, H^{(k)}\}$ denoting the set of hypotheses to be tested, given $\theta = (\theta^{(1)}, \ldots, \theta^{(k)}) \in \Theta$, we let

$$\mathcal{T}(\theta) = \left\{ H^{(j)} \in \mathcal{H} : \theta^{(j)} \in H^{(j)} \right\}$$

denote the collection of true hypotheses when P_θ is the underlying probability measure, and

$$\mathcal{F}(\theta) = \left\{ H^{(j)} \in \mathcal{H} : \theta^{(j)} \notin H^{(j)} \right\} = \mathcal{H} \setminus \mathcal{T}(\theta) \tag{2}$$

the false hypotheses. The FWER is the probability of rejecting any true hypothesis,

$$\text{FWER} = \text{FWER}(\theta) = P_\theta(\text{any } H^{(j)} \in \mathcal{T}(\theta) \text{ rejected}). \tag{3}$$

In what follows, we will frequently drop the argument θ from these expressions for brevity.

At any point during sampling, we shall refer to the *active* hypotheses as those that have not yet been accepted or rejected and *active* data streams as those corresponding to active hypotheses. A *sequential multiple testing procedure* for the data streams (1) is simply a sampling and decision procedure that maps the current data from all the data streams and the list of active hypotheses to one of the following:

(i) A list of one or more active hypotheses to reject;
(ii) A list of one or more active hypotheses to accept;
(iii) An additional sample size to draw from each active data stream before reevaluation.

We note that the additional sample size in (c) can be 1, as in full sequential sampling. As a simplistic example, suppose there are $k = 2$ data streams (1) and it is desired to test the respective hypotheses $H^{(1)}$ and $H^{(2)}$. A multiple testing procedure may first decide to sample 10 observations from the streams, yielding

$$\begin{array}{l} X_1^{(1)}, X_2^{(1)} \ldots, X_{10}^{(1)} \\ X_1^{(2)}, X_2^{(2)} \ldots, X_{10}^{(2)}. \end{array}$$

Based on these data, the procedure may decide to reject $H^{(2)}$ and then sample a single additional observation from stream 1 (the lone remaining active stream), yielding

$$\begin{array}{l} X_1^{(1)}, X_2^{(1)} \ldots, X_{10}^{(1)}, X_{11}^{(1)} \\ X_1^{(2)}, X_2^{(2)} \ldots, X_{10}^{(2)}. \end{array}$$

At this point, the procedure may decide not to accept or reject $H^{(1)}$ but rather sample an additional seven observations from stream 1, yielding

$$\begin{array}{l} X_1^{(1)}, X_2^{(1)} \ldots, X_{10}^{(1)}, \ldots, X_{18}^{(1)} \\ X_1^{(2)}, X_2^{(2)} \ldots, X_{10}^{(2)}, \end{array}$$

at which time, the procedure may decide to accept $H^{(1)}$. Examples of sequential multiple testing procedures will be given in Section 3.

Our main result, given in Theorem 1, extends a result of Goeman & Solari (2010) for fixed sample size multiple testing procedures to the sequential setting (i.e. for testing on data streams), in which sequential sampling may occur between acceptances/rejections of hypotheses. Given any sequential multiple testing procedure meeting the aforementioned general definition, its rejection behavior (and hence its FWER, as we will see) can be described by its *rejection function*, which we denote by ρ, which is a possibly random function mapping the set $\mathcal{R} \subseteq \mathcal{H}$ of already rejected hypotheses, the set $\mathcal{A} \subseteq \mathcal{H}$ (disjoint from $\mathcal{R}$) of already accepted hypotheses, the current sample size $n \in \mathcal{N}$ and the set $\mathcal{D}_n$ of all data available at time n to a set $\rho(\mathcal{R}, \mathcal{A}, n, \mathcal{D}_n) \subseteq \mathcal{H} \setminus (\mathcal{R} \cup \mathcal{A})$ of hypotheses to reject. Here, $\mathcal{N}$ is the set of all possible streamwise sample sizes of the procedure. Because the last argument $\mathcal{D}_n$ of $\rho(\mathcal{R}, \mathcal{A}, n, \mathcal{D}_n)$ will always be the available data at time n, in what follows, we denote $\rho(\mathcal{R}, \mathcal{A}, n, \mathcal{D}_n)$ simply by $\rho(\mathcal{R}, \mathcal{A}, n)$. Letting $\varnothing$ denote the empty set, the value $\rho(\mathcal{R}, \mathcal{A}, n) = \varnothing$ indicates that either additional sampling will be performed or that the testing procedure is terminated, which occurs if $n = \max \mathcal{N}$ or $\mathcal{R} \cup \mathcal{A} = \mathcal{H}$. By convention, define $\rho(\mathcal{R}, \mathcal{A}, \infty) = \varnothing$. Therefore, iterations of ρ describe the procedure's successive rejections of hypotheses, and we will keep track of all the hypotheses that have been rejected after each iteration of ρ in sets $\mathcal{R}_i \subseteq \mathcal{H}$ and the hypotheses that have been accepted after each iteration in $\mathcal{A}_i \subseteq \mathcal{H}$. To this end, define $\mathcal{R}_0 = \varnothing$, $n_0 = 0$, and

$$\begin{aligned} \mathcal{R}_{i+1} &= \mathcal{R}_i \cup \rho(\mathcal{R}_i, \mathcal{A}_i, n_{i+1}) \quad \text{for } i = 0, 1, 2, \ldots, \text{ where} \\ n_{i+1} &= \inf\{n \in \mathcal{N}, n \geq n_i : \rho(\mathcal{R}_i, \mathcal{A}_i, n) \neq \varnothing\}, \end{aligned} \tag{4}$$

letting $\inf \varnothing = \infty$ as usual. In what follows, we do not need to define the sets $\mathcal{A}_i$ of accepted hypotheses explicitly as we did for the $\mathcal{R}_i$ in (4); we only assume that the $\mathcal{A}_i$ are defined in some way, such that $\varnothing = \mathcal{A}_0 \subseteq \mathcal{A}_1 \subseteq \ldots$ and $\mathcal{A}_i \cap \mathcal{R}_i = \varnothing$ for all i. There are at most k non-trivial iterations of ρ in the sense that $\rho \neq \varnothing$, by virtue of the fact that there are k hypotheses and hence at most k hypotheses that could be rejected. Consequently, $\mathcal{R}_k = \mathcal{R}_{k+1} = \ldots$ and this common set is the totality of all hypotheses rejected by the procedure, and FWER(θ) can thus be written $P_\theta(\mathcal{R}_k \not\subseteq \mathcal{F}(\theta))$.

The following theorem gives a rejection principle for sequential tests and establishes its sufficiency for FWER control.

Theorem 1. *Let $\theta \in \Theta$ denote the true value of the global parameter, $\alpha \in (0,1)$, and $\mathcal{H}$ and $\mathcal{R}_k$ as defined earlier. If ρ and $\mathcal{N}$ are the rejection function and sample size set, respectively, of a sequential multiple testing procedure, such that*

(1) for any subsets $\mathcal{R}, \mathcal{R}', \mathcal{A}$ of $\mathcal{H}$ with $\mathcal{R} \subseteq \mathcal{R}'$ and $\mathcal{A} \cap \mathcal{R} = \varnothing$, and any $n \in \mathcal{N}$, we have

$$\rho(\mathcal{R}, \mathcal{A}, n) \subseteq \rho(\mathcal{R}', \varnothing, n) \cup \mathcal{R}' \tag{5}$$

with P_θ-probability 1, and

(2)

$$P_\theta(\rho(\mathcal{F}(\theta), \varnothing, n) \subseteq \mathcal{F}(\theta) \textit{ for all } n \in \mathcal{N}) \geq 1 - \alpha, \tag{6}$$

then

$$P_\theta(\mathcal{R}_k \not\subseteq \mathcal{F}(\theta)) \leq \alpha, \tag{7}$$

that is, FWER(θ) is no greater than α.

Proof. Let $\mathcal{F} = \mathcal{F}(\theta)$, $V = \{\rho(\mathcal{F}, \varnothing, n) \subseteq \mathcal{F}$ for all $n \in \mathcal{N}\}$ and $W_i = \{\mathcal{R}_i \subseteq \mathcal{F}\}$, $i = 0, 1, \ldots$. We will prove by induction that, as events, $V \subseteq W_i$ for all $i \geq 0$; the result (7) then follows from the $i = k$ case and (6). The $i = 0$ case is trivial because $\mathcal{R}_0 = \varnothing \subseteq \mathcal{F}$. Suppose that $V \subseteq W_i$. Then on V, we have

$$\begin{aligned}
\mathcal{R}_{i+1} &= \mathcal{R}_i \cup \rho(\mathcal{R}_i, \mathcal{A}_i, n_{i+1}) \quad \text{[by (4)]} \\
&\subseteq \mathcal{R}_i \cup (\rho(\mathcal{F}, \varnothing, n_{i+1}) \cup \mathcal{F}) \quad \text{[by (5) and the inductive hypothesis]} \\
&= \rho(\mathcal{F}, \varnothing, n_{i+1}) \cup \mathcal{F} \quad \text{[by the inductive hypothesis]} \\
&= \mathcal{F}.
\end{aligned}$$

□

The rejection principle for fixed sample size procedures presented in Goeman & Solari (2010) can be regarded as a special case of Theorem 1 because all fixed sample size procedures are sequential procedures, with the fixed sample size being the only element of $\mathcal{N}$.

3. Applications of this rejection principle

In this section, we apply the rejection principle in Theorem 1 in a number of settings, some with special structures and some without.

3.1. Testing hypotheses without assumed special structure

A sequential step-down procedure. Bartroff & Lai (2010) proposed a sequential multiple testing procedure, extending Holm's (1979) fixed sample size step-down procedure, which controls FWER regardless of between-stream dependence, requiring only that each hypothesis has a sequential test statistic that marginally controls the conventional type I error probability. After briefly introducing Bartroff and Lai's procedure, we show that its error control is a special case of Theorem 1.

Here, we present the Bartroff–Lai procedure in slightly more generality than in their original paper. In particular, here we remove the need for (a) common critical values among the k sequential test statistics by using standardizing functions, in the following, introduced by Bartroff & Song (2014) and (b) critical values for all possible significance levels; here, we only need critical values corresponding to certain fractions of the desired FWER bound α. Given a set $\mathcal{N}$ of possible per-stream sample sizes, assume that for each $j = 1, \ldots, k$, associated with the jth hypothesis $H^{(j)}$ and data stream $X_1^{(j)}, X_2^{(j)}, \ldots$ is a scalar-valued sequential test statistic $T_n^{(j)} = T_n^{(j)}(X_1^{(j)}, \ldots, X_n^{(j)})$ with k critical values $B_1^{(j)} \geq \ldots \geq B_k^{(j)}$, such that

$$P_{\theta^{(j)}}\left(T_n^{(j)} \geq B_s^{(j)} \quad \text{for some } n \in \mathcal{N}\right) \leq \frac{\alpha}{k-s+1} \quad \text{for all} \quad \theta^{(j)} \in H^{(j)}, \tag{8}$$

for all $s = 1, \ldots, k$. The inequality (8) just says that the sequential test that stops and rejects $H^{(j)}$ at the first $n \in \mathcal{N}$, such that $T_n^{(j)} \geq B_s^{(j)}$, and accepts $H^{(j)}$; otherwise, it has type I error probability $\alpha/(k-s+1)$. For $j = 1, \ldots, k$, define the *standardizing function*

$$\varphi^{(j)}(x) = \begin{cases} x - B_k^{(j)} + 1, & \text{for } x \leq B_k^{(j)} \\ \frac{x - B_s^{(j)}}{B_{s-1}^{(j)} - B_s^{(j)}} + k - s + 1, & \text{for } B_s^{(j)} \leq x \leq B_{s-1}^{(j)} \text{ if } B_{s-1}^{(j)} > B_s^{(j)}, \quad 1 < s \leq k \\ x - B_1^{(j)} + k, & \text{for } x \geq B_1^{(j)}, \end{cases} \tag{9}$$

which is an increasing, piecewise-linear function, such that $\varphi^{(j)}(B_s^{(j)}) = k - s + 1$ for $s = 1, \dots, k$, and thus

$$T_n^{(j)} \geq B_s^{(j)} \quad \Leftrightarrow \quad \varphi^{(j)}(T_n^{(j)}) \geq k - s + 1. \tag{10}$$

The standardizing functions will be applied to the test statistics before ranking them, and they allow us to compare the test statistics $T_n^{(1)}, \dots, T_n^{(k)}$, which may be on different scales. In general, the standardizing function can be any increasing function, such that $\varphi^{(j)}(B_s^{(j)})$ does not depend on j. To use a different standardizing function, all that would need to be adjusted in what follows is the right hand side of the inequality in (11), as shown in the following.

Letting $\mathcal{I}_1 = \{1, 2, \dots, k\}$, $r_1 = 0$ and $n_0 = 0$, the ith stage ($i = 1, \dots, k$) of the Bartroff–Lai procedure proceeds as follows.

(1) Sample each active data stream $\{X_n^{(j)}\}_{j \in \mathcal{I}_i}$ up to sample size

$$n_i = \inf\left\{n \in \mathcal{N} : n > n_{i-1} \quad \text{and} \quad T_n^{(j)} \geq B_{r_i+1}^{(j)} \quad \text{for some} \quad j \in \mathcal{I}_i\right\}.$$

(2) With $\varphi^{(j)}$ given by (9), standardize and order the active test statistics $\widetilde{T}_{n_i}^{(j)} = \varphi^{(j)}(T_{n_i}^{(j)})$, $j \in \mathcal{I}_i$, as follows:

$$\widetilde{T}_{n_i}^{(j(i,1))} \geq \widetilde{T}_{n_i}^{(j(i,2))} \geq \dots \geq \widetilde{T}_{n_i}^{(j(i,|\mathcal{I}_i|))}.$$

(3) Reject $H^{(j(i,1))}, H^{(j(i,2))}, \dots, H^{(j(i,m_i))}$, where

$$m_i = \min\left\{m \geq 1 : \widetilde{T}_{n_i}^{(j(i,m+1))} < k - r_i - m\right\}. \tag{11}$$

(4) If $i = k$ or $n_i = \max\mathcal{N}$, stop and accept all remaining active hypotheses. Otherwise, let $\mathcal{I}_{i+1}$ be the indices of the remaining hypotheses, set $r_{i+1} = r_i + m_i$ and continue on to stage $i + 1$.

An important caveat is that, at any point, any of the active hypotheses may be accepted without violating the FWER control proved in the following, as long as the set $\mathcal{I}_i$ of active hypotheses is appropriately updated. To maintain generality, here, we do not specify an acceptance rule for the Bartroff–Lai procedure. Sequential multiple testing procedures with explicit acceptance, as well as rejection, rules are considered below, which control both the type I and II FWERs, the latter defined there.

Because the Bartroff–Lai procedure presented here is slightly more general than the one proved to control FWER in Bartroff & Lai (2010, Theorem 2.1), we record this procedure's FWER control in Corollary 1, which we prove by applying the rejection principle in Theorem 1.

Corollary 1. *If (8) holds, then the procedure defined earlier in steps 1-4 satisfies FWER*$(\theta) \leq \alpha$ *for all* $\theta \in \Theta$.

Proof. It is not hard to see that the rejection function of the aforementioned procedure is given by

$$\rho(\mathcal{R}, \mathcal{A}, n) = \left\{H^{(j)} \in \mathcal{H} \setminus (\mathcal{R} \cup \mathcal{A}) : \quad \widetilde{T}_n^{(j)} \geq k - |\mathcal{R}|\right\}, \tag{12}$$

about which we verify (5) and (6). For (5), given $\mathcal{R}, \mathcal{R}', \mathcal{A}$ as described there and $n \in \mathcal{N}$, if $H^{(j)} \in \rho(\mathcal{R}, \mathcal{A}, n) \setminus \mathcal{R}'$, then $\widetilde{T}_n^{(j)} \geq k - |\mathcal{R}| \geq k - |\mathcal{R}'|$ because $\mathcal{R} \subseteq \mathcal{R}'$, hence, $H^{(j)} \in \rho(\mathcal{R}', \varnothing, n)$, so ρ satisfies (5). For (6), without loss of generality, assume that $\mathcal{T} \neq \varnothing$ because the following probability is zero otherwise. Let $V_j = \{\widetilde{T}_n^{(j)} \geq k - |\mathcal{F}| \text{ for some } n \in \mathcal{N}\}$. Using the Bonferroni inequality, (10) and (8),

$$P_\theta(\rho(\mathcal{F}, \varnothing, n) \nsubseteq \mathcal{F} \quad \text{for some } n \in \mathcal{N}) = P_\theta\left(\bigcup_{j:\, H^{(j)} \in \mathcal{T}} V_j\right) \leq \sum_{j:\, H^{(j)} \in \mathcal{T}} P_{\theta^{(j)}}(V_j)$$

$$= \sum_{j:\, H^{(j)} \in \mathcal{T}} P_{\theta^{(j)}}\left(T_n^{(j)} \geq B_{|\mathcal{F}|+1}^{(j)} \text{ for some } n \in \mathcal{N}\right) \leq \sum_{j:\, H^{(j)} \in \mathcal{T}} \frac{\alpha}{k - |\mathcal{F}|} = |\mathcal{T}| \cdot \frac{\alpha}{|\mathcal{T}|} = \alpha.$$

□

Tests that simultaneously control type I and II familywise error rates. Extending the procedure in the previous section, Bartroff & Song (2014) proposed a sequential test that simultaneously controls both the type I and II FWERs, the latter defined in the following in (13) analogously to the type I version (3). The error control of this procedure can also be seen as a special case of the rejection principle. Adding to the setup in Section 2, suppose one also has alternative hypotheses $G^{(1)}, \ldots, G^{(k)}$, such that $G^{(j)} \subseteq \Theta^{(j)}$ and $G^{(j)} \cap H^{(j)} = \varnothing$ for all $j = 1, \ldots, k$. With this, we redefine the false hypotheses from (2) to be

$$\mathcal{F}(\theta) = \{H^{(j)} \in \mathcal{H} : \theta^{(j)} \in G^{(j)}\}$$

and define the type II FWER as

$$\text{FWER}_{II}(\theta) = P_\theta(\text{any } H^{(j)} \in \mathcal{F}(\theta) \text{ accepted}). \tag{13}$$

Given desired FWER bounds α and β, the procedure requires only that each data stream $X_1^{(j)}, X_2^{(j)}, \ldots$ has a scalar-valued sequential test statistic $T_n^{(j)} = T_n^{(j)}(X_1^{(j)}, \ldots, X_n^{(j)})$ with critical values $A_1^{(j)}, \ldots, A_k^{(j)}, B_1^{(j)}, \ldots, B_k^{(j)}$, such that

$$P_{\theta^{(j)}}(T_n^{(j)} \geq B_s^{(j)} \text{ some } n,\ T_{n'}^{(j)} > A_1^{(j)} \text{ all } n' < n) \leq \frac{\alpha}{k - s + 1} \quad \text{for all} \quad \theta^{(j)} \in H^{(j)} \tag{14}$$

$$P_{\theta^{(j)}}(T_n^{(j)} \leq A_s^{(j)} \text{ some } n,\ T_{n'}^{(j)} < B_1^{(j)} \text{ all } n' < n) \leq \frac{\beta}{k - s + 1} \quad \text{for all} \quad \theta^{(j)} \in G^{(j)} \tag{15}$$

for all $j, s = 1, \ldots, k$. These inequalities simply guarantee that each sequential test marginally controls the conventional type I and II error probabilities at desired fractions of α, β.

For brevity, we do not restate Bartroff and Song's (2014) procedure here, but rather just say that it has a similar flavor to the one above but is more complex in that it interweaves rejections and acceptances of the $H^{(j)}$ at each stage. It also utilizes a standardizing function, mapping $B_s^{(j)}$ to $k - s + 1$ as earlier in (9) and mapping $A_s^{(j)}$ to $-(k - s + 1)$. The procedure controls the type I and II FWERs, regardless of dependence between the data streams, as long as (14)–(15) hold. This can be easily proved using the rejection principle, whose application here is interesting because it is used to prove control of type II FWER as well as type I. The proof proceeds by defining the procedure's *acceptance function* $\widetilde{\rho}(\mathcal{R}, \mathcal{A}, n)$, analogous to the rejection function ρ in Section 2, and these two are alternated to give the procedure's accept/reject decisions. Then Theorem 1 is applied to both ρ and $\widetilde{\rho}$ separately to prove type I and II FWER control, respectively. For this procedure, ρ takes the same form (12) and the acceptance function is similar,

$$\widetilde{\rho}(\mathcal{R}, \mathcal{A}, n) = \left\{H^{(j)} \in \mathcal{H} \setminus (\mathcal{R} \cup \mathcal{A}) : \quad \widetilde{T}_n^{(j)} \leq -(k - |\mathcal{A}|)\right\}.$$

3.2. Testing hypotheses with special structure

Whereas the previous sections assumed no special structure of the hypotheses being tested, in some settings, logical relationships or priorities exist among the hypotheses, which can be exploited by testing the hypotheses in a certain order and allowing less stringent (i.e. more powerful) tests to be used. In this section, we consider sequentially testing hypotheses in order (Rosenbaum, 2008), and later, the special case of sequentially testing closed hypotheses (Marcus *et al.*, 1976).

Testing hypotheses in order. In many multiple testing situations, it is natural to only test a certain hypothesis if certain other hypotheses have already been rejected; two real examples are given in Section 4. Rosenbaum (2008) considered various ordering schemes and gave fixed sample size tests that control the FWER. The most general ordering scheme Rosenbaum (2008) considers is the following, although his results apply to hypotheses and partitions with more general (e.g. infinite) index sets, whereas here, we simply consider hypotheses indexed by $\{1,\dots,k\}$ for coherence with the previous sections. Let $\mathcal{H}_1,\dots,\mathcal{H}_s$ be a partition of $\mathcal{H}$, such that it is desired to only test the hypotheses in $\mathcal{H}_i$ if all the hypotheses in $\bigcup_{i'<i}\mathcal{H}_{i'}$ have already been rejected. Recall that $\mathcal{H}_1,\dots,\mathcal{H}_s$ being a partition of $\mathcal{H}$ means that the $\mathcal{H}_i$ is disjoint and its union is $\mathcal{H}$. For $j=1,\dots,k$, let i_j denote the unique index i of the $\mathcal{H}_i$ containing $H^{(j)}$, that is, $H^{(j)}\in\mathcal{H}_{i_j}$. Rosenbaum (2008) calls a subset $\mathcal{H}'\subseteq\mathcal{H}$ *exclusive* if, at most, one hypothesis $H^{(j)}\in\mathcal{H}'$ is true, and $\mathcal{H}_1,\cdots,\mathcal{H}_s$ is *sequentially exclusive* if all the hypotheses in $\bigcup_{i'<i}\mathcal{H}_{i'}$ being false implies that $\mathcal{H}_i$ is exclusive, for all $i=1,\dots,s$. In the fixed sample size setting with valid p-values $p^{(1)},\dots,p^{(k)}$ for testing $H^{(1)},\dots,H^{(k)}$, respectively, Rosenbaum (2008, Proposition 3) shows that if $\mathcal{H}_1,\cdots,\mathcal{H}_s$ are sequentially exclusive, then the following test controls the FWER at level α: Reject $H^{(j)}$ if and only if $p^{(j)}\le\alpha$ and all hypotheses in $\bigcup_{i<i_j}\mathcal{H}_i$ have already been rejected. Rosenbaum's (2008) requirement that the $\mathcal{H}_i$ be 'intervals' is not needed if one does not require that the hypotheses $H^{(1)},\dots,H^{(k)}$ be strictly tested in the indexed order because the hypotheses can simply be re-indexed within each subset $\mathcal{H}_i$ to make it an interval.

Here, we present a sequential multiple testing procedure for testing hypotheses in order and use the rejection principle in Theorem 1 to prove its FWER control. We adopt the notation for data streams, parameters and hypotheses given in Section 2, and we assume that there is a sequentially exclusive partition $\mathcal{H}_1,\cdots,\mathcal{H}_s$ of $\mathcal{H}$ representing the desired order of testing. Given a desired FWER bound α and a set $\mathcal{N}$ of possible streamwise sample sizes, we also assume that, for each $j=1,\dots,k$, associated with the data stream $X_1^{(j)},X_2^{(j)},\dots$ and hypothesis $H^{(j)}$ is a scalar-valued sequential test statistic $T_n^{(j)}=T_n^{(j)}(X_1^{(j)},\dots,X_n^{(j)})$ with a critical value $B^{(j)}$ satisfying

$$P_{\theta^{(j)}}\left(T_n^{(j)}\ge B^{(j)}\quad\text{for some } n\in\mathcal{N}\right)\le\alpha\quad\text{for all}\quad\theta^{(j)}\in H^{(j)}.\tag{16}$$

Note that here, we only need a single critical value $B^{(j)}$ for each test statistic rather than the k critical values needed in the more general, unstructured setup in Section 3.1, which our exploitation of the sequential exclusivity property here will allow us to sidestep.

Let $\mathcal{I}_1=\{1,\dots,k\}$, $\ell_1=1$ and $n_0=0$. The ith stage ($i=1,\dots,k$) of the sequential procedure for testing hypotheses in order proceeds as follows.

(1) Sample each active data stream $\{X_n^{(j)}\}_{j\in\mathcal{I}_i}$ up to sample size

$$n_i=\inf\left\{n\in\mathcal{N}:n>n_{i-1}\quad\text{and}\quad T_n^{(j)}\ge B^{(j)}\quad\text{for some}\quad j:H^{(j)}\in\mathcal{H}_{\ell_i}\right\}.$$

(2) Reject $H^{(j)} \in \mathcal{H}_{\ell_i}$ if all hypotheses in $\bigcup_{\ell' < \ell_i} \mathcal{H}_{\ell'}$ have been rejected and $T_{n_i}^{(j)} \geq B^{(j)}$.
(3) If $i = k$ or $n_i = \max \mathcal{N}$, stop and accept all remaining hypotheses. Otherwise, set

$$\ell_{i+1} = \begin{cases} \ell_i + 1, & \text{if all hypotheses in } \mathcal{H}_{\ell_i} \text{ have been rejected,} \\ \ell_i, & \text{otherwise,} \end{cases}$$

$$\mathcal{I}_{i+1} = \left\{ j : H^{(j)} \in \bigcup_{\ell' \geq \ell_{i+1}} \mathcal{H}_{\ell'} \quad \text{and } H^{(j)} \text{ has not been rejected} \right\},$$

and continue on to stage $i + 1$.

The FWER control of this procedure is easily established using the rejection principle of Theorem 1.

Corollary 2. *If $\mathcal{H}_1, \ldots, \mathcal{H}_s$ is a sequentially exclusive partition of $\mathcal{H}$ and (16) holds, then the procedure defined earlier in steps 1-3 satisfies FWER$(\theta) \leq \alpha$ for all $\theta \in \Theta$.*

Proof. The rejection function is given by

$$\rho(\mathcal{R}, \mathcal{A}, n) = \left\{ H^{(j)} \in \mathcal{H} \setminus (\mathcal{R} \cup \mathcal{A}) : T_n^{(j)} \geq B^{(j)} \quad \text{and} \quad \mathcal{H}_\ell \subseteq \mathcal{R} \quad \text{for all} \quad \ell < i_j \right\},$$

about which we verify (5) and (6). For (5), with $\mathcal{R}, \mathcal{R}', \mathcal{A}$ as described there and $n \in \mathcal{N}$, if $H^{(j)} \in \rho(\mathcal{R}, \mathcal{A}, n) \setminus \mathcal{R}'$, then all conditions for $H^{(j)}$ to be in $\rho(\mathcal{R}', \varnothing, n)$ are satisfied, the latter because $\mathcal{H}_\ell \subseteq \mathcal{R} \subseteq \mathcal{R}'$ for all $\ell < i_j$. For (6), without loss of generality, assume $\mathcal{T} \neq \varnothing$ and let ℓ^* be the smallest index of a subset $\mathcal{H}_{\ell^*}$ containing a true hypothesis, that is, $\ell^* = \min\{\ell : \mathcal{H}_\ell \cap \mathcal{T} \neq \varnothing\}$, and let $H^{(j^*)}$ be an arbitrarily chosen but fixed hypothesis in $\mathcal{H}_{\ell^*} \cap \mathcal{T}$. On $V := \{\rho(\mathcal{F}, \varnothing, n) \not\subseteq \mathcal{F}$ for some $n \in \mathcal{N}\}$, there is some true $H^{(j)} \in \rho(\mathcal{F}, \varnothing, n)$ with $T_n^{(j)} \geq B^{(j)}$ and $\mathcal{H}_\ell \subseteq \mathcal{F}$ for all $\ell < i_j$. It follows from the latter that $i_j = \ell^*$ and by this and sequential exclusivity, $j = j^*$. Using these facts and (16), we have

$$P_\theta(V) \leq P_{\theta^{(j^*)}}\left(T_n^{(j^*)} \geq B^{(j^*)} \text{ for some } n \in \mathcal{N} \right) \leq \alpha,$$

showing that ρ satisfies (6). □

Closed Testing. A frequently encountered special case of testing hypotheses in order is closed testing. The set of hypotheses $\mathcal{H} = \{H^{(1)}, \ldots, H^{(k)}\}$ is *closed* if it is closed under intersection. Marcus *et al.* (1976) introduced a fixed sample size method of testing a closed set $\mathcal{H}$ that controls the FWER and only requires a level-α test of each intersection hypothesis $\bigcap_{j \in J} H^{(j)}$, $J \subseteq \{1, \ldots, k\}$. Beginning with the *global hypothesis* $\bigcap_{j=1}^k H^{(j)}$, their procedure tests the elements of $\mathcal{H}$ in order of decreasing *dimension* (the maximum number of $H^{(j)}$ being intersected), and $H \in \mathcal{H}$ is tested if and only if all elements of $\mathcal{H}$ contained in H have been rejected. Fixed sample size closed testing is a special case of Rosenbaum's (2008) testing in order formulation.

In the sequential realm, Tang & Geller (1999) gave a group sequential procedure for closed testing of hypotheses about multivariate normal data. A more general sequential procedure for closed testing can be derived using the rejection principle via the sequential procedure in Section 3.2 and Corollary 2. The relevant partition of $\mathcal{H}$ is the following, defined inductively for $i = 1, \ldots, k$:

$$\mathcal{H}_i = \left\{ H = \bigcap_{j \in J} H^{(j)} : |J| = k - i + 1, \quad H \notin \mathcal{H}_{i'} \quad \text{any } i' < i \right\}. \tag{17}$$

The subset $\mathcal{H}_i$ contains all hypotheses of dimension $k - i + 1$, as is guaranteed by the last condition in (17). For example, $\mathcal{H}_1$ contains only the global hypothesis, $\mathcal{H}_2$ contains all intersections of dimension $k - 1$, and so on. Applying the sequential procedure in Section 3.2 to this partition results in a sequential procedure that tests the hypotheses in order of decreasing dimension, using a level-α test for each, with sampling of the active data streams occurring between rejection decisions. After establishing that the partition (17) is sequentially exclusive, it follows immediately from Corollary 2 that this procedure controls the FWER.

Corollary 3. *If (16) holds and $\mathcal{H}$ is closed, then the partition (17) is sequentially exclusive, hence, the procedure defined in steps 1-3 of Section 3.2 applied to (17) satisfies $FWER(\theta) \le \alpha$ for all $\theta \in \Theta$.*

Proof. To establish sequential exclusivity, suppose there are distinct hypotheses $H, H' \in \mathcal{H}_i$ that are both true, that is, $\theta \in H$ and $\theta \in H'$. Then $\theta \in H \cap H'$, so $H \cap H'$ is true, and $H \cap H' \in \mathcal{H}_{i'}$ for some $i' < i$ by virtue of H, H' being distinct. □

A similar but distinct formulation of sequentially testing closed hypotheses is given in Bartroff & Lai (2010, Theorem 2.2), which does not explicitly force test in order of decreasing dimension but rather gives a sufficient condition on the test statistics under which the procedure in Section 3.1 would test in this order anyway.

4. Examples

In this section, we give two examples of the sequential procedures in Section 3.2 applied to real testing situations. In Section 4.1, we apply the sequential procedure for testing in order to an observational study involving chromosome aberration data, and in Section 4.2, we apply the sequential closed testing procedure to estimate the maximum safe dose of a treatment. In both cases, the performance of the sequential procedure is compared with the corresponding fixed sample size procedure, and the efficiency gain in terms of savings in average sample size of the sequential procedure is highlighted.

4.1. Chromosome aberrations of patients exposed to anti-tuberculosis drugs

In non-randomized testing situations, such as observational studies, it is common for treatment responses to be compared with more than one 'control' response, such as baseline and non-treatment, because this can provide information on differences due to non-random treatment assignment (e.g. see Rosenbaum, 2002, Section 8).

Masjedi *et al.* (2000) studied possible mutagenic effects of anti-tuberculosis drugs by comparing the frequency of chromosome aberrations including gaps per 100 cells in an observational study of $n = 36$ patients before (denoted b) and after (denoted a) treatment and 36 healthy controls (denoted c), who matched the treatment group by sex and age and were selected from relatives of the treatment group when possible. The response triples (y_{ci}, y_{bi}, y_{ai}), $i = 1, \ldots, n$, are given in Table 1, where larger numbers indicate more chromosome damage.

Rosenbaum (2008) considers the model

$$y_{ai} = \mu_a + \pi_i + \lambda_i + \varepsilon_i,$$

$$y_{bi} = \mu_b + \pi_i + \lambda_i + \zeta_i,$$

$$y_{ci} = \mu_c + \pi_i + \eta_i,$$

Table 1. *Masjedi et al. (2000) data on total chromosome aberrations per 100 cells including gaps*

Control y_{ci}	Before treatment y_{bi}	After treatment y_{ai}
1.00	0.50	3.00
1.50	4.50	5.50
0.50	3.50	5.00
0.50	2.66	3.33
0.66	1.50	4.50
1.00	5.00	7.00
1.00	1.33	5.33
0.66	1.50	2.50
0.00	2.00	5.33
1.33	1.50	3.00
1.50	1.33	3.33
2.00	2.00	2.00
1.33	2.00	4.66
0.00	2.66	10.00
3.00	1.33	3.33
0.50	3.50	5.00
0.66	3.00	5.00
1.33	2.66	3.33
3.00	0.00	4.00
0.66	1.50	7.00
0.50	1.00	3.00
0.66	4.00	4.00
2.00	1.33	2.66
1.33	0.66	3.33
0.00	1.50	3.50
1.00	0.66	2.00
0.50	2.00	3.33
1.33	1.00	3.50
0.50	1.33	2.66
1.00	2.00	2.00
0.66	2.00	4.00
1.50	1.50	1.50
2.66	2.00	3.50
1.33	0.66	3.33
0.66	0.00	2.66
1.00	1.50	1.50

$i = 1, \ldots, n$, where π_i, λ_i, ε_i, ζ_i and η_i are independent with continuous distributions F_π, F_λ, F_ε, F_ζ and F_η, assumed to be symmetric about zero. Here, μ represents the group effect; the random variables π_i and λ_i reflect the correlation due to matching and treatment versus control, respectively; and ε_i, ζ_i and η_i are error terms. The primary scientific question was to ascertain whether the post-treatment responses exceed the baseline and control responses by more than the baseline and control responses differ from each other, that is,

$$\mu_a - \max(\mu_b, \mu_c) > \max(\mu_b, \mu_c) - \min(\mu_b, \mu_c),$$

hence, the null hypothesis of primary interest is the negation of this,

$$H_0 : \mu_a - \max(\mu_b, \mu_c) \leq \max(\mu_b, \mu_c) - \min(\mu_b, \mu_c).$$

However, even if H_0 cannot be rejected, it may be beneficial to be able to draw some weaker conclusions about the treatment. Namely, Rosenbaum (2008) defines the five additional hypotheses as follows:

$$\begin{aligned}
&H_+ : \mu_a \le (\mu_b + \mu_c)/2,\\
&H_b : \mu_a \le \mu_b,\\
&H_c : \mu_a \le \mu_c,\\
&H_* : \mu_a - \mu_c \le \mu_c - \mu_b,\\
&H_\sharp : \mu_a - \mu_b \le \mu_b - \mu_c.
\end{aligned}$$

The logical implications of these hypotheses suggest the sequentially exclusive partition $\mathcal{H}_1 = \{H_+\}$, $\mathcal{H}_2 = \{H_b, H_c\}$, $\mathcal{H}_3 = \{H_*, H_\sharp\}$ and $\mathcal{H}_4 = \{H_0\}$ of $\mathcal{H} = \{H_0, H_+, H_b, H_c, H_*, H_\sharp\}$, and Fig. 1 describes the procedure for testing these hypotheses in order, which applies to both the fixed sample size procedure of Rosenbaum (2008) as well as the sequential procedure defined earlier in which sampling may occur between decisions, and Corollary 2 shows that the FWER is controlled.

Because the Masjedi *et al.* (2000) data (y_{ai}, y_{bi}, y_{ci}) exhibit strong non-normality, Rosenbaum (2008) applied the one-sided version of Wilcoxon's signed-rank test to obtain the fixed sample size p-values, using $y_{ai} - (y_{bi} + y_{ci})/2$ to test H_+, $y_{ai} - y_{bi}$ to test H_b, and so on. Following this approach, we apply a sequential version of these tests in which the sequential test statistics $T_n^{(j)}$ $(j = 1, \ldots, 6)$ are taken to be the repeatedly calculated p-values for Wilcoxon's test, with the common critical value $B = B^{(1)} = \ldots = B^{(6)}$ adjusted to control the marginal type I error probability (16) at $\alpha = 0.05$, and was determined by the Monte Carlo method. These sequential tests were then applied to the Masjedi et al. data in Table 1, after setting aside the five tied observations for a total of 31 triples. To assess the performance of this sequential test, Monte Carlo simulations were performed in which the order of the triples were permuted 50 000 times, and the sample size needed to reach an accept/reject decision on all hypotheses was recorded each time. In order to see the effect of different sequential sampling schemes, this was carried out for the five different choices of sample size sets $\mathcal{N}$ given in Table 2: fully

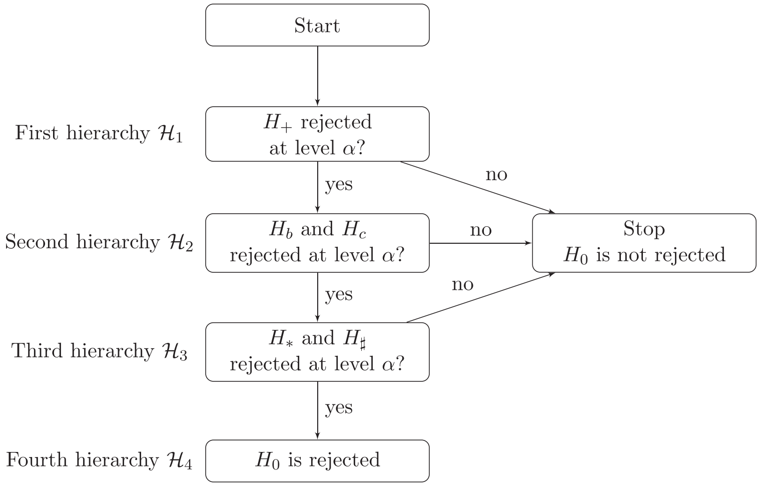
Start
First hierarchy $\mathcal{H}_1$
H_+ rejected at level α?
yes
no
Second hierarchy $\mathcal{H}_2$
H_b and H_c rejected at level α?
no
Stop
H_0 is not rejected
yes
no
Third hierarchy $\mathcal{H}_3$
H_* and $H_\sharp$ rejected at level α?
yes
Fourth hierarchy $\mathcal{H}_4$
H_0 is rejected

Fig. 1. Flow chart of the testing procedure for the chromosome aberration example.

Table 2. *The sample size set $\mathcal{N}$, critical value $\boldsymbol{B}$ and average sample size of various Wilcoxon signed-rank tests for testing hypotheses about the chromosome aberration data*

$\mathcal{N}$	$\boldsymbol{B}$	Average sample size
$\{1,\ldots,31\}$	0.0031	18.9
$\{10,15,20,25,31\}$	0.0068	18.9
$\{10,20,25,31\}$	0.0082	20.1
$\{10,20,31\}$	0.0098	20.8
$\{15,31\}$	0.0140	21.3

sequential sampling in which $\mathcal{N} = \{1,\ldots,31\}$ and four group-sequential schemes with five, four, three and two groups, respectively, with evenly sized groups (until the final group, which is one larger). Table 2 also reports the common critical value B and the average total sample size over the 50 000 simulated paths for each scheme. In all cases, the average sample size was dramatically reduced from 31. Even in the worst case of two-stage sampling, with $\mathcal{N} = \{15, 31\}$, more than nine observations were saved on average, nearly one-third of the largest possible sample size, 31. Interestingly, the five-group scheme with $\mathcal{N} = \{10, 15, 20, 25, 31\}$ has nearly the same average sample size as the fully sequential scheme, which may be appealing in applications where fully sequential testing is not practical.

4.2. Identifying the maximum safe dose in toxicological studies

Tamhane *et al.* (2001) describe a novel multiple testing approach to determine the maximum safe dose (MAXSD) of crop protection products such as pesticides, herbicides and fungicides, which are tested for safety on non-target species and for which the multiple testing error control guarantees a prescribed bound on recommending an unsafely high dose; their approach is equally applicable to clinical trials for safety with human subjects. In this section, we apply the rejection principle developed earlier to derive a sequential version of this procedure.

Assume there are k discrete non-zero dose levels, which we label $1,\ldots,k$ in increasing order, and include the level 0 to denote 'no treatment'. Following Tamhane *et al.* (2001; Section 3), we adopt an ANOVA setup in which there are $k+1$ groups of subjects, each treated at one of these dose levels. Let y_{ij} $(i = 1,\ldots,n_i,\ j = 0,\ldots,k)$ denote the response of the ith subject in the jth group and let $\mu_j = E(y_{ij})$ $(j = 0,\ldots,k)$. Large values of y_{ij} indicate safety of the treatment. For example, in the crop protection setting mentioned earlier, y_{ij} may be the growth of a non-target organism when exposed to the potentially toxic treatment. Define the hypotheses

$$H^{(j)} : \mu_j \le \lambda\mu_0 \quad \text{versus} \quad G^{(j)} : \mu_j > \lambda\mu_0, \quad j = 1,\ldots,k,$$

in which $\lambda \in (0, 1]$ is a fixed, agreed-upon response threshold for safety and, therefore, the null hypothesis $H^{(j)}$ means that the jth dose is unsafe; the MAXSD is defined as the largest $j \in \{0,\ldots,k\}$, such that $H^{(j)}$ is false. Tamhane *et al.* (2001) propose a multiple testing approach to encode the natural ordering $\mu_0 \ge \mu_1 \ge \ldots \ge \mu_k$, as follows. Replacing $H^{(j)}$ by

$$\widetilde{H}^{(j)} = \bigcap_{j'=j}^{k} H^{(j')},$$

we see that $\widetilde{H}^{(1)}, \ldots, \widetilde{H}^{(k)}$ form a closed family of hypotheses, and the test for closed hypotheses in Section 3.2 and Corollary 3 can be used to test these hypotheses sequentially. Because this procedure tests the hypotheses in order of decreasing dimension, as discussed in Section 3.2, the hypotheses will be tested in the order $\widetilde{H}^{(k)}, \widetilde{H}^{(k-1)}, \ldots, \widetilde{H}^{(1)}$. Here, we focus on sequential control of only type I FWER but, alternatively, simultaneous control of type I and II FWERs could be considered using the test of Bartroff & Song (2014) discussed in Section 3.1.

The following simulation study was performed to explore the operating characteristics of the sequential testing procedure. Setting $k = 4$ and taking y_{ij} to be i.i.d. $N(\mu_i, 1)$ observations, the mean responses μ_i were chosen to be those in the second column of Table 3, making the true MAXSD 1, indicated by an asterisk in the table. This choice of mean responses, with $\mu_1 = 0$, represents the commonly encountered but confounding testing situation in which the smallest non-zero dose actually not only has mean response zero but is also the correct dose; we further note that recommending dose level 0 as the MAXSD, although it has the same mean response as dose level 1, has much different and possibly dangerous implications for the utilization of the MAXSD in future scientific work. An $\alpha = 0.05$ version of both the sequential test in Section 3.2 and the fixed sample size version was implemented, and Table 3 gives the estimated average sample size (denoted Avg. SS) of each group and the probability (denoted by $P(\text{MAXSD} = j)$) of choosing each dose level as the estimated MAXSD for both the sequential and fixed sample size procedures, based on 50 000 Monte Carlo simulated data sets. For both of these tests, standard two-sample t-tests were used as the individual test statistics to test the $H^{(j)}$, with λ taken to be 1 and the critical values were determined by Monte Carlo in the sequential case. The maximum sample size of the sequential procedure was 50, to mirror the sample size of the fixed sample size procedure. At the three dose levels $j = 2, 3, 4$ exceeding the MAXSD, the sequential procedure only required on average 28.6, 8.9 and 2.6 observations, respectively, which is a dramatic reduction from the fixed sample size procedure that used 50 observations at each of these dose levels. While dosing subjects at levels above the MAXSD may not be a concern in some studies involving plants, it would be of chief concern in clinical trials with human patients who are likely to experience toxicity, or even possibly death, at those levels. Both the sequential and fixed sample size procedures correctly identified the true MAXSD more than 80 *per cent* of the time, with the sequential procedure less likely to identify the true MAXSD than the fixed sample size procedure. On the other hand, the fixed sample size procedure was more likely to underestimate the MAXSD (4.70 *per cent*) compared with the sequential procedure (1.57 *per cent*), which is also undesirable but for different reasons, such as an ineffective crop protection plan being implemented or sick human patients receiving ineffective treatment. The last line of the table contains a weighted average estimated MAXSD for both tests.

Table 3. *Performance of the sequential and fixed sample size procedures for identifying the maximum safe dose*

		Sequential		Fixed sample	
Level = j	μ_j	Avg. SS	$P(\text{MAXSD} = j)$	Avg. SS	$P(\text{MAXSD} = j)$
0	0	50.0	1.57%	50	4.70%
1*	0	49.7	82.43%	50	89.68%
2	0.5	28.6	16.00%	50	5.62%
3	1.0	8.9	0%	50	0%
4	2.0	2.6	0%	50	0%
Weighed average MAXSD		1.1		1.0	

MAXSD, maximum safe dose; Avg. SS, average sample size.

5. Discussion

We have given sufficient conditions, in the form of (5)–(6), for a sequential procedure to control the FWER and have shown that they can be applied to testing situations where a special structure is assumed or not assumed. In addition to the rejection principle's utility in deriving new sequential procedures, it also provides a unified view of sequential FWER control. Although it remains an open question whether this rejection principle is also a necessary condition for FWER control, in any case, the principle may be useful for developing an optimality theory of sequential multiple testing procedures, in which little is known, even in the case of independent data streams. It seems likely that finding the most sequentially efficient procedure satisfying (5)–(6) may be more attainable than finding the best sequential procedure controlling the FWER.

We have not gone into detail about applying the general sequential procedures discussed in Section 3.1, and we refer interested readers to the respective references (Bartroff & Lai, 2010; Bartroff & Song, 2014) for details. We will say here that both of these procedures can handle testing, in a given data stream j, the commonly encountered hypotheses of the form

$$H^{(j)} : \theta^{(j)} = \theta_0^{(j)} \quad \text{versus} \quad G^{(j)} : \theta^{(j)} \neq \theta_0^{(j)}, \tag{18}$$

where $\theta^{(j)}$ is a possibly vector-valued parameter and $\theta_0^{(j)}$ is some fixed value of interest. For example, in a parametric setup, sequential generalized log-likelihood ratio statistics can be used and signed-root normal approximations (Jennison & Turnbull, 1997) or Monte Carlo can be used to compute critical values; see Bartroff & Song (2014) for details. Which of the procedures to use in practice may depend on the application. The first procedure in Section 3.1 does not explicitly specify an acceptance rule (although, as mentioned there, acceptances can be incorporated without violating the FWER control), and it therefore may be more appropriate in situations where 'early stopping' is not as high a priority under the null $H^{(j)}$ as under $G^{(j)}$, such as when $H^{(j)}$ represents a drug being safe or a process being 'in control'. On the other hand, if early stopping is desirable under both $H^{(j)}$ and $G^{(j)}$, then the procedure of Bartroff & Song (2014) discussed second in Section 3.1 that controls both the type I and II FWERs may be more appropriate. In order to control the type II FWER, this procedure naturally requires control of the marginal type II error rate in the form of (15), which may not be possible with $G^{(j)}$ as written in (18), because values of $\theta^{(j)} \in G^{(j)}$ arbitrarily close to $\theta_0^{(j)}$ can make bounds like (15) impossible for any test. This can be remedied by constructing a surrogate alternative hypothesis $\widetilde{G}^{(j)}$ under which type II error control is possible; for example, $\widetilde{G}^{(j)} : ||\theta^{(j)} - \theta_0^{(j)}|| \geq \delta$ for some $\delta > 0$ and norm $|| \cdot ||$. Here, δ may represent the minimum significant separation of the parameter, or similar, and be well motivated by the domain of application. On the other hand, the statistician could treat δ as a parameter to choose before testing in order to attain a sequential procedure with desirable operating characteristics, such as expected sample size. A similar analysis pertains to null hypotheses of the form $H^{(j)} : \theta^{(j)} \leq \theta_0^{(j)}$ for scalar-valued parameter $\theta^{(j)}$ and, more generally, of the form $H^{(j)} : u(\theta^{(j)}) \leq u_0^{(j)}$ for vector-valued $\theta^{(j)} \in \mathbb{R}^d$ and given smooth function $u : \mathbb{R}^d \to \mathbb{R}$ and fixed scalar value u_0; see Bartroff & Lai (2008a), Bartroff & Lai (2008b) and Bartroff *et al.* (2013) for examples of sequential generalized likelihood ratio tests for these situations.

In addition to the aforementioned optimality theory, another area of further work is to generalize the sequential sampling schemes. Earlier, we assumed that the set of possible streamwise sample sizes $\mathcal{N}$ is fixed in advance, but an alternative approach is to incorporate an efficient adaptive scheme wherein the next sampling increment can be chosen as a function of the data, making the resulting procedures 'adaptive' in yet another sense. Adaptive sequential sampling schemes for hypothesis testing have been considered by many authors including Jennison & Turnbull (2006) and Bartroff (2006a, 2006b, 2007). Incorporating adaptive sampling schemes such as these into the multiple testing procedures is an exciting area of future research.

Acknowledgements

Bartroff's work was partially supported by grant DMS-1310127 from the National Science Foundation and grant R01 GM068968 from the National Institutes of Health.

References

Baillie, D. (1987). Multivariate acceptance sampling – some applications to defense procurement. *J. Roy. Stat. Soc. D-Sta.* **36**, (5), 465–478.

Bartroff, J. (2006a). Efficient three-stage t-tests. In *Recent developments in nonparametric inference and probability: Festschrift for Michael Woodroofe*, vol. 50, IMS Lecture Notes Monograph Series. Institute of Mathematical Statistics, Hayward, CA; 105–111.

Bartroff, J. (2006b). Optimal multistage sampling in a boundary-crossing problem. *Seq. Anal.* **25**, 59–84.

Bartroff, J. (2007). Asymptotically optimal multistage tests of simple hypotheses. *Ann. Stat.* **35**, 2075–2105.

Bartroff, J. & Lai, T. L. (2008a). Efficient adaptive designs with mid-course sample size adjustment in clinical trials. *Stat. Med.* **27**, 1593–1611.

Bartroff, J. & Lai, T. L. (2008b). Generalized likelihood ratio statistics and uncertainty adjustments in adaptive design of clinical trials. *Seq. Anal.* **27**, 254–276.

Bartroff, J. & Lai, T. L. (2010). Multistage tests of multiple hypotheses. *Communications in Statistics – Theory and Methods* **39**, 1597–1607.

Bartroff, J., Lai, T. L. & Shih, M. (2013). *Sequential experimentation in clinical trials: design and analysis*, Springer, New York.

Bartroff, J. & Song, J. (2014). Sequential tests of multiple hypotheses controlling type I and II familywise error rates. *J. Stat. Plan. Infer.* **153**, 100–114.

Cook, R. & Farewell, V. (1994). Guidelines for monitoring efficacy and toxicity responses in clinical trials. *Biometrics* **50**, 1146–1152.

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

De, S. & Baron, M. (2012b). Step-up and step-down methods for testing multiple hypotheses in sequential experiments. *J. Stat. Plan. Infer.* **142**, 2059–2070.

Fisher, S. (1932). *Statistical methods for research workers*, Oliver and Boyd, Edinburgh.

Goeman, J. & Solari, A. (2010). The sequential rejection principle of familywise error control. *Ann. Stat.* **38**, (6), 3782–3810.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70.

Hommel, G., Bretz, F. & Maurer, W. (2007). Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies. *Stat. Med.* **26**, (22), 4063–4073.

Jennison, C. & Turnbull, B. (1993). Group sequential tests for bivariate response: interim analyses of clinical trials with both efficacy and safety endpoints. *Biometrics* **49**, 741–752.

Jennison, C. & Turnbull, B. W. (1997). Group sequential analysis incorporating covariate information. *J. Am. Stat. Assoc.* **92**, 1330–1341.

Jennison, C. & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*, Chapman & Hall/CRC, New York.

Jennison, C. & Turnbull, B. W. (2006). Adaptive and nonadaptive group sequential tests. *Biometrika* **93**, 1–21.

Marcus, R., Peritz, E. & Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.

Masjedi, M. R., Heidary, A., Mohammadi, F., Velayati, A. A. & Dokouhaki, P. (2000). Chromosomal aberrations and micronuclei in lymphocytes of patients before and after exposure to anti-tuberculosis drugs. *Mutagenesis* **15**, 489–494.

Mei, Y. (2010). Efficient scalable schemes for monitoring a large number of data streams. *Biometrika* **97**, (2), 419–433.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079–1087.

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

Paulson, E. (1964). A sequential procedure for selecting the population with the largest mean from k normal populations. *Ann. Math. Stat.* **35**, 174–180.

Romano, J. P. & Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Am. Stat. Assoc.* **100**, (469), 94–108.

Rosenbaum, P. (2002). *Observational studies*, Springer, New York.

Rosenbaum, P. (2008). Testing hypotheses in order. *Biometrika* **95**, (1), 248–252.

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

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika* **40**, 87–110.

Seber, G. A. F. & Lee, A. J. (2003). *Linear regression analysis*, (second edition)., Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

Siegmund, D. (1985). *Sequential analysis: tests and confidence intervals*, Springer-Verlag, New York.

Siegmund, D. (1993). A sequential clinical trial for comparing three treatments. *Ann. Stat.* **21**, (1), 464–483.

Tamhane, A., Dunnett, C., Green, J. & Wetherington, J. (2001). Multiple test procedures for identifying the maximum safe dose. *J. Am. Stat. Assoc.* **96**, (455), 835–843.

Tang, D.-I. & Geller, N. L. (1999). Closed testing procedures for group sequential clinical trials with multiple endpoints. *Biometrics* **55**, 1188–1192.

Tang, D.-I., Geller, N. L. & Pocock, S. J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics* **49**, 23–30.

Tang, D.-I., Gnecco, C. & Geller, N. L. (1989). Design of group sequential clinical trials with multiple endpoints. *J. Am. Stat. Assoc.* **84**, 776–779.

Tartakovsky, A., Li, X. & Yaralov, G. (2003). Sequential detection of targets in multichannel systems. *IEEE T. Inform. Theory.* **49**, (2), 425–445.

Wald, A. (1947). *Sequential analysis*, Wiley, New York. Reprinted by Dover.

Ye, Y., Li, A., Liu, L. & Yao, B. (2013). A group sequential Holm procedure with multiple primary endpoints. *Stat. Med.* **32**, 1112–1124.

Received April 2014, in final form March 2015

Jay Bartroff, Department of Mathematics, 3620 South Vermont Avenue, KAP 104, University of Southern California, Los Angeles, CA 90089, USA.
E-mail: bartroff@usc.edu