Sequential tests controlling generalized familywise error rates

Shyamal K. De[a], Michael Baron[b,c,*]

[a]*School of Mathematical Sciences, National Institute of Science Education and Research, Bhubaneswar, OD, India*
[b]*Department of Mathematics and Statistics, American University, Washington, DC, USA*
[c]*Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX, USA*

Abstract

Sequential methods are developed for conducting a large number of simultaneous tests while controlling the Type I and Type II *generalized* familywise error rates. Namely, for the chosen values of α, β, k, and m, we derive simultaneous tests of d individual hypotheses, based on sequentially collected data, that keep the probability of at least k Type I errors not exceeding level α and the probability of at least m Type II errors no greater than β. This generalization of the classical notions of familywise error rates allows substantial reduction of the expected sample size of the multiple testing procedure.

Keywords: Generalized familywise error rate, Holm-Bonferroni procedure, Likelihood ratio, Stopping boundary, Stopping time

2010 MSC: 62L, 62F03, 62H15.

1. Introduction

A large number of sequential statistical experiments are designed to answer many questions, that is, to test a set of hypotheses. Moreover, an answer is needed for each question, and thus, each individual hypothesis has to be tested instead of one composite hypothesis. Such problems arise in clinical trials for testing multiple efficacy and safety endpoints

*Corresponding author
Email addresses: sde@niser.ac.in (Shyamal K. De), baron@american.edu,mbaron@utdallas.edu (Michael Baron)

[11, 14, 26, 27], DNA and protein sequence analysis [24, 30], epidemiology [10], cybersecurity [20], and so on.

For fixed-size samples, the methodology of testing multiple hypotheses is very well developed over the last two decades or so. Efficient procedures have been proposed to control the familywise error rate or the false discovery rate, see e.g., [2, 8], or [7] for the overview or [5] for the bibliography.

For sequentially collected data, a few matching methods have recently been proposed for testing multiple hypotheses. An adaptive multistage step-down procedure proposed in [1] controls the *Type I familywise error rate*, defined as the probability of rejecting at least one true hypothesis. Generalizing the concept of Wald's sequential probability ratio test (SPRT) to multiple hypotheses, [3] develops a testing procedure that controls both Type I and Type II familywise error rates in the strong sense. By analogy, the latter is defined as the probability of accepting at least one false null hypothesis. Control of both error rates appears possible due to the flexibility of sequential designs, similarly to the single-hypothesis SPRT attaining both desired probabilities of Type I and Type II errors. A modification of this sequential procedure is proposed in [3], combining the ideas of Wald's SPRT and Holm-type stepwise testing. Improving the plain Bonferroni methods, this new algorithm requires a smaller expected sample size, reducing the overall expected costs of the experiment and at the same time controlling both familywise error rates.

Here and in the sequel, by the *sample size* we understand the *number of sampled units*, such as patients, computer parts, etc. We assume that each sampled unit i contributes to the total cost of an experiment regardless of how many components X_{ij} (such as vital signs of patients or electronic measurements of manufactured parts) are recorded on unit i. This is quite common in many experiments (e.g., [3, 11, 15, 26]). For example, in clinical trials,

certain amount is budgeted for each participating patient, covering the cost of a treatment, service, insurance, incentive, and possibly, accommodation and transportation. However, once a patient participates in the trial, the individual measurements such as items in a written or oral questionnaire, blood work, or other analysis, usually require an incomparably lower additional cost, if any at all.

Thus, we consider the cost function that is proportional to or monotonically dependent on *the number of sampling units*. It is to be distinguished from *the total number of recorded measurements* X_{ij} that is used in [1] as a *sample size* (and misinterpreted in [3, 4]).

Sampling strategies should be different under these two cost functions. Under a cost function *per sampling unit*, all measurements X_{ij} will be recorded for all the sampled units. However, if a cost *per measurement* is considered, recording the j-th component will probably be terminated once the answer to the corresponding j-th test is obtained. The difference may be quite substantial when one of the tests requires a much larger sample size than the others.

It has been noted that the strong control of familywise error rates is an overly stringent condition in practical situations where the number of tested null hypotheses is large, such as hundreds or thousands. Examples are found in biology, genomics, computer science, communications, and many other areas, see e.g., [6, 16, 18, 24], and many examples in [7] and [9]. Indeed, a few erroneous decisions among a large number of tests, false rejections or missed discoveries, can be tolerated. Studies show that a slight relaxation of FWER related constraints can result in a significant reduction of the required sample size.

For these reasons, [19] introduced a concept of *generalized familywise error rates* or k-FWER which is the probability of rejecting at least k true null hypotheses. Controlling k-FWER at the given desired level is a weaker constraint (for $k \geq 2$) than controlling the

standard FWER, and therefore, this condition can be satisfied by a smaller sample. Several non-sequential testing procedures controlling k-FWER were developed in [12, 21, 22, 28].

In this article, we construct sequential multiple testing procedures that control the generalized familywise error rate. Inheriting the control of both Type I and Type II errors from the original Wald's SPRT and the sequential multiple testing procedures, the proposed schemes control both *Type I k-FWER* and *Type II m-FWER*. These sequential tests are constructed by a suitable modification of sequential procedures of [4].

The concepts are formalized in the next section, and the sequential testing procedure controlling Type I and Type II FWER is reviewed. Sections 2 and 3 contain two modifications of this sequential test that are based on different methods of controlling the generalized FWER, reducing the required sample size in both cases. These two approaches are compared in Section 4.

Proposed tests are applicable to simple-vs-simple hypotheses of type (1) or one-sided tests in the case of a monotone likelihood ratio (2). What our methods do not cover are situations with nuisance parameters such as t-tests, and this is a subject of future work.

2. The Intersection scheme and control of familywise error rates I and II

Stepwise testing method of [13] is based on the ordered p-values. Comparing with the plain Bonferroni adjustment for multiplicity, Holm's method yields higher power, and it requires a smaller sample size in order to attain the given power. Applying Holm's concept to sequential experiments, [4] developed a sequential stepwise procedure for multiple testing that attains simultaneous control of familywise error rates of Types I and II.

Consider a sequence of independent vector-valued observations $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \in \mathbb{R}^d$. Each observed random vector $\boldsymbol{X}_n = \left(X_n^{(1)}, \ldots, X_n^{(d)}\right) \in \mathbb{R}^d$ represents d measurements, possibly and likely to be dependent, on a sampling unit n. This may be a questionnaire or

results of d medical tests for the n-th patient during a sequential clinical trial or a set of d markers in genome scans. We assume that the j^{th} components of $\boldsymbol{X}_n$ has a marginal density (pmf or pdf) $f_j(x|\theta^{(j)})$ with respect to a reference measure μ_j, $j = 1, \ldots, d$, parameterized by a scalar or vector parameter $\theta^{(j)}$. Suppose that the goal of this sequential experiment is to conduct a battery of tests on parameters $\theta^{(1)}, \ldots, \theta^{(d)}$,

$$H_0^{(j)} : \theta^{(j)} = \theta_0^{(j)} \text{ vs. } H_A^{(j)} : \theta^{(j)} = \theta_1^{(j)}, \text{ for } j = 1, \ldots, d. \tag{1}$$

For those components j, where $\theta^{(j)} \in \mathbb{R}$ and $f_j(x|\theta^{(j)})$ have monotone likelihood ratio property, (1) is equivalent to testing

$$H_0^{(j)} : \theta^{(j)} \leq \theta_0^{(j)} \text{ vs. } H_A^{(j)} : \theta^{(j)} \geq \theta_1^{(j)}, \text{ for } j = 1, \ldots, d, \tag{2}$$

where $\theta_0^{(j)} < \theta_1^{(j)}$. These tests are required to control the *Type I and Type II familywise error rates* in the strong sense at the assigned levels α and β. That is, we need

$$\textit{FWER}_I = \max_{\substack{(\theta^{(1)},\ldots,\theta^{(d)}) \\ \exists j \,:\, H_0^{(j)} \text{is true}}} \boldsymbol{P}\left\{ \bigcup_{\substack{j=1,\ldots,d \\ H_0^{(j)} \text{is true}}} \text{reject } H_0^{(j)} \right\} \leq \alpha; \tag{3}$$

$$\textit{FWER}_{II} = \max_{\substack{(\theta^{(1)},\ldots,\theta^{(d)}) \\ \exists j \,:\, H_0^{(j)} \text{is false}}} \boldsymbol{P}\left\{ \bigcup_{\substack{j=1,\ldots,d \\ H_0^{(j)} \text{is false}}} \text{accept } H_0^{(j)} \right\} \leq \beta. \tag{4}$$

In other words, we control the probability of committing at least one Type I error for any non-empty combination of true null hypotheses, and similarly, the probability of committing at least one Type II error for any non-empty combination of false null hypotheses. Automatically, the *familywise power* is controlled at the level $(1 - \beta)$,

$$\textit{FWP} = \max_{\substack{(\theta^{(1)},\ldots,\theta^{(j)}) \\ \exists j \,:\, H_0^{(j)} \text{is false}}} \boldsymbol{P}\left\{ \bigcap_{\substack{j=1,\ldots,d \\ H_0^{(j)} \text{is false}}} \text{reject } H_0^{(j)} \right\} \geq 1 - \beta.$$

This is one of several definitions of familywise power found in the literature on multiple comparisons.

A battery of tests satisfying (3,4) can easily be obtained by the Bonferroni or generalized Bonferroni methods where the individual hypothesis $H_0^{(j)}$ is tested at the nominal significance level α_j and power at least $(1-\beta_j)$ with $\sum \alpha_j \leq \alpha$ and $\sum \beta_j \leq \beta$ ([3]). Although one can attempt to optimize the choice of α_j and β_j, Bonferroni's inequality is known to be rather crude for a large number of hypotheses d. Inequalities (3,4) will be satisfied with a wide margin in this case, and as a result, this multiple testing procedure will require an unnecessarily large sample size.

As an improvement of Bonferroni procedure, [4] combined the Holm-type stepwise testing principle with Wald's sequential probability ratio testing of individual hypotheses and derived a sequential scheme satisfying (3,4) with a uniformly lower expected sample size than the Bonferroni procedure. The algorithm is based on log-likelihood ratio statistics

$$\Lambda_n^{(j)} = \log \prod_{k=1}^{n} \frac{f_j(X_k^{(j)} \mid \theta_1^{(j)})}{f_j(X_k^{(j)} \mid \theta_0^{(j)})},$$

arranged in their non-increasing order,

$$\Lambda_n^{[1]} \geq \Lambda_n^{[2]} \geq \ldots \geq \Lambda_n^{[d]}.$$

The stopping rule

$$T_{1,1} = \inf \left\{ n \,:\, \bigcap_{j=1}^{d} \Lambda_n^{[j]} \notin (b_j, a_j) \right\} \tag{5}$$

with stopping boundaries

$$a_j = \log \frac{d-j+1}{\alpha}, \quad b_j = \log \frac{\beta}{j}, \quad \text{for } j = 1, \ldots, d, \tag{6}$$

followed by the decision rule rejecting all the null hypotheses corresponding to $\Lambda_{T_{1,1}}^{[j]} \geq a_j$ and accepting the ones corresponding to $\Lambda_{T_{1,1}}^{[j]} \leq b_j$ is shown to satisfy (3,4) controlling both familywise error rates in the strong sense ([4] and Fig. 1).

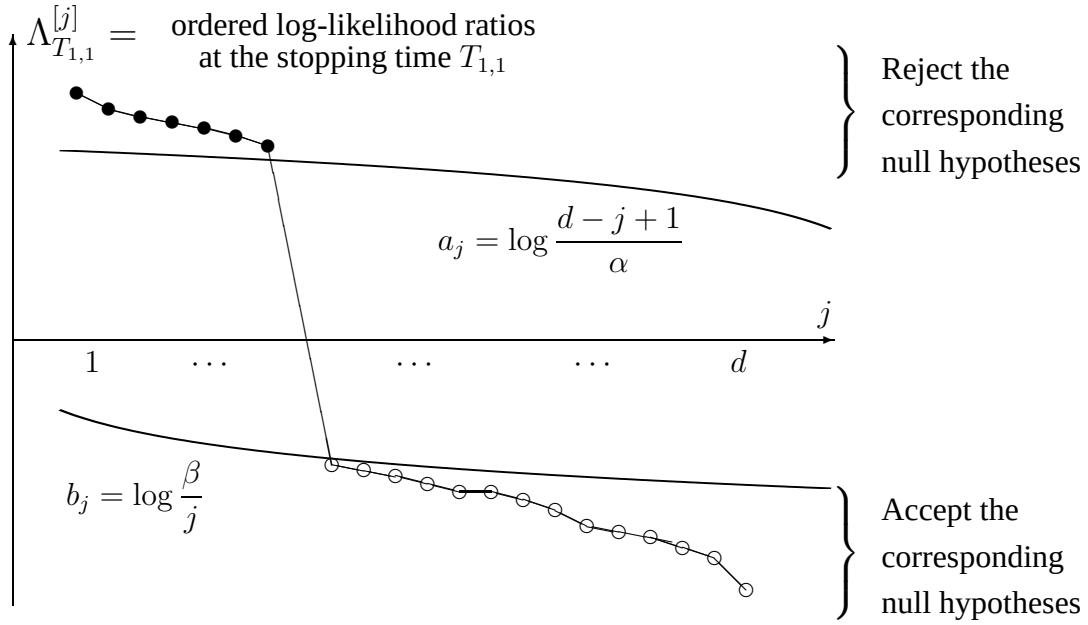

Figure 1: Sequential scheme controlling Type I and Type II familywise error rates. At this moment, sampling stops; null hypotheses corresponding to the log-likelihood ratios above a_j are rejected.

Control of $FWER_I$ and $FWER_{II}$ is split between the rejection and acceptance boundaries. It was proven that the chosen upper boundary a_j controls $FWER_I$ at level α for any $b_j < 0$, and the lower boundary b_j controls $FWER_{II}$ at level β for any $a_j > 0$. Hence, simultaneous use of the stopping boundaries (6) provides the intersection of conditions (3) and (4). For this reason, (5) was named the *Intersection stopping rule.*

2.1. Discussion of the intersection rule

It can be noticed that the use of the Intersection stopping rule (5) can produce the following scenario. Some of the tested hypotheses may already be resolved at some moment, accepted or rejected, but sampling continues due to the unresolved tests. Since sampling of *all* components continues, further data may force to revert the decisions obtained earlier. A test statistic, after visiting an acceptance or rejection region, can potentially swing back into the continue-sampling area, and further, it may cross the opposite boundary. A few alternative strategies may be considered in this case:

(a) Sampling stops at the first time when decision is made on at least one test (rule $T^{\min}$ in [4], used earlier in [15], chap. 15). All hypotheses that are not rejected at this time are accepted controlling $FWER_I$ but not $FWER_{II}$.

(b) Sampling continues until each test reaches decision once, after which the corresponding components are discarded from further testing (called incomplete $T^{\max}$ in [4], essentially used earlier in [25]). This rule is called "incomplete" because it is based on incomplete data, in violation to the sufficiency principle. At no additional cost, one cannot lose from sampling all the components for each sampled unit. Then, all of them should be used for the final decision, as long as it is based on a sufficient statistic. Thus, this rule leaves room for improvement.

(c) The previous rule can be improved by Rao-Blackwellization that results in a randomized "complete $T^{\max}$" procedure (Theorem 2.1 of [4]). It controls both $FWER_I$ and $FWER_{II}$ in the strong sense, satisfies the sufficiency principle, and uses the same stopping rule as (b). However, it allows some of the likelihood ratios of rejected hypotheses to appear in the continue-sampling or even the acceptance region at the stopping time.

(d) The non-randomized intersection rule (5) requires sampling until all d tests reach decision simultaneously. This may require more sampling units than any of the rules in (a,b,c). The benefit is in making each decision according to the final value of the test statistic. Indeed, if a larger sample claims reversion of some earlier small-sample decisions, a statistician should be thankful for having sampled additional data which allowed to correct the earlier Type I or Type II errors. At the same time, the difference in sampling costs between rules (b,c,d) is very small (Tables 1,2 in [4]).

Similar dilemma appears in group sequential or sequentially planned statistical experiments. What should one do if a test statistic crosses a stopping boundary in the middle of

a sampled group and returns to the continue-sampling region or even crosses the opposite boundary by the end of the group? According to the sufficiency principle, only the final test statistic should be taken into account. Indeed, if the same data were collected in a different order, the boundary would not have been crossed at all. Since the set of order statistics is sufficient in the i.i.d. case, the order of sampled observations should not affect the terminal decision, and therefore, crossing any boundary in the middle of a group should simply be ignored (a relevant discussion of this phenomenon is on p.20 of [23]).

3. Control of generalized familywise error rates

The required sample size (5) can be reduced further by replacing (3,4) with a less stringent condition. This is particularly important for large-scale statistical inferences that are likely to require unreasonably large sample sizes and high costs unless conditions (3,4) are relaxed.

In this article, sequential testing procedures are developed that control *generalized familywise Type I and Type II error rates* instead of (3,4). Suppose that the set of tested d null hypotheses contains at least k true and at least m false ones. Then the generalized familywise Type I error rate, $GFWER_I(k)$, is defined as the probability of rejecting at least k true null hypotheses, and the corresponding Type II rate $GFWER_{II}(m)$ is the probability of accepting at least m false null hypotheses. That is,

$$\begin{aligned} GFWER_I(k) &= \max_{\theta^{(1)},\ldots,\theta^{(d)}} \boldsymbol{P}\left\{ \sum_{\substack{j=1,\ldots,d \\ H_0^{(j)}\text{is true}}} I\left\{ \text{reject } H_0^{(j)} \right\} \geq k \right\}; \\ GFWER_{II}(m) &= \max_{\theta^{(1)},\ldots,\theta^{(d)}} \boldsymbol{P}\left\{ \sum_{\substack{j=1,\ldots,d \\ H_0^{(j)}\text{is false}}} I\left\{ \text{accept } H_0^{(j)} \right\} \geq m \right\}; \end{aligned} \tag{7}$$

The goal is to control the generalized FWE rates at levels α and β, respectively, i.e.,

$$\textit{GFWER}_{I}(k) \leq \alpha \quad \text{and} \quad \textit{GFWER}_{II}(m) \leq \beta. \tag{8}$$

In other words, while conducting a large-scale battery of d tests, we agree to tolerate at most $(k-1)$ Type I errors and at most $(m-1)$ Type II errors. Committing more Type I (II) errors should only happen with probability α (β). Similarly, we can define *generalized power* as the probability of detecting all the deviations of parameters from their corresponding null values, except, perhaps, $(m-1)$ of them.

For $k = 1$ and $m = 1$, we have $\textit{GFWER}_{I}(1) = \textit{FWER}_{I}$ and $\textit{GFWER}_{II}(m) = \textit{FWER}_{II}$, and condition (8) reduces to (3,4). Larger k and m yield weaker constraints on the testing algorithm.

Control of $\textit{GFWER}_{I}(k)$ and $\textit{GFWER}_{II}(m)$ is attained by a suitable modification of the Intersection stopping rule (5). Two approaches are proposed here. The first approach, *method of reduced boundaries*, is a sequential extension of the step-down procedure proposed in [19]. The second approach, *the curtailed procedure*, obtained by a generalization of the "intersection scheme" of [4], has an enhanced performance as it controls the error rates at the same levels at the expense of a smaller expected sample size.

3.1. Method of Reduced Boundaries

For non-sequential statistical experiments, [19] proposed a step-down algorithm that controls the generalized familywise Type I error rate. The idea is to increase significance levels α_j appearing in the Holm's method and to follow a step-down algorithm. Accordingly, the rejection region is expanded allowing for more rejections of null hypotheses.

Applying this approach to sequential experiments, we adjust levels α_j and β_j and modify the stopping boundaries to shrink the continue-sampling region in a way that satisfies

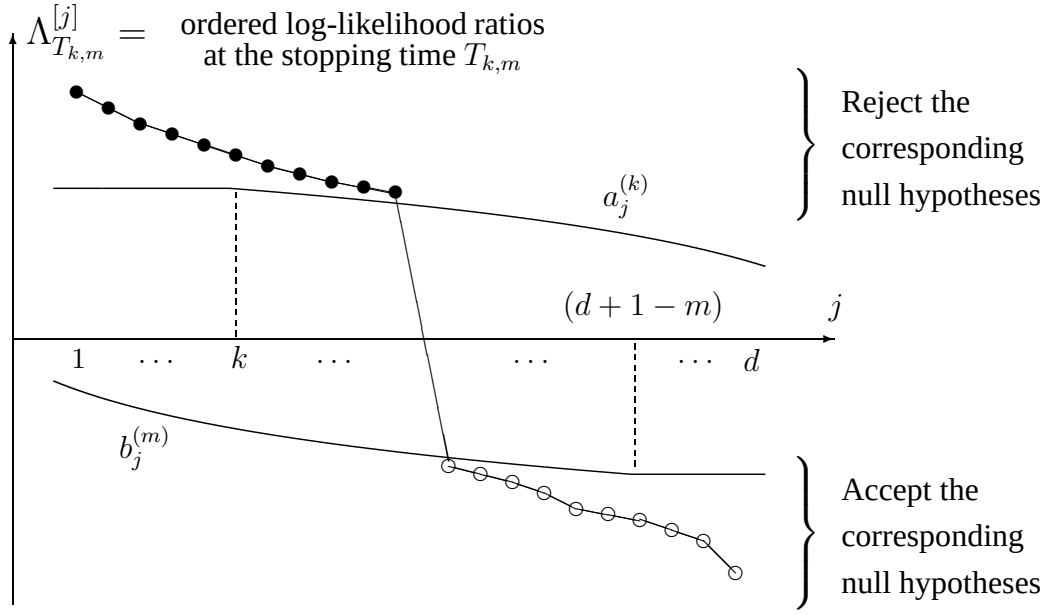

Figure 2: Reduced boundaries sequential testing procedure controlling generalized familywise error rates *GFWER$_I$*(k) and *GFWER$_{II}$*(m).

both constraints (8) on the generalized familywise error rates. This test is also based on the ordered log-likelihood ratio statistics $\Lambda_n^{[j]}$, $j = 1, \ldots, d$. With the new stopping boundaries

$$\begin{aligned} a_j^{(k)} &= \ln \frac{d-(j-k)^+}{\alpha k} = \begin{cases} \ln \dfrac{d+k-j}{\alpha k} & \text{for } j > k, \\ \ln \dfrac{d}{\alpha k} & \text{for } j \le k, \end{cases} \\ b_j^{(m)} &= \ln \frac{m\beta}{d-(d-m+1-j)^+} = \begin{cases} \ln \dfrac{m\beta}{m-1+j} & \text{for } j < d+1-m, \\ \ln \dfrac{m\beta}{d} & \text{for } j \ge d+1-m, \end{cases} \end{aligned} \tag{9}$$

sampling stops at the *Reduced Boundary stopping time*

$$T_{k,m} = \inf \left\{ n : \bigcap_{j=1}^{d} \Lambda_n^{[j]} \notin (b_j^{(m)}, a_j^{(k)}) \right\}. \tag{10}$$

At this moment, the null hypotheses $H_0^{[j]}$ corresponding to $\Lambda^{[j]}_{T_{k,m}} \ge a_j^{(k)}$ are rejected, and the ones corresponding to $\Lambda^{[j]}_{T_{k,m}} \le b_j^{(m)}$ are accepted, as on Fig. 2.

One can see that the rejection boundary $a_j^{(k)}$ is decreasing in k while the acceptance

boundary $b_j^{(m)}$ is increasing in m. Therefore, increasing k and m and making the condition (8) less stringent forces the sampling region $(a_j^{(k)}, b_j^{(m)})$ to shrink, and sampling stops earlier. In particular, the sample size required by the Reduced Boundary algorithm is no greater than that required by the Intersection stopping rule of Section 2 with probability one, and the expected sample size is strictly smaller.

Theorem 1. *The stopping time $T_{k,m}$ is proper for any k and m, and the introduced Reduced Boundary sequential testing procedure controls GFWER$_I(k)$ and GFWER$_{II}(m)$ in the strong sense, satisfying (8).*

PROOF. The stopping rule $T_{1,1}$ is proper by Theorem 4 of [4]. For any $k, m \geq 1$, $a_j^{(k)} \leq a_j$ and $b_j^{(m)} \geq b_j$, therefore, $T_{k,m} \leq T_{1,1}$ with probability one, under any combination of null and alternative hypotheses. Hence $T_{k,m}$ is also proper, i.e., $P\{T_{k,m} < \infty\} = 1$.

To prove control of generalized familywise error rates, consider the index set $I_0 \subset \{1, \ldots, d\}$ of *true null hypotheses.* Let $\lambda_{T_{k,m}}^{[j]}$ denote the j-th largest log-likelihood ratios (LLR) within I_0 at the stopping time $T_{k,m}$, i.e.,

$$\lambda_{T_{k,m}}^{[1]} \geq \cdots \geq \lambda_{T_{k,m}}^{[j]} \geq \cdots \geq \lambda_{T_{k,m}}^{[|I_0|]}.$$

Then, suppose that the k-th largest LLR among I_0 is the j_0-th largest LLR among all $\left\{\Lambda_{T_{k,m}}^{[1]}, \ldots, \Lambda_{T_{k,m}}^{[d]}\right\}$ at time $T_{k,m}$, i.e., $\lambda_{T_{k,m}}^{[k]} = \Lambda_{T_{k,m}}^{[j_0]}$. Note that $j_0 \geq k$ and $j_0 - k \leq d - |I_0|$ because the number of LLR among the largest j_0 of them that do *not* correspond to true null hypotheses cannot exceed the number of false null hypotheses. Therefore,

$$k \leq j_0 \leq d - |I_0| + k. \tag{11}$$

Next, all the rejections have to be caused by the largest LLRs. Indeed, at the time $T_{k,m}$, event $\Lambda_{T_{k,m}}^{[j]} < a_j^{(k)}$ implies $\Lambda_{T_{k,m}}^{[j]} \leq b_j^{(m)}$. Since $\min_j a_j^{(k)} = a_d^{(k)} > 0 > b_1^{(m)} = \max_j b_j^{(m)}$, it is impossible to have $\Lambda_{T_{k,m}}^{[i]} \geq a_i^{(k)}$ and $\Lambda_{T_{k,m}}^{[j]} < a_j^{(k)}$ for any $i < j$.

Therefore, in the case of at least k *false* rejections of true null hypotheses, the largest k LLRs within I_0, and accordingly, the largest j_0 among all LLRs are above the upper

stopping boundary, i.e.,

$$\Lambda^{[1]}_{T_{k,m}} \geq a^{(k)}_1, \ldots, \Lambda^{[j_0]}_{T_{k,m}} \geq a^{(k)}_{j_0}.$$

Hence,

$$\textit{GFWER}_I(k) \leq P\left(\Lambda^{[j_0]}_{T_{k,m}} \geq a_{j_0}\right) = P\left(\Lambda^{[j_0]}_{T_{k,m}} \geq \ln \frac{d+k-j_0}{\alpha k}\right) \leq P\left(\lambda^{[k]}_{T_{k,m}} \geq \ln \frac{|I_0|}{\alpha k}\right),$$

using (11). Inequality $\lambda^{[k]}_{T_{k,m}} \geq \ln \frac{|I_0|}{\alpha k}$ implies that at least k log-likelihood ratios within the set I_0 are greater than or equal to $\ln \frac{|I_0|}{\alpha k}$. Hence, using Markov inequality,

$$\begin{aligned}
\textit{GFWER}_I(k) &\leq P\left(\sum_{j\in I_0} I\left\{\lambda^{[j]}_{T_{k,m}} \geq \ln \frac{|I_0|}{\alpha k}\right\} \geq k\right) \leq \frac{1}{k}\sum_{j\in I_0} P\left(\lambda^{(j)}_{T_{k,m}} \geq \ln \frac{|I_0|}{\alpha k}\right) \\
&\leq \frac{1}{k}\sum_{j\in I_0} \exp\left(-\ln \frac{|I_0|}{\alpha k}\right) = \alpha.
\end{aligned}$$

The last inequality is proved in Lemma 2 of [4].

Control of the Type I GFWER is proved. Using a similar logic to show control of the Type II GFWER, let $L^{\{m\}}_{T_{k,m}}$ denote the m-th smallest LLR among the set I_A of *false* null hypotheses, and suppose that the m-th smallest LLR within I_A is the j_A-th smallest LLR in $\{1, \ldots, d\}$, i.e., $L^{\{m\}}_{T_{k,m}} = \Lambda^{\{j_A\}}_{T_{k,m}}$. In I_A, $L^{\{m\}}_{T_{k,m}}$ is also the $(|I_A| - m + 1)$-th largest LLR, and in $\{1, \ldots, d\}$, $\Lambda^{\{j_A\}}_{T_{k,m}}$ is the $(d - j_A + 1)$-th largest, i.e.,

$$L^{\{m\}}_{T_{k,m}} = L^{[|I_A|-m+1]}_{T_{k,m}} = \Lambda^{\{j_A\}}_{T_{k,m}} = \Lambda^{[d-j_A+1]}_{T_{k,m}}.$$

Therefore,

$$m \leq j_A \leq d - |I_A| + m. \tag{12}$$

If at least m false hypotheses are accepted causing Type II errors, then $\Lambda^{\{1\}}_{T_{k,m}}, \ldots, \Lambda^{\{j_A\}}_{T_{k,m}}$ appear below the acceptance boundary, i.e.,

$$\Lambda^{[d]}_{T_{k,m}} \leq b^{(m)}_d, \ldots, \Lambda^{[d-j_A+1]}_{T_{k,m}} \leq b^{(m)}_{d-j_A+1}.$$

Then, the Type II GFWER can be bounded using (12) and Markov inequality as follows,

$$\begin{aligned}
\textit{GFWER}_{II}(m) &\leq P\left(\Lambda^{[d-j_A+1]}_{T_{k,m}} \leq \ln \frac{m\beta}{d+m-j_A}\right) \leq P\left(L^{\{m\}}_{T_{k,m}} \leq \ln \frac{m\beta}{|I_A|}\right) \\
&\leq P\left(\sum_{j\in I_A} I\left(L^{(j)}_{T_{k,m}} \leq \ln \frac{m\beta}{|I_A|}\right) \geq m\right) \leq \frac{1}{m}\sum_{j\in I_A} P\left(L^{(j)}_{T_{k,m}} \leq \ln \frac{m\beta}{|I_A|}\right) \\
&\leq \frac{1}{m}\sum_{j\in I_A} \exp\left(\ln \frac{m\beta}{|I_A|}\right) = \beta.
\end{aligned}$$

Again, the last inequality is obtained using Lemma 2 of [4].

Shrinking the stopping boundaries inevitably results in the reduction of the sample size. This saving appears rather substantial for testing a large number of hypotheses, as seen in Section 4. However, the stopping rule (10) is not activated until *the last* log-likelihood ratio leaves the continue-sampling region. Intuitively, the maximum of d random times is large when d is large, and thus, procedure (10) is not optimal. The procedure introduced in the next section yields a lower expected sample size while still controlling for the Type I and Type II GFWER.

3.2. Curtailed Intersection Scheme

The Intersection scheme described in Section 2 with stopping rule (5) can be modified in a different, more efficient, and perhaps, a simpler way.

Continue using stopping boundaries (6) but stop earlier, curtailing the Intersection scheme of Section 2. The Intersection procedure continues until all the log-likelihood ratios leave the continue-sampling region (b_j, a_j). At the new stopping time, we allow some LLRs to remain in (b_j, a_j) as long as the number of LLRs between the smallest and the largest LLR in (b_j, a_j) does not exceed $(k + m - 2)$. The Intersection algorithm is now a special case of the Curtailed procedure with $k = m = 1$.

Formally, this Curtailed stopping rule will be defined as

$$\tau = \tau_{k,m} = \inf \left\{ n : \sum_{j=1}^{d} I\left(\Lambda_n^{[j]} \in \left[B_n, A_n\right]\right) \leq k + m - 2 \right\} \tag{13}$$

where $\Lambda_n^{[j]}$ is the j-th largest LLR, A_n and B_n are random variables defined as $A_n = \sup\left\{\Lambda_n^{[j]} : b_j < \Lambda_n^{[j]} < a_j\right\}$ and $B_n = \inf\left\{\Lambda_n^{[j]} : b_j < \Lambda_n^{[j]} < a_j\right\}$, and the boundaries a_j and b_j are given by (6). If there is no $\Lambda_n^{[j]} \in (b_j, a_j)$ at time n, then $A_n = -\infty$, $B_n = +\infty$, the interval $[B_n, A_n]$ is an empty set, and sampling terminates according to (13).

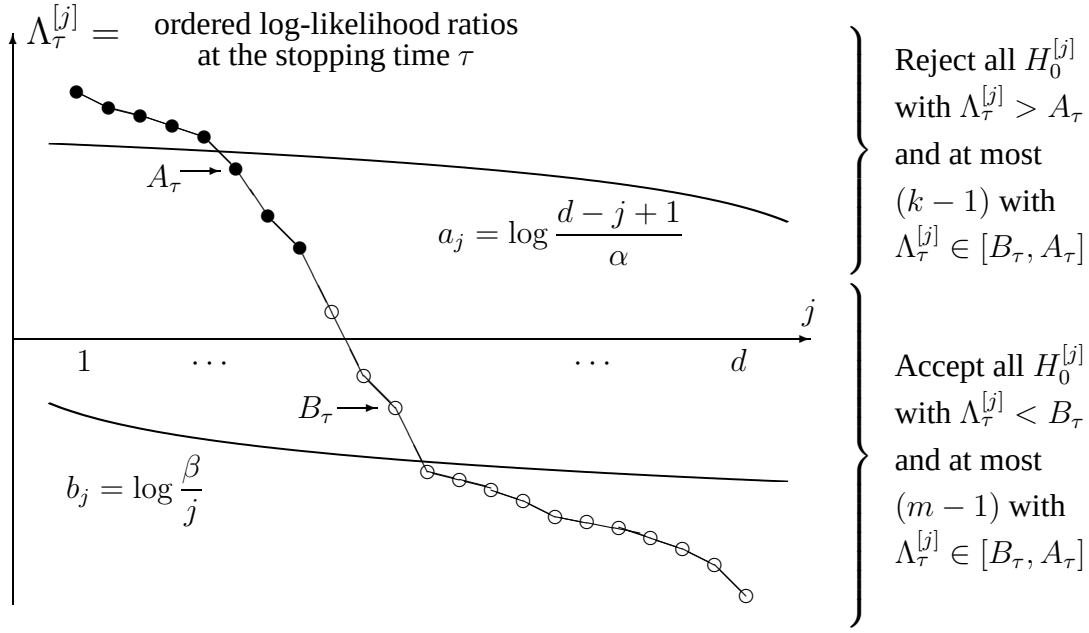

Figure 3: The curtailed intersection scheme controls $GFWER_I(k)$ and $GFWER_{II}(m)$ and does not wait until all LLRs leave the continue-sampling region.

At the stopping time τ, the *Curtailed terminal decision rule* is

(1) to reject all the null hypotheses $H_0^{[j]}$ corresponding to the ordered LLRs $\Lambda_\tau^{[j]} > A_\tau$ *and* at most $(k-1)$ largest LLRs in $[B_\tau, A_\tau]$;

(2) to accept all $H_0^{[j]}$ corresponding to $\Lambda_\tau^{[j]} < B_\tau$ *and* at most $(m-1)$ smallest LLRs in $[B_\tau, A_\tau]$.

Since the total number of LLRs in the interval $[B_\tau, A_\tau]$ cannot exceed $(k+m-2)$ at the stopping time, all the hypotheses are either rejected or accepted (Fig. 3).

Theorem 2. *The stopping time $\tau_{k,m}$ is proper for all $k, m \geq 1$, and the Curtailed Intersection scheme controls the generalized familywise error rates $GFWER_I(k)$ and $GFWER_{II}(m)$ in the strong sense, i.e.,*

$$GFWER_I(k) \leq \alpha \quad \textit{and} \quad GFWER_{II}(m) \leq \beta.$$

PROOF. At the time (5), the stopping condition (13) is always satisfied. Therefore, the Curtailed scheme cannot run beyond the Intersection stopping time $T_{1,1}$, i.e., $P(\tau_{k,m} \leq$

$T_{1,1}) = 1$ for any $k, m \geq 1$ and under any combination of null and alternative hypotheses. Since the stopping time $T_{1,1}$ is proper by Theorem 4 of [4], the Curtailed stopping rule is also proper, i.e., $\boldsymbol{P}\{\tau < \infty\} = 1$.

Let us now prove that the Curtailed testing procedure controls *GFWER*$_I(k)$ at level α. We denote the index sets of true and false null hypotheses by I_0 and I_A respectively where $I_0 \cup I_A = \{1, \ldots, d\}$. According to the definition (7), only the case $|I_0| \geq k$ needs to be considered.

Arrange LLRs in non-increasing order at the stopping time τ as $\Lambda_\tau^{[1]} \geq \ldots \geq \Lambda_\tau^{[d]}$. Let j_0 be the first index (in that order) that belongs to I_0. In other words, j_0 is such that the null hypothesis $H_0^{[j_0]}$ corresponding to the j_0-th largest LLR $\Lambda_\tau^{[j_0]}$ is true, but all the null hypotheses corresponding to larger LLRs are false. Thus, there are at least $(j_0 - 1)$ false nulls, i.e., $j_0 - 1 \leq |I_A| = d - |I_0|$, and therefore,

$$a_{j_0} = \log\left(\frac{d + 1 - j_0}{\alpha}\right) \geq \log\frac{|I_0|}{\alpha}. \tag{14}$$

Type I errors cannot be made on the false null hypotheses $H_0^{[1]}, \ldots, H_0^{[j_0-1]}$. Next, if $\Lambda_\tau^{[j_0]} < a_j$, then $\Lambda_\tau^{[j_0]}$ is in $[B_\tau, A_\tau]$, and at most $(k-1)$ hypotheses $H_0^{[j]}$ can be rejected for $j \geq j_0$, by the definition of the Curtailed decision rule (also see Fig. 3). Thus, at most $(k-1)$ Type I errors can be committed if $\Lambda_\tau^{[j_0]} < a_j$.

For any non-random $N > 0$ and any I_0 with $|I_0| \geq k$, using (14), we have

$$\begin{aligned}
\boldsymbol{P}\{\text{at least } k \text{ Type I errors}\} \leq & \boldsymbol{P}\left\{\Lambda_\tau^{[j_0]} \geq a_{j_0}\right\} \leq \boldsymbol{P}\left\{\Lambda_\tau^{[j_0]} \geq \log\frac{|I_0|}{\alpha}\right\} \\
= & \boldsymbol{P}\left\{\max_{j \in I_0} \Lambda_\tau^{(j)} \geq \log\frac{|I_0|}{\alpha}\right\} \\
\leq & \boldsymbol{P}\{\tau > N\} + \boldsymbol{P}\left\{\max_{1 \leq n \leq N} \max_{j \in I_0} \Lambda_n^{(j)} \geq \log\frac{|I_0|}{\alpha}\right\} \\
\leq & \boldsymbol{P}\{\tau > N\} + \sum_{j \in I_0} \boldsymbol{P}\left\{\max_{1 \leq n \leq N} \exp\left(\Lambda_n^{(j)}\right) \geq \frac{|I_0|}{\alpha}\right\}.
\end{aligned} \tag{15}$$

The likelihood ratio $\exp\left(\Lambda_n^{(j)}\right)$ is a non-negative martingale for $j \in I_0$ with respect to the filtration generated by $(X_1^{(j)}, X_2^{(j)}, \ldots)$. Then, by Doob's maximal inequality for sub-martingales (e.g., [29], Sect. 14.6; [17], Sect. 4.5),

$$\boldsymbol{P}\left\{\max_{1 \leq n \leq N} \exp\left(\Lambda_n^{(j)}\right) \geq \frac{|I_0|}{\alpha}\right\} \leq \frac{\alpha}{|I_0|}\,\mathbf{E}\left\{\exp\left(\Lambda_N^{(j)}\right)\right\} = \frac{\alpha}{|I_0|}.$$

Using this inequality to bound probabilities in (15) and taking limit as $N \to \infty$, we obtain

$$\textit{GFWER}_I(k) = \boldsymbol{P}\{\text{at least } k \text{ Type I errors}\} \le \lim_{N\to\infty} \boldsymbol{P}\{\tau > N\} + \sum_{j\in I_0} \frac{\alpha}{|I_0|} = \alpha.$$

The other inequality, $\textit{GFWER}_{II}(m) \le \beta$, is proven using similar arguments.

This curtailing approach essentially reflects no penalty for committing up to $(k-1)$ Type I errors and $(m-1)$ Type II errors. To protect both generalized familywise error rates, it suffices to make any decision on the nulls corresponding to LLRs in $[B_\tau, A_\tau]$, rejecting fewer than k and accepting fewer than m nulls.

3.3. Further guidance. Decision making between the boundaries

The Curtailed intersection scheme of Section 3.2 allows several different ways to reject $n_0 \le k-1$ and to accept $n_1 \le m-1$ null hypotheses for those LLRs that belong to the interval $[B_\tau, A_\tau]$ at time τ. Any choice between them guarantees control of both $\textit{GFWER}_I(k)$ and $\textit{GFWER}_{II}(m)$ in the strong sense, according to Theorem 2. At the same time, a few optimizing strategies can be considered for these LLRs.

Let s be the number of LLRs in the interval $[B_\tau, A_\tau]$ at the stopping time τ. According to (13), we have $s \le k+m-2$. Moreover, among the corresponding s null hypotheses, at most $(k-1)$ are to be rejected and at most $(m-1)$ are to be accepted. Therefore, we must accept at least $(s-k+1)$ and reject at least $(s-m+1)$ of them. Under these conditions, any decision rule controls both GFWERs.

Deciding on the s tests whose corresponding LLRs landed between the stopping boundaries at time τ, it is natural to reject those that are closest to rejection and to accept those that are closest to acceptance. The remainder of this section offers several ways to measure this "closeness". These are just guidelines, and none of them is claimed to control error probabilities in any stronger sense than controlling GFWER.

A quick solution is to accept the null hypotheses $H_0^{(j)}$ corresponding to the smallest $(s-k+1)^+$ LLRs $\Lambda_\tau^{(j)} \in [B_\tau, A_\tau]$ and to reject the remaining "undecided" null hypotheses, if Type I errors are deemed (infinitely) more costly. Or, if Type II errors are costlier, then reject the nulls corresponding to the largest $(s-m+1)^+$ LLRs in $[B_\tau, A_\tau]$ and accept the remaining ones whose LLRs land between the boundaries B_τ and A_τ.

Between these two extremes, we can let the data suggest which hypotheses should be rejected. Quantifying the strength of support of $H_0^{(j)}$ and $H_A^{(j)}$ by the corresponding $\Lambda_\tau^{(j)}$-based p-values, the decision on the j-th test will depend on

$$\begin{aligned} p_I^{(j)} = \text{Type I p-value} &= \inf\left\{\alpha:\ \text{level } \alpha \text{ test rejects } H_0^{(j)}\right\} \quad \text{and} \\ p_{II}^{(j)} = \text{Type II p-value} &= \inf\left\{\beta:\ \text{test with power } (1-\beta) \text{ accepts } H_0^{(j)}\right\}. \end{aligned}$$

Assigning equal weights to Type I and Type II errors, reject $H_0^{(j)}$ if $p_I^{(j)} < p_{II}^{(j)}$. However, one may consider different weights and reject when $Cp_I^{(j)} < p_{II}^{(j)}$ for some chosen C. These decisions are still based on the marginal distributions of $X_n^{(j)}$.

Example. Consider testing d hypotheses $H_0^{(j)}$ vs $H_A^{(j)}$ about the means of normally distributed components, normalized to have $\theta_0^{(j)} = 0$, $\theta_1^{(j)} = 0.2$, and variances $\sigma_j^2 = 1$ for all j. Controlling the probabilities of at least $k = 10$ Type I errors and at least $m = 10$ Type II errors at levels $\alpha = 0.05$ and $\beta = 0.10$, sampling stopped after $\tau = 600$ observations. At this moment, $B_\tau = -5.1$, $A_\tau = 7.2$, and there are $s = 15$ LLRs between them. Among them, the 6th, 7th, ..., 10th largest LLRs are observed to be 6.3, 2.2, 1.1, 0.4, and (-0.6). Only these values matter for our decision because $s-(m-1) = 6$ nulls with the highest LLR have to be rejected, and $s-(k-1) = 6$ with the lowest LLR must be accepted.

For this sample size, each LLR has Normal$(\pm 12, 24)$ distribution, restricted to $[B_\tau, A_\tau]$, with a negative mean under H_0 and positive under H_A. The corresponding marginal p-values are $p_I = .001, .024, .047, .072, .126$ and $p_{II} = .749, .139, .080, .055, .031$. Thus,

with no preference between Type I and Type II errors, reject $H_0^{[6,7,8]}$ and accept $H_0^{[9,10]}$. But assume that Type I errors are twice as costly as Type II errors and take $C = 2$. Then we should be more careful rejecting the null hypotheses, and we reject $H_0^{[6,7]}$ only because $2p_I^{[j]}/p_{II}^{[j]} > 1$ for $j = 8, 9, 10$.

Summarizing, the Curtailed scheme may leave several decision options for those LLR that appear between the stopping boundaries at the stopping time. We recommend to reject $H_0^{(j)}$ whose LLR are "close" to rejection and accept those whose LLR are "close" to acceptance, where "closeness" can be measured in different ways. Either way, the scheme controls both generalized error rates in the strong sense.

4. Performance evaluation and comparison

In this section, performance of the proposed sequential multiple testing schemes is evaluated and compared by a simulation study. Our main goals are:

(i) to compare the two proposed sequential schemes, the Reduced boundary intersection scheme (RBIS, Section 3.1) and the Curtailed intersection scheme (CIS, Section 3.2), controlling $GFWER_I$ and $GFWER_{II}$;

(ii) to compare their performance with the Intersection scheme (IS, Section 2) controlling $FWER_I$ and $FWER_{II}$ in order to evaluate the gain in terms of a smaller expected sample size at the expense of a weaker error control;

(iii) To compare the named sequential procedures against the non-sequential Lehmann-Romano multiple testing procedure, also designed to control $GFWER_I$, in order to assess the benefit of conducting the experiment sequentially.

We consider sequences of vector-valued observations with Normal and Bernoulli components, which are the two most common distributions in clinical trials. Vectors $\boldsymbol{X}_1$, $\boldsymbol{X}_2, \ldots \in \mathbb{R}^d$ are generated, where $\boldsymbol{X}_i = (X_i^{(1)}, \ldots, X_i^{(d)})$ for $i = 1, 2, \ldots$, with $X_i^{(j)}$

having *Normal*$(\theta^{(j)}, 1)$ distribution for $j = 1, \ldots, d/2$ and *Bernoulli*(p_j) distribution for $j = d/2 + 1, \ldots, d$ (let d be divisible by four). As examples, we consider the cases of mutually independent components, correlation coefficient 0.5 between any two Normal components, and correlation coefficient 0.9 between any two of them. Bernoulli components remain independent.

The following d hypotheses are being tested simultaneously at the nominal familywise error rates $\alpha = 0.05$ and $\beta = 0.10$,

$$H_0^{(j)}: \ \theta^{(j)} = 0 \quad \text{vs} \quad H_A^{(j)}: \ \theta^{(j)} = 0.5 \qquad \text{for } j = 1, \ldots, d/2$$
$$H_0^{(j)}: \ p_j = 0.5 \quad \text{vs} \quad H_A^{(j)}: \ p_j = 0.75 \qquad \text{for } j = d/2 + 1, \ldots, d.$$

In the simulation study, the first $d/4$ null hypotheses about the Normal means are false, and the next $d/4$ are true. Among Bernoulli components, again, the first $d/4$ nulls are false, and the last $d/4$ are true.

4.1. Performance of Reduced boundaries and Curtailed schemes. Cost saving at the expense of more errors

How much cost is being saved by allowing a few Type I and/or Type II errors? Which of the proposed sequential testing schemes is more efficient?

Performance of Reduced boundaries and Curtailed sequential schemes (RBIS and CIS) is compared to the Intersection scheme (IS) in Table 1. RBIS and CIS control *GFWER*$_I$ (k) and *GFWER*$_{II}$ (m) for the specified values of k and m, whereas IS satisfies a stronger condition and controls *FWER*$_I$ and *FWER*$_{II}$.

Reported estimates are based on 20,000 simulated sequences. Standard errors of the estimated expected sample size $\widehat{ET}$ are included.

Whenever the Curtailed scheme has several decision options at its stopping time because of $s < k + m - 2$ LLR between the stopping boundaries, we choose the first option

Table 1: Performance of the Reduced boundaries intersection scheme (RBIS), the Curtailed intersection scheme (CIS), and the original Intersection scheme (IS).

No. of tests	ρ	Testing scheme	k, m	$\widehat{ET}$	St. err. of $\widehat{ET}$	*GFWER*$_I(k)$ (%)	*GFWER*$_{II}(m)$ (%)
$d = 12$	0	IS	1, 1	91.8	0.2	0.62	0.91
		RBIS	2, 2	84.5	0.2	< 0.01	0.01
		CIS	2, 2	57.9	0.1	0.17	0.45
	0.5	IS	1, 1	91.0	0.2	0.50	0.54
		RBIS	2, 2	83.6	0.2	0.04	0.06
		CIS	2, 2	58.6	0.1	0.49	0.85
	0.9	IS	1, 1	85.1	0.2	0.39	0.58
		RBIS	2, 2	78.3	0.2	0.11	0.29
		CIS	2, 2	61.0	0.1	0.47	0.92
$d = 40$	0	IS	1, 1	137.9	0.2	0.24	0.29
		RBIS	2, 2	129.7	0.2	< 0.01	< 0.01
		CIS	2, 2	101.0	0.1	0.09	0.19
	0.5	IS	1, 1	132.7	0.2	0.09	0.20
		RBIS	2, 2	124.1	0.2	< 0.01	< 0.01
		CIS	2, 2	100.2	0.1	0.09	0.17
	0.9	IS	1, 1	122.1	0.2	0.18	0.22
		RBIS	2, 2	113.7	0.2	< 0.01	0.02
		CIS	2, 2	93.6	0.2	0.05	0.09
$d = 60$	0	IS	1, 1	154.2	0.3	0.21	0.19
		RBIS	3, 3	141.3	0.2	< 0.01	< 0.01
		CIS	3, 3	101.7	0.1	0.01	0.04
	0.5	IS	1, 1	146.8	0.2	0.10	0.13
		RBIS	3, 3	134.3	0.2	< 0.01	< 0.01
		CIS	3, 3	100.8	0.1	0.08	0.05
	0.9	IS	1, 1	135.3	0.2	0.16	0.14
		RBIS	3, 3	122.9	0.2	< 0.01	< 0.01
		CIS	3, 3	94.7	0.2	0.03	0.07
$d = 100$	0	IS	1, 1	175.7	0.2	0.09	0.12
		RBIS	5, 5	155.4	0.2	< 0.01	< 0.01
		CIS	5, 5	104.5	0.1	< 0.01	< 0.01
	0.5	IS	1, 1	165.6	0.2	0.07	0.07
		RBIS	5, 5	146.1	0.2	< 0.01	< 0.01
		CIS	5, 5	104.2	0.1	0.01	0.03
	0.9	IS	1, 1	153.4	0.2	0.09	0.09
		RBIS	5, 5	134.3	0.2	< 0.01	< 0.01
		CIS	5, 5	98.5	0.2	0.01	0.03

listed in Section 3.3), accept $(s-k+1)^+$ of the corresponding null hypotheses, and reject the remaining "undecided" ones.

Conclusions.

(1) For sequential testing, the Curtailed method appears to be more efficient than the Lehmann-Romano reduced boundaries approach. Controlling $\mathit{GFWER}_I \le \alpha = 0.05$ and $\mathit{GFWER}_{II} \le \beta = 0.10$, the Curtailed scheme requires 20%-30% fewer sampling units than the Reduced boundaries procedure.

(2) The study confirms that allowing some Type I and/or Type II error brings considerable savings in the expected sample size. Comparing with the Intersection scheme in the considered cases, it reduces the expected sample size by about 30% when 10% of d tests are allowed to result in error with probabilities α and β. This saving increases with the number of tests d and reaches 40% saving for $d = 100$ and $k = m = 5$.

(3) Although the nominal error rates are $\alpha = 0.05$ and $\beta = 0.10$, and they are guaranteed by our results, the actual FWER and GFWER are substantially lower due to the use of Markov inequality in the proofs of Theorems 1 and 2. The inequality is especially crude for a large number of tests. Notice that the GFWER for both sequential procedures is given as a *percent*. For example, the actual $\mathit{GFWER}_I(1) = \mathit{FWER}_I$ in the case $d = 12, \rho = 0, k = m = 1$ equals 0.0062. Such a difference between the nominal guaranteed error rates and the actual error rates indicates a room for improvement and possible existence of more efficient sequential procedures.

In the case of $k = m = 1$, GFWER reduces to FWER, and both schemes RBIS and CIS become equivalent to the Intersection scheme (5).

(4) The expected sample size of all three sequential procedures decreases with correlation. Certainly, we considered too few situations to make a global conclusion, and the

study of the effect of correlation on the performance of sequential multiple testing procedures is beyond the scope of this paper, but in all the considered testing scenarios, all three procedures required a lower expected sample size with $\rho = 0.5$ and even lower with $\rho = 0.9$.

4.2. Comparison against a non-sequential procedure

What is the benefit of conducting a multiple testing experiment sequentially?

Lehmann and Romano in [19] proposed the Reduced boundaries approach for controlling $GFWER_I$ in non-sequential testing of multiple hypotheses. Table 2 evaluates performance of their procedure for the same testing problems as in Section 4.1. For the fair comparison, its (non-random) sample size n is chosen to be the expected sample size of sequential scheme in Table 1, $\widehat{ET}$, rounded up to the nearest integer.

Conclusions. Under the same expected costs (in terms of the sample size), sequential procedures in Table 1 yield substantially lower generalized familywise error rates in most cases. Both sequential and non-sequential procedures yield much lower than nominal GFWERs.

4.3. Empirical matching of GFWER; eliminating the overkill

Tables 1 and 2 clearly show that a multiple testing procedure that is conducted at nominal levels α and β will actually yield the GFWER lower (much lower for large d) than α and β. This includes the non-sequential Lehmann-Romano procedure (LR), and it occurs due to the use of either Markov or Bonferroni inequality that guarantees control of GFWER. Except for extreme cases, either inequality is not sharp, especially for large d, and thus, there is a room for improvement. As a result of this "overkill", the eventual sample size appears larger than it is needed to satisfy the GFWER conditions.

Table 2: Performance of the non-sequential Lehmann-Romano procedure with $\alpha = 0.05$ and approximately the same sample size as the corresponding sequential schemes, RBIS and CIS.

Number of tests	k, m	ρ	n	Lehmann-Romano		Sequential scheme		
				$GFWER_I$ (%)	$GFWER_{II}$ (%)	Scheme	$GFWER_I$ (%)	$GFWER_{II}$ (%)
$d = 12$	2, 2	0	85	0.47	0.05	RBIS	< 0.01	0.01
			58	0.53	3.33	CIS	0.17	0.45
		0.5	84	0.85	0.34	RBIS	0.04	0.06
			59	0.89	4.03	CIS	0.49	0.85
		0.9	78	1.61	1.46	RBIS	0.11	0.29
			61	1.56	5.19	CIS	0.47	0.92
$d = 40$	2, 2	0	130	0.42	<0.01	RBIS	< 0.01	<0.01
			101	0.80	0.14	CIS	0.09	0.19
		0.5	125	1.15	0.17	RBIS	<0.01	<0.01
			101	1.49	1.82	CIS	0.09	0.17
		0.9	114	1.19	0.58	RBIS	<0.01	0.02
			94	1.22	2.33	CIS	0.05	0.09
$d = 60$	3, 3	0	142	0.11	<0.01	RBIS	<0.01	<0.01
			102	0.07	0.05	CIS	0.01	0.04
		0.5	134	0.74	0.06	RBIS	<0.01	<0.01
			101	0.73	1.2	CIS	0.08	0.05
		0.9	123	0.905	0.37	RBIS	<0.01	<0.01
			95	1.01	2.26	CIS	0.03	0.07
$d = 100$	5, 5	0	156	< 0.01	< 0.01	RBIS	<0.01	< 0.01
			105	< 0.01	< 0.01	CIS	< 0.01	< 0.01
		0.5	147	0.48	0.02	RBIS	< 0.01	< 0.01
			105	0.52	0.69	CIS	0.01	0.03
		0.9	135	0.85	0.12	RBIS	< 0.01	< 0.01
			99	0.93	1.61	CIS	0.01	0.03

One possibility is to increase the initial levels of α and β so that the resulting procedure has the desired GFWER. This is done in Table 3. The adjusted values of α and β are found purely *empirically* for the given testing problems, so the control of GFWER is not supported theoretically. Simulations show that the adjusted stopping boundaries yield the desired $GFWER_I \approx 0.05$ and $GFWER_{II} \approx 0.10$. Essentially, the resulting error rates are functions of the chosen levels of α and β, which are found numerically by solving two simultaneous equations, $GFWER_I(\alpha, \beta) = \alpha^*$ and $GFWER_{II}(\alpha, \beta) = \beta^*$, by means of Monte Carlo simulations.

Conclusions.

(1) Roughly 50% of the sample size can be saved by adjusting the initial α and β values in order to eliminate the "overkill" and obtain approximately the desired generalized familywise error rates. This includes sequential and non-sequential procedures.

(2) Sequential schemes require smaller expected sample size than the Lehmann-Romano non-sequential scheme. The cost saving, in terms of the sample size, is between 10% and 28% for $d \leq 100$.

(3) Efficient multiple testing procedures that guarantee control of (generalized) FWER without substantial overkill is a good subject for future research.

Acknowledgements

The authors are grateful to the reviewers, the Associate Editor, and the Editor for their thoughtful comments that led to a major revision of this article. This research is funded by the National Science Foundation and the National Security Agency.

Table 3: Performance of adjusted sequential (RBIS and CIS) and non-sequential (LR) procedures that yield the desired $GFWER_I \approx 0.05$ and $GFWER_{II} \approx 0.10$.

No. of tests	k, m	ρ	Testing scheme	$\widehat{ET}$	St. err. of $\widehat{ET}$	$GFWER_I(k)$ (%)	$GFWER_{II}(m)$ (%)
$d = 12$	$k = m = 2$	0	LR	35	0	5.06	11.47
			RBIS	30.2	0.1	4.99	9.85
			CIS	28.4	0.1	4.92	9.72
		0.9	LR	42	0	4.83	9.68
			RBIS	34.5	0.1	4.97	9.82
			CIS	34.0	0.1	4.99	9.88
$d = 40$	$k = m = 2$	0	LR	65	0	5.09	10.9
			RBIS	57.9	0.1	4.95	9.78
			CIS	58.0	0.1	4.80	9.95
		0.9	LR	62	0	4.90	9.56
			RBIS	47.4	0.12	4.71	9.97
			CIS	46.9	0.1	4.87	10.0
$d = 60$	$k = m = 3$	0	LR	55	0	4.03	8.56
			RBIS	51.3	0.1	5.01	9.98
			CIS	49.7	0.1	4.94	10.0
		0.9	LR	55	0	4.84	9.5
			RBIS	42.4	0.11	4.88	9.68
			CIS	41.9	0.09	4.69	9.99
$d = 100$	$k = m = 5$	0	LR	48	0	4.39	9.44
			RBIS	46.7	0.07	4.80	9.81
			CIS	43.0	0.04	4.93	9.71
		0.9	LR	51	0	5.01	10.16
			RBIS	42.9	0.11	4.95	9.98
			CIS	36.7	0.08	4.87	10.01

References

[1] Bartroff, J., Lai, T.-L., 2010. Multistage tests of multiple hypotheses. Communications in Statistics - Theory and Methods 39, 1597–1607.

[2] Benjamini, Y., Bretz, F., S. Sarkar, e., 2004. Recent Developments in Multiple Comparison Procedures. IMS Lecture Notes - Monograph Series, Beachwood, Ohio.

[3] De, S., Baron, M., 2012. Sequential Bonferroni methods for multiple hypothesis testing with strong control of familywise error rates I and II. Sequential Analysis 31 (2), 238–262.

[4] De, S., Baron, M., 2012. Step-up and step-down methods for testing multiple hypotheses in sequential experiments. J. Statist. Plann. Inference 142, 2059–2070.

[5] De, S. K., 2012. Simultaneous Testing of Multiple Hypotheses in Sequential Experiments. Ph. D. Dissertation, The University of Texas at Dallas.

[6] Devlin, B., Roeder, K., 1999. Genomic control for association studies. Biometrics 55 (4), 997–1004.

[7] Dmitrienko, A., Tamhane, A. C., F. Bretz, e., 2010. Multiple Testing Problems in Pharmaceutical Statistics. CRC Press, Boca Raton, FL.

[8] Dudoit, S., van der Laan, M. J., 2008. Multiple Testing Procedures with Applications to Genomics. Springer, New York.

[9] Efron, B., 2010. Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Cambridge University Press, New York.

[10] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. Nature 457, 1012–1015.

[11] Glimm, E., Maurer, W., Bretz, F., 2010. Hierarchical testing of multiple endpoints in group-sequential trials. Statistics in Medicine 29, 219–228.

[12] Guo, W., Romano, J. P., 2007. A generalized Sidak-Holm procedure and control of generalized error rates under independence. Statist. App. Gen. Mol. Bio. 6, Article 3.

[13] Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6, 65–70.

[14] Jennison, C., Turnbull, B. W., 1993. Group sequential tests for bivariate response: Interim analyses of clinical trials with both efficacy and safety endpoints. Biometrics 49, 741–752.

[15] Jennison, C., Turnbull, B. W., 2000. Group sequential methods with applications to clinical trials. Chapman & Hall, Boca Raton, FL.

[16] Korn, E. L., Troendle, J. F., McShane, L. M., Simon, R., 2004. Controlling the number of false discoveries: application to high-dimensional genomic data. J. Statist. Plann. Inference 124 (1), 379–398.

[17] Kuo, H. H., 2006. Introduction to Stochastic Integration. Springer, New York.

[18] Lazzeroni, L. C., Ray, A., 2012. The cost of large numbers of hypothesis tests on power, effect size and sample size. Molecular Psychiatry 17 (1), 108–114.

[19] Lehmann, E. L., Romano, J. P., 2005. Generalizations of the familywise error rate. Ann. Stat. 33, 1138–1154.

[20] Lévy-Leduc, C., Roueff, F., 2009. Detection and localization of changepoints in high-dimensional network traffc data. Ann. Appl. Statist. 3, 637–662.

[21] Romano, J. P., Wolf, M., 2007. Control of generalized error rates in multiple testing. Ann. Stat. 35, 1378–1408.

[22] Sarkar, S. K., 2007. Step-up procedures controlling generalized FWER and generalized FDR. Ann. Stat. 35, 2405–2420.

[23] Schmegner, C., Baron, M., 2004. Principles of optimal sequential planning. Sequential Analysis 23 (1), 11–32.

[24] Siegmund, D. O., Yakir, B., Zhang, N. R., 2011. Detecting simultaneous variant intervals in aligned sequences. Ann. Appl. Statist. 5, 645–668.

[25] Sobel, M., Wald, A., 1949. A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. Ann. Math. Stat. 20 (4), 502–522.

[26] Tamhane, A. C., Mehta, C. R., Liu, L., 2010. Testing a primary and a secondary endpoint in a group sequential design. Biometrics 66, 1174–1184.

[27] Tang, D.-I., Geller, N. L., Pocock, S. J., 1993. On the design and analysis of randomized clinical trials with multiple endpoints. Biometrics 49 (1), 23–30.

[28] van der Laan, M. J., Dudoit, S., Pollard, K. S., 2004. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. Statistical Applications in Genetics and Molecular Biology 3 (1), Article 15.

[29] Williams, D., 1991. Probability with Martingales. Cambridge University Press, Cambridge, UK.

[30] Xie, Y., Siegmund, D., 2013. Sequential multi-sensor change-point detection. Annals Statist. 41 (2), 670–692.