JACOBS PUBLISHERS

Jacobs Journal of Biostatistics

OPEN ACCESS

Research article

Multiple Tests in Group Sequential Clinical Trials

Zhao T.[1], Baron M.[2*]

[1] *Merck, Rahway NJ, USA*

[2] *American University, Washington DC, USA*

[*] *Corresponding author: Baron M, American University, USA. Email: baron@american.edu*

Received: 08-13-2016

Accepted: 08-29-2016

Published: 09-29-2016

Abstract

Efficient methods are elaborated for the simultaneous testing of multiple hypotheses in group sequential clinical trials. Proposed tests control the Type I and Type II familywise error rates at the levels of α and β and require sampling at most K groups of patients, where α, β, and K are pre-assigned. The new step-down sequential technique allow to reduce the overall sample size under these constraints. It results in a substantial cost saving over the Bonferroni-corrected Pocock and O'Brien-Fleming tests. Optimization of the truncated single-hypothesis sequential probability ratio test appears more efficient than the Pollock-Golhar sequential rule proposed earlier for the same problem.

Keywords: Familywise error rates; Sequential clinical trials; Sequential probability ratio test; Stepwise testing; Stopping boundaries.

Abbreviations

FWER: Familywise Error Rate;
SPRT: Sequential Probability Ratio Test;
TSPRT: Truncated Sequential Probability Ratio Test

1 Introduction

Rapid growth of the pharmaceutical industry has led to an increasing need of accurate testing of medical procedures for their efficacy and safety, which includes simultaneous tests of several hypotheses. Majority of clinical trials are conducted to answer more than one question. They may involve several treatments, multiple endpoints, or more than one population.

For example, in a study of chronic pain [16], the treatment's efficiency was assessed along six domains – pain, physical functioning, emotional functioning, participants' improvement and satisfaction with treatment, symptoms and adverse events, and participants' adherence to the treatment regimen. In asthma trials [11], several endpoints were merged into two groups, the pulmonary functions and the patient outcomes. The treatment was considered efficient if at least one positive effect was found in each group. A recent clinical trial of bucindolol for chronic heart patients included survival time as a

primary endpoint as well as several secondary endpoints [3, 4]. Similarly, a recent clinical trial of Prometa for methamphetamine addiction tested the time to relapse as the primary endpoint and many secondary endpoints grouped into measures of craving, cognitive function, and speed of reaction [17, 18]. In addition, efficiency was tested within subpopulations according to gender, age, socio-economic status, and the mode of drug use. Often clinical trials involve comparison of multiple treatments, ([7, 12]; [10], ch. 16; [21], chap. 8; and others).

Multiplicity often arises in genetic experiments. For example, a number of biomarkers were tested for their association with several clinical and therapeutic measures related to breast cancer [14].

In all these trials, it was important to receive a conclusive answer to each individual test, instead of grouping them into one composite hypothesis.

While *non-sequential* methods of multiple testing have been well developed during the recent two decades or so, little has been done for *sequential* clinical trials. An important requirement in such trials is the given fixed *truncation point* which is the maximum allowed number of sampled groups or the time when the experiment has to be terminated. Unlike the Wald's sequential probability ratio test (SPRT; [19]; [8], ch. 2; [15], ch. 3) and other classical sequential procedures, clinical trials require stopping times that are *bounded* with probability one.

Therefore, we aim to design *truncated sequential algorithms* for multiple hypothesis testing that control the *familywise error rate*, the *familywise power*, and complete data sampling at or before the given time point. Under these three constraints, we aim to minimize the expected sample size by using the optimal error spending approach and a recently developed methodology for stepwise sequential testing. A smaller expected sample size inevitably implies a lower overall expected cost of a clinical trial. This, in turn, yields lower treatment costs, and ultimately, it reduces the overall cost of the health care.

To fix ideas, consider testing multiple hypotheses,

$$H_0^{(1)} \text{ vs. } H_A^{(1)},\ \ H_0^{(2)} \text{ vs. } H_A^{(2)},\ \ldots,\ H_0^{(d)} \text{ vs. } H_A^{(d)},$$

and let J_0 be the index set of true null hypotheses. The *Type I familywise error rate (FWER$_I$)* is defined as

$$\begin{aligned} &FWER_I(J_0) \\ &\quad = \max_{J_0 \neq \varnothing} \mathbf{P}\{\text{at least one Type I error} \mid J_0\} \\ &\quad = \max_{J_0 \neq \varnothing} \mathbf{P}\Big\{ \bigcup_{j \in J_0} \Big(H_0^{(j)} \text{is rejected}\Big) \mid J_0\Big\}. \end{aligned}$$

Controlling the Type I familywise error rate *in the strong sense* implies guaranteeing that the probability of making at least one Type I error does not exceed the given significance level α for *all* possible combinations of true null and alternative hypotheses, i.e., that $FWER_I \leq \alpha$. Clearly, controlling the Type I error probability at a level α for each individual test may result in a much higher probability to make *at least one* Type I error.

By analogy, [5, 6] consider the *Type II familywise error rate*, which is the probability of not rejecting (or accepting) at least one false null hypothesis,

$$\begin{aligned} &FWER_{II}(J_1) \\ &\quad = \max_{J_0 \neq \varnothing} \mathbf{P}\{\text{at least one Type II error} \mid J_0\} \\ &\quad = \max_{J_0 \neq \varnothing} \mathbf{P}\Big\{ \bigcup_{j \notin J_1} \Big(H_0^{(j)} \text{is not rejected}\Big) \mid J_0\Big\}. \end{aligned}$$

Controlling $FWER_{II}$ in the strong sense means guaranteeing $FWER_{II} \leq \beta$ for a pre-determined level β, or *familywise power* $FWP \geq 1 - \beta$.

Our problem is to design group sequential clinical trials for testing multiple hypotheses that

(a) Control the Type I familywise error rate at the given level α;
(b) Control the Type II familywise error rate at the given level β;
(c) Require at most the given number of groups K;
(d) Minimize the overall expected sample size under conditions (a)-(c).

For a clinical trial satisfying (a) and (c) only, [10], chap. 15, proposes to run individual single-hypothesis level α group sequential tests until at least one of them results in a rejection. At this moment, accept all the null hypotheses that are not rejected. This testing procedure controls FWER$_I$at the

level α rather conservatively, and it does not preserve FWER_{II}.

Open ended sequential designs satisfying (a), (b), and (d) were considered in [1, 5, 6]. In this paper, we are adding the *truncation condition* (c) a typical requirement for clinical trials.

The *Bonferroni method* for multiple comparisons is always available to solve (a)-(d), and we consider it in Section 2. To apply it, one conducts the j-th test at a significance level α_j and power $(1-\beta_j)$ with $\sum_j \alpha_j = \alpha$ and $\sum_j \beta_j = \beta$. This approach is universal, and its application to truncated group sequential experiments is rather straightforward. However, the underlying Bonferroni (Boole) inequality $\mathbf{P}\left\{\cup_1^d A_j\right\} \le \sum_1^d \mathbf{P}\{A_j\}$ is rather crude, especially for large d. As a result, the actual significance levels appear much lower than α_j, creating an overkill and requiring unnecessarily large samples.

Our goal is *to improve performance* of multiple testing procedures and to design tests that perform more efficiently than the Bonferroni procedure. This is possible by a Holm-type approach [9] based on the ordering of marginal test statistics according to their significance. A sequential version of the Holm test is elaborated in [6], and being sequential, it is able to control both Type I and Type II error rates. This procedure requires closed-form analytic expressions for the stopping boundaries, unlike Pocock, O'Brien-Fleming, and Wang-Tsiatis tests [10] that are common in group sequential clinical trials.

Extending the method of [6] to *truncated* sequential experiments, we truncate the Wald's SPRT for each individual test and adjust its stopping boundaries to preserve its control of Type I and Type II error probabilities. Taking advantage of the optimal *error spending* between the truncation point K and the previous interim points of a clinical trial, we obtain a test that outperforms the truncated SPRT of [13], in terms of a smaller expected sample size. This test is derived in Section 3.

The developed truncated SPRT (TSPRT) is then used in a stepwise sequential scheme for the simultaneous testing of multiple hypotheses in Section 4. The resulting procedure strongly controls both familywise error rates, samples no more than K groups, and requires a smaller sample size than the Bonferroni method.

2 Bonferroni approach in group sequential clinical trials

Bonferroni method is arguably the most popular approach for multiple significance testing that yields a familywise error rate no higher than the desired given level.

2.1 Multiple testing problem

Consider a clinical trial that results in observing a sequence of independent and identically distributed random vectors $\{\boldsymbol{X}_n, n = 1, 2, ...\}$, where $\boldsymbol{X}_n = (X_1^{(1)}, \dots, X_n^{(d)}) \in \mathbb{R}^d$ is a set of measurements on the n-th patient. The j-th component of $\boldsymbol{X}_n$ follows a distribution from a known family $\mathcal{F}^{(j)} = \left\{\mathcal{F}^{(j)}(\cdot \mid \theta^{(j)}),\ \theta^{(j)} \in \Theta^{(j)}\right\}$ with a density or probability mass function $f(\cdot \mid \theta^{(j)})$.

All $\theta^{(1)}, \dots, \theta^{(d)}$ are *parameters of interest.* These may be indicators of efficacy and safety of a treatment such as the mean blood pressure or pulse, the mean cholesterol level, the mean craving, the mean weight loss, the mean hospitalization time, proportion of patients who experienced headache, nausea, or other side effects.

To determine existence of significant effects in *each* of the d dimensions, the following d tests are conducted simultaneously,

$$\begin{aligned} H_0^{(1)} &: \theta_1 \le \theta_0^{(1)} \quad \text{vs.} \quad H_A^{(1)} : \theta_1 \ge \theta_1^{(1)} \\ H_0^{(2)} &: \theta_2 \le \theta_0^{(2)} \quad \text{vs.} \quad H_A^{(2)} : \theta_2 \ge \theta_1^{(2)} \\ &\dots \\ H_0^{(d)} &: \theta_d \le \theta_0^{(d)} \quad \text{vs.} \quad H_A^{(d)} : \theta_d \ge \theta_1^{(d)}, \end{aligned} \tag{1}$$

(other combination of left-tail and right-tail one-sided tests are obtained by a straightforward reparameterization).

In most common situations, probabilities of Type I and Type II errors in each of these one-sided tests are attained when $\theta^{(j)} = \theta_0^{(j)}$ and when $\theta^{(j)} = \theta_1^{(j)}$. For example, when the likelihood ratio test statistic is used to test (1) for each j, and its marginal distribution has the *monotone likelihood ratio* property

under $\mathcal{F}^{(j)}$, we have

$$\sup_{H_0^{(j)}} \mathbf{P}\{\text{Type I error}\} = \mathbf{P}\left\{\text{Type I error} \mid \theta_0^{(j)}\right\},$$

and

$$\sup_{H_A^{(j)}} \mathbf{P}\{\text{Type II error}\} = \mathbf{P}\left\{\text{Type II error} \mid \theta_1^{(j)}\right\},$$

by the arguments similar to the Karlin-Rubin Theorem (e.g., Theorem 8.3.17 of [2]). Thus, we can equivalently consider testing

$$H_0^{(j)}: \theta_j = \theta_0^{(j)} \text{ vs. } H_A^{(j)}: \theta_j = \theta_1^{(j)}, \quad (2)$$

for $j = 1, \ldots, d$. This also applies to all the one-sided Z-tests by Pocock, O'Brien and Fleming, or Wang and Tsiatis, under the canonical conditions ([10], sec. 3.1)

No assumption about the joint distribution of d components of $\boldsymbol{X}_n$ is made. Special methods based on the joint distribution of $(X_n^{(1)}, \ldots, X_n^{(d)})$ may be designed to improve performance of the proposed tests, however, we generally assume that this joint distribution is unknown.

Our group sequential testing solution $\mathcal{T} = (m, T, \delta)$ consists of three components:

(1) m, the number of sampling units in each sampled group;

(2) T, the stopping time;

(3) $\delta = (\delta_1, \ldots, \delta_d)$, the decision rule on the acceptance or rejection of each null hypothesis. That is, we need to decide how we sample, when we stop, and what decisions we take after that on each of the tested hypothesis.

For any $\alpha \in (0,1)$, $\beta \in (0,1)$, and $K = 1, 2, \ldots$, our goal to guarantee

$$FWER_I \le \alpha, \quad FWER_{II} \le \beta, \quad \mathbf{P}\{T \le K\} = 1, \quad (3)$$

with an objective to minimize the overall expected cost $\mathbf{E}(mT)$ of the experiment. Since α-level tests of significance in group sequential experiments are already well developed, we start by building a simple Bonferroni multiple testing procedure based on these tests.

2.2 (Generalized) Bonferroni procedure

Assume that there exists a group sequential test $\mathcal{T}_j = (m_j, T_j, \delta_j)$ for testing $H_0^{(j)}$ versus $H_A^{(j)}$ with Type I and Type II error probabilities α_j and β_j and a maximum number of groups K. These can be Pocock, O'Brien-Fleming, or Wang-Tsiatis one-sided tests [10]. To construct a generalized Bonferroni procedure for testing (1), choose α_j and β_j such that $\sum \alpha_j = \alpha$ and $\sum \beta_j = \beta$. A typical choice is $\alpha_j = \alpha/d$ and $\beta_j = \beta/d$ for all $j = 1, \ldots, d$, but this is not necessary. Then, let $m = \max\{m_1, \ldots, m_d\}$ and $T = \max\{T_1, \ldots, T_d\}$.

Upon termination of the sampling procedure at time T, the decision on the j-th null hypothesis $H_0^{(j)}$ is taken according to the corresponding j-th test statistic at time T_j. If it belongs to the rejection region for the j-th test, reject $H_0^{(j)}$. If it belongs to the acceptance region, accept $H_0^{(j)}$. Notice that at time T_j, it this statistic to land in either one of this regions.

Theorem 1. *The generalized Bonferroni test $\mathcal{T}_{BONF} = (m, T, \delta)$ satisfies the three conditions in (3).*

Proof. For the introduced test, $T \le T_j \le K$ for all $j = 1, \ldots, d$. Also, using the Bonferroni (Boole) inequality,

$$\begin{aligned} FWER_I &= \mathbf{P}\left(\bigcup_{j=1}^{d} \left\{\text{Type I error on } H_0^{(j)}\right\} \right) \\ &\le \sum_{j=1}^{d} \mathbf{P}\left\{ \text{Type I error on } H_0^{(j)} \right\} \\ &= \sum_{j=1}^{d} \alpha_j = \alpha, \end{aligned}$$

and similarly, $FWER_{II} \le \beta$. □

Table 1 shows performance of the proposed Bonferroni group sequential with $d = 4$ tests, error rates limited by $\alpha = 0.05$, $\beta = 0.10$, and a maximum of $K = 6$ sampled groups. Null hypotheses $\theta^{(j)} = 0$ are tested against $\theta^{(j)} = 0.5$ for sequences of Normal$(\theta^{(j)}, \sigma = 1.2)$ responses.

Number of true null hypotheses	Expected number of groups	Expected sample size	Type I familywise error rate	Type II familywise error rate
Pocock based group sequential design				
0	4.31	86.19	0	0.0937
1	5.98	119.57	0.0131	0.0710
2	6	120	0.0248	0.0484
3	6	120	0.0377	0.0270
4	6	120	0.0481	0
O'Brien-Fleming based group sequential design				
0	4.90	88.27	0	0.0779
1	5.99	107.91	0.0109	0.0621
2	6	107.99	0.0231	0.0410
3	6	108	0.0368	0.0190
4	6	108	0.0486	0

Table 1: Performance characteristics of the Bonferroni group sequential multiple testing procedure with $d = 4$ tests. Nominal familywise error rates are $\alpha = 0.05$ and $\beta = 0.10$.

It can be seen that for a small number of simultaneous tests $j = 4$, the familywise error rates are close to their nominal desired values $\alpha = 0.05$ and $\beta = 0.10$ only when either all the null hypotheses are true or all the alternative hypotheses are true. When only a portion of $H_0^{(j)}$ or $H_A^{(j)}$ are true, there are fewer opportunities to commit a Type I or a Type II error, and therefore, the familywise error rates are substantially lower than the nominal α and β.

Results for $d = 25$ simultaneous tests are summarized in Table 2. Certainly, testing a larger number of hypotheses requires a larger group size. However, we can see that with at least two true null hypotheses, the expected sample size $\mathbf{E}(mT)$ is very close to its allowed limit (mK), and $\mathbf{E}(T) \approx K = 6$. Indeed, in this case, there is a very high probability to accept at least one of these hypotheses. Since acceptance can only occur at the last interim point $T = K$ for the considered Pocock and O'Brien-Fleming tests, there is a very high chance to sample all K groups until the last allowed interim point. For example, when responses are independent, and just two of $d = 25$ null hypotheses are true, sampling does not stop until the last interim point with probability

$$\mathbf{P}\{T = K\} \geq \mathbf{P}\left\{\begin{array}{c}\text{at least one true } H_0^{(j)} \\ \text{is accepted}\end{array}\right\}$$
$$\geq 1 - (\beta/d)^2 = 1 - (0.10/25)^2 = 0.999984.$$

Number of true null hypotheses	Expected number of groups	Expected sample size	Type I familywise error rate	Type II familywise error rate
Pocock based group sequential design				
0	5.04	171.28	0	0.0847
1	5.99	203.94	0.0022	0.0757
2	6	204	0.0041	0.0685
...	...	...	...	...
24	6	204	0.0456	0.0026
25	6	204	0.0470	0
O'Brien-Fleming based group sequential design				
0	5.50	164.89	0	0.0868
1	6	179.99	0.0021	0.0841
2	6	180	0.0039	0.0831
...	...	...	...	...
24	6	180	0.0479	0.0035
25	6	180	0.0520	0

Table 2: Performance characteristics of the Bonferroni group sequential multiple testing procedure with $d = 25$ tests. Nominal familywise error rates are $\alpha = 0.05$ and $\beta = 0.10$.

2.3 Conclusions and Further Steps

Bonferroni inequality is known to be rather crude for moderate to large number of components d. Therefore, although the Bonferroni method remains the most popular adjustment for multiplicity in simultaneous inferences, there is a room for improvement.

For the case of independent components of $\boldsymbol{X}_n$, the familywise error rates are not far from their nominal levels when $H_0^{(j)}$ are either all true or all false. For other situations, performance of Bonferroni procedures can be noticeably improved. Also, for the

responses collected from the same patient, the independence assumption is not valid, and thus, we do not assume fully known joint distributions.

In the next sections, advanced methods are developed that outperform the Bonferroni procedure in terms of requiring a smaller expected sample size under conditions (3). This improvement is attained by utilizing Holm-type stepwise procedures of [6] in group sequential framework. Implementation of this method requires the use of likelihood ratios and analytically tractable stopping boundaries. Thus, we start by developing a truncated sequential probability ratio test for testing a single hypothesis ($d = 1$) that controls both probabilities of Type I and Type II errors.

3 Truncated SPRT and Optimal Error Spending

In this section, we derive a truncated version of Wald's single-hypothesis sequential probability ratio test. Later, we extend this test to the case of multiple hypotheses and build a truncated sequential stepwise procedure for multiple testing.

3.1 A single-hypothesis truncated test

Suppose we observe a sequence of independent random variables $X_1, X_2, \ldots$ with the common density $f(x|\theta)$, where θ is the parameter of interest. Our immediate goal is to test

$$H_0 : \theta \le \theta_0 \text{ vs. } H_A : \theta \ge \theta_1 \qquad (4)$$

sequentially, where the data are collected in groups of size m, with $\boldsymbol{X}_k = (X_{(k-1)m+1}, \ldots, X_{km})$ being the kth group. The test has to control probabilities of Type I and Type II errors so that

$$\begin{aligned} &\mathbf{P}\{\text{Reject } H_0 \mid H_0\} \le \alpha \\ &\mathbf{P}\{\text{Do not reject } H_0 \mid H_A\} \le \beta \end{aligned}$$

for given $\alpha, \beta \in (0, 1)$.

The classic Wald's sequential probability ratio test (SPRT) is based on log-likelihood ratios $\Lambda_k = \log\{f(\boldsymbol{X}_k \mid \theta_1)/f(\boldsymbol{X}_k \mid \theta_0)\}$ and stopping boundaries $a' = \log\{(1-\beta)/\alpha\}$ and $b' = \log\{\beta/(1-\alpha)\}$. The SPRT stopping time is

$$T_{\text{SPRT}} = \min\{n \ge 1 : \ \Lambda_k \notin (b', a')\} \qquad (5)$$

with the decision rule

$$\delta_{\text{SPRT}} = \begin{cases} \text{reject } H_0 & \text{if } \Lambda_{T_{\text{SPRT}}} \ge a', \\ \text{accept } H_0 & \text{if } \Lambda_{T_{\text{SPRT}}} \le b'. \end{cases}$$

The SPRT controls the error probabilities approximately, subject to the *Wald approximation* that ignores the overshoot and assumes the test statistic Λ_k stops precisely at the stopping boundary (e.g., [8, 15]). In multiple testing, however, not all the log-likelihood ratio statistics are near their respective boundaries at their common stopping time, and the Wald approximation becomes inaccurate. Therefore, we correct the stopping boundaries to $a = -\log\alpha$ and $b = \log\beta$, which guarantees *exact* control of both error probabilities, without an approximation; see [6] for a proof.

Wald's SPRT is *open-ended*, the stopping time T is not a bounded random variable. We modify it in such a way that $T \le K$ with probability one, and the probabilities of Type I and Type II errors are still controlled at the given levels of α and β. If sampling continues until the last allowed interim point K, then a decision has to be made at K. Thus, in addition to the *stopping boundaries*, a new *decision boundary* is introduced, which defines acceptance and rejection regions at time K.

Our proposed procedure, the *truncated sequential probability ratio test* or TSPRT, is derived from the Wald's SPRT for a single hypothesis. It is also based on log-likelihood ratios

$$\Lambda_k = \log \frac{f(\boldsymbol{X}_k|\theta_1)}{f(\boldsymbol{X}_k|\theta_0)}.$$

The stopping time is now the first time the test statistic crosses the boundary, or time K, whichever occurs first,

$$T = \min\left(K, \ \min\{k : \ \Lambda_k \notin (b, a)\}\right),$$

and now it is bounded by K, i.e. $T \le K$ almost surely.

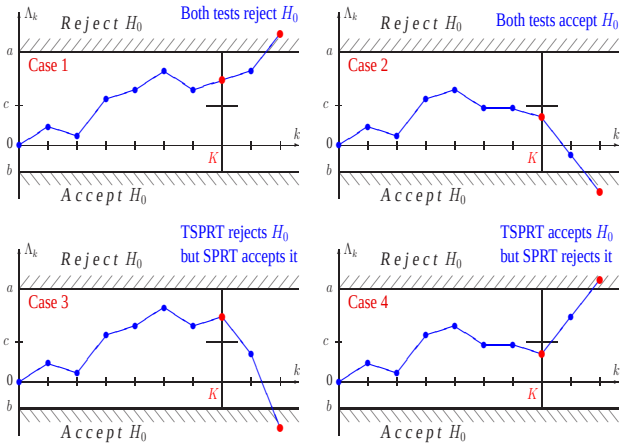

Figure 1: Truncated SPRT and open-ended SPRT. Cases 1-2: same conclusion (top row). Cases 3-4: different conclusions (bottom row)

Before the truncated time K, the terminal decision coincides with the Wald's SPRT. A new decision rule is only needed at K so that the choice between H_0 and H_A is made at time K even if the SPRT has not stopped.

The *decision rule* at the truncation point K will be determined by a *decision boundary* $c \in (b, a)$. At this time, the null hypothesis is rejected if Λ_K is between c and a, and it is accepted if Λ_K is between b and c. The decision rule δ of the truncated SPRT is then defined as follows,

$$\delta = \begin{cases} \text{reject } H_0 & \text{if } \Lambda_T \geq a \text{ or } \Lambda_T \geq c \cap T = K, \\ \text{accept } H_0 & \text{if } \Lambda_T \leq b \text{ or } \Lambda_T < c \cap T = K. \end{cases}$$

3.2 Comparison of SPRT and TSPRT

Derivation of stopping and decisions boundaries of TSPRT that guarantee control of Type I and Type II error probabilities will be done through its comparison with the standard Wald SPRT. Using the same stopping boundaries for both tests, we identify five possible situations.

Case 0, early stopping. When $T_{\text{SPRT}} < K$, regardless of the boundary that is crossed, TSPRT has the same stopping time and the same terminal decision as the SPRT.

The other four cases correspond to the case when SPRT continues beyond time K, and therefore, $T = K < T_{\text{SPRT}}$. These cases are shown on Figure 1.

Case 1, the same decision. On Figure 1, top left, the two graphs show that SPRT and TSPRT result in the same decision. Truncated at K, with $\Lambda_K > c$ implies that H_0 is rejected by the TSPRT. The SPRT continues sampling until Λ_k crosses the boundary. Here $\Lambda_{T_{\text{SPRT}}} \geq a$, so the SPRT also rejects H_0.

Case 2, the same decision. Similarly, the top-right graph on Figure 1 shows that at time K, $\Lambda_K < c$ and TSPRT accepts H_0 whereas the SPRT accepts H_0 finally, at time T_{SPRT}. Therefore, both the truncated SPRT and the classical SPRT fail to reject H_0.

Cases 0-2 do not generate any difference between SPRT and TSPRT in their probabilities of Type I error or Type II error. The difference can only be generated by the situations shown in the bottom row of Figure 1.

Case 3, TSPRT has a higher probability of Type I error. On Figure 1, bottom left, the $\Lambda_{T_{\text{SPRT}}} \leq b$, so SPRT accepts the null hypothesis. At the same time, Λ_K falls between c and a, therefore, the TSPRT rejects H_0. This case represents all the realizations of the stochastic process $\{\Lambda_k\}$ where the TSPRT rejects H_0 whereas the SPRT does not.

Case 4, TSPRT has a higher probability of Type II error. Similarly, on Figure 1, bottom right, $\Lambda_{T_{\text{SPRT}}} \geq a$, SPRT rejects H_0, while $\Lambda_K \in (b, c)$, and thus, the TSPRT accepts H_0. This case results in a lower probability of Type I error and a higher probability of Type II error by the TSPRT, relative to the SPRT.

We see that only cases 3-4 cause difference between error probabilities of SPRT and TSPRT. Estimating probabilities of these cases, we derive the stopping and decision boundaries for the TSPRT that control both Type I and Type II errors.

3.3 Decision boundaries that control error probabilities

The plan is to split the overall significance level α into α^*, the probability of committing a Type I error while stopping before time K, and $(\alpha - \alpha^*)$, the probability of a Type I error at the termination point K. Depending on α^*, a rejection boundary c_1 at time K is chosen as a function of α, α^*, and the group size

m.

Similarly, we split the desired probability of Type II error β into β^* and $(\beta-\beta^*)$ and choose the acceptance boundary c_2 at time K. From these, we deduce a decision boundary c that serves at the same time as the rejection and the acceptance boundary and controls both error probabilities simultaneously. This requires a certain minimum group size m that will then be calculated.

The last step is to optimize the test in terms of α^* and β^*. That is, we compute the optimal *error spending* that minimizes the expected sample size required for this group sequential test.

Let us derive the decision boundary c_1 that controls the Type I error. Derivation of c_2 for the Type II error is similar. We start by constructing an SPRT with a significance level $\alpha^* \in (0,\alpha)$ which will be selected later. Comparison of TSPRT and SPRT in Section 3.3 shows that

$$\begin{aligned}
&\mathbf{P}\{\text{Type I error by TSPRT}\}\\
&= \mathbf{P}_{H_0}\{\text{Type I error by SPRT}\}\\
&\quad + \mathbf{P}_{H_0}(\{\Lambda_1,...,\Lambda_{K-1}\in(b,a)\}\cap\{\Lambda_K\in[c_1,a)\}\\
&\qquad \cap\{\Lambda_k\le b \text{ before } \Lambda_k\ge a\})\\
&\quad - \mathbf{P}_{H_0}(\{\Lambda_1,...,\Lambda_{K-1}\in(b,a)\}\cap\{\Lambda_K\in(b,c_1)\}\\
&\qquad \cap\{\Lambda_k\ge a \text{ before } \Lambda_k\le b\})\\
&\le \alpha^* + \mathbf{P}_{H_0}\{\text{Case 3}\} - \mathbf{P}_{H_0}\{\text{Case 4}\}\\
&\le \alpha^* + \mathbf{P}_{H_0}\{\text{Case 3}\}.
\end{aligned} \tag{6}$$

In order to bound $\mathbf{P}\{\text{Case 3}\}$, consider the sequence of SPRT statistics that starts at time $k=K$ with the initial value of Λ_K, marked with a red dot on Figure 1, bottom left. The probability that it results in the Type I error for any $\Lambda_K\in[c_1,a)$ is bounded from above by the case when $\Lambda_K=c_1$.

This is equivalent to the SPRT that starts at the value $\Lambda_0=c_1$ and has stopping boundaries $(a-c_1)$ and $(b-c_1)$. For the latter SPRT, we find the probability of Type I error from the classical results of [19, 20] as

$$\begin{aligned}
&\mathbf{P}_{H_0}\{\Lambda_k\ge a-c_1 \text{ before } \Lambda_k\le b-c_1\}\\
&\quad = \frac{\alpha^* e^{c_1}-\alpha^*\beta}{1-\alpha^*\beta},
\end{aligned}$$

from where

$$\begin{aligned}
&\mathbf{P}_{H_0}\{\text{Case 3} \mid \Lambda_K\in[c_1,a)\}\\
&\quad\le \mathbf{P}_{H_0}\{\text{Case 3} \mid \Lambda_K=c_1\}\\
&\quad\le 1-\frac{\alpha^* e^{c_1}-\alpha^*\beta}{1-\alpha^*\beta}=\frac{1-\alpha^* e^{c_1}}{1-\alpha^*\beta}.
\end{aligned} \tag{7}$$

In general, the probability of the condition in (7) can be bounded by large-deviation inequalities for random walks such as the Chernoff inequality,

$$\begin{aligned}
&\mathbf{P}_{H_0}\{\Lambda_K\in[c_1,a)\}\le \mathbf{P}_{H_0}\{\Lambda_K\ge c_1\}\\
&\quad\le \mathbf{P}_{H_0}\left\{\sum_{i=1}^{Km}\log\frac{f(X_i\mid\theta_1)}{f(X_i\mid\theta_0)}\ge c_1\right\}\\
&\quad\le \inf_{t>0} e^{-tc_1}\phi_0^{Km}(t),
\end{aligned} \tag{8}$$

where $\phi_0(t)$ is the moment generating function of log-likelihood ratios $\log\{f(X_i\mid\theta_1)/f(X_i\mid\theta_0)\}$ under H_0. Use (7) and inequality (8) to bound the probability $\mathbf{P}\{\text{Case 3}\}$ and then obtain the *minimum rejection boundary* c_1 as a solution of the equation

$$\frac{1-\alpha^* e^{c_1}}{1-\alpha^*\beta}\inf_{t>0} e^{-tc_1}\phi_0^{Km}(t)=\alpha-\alpha^*. \tag{9}$$

Sharper bounds can be obtained for specific models. For example, the Chernoff inequality is redundant when the data follow a Normal(θ,σ) distribution. In this case, Λ_K is Normal with $\mathbf{E}(\Lambda_K)=-Km(\theta_1-\theta_0)^2/(2\sigma^2)$ and variance $\mathrm{Var}(\Lambda_K)=Km(\theta_1-\theta_0)^2/\sigma^2$ under H_0, so that

$$\begin{aligned}
&\mathbf{P}_{H_0}\{\Lambda_K\ge c_1\}\\
&\quad = \Phi\left(-\frac{c_1+Km(\theta_1-\theta_0)^2/(2\sigma^2)}{\sqrt{Km}|\theta_1-\theta_0|/\sigma}\right),
\end{aligned}$$

where $\Phi(\cdot)$ is the Standard Normal cumulative distribution function. The minimum rejection boundary c_1 in this case is found as a solution of equation

$$\frac{1-\alpha^* e^{c_1}}{1-\alpha^*\beta}\,\Phi\left(-\frac{2\sigma^2 c_1+Km(\theta_1-\theta_0)^2}{2\sigma\sqrt{Km}|\theta_1-\theta_0|}\right)=\alpha-\alpha^*. \tag{10}$$

Theorem 2. (Control of the Type I error probability) *Consider the truncated group sequential procedure for testing (4) with the stopping time (5),*

group size m, truncation point K, rejection stopping boundary $a = -\log\alpha^*$ for $k < K$, acceptance stopping boundary $b = \log\beta$ for $k < K$, with arbitrary $\beta \in (0,1)$, and decision boundary c_1 for $k = K$ given as a solution of (9) in the general case and (10) for the case of a Normal distribution.

The probability of Type I error of this testing procedure is controlled at level α.

Proof. The Type I error can occur either before the truncation point K or at K. If $T < K$, the TSPRT stops at the same time as the SPRT and results in the same decision. In the case of $T = K$, the probability of sampling K groups and making a Type I error is bounded by $(\alpha - \alpha^*)$, according to (9) and (10). Therefore,

$$\begin{aligned}
&\mathbf{P}\{\text{Type I error}\}\\
&= \mathbf{P}\{T < K \cap \text{Type I error}\}\\
&\quad + \mathbf{P}\{T = K \cap \text{Type I error}\}\\
&\le \alpha^* + (\alpha - \alpha^*) = \alpha.
\end{aligned}$$

□

A similar method can be used to control the *probability of Type II error.* This time, we compare the proposed TSPRT against the SPRT with stopping boundaries $a = -\ln\alpha$ and $b = \ln\beta^*$. This SPRT controls the probability of Type I error at the level α and the probability of Type II error at the level β^*. Hence,

$$\begin{aligned}
&\mathbf{P}\{\text{Type II error by TSPRT}\}\\
&= \mathbf{P}_{H_A}\{\text{Type II error by SPRT}\}\\
&\quad + \mathbf{P}_{H_A}\{\{\Lambda_1, ..., \Lambda_{K-1} \in (b,a)\} \cap \{\Lambda_K \in (b, c_2)\}\\
&\qquad \cap \{\Lambda_k \ge a \text{ before } \Lambda_k \le b\}\}\\
&\quad - \mathbf{P}_{H_A}\{\{\Lambda_1, ..., \Lambda_{K-1} \in (b,a)\} \cap \{\Lambda_K \in [c_2, a)\}\\
&\qquad \cap \{\Lambda_k \le b \text{ before } \Lambda_k \ge a\}\}\\
&\le \beta^* + \mathbf{P}_{H_A}\{\text{Case 4}\} - \mathbf{P}_{H_A}\{\text{Case 3}\}\\
&\le \beta^* + \mathbf{P}_{H_A}\{\text{Case 4}\}.
\end{aligned}$$

Our goal is to choose the acceptance boundary c_2 to guarantee $P_{H_A}(\text{Case 4}) \le \beta - \beta^*$. Then, similarly to Theorem 2, we obtain

$$\mathbf{P}\{\text{Type II error by TSPRT}\} \le \beta.$$

The probability of Case 4 can be bounded as

$$\begin{aligned}
&\mathbf{P}_{H_A}\{\text{Case 4}\}\\
&\le \mathbf{P}_{H_A}\{\Lambda_K \in (b, c_2)\}\\
&\quad \times \mathbf{P}_{H_A}\{\Lambda_k \ge a \text{ before } \Lambda_k \le b \mid \Lambda_K < c_2\}\\
&\le \mathbf{P}_{H_A}\{\Lambda_K < c_2\}\\
&\quad \times \mathbf{P}_{H_A}\{\Lambda_k \ge a \text{ before } \Lambda_k \le b \mid \Lambda_K = c_2\},
\end{aligned}$$

where, similarly to the bounding of the Type I error probability,

$$\begin{aligned}
&\mathbf{P}_{H_A}\{\Lambda_k \ge a \text{ before } \Lambda_k \le b \mid \Lambda_K = c_2\}\\
&= 1 - \frac{\beta^* e^{-c_2} - \alpha\beta^*}{1 - \alpha\beta^*} = \frac{1 - \beta^* e^{-c_2}}{1 - \alpha\beta^*}
\end{aligned}$$

from the SPRT that starts from (K, Λ_K), and

$$\mathbf{P}_{H_A}\{\Lambda_K \le c_2\} \le \inf_{t>0} e^{tc_2}\phi_1^{Km}(t), \qquad (11)$$

by the Chernoff inequality, with $\phi_1(t)$ being the moment generating function of the marginal log-likelihood ratio under the alternative hypothesis H_A. The maximum acceptance boundary c_2 is then computed as the solution of the equation,

$$\frac{1 - \beta^* e^{c_2}}{1 - \alpha\beta^*} \inf_{t>0} e^{tc_2}\phi_1^{Km}(t) = \beta - \beta^*. \qquad (12)$$

For the case of Normal distributions,

$$\mathbf{P}_{H_A}\{\Lambda_K < c_2\} = \Phi\left(\frac{c_2 - Km(\theta_1 - \theta_0)^2/(2\sigma^2)}{\sqrt{Km}|\theta_1 - \theta_0|/\sigma}\right),$$

where $\Phi(\cdot)$ is the Standard Normal cumulative distribution function. In this case, c_2 solves

$$\frac{1 - \beta^* e_2^c}{1 - \alpha\beta^*}\,\Phi\left(\frac{2\sigma^2 c_2 - Km(\theta_1 - \theta_0)^2}{2\sigma\sqrt{Km}|\theta_1 - \theta_0|}\right) = \beta - \beta^*. \qquad (13)$$

Then, the control of Type II error probability by TSPRT is guaranteed by the following Theorem.

Theorem 3. (Control of the Type II error probability) *Consider the truncated group sequential procedure for testing (4) with the stopping time (5), group size m, truncation point K, rejection stopping boundary $a = -\log\alpha$ for $k < K$, with arbitrary $\alpha \in (0,1)$, acceptance stopping boundary $b = \log\beta^*$ for $k < K$, and decision boundary c_2 for $k = K$, given as a solution of (12) in the general case and (13) for the case of Normal distributions.*

The probability of Type II error of this testing procedure is controlled at level β.

Proof. This theorem is proven along the steps similar to the proof of Theorem 2. Details are omitted. □

3.4 Group size and simultaneous control of both error probabilities

It remains to combine the testing schemes defined in Theorems 2 and 3 to obtain a truncated sequential procedure that control *both* Type I and Type II error probabilities.

To achieve the simultaneous control of both error probabilities, we split $\alpha = \alpha^* + (\alpha - \alpha^*)$ and $\beta = \beta^* + (\beta - \beta^*)$ and apply the TSPRT described above with stopping boundaries $a = -\log \alpha^*$ and $b = \log \beta^*$ and decision boundaries c_1 and c_2.

However, the smallest rejection boundary c_1 guaranteeing $\mathbf{P}\{A\} \le \alpha - \alpha^*$ may still exceed the largest acceptance boundary c_2 that guarantees $\mathbf{P}\{B\} \le \beta - \beta^*$. We cannot afford $c_1 > c_2$ because in this case the test results in no decision in the case when $\Lambda_k \in (b, a)$ for $k < K, \Lambda_K \in (c_2, c_1)$.

Therefore, we choose the suitable *group size* m in order to have $c_1 \le c_2$. This is similar to finding the fixed-sample size required to achieve a certain power of a test. But does such m exist?

Lemma 1. *For a sufficiently large group size m, we have $c_1 \le c_2$, and therefore, the union of the rejection region and the acceptance region for the statistic Λ_K at the truncation point K is $\mathbb{R}$.*

Proof. Under H_0, the log-likelihood ratio statistic Λ_K has a negative expected value $\mathbf{E}_{H_0}\Lambda_K = -mK \cdot K(\theta_0, \theta_1)$, where $K(\theta_0, \theta_1)$ is the Kullback-Leibler information number. Then, by the (weak) Law of Large Numbers, as the group size m increases to ∞, we have

$$\mathbf{P}\{\text{Type I error at } T = K\} \le \mathbf{P}_{H_0}\{\Lambda_K \ge 0\} \to 0$$

for any $c_1 \ge 0$. Similarly, under H_A, the expected value $\mathbf{E}_{H_A}\Lambda_K = mK \cdot K(\theta_1, \theta_0) > 0$, and

$$\mathbf{P}\{\text{Type II err. at } T = K\} \le \mathbf{P}_{H_A}\{\Lambda_K \le 0\} \to 0$$

for any $c_2 \ge 0$. Therefore, there exists a group size m that is so large that for $c_1 = c_2 = 0$,

$$\begin{aligned} \mathbf{P}\{\text{Type I error at } T = K\} &\le \alpha - \alpha^*, \\ \mathbf{P}\{\text{Type II error at } T = K\} &\le \beta - \beta^*. \end{aligned}$$

This is a possible (but not necessarily optimal) choice of $c_1 \le c_2$. □

We summarize the obtained procedure in the following theorem.

Theorem 4. *The truncated group sequential procedure satisfying Theorems 2 and 3 with error spending α^*, $(\alpha - \alpha^*)$, β^*, and $(\beta - \beta^*)$ and the group size m guaranteed by Lemma 1 controls the probability of Type I error at level α and the probability of Type II error at level β.*

To optimize even further, the expected overall sample size $\mathbf{E}(mT)$ can now be numerically minimized in terms of $\alpha^* \in (0, \alpha)$ and $\beta^* \in (0, \beta)$.

3.5 Two-stage design and the Pollock-Golhar method

As an illustrative example, we consider the case of Normal responses and $K = 2$, which becomes the classical *two-stage group sequential design.*

Stage 1, interim point. After sampling the first group of m units, the decision rule is:

$$\begin{cases} \text{Stop and reject } H_0 & \text{if } \Lambda_1 \ge a, \\ \text{Stop and accept } H_0 & \text{if } \Lambda_1 \le b, \\ \text{Sample the 2nd group} & \text{if } \Lambda_1 \in (b, a). \end{cases}$$

One can control the error probabilities at the first stage at levels α^* and β^* by choosing $a = -\log(\alpha^*)$ and $b = \log(\beta^*)$, as in the general case above. However, *precise boundaries*

$$\begin{aligned} a &= -\Phi^{-1}(\alpha^*)\delta\sqrt{m} - m\delta^2/2, \\ b &= \Phi^{-1}(\beta^*)\delta\sqrt{m} + m\delta^2/2 \end{aligned} \qquad (14)$$

can be obtained from the Normal$(\pm m\delta^2/2, m\delta^2)$ distribution of Λ_1 under $H_{0,A}$, where $\delta = |\theta_1 - \theta_0|/\sigma$ is the standardized difference between the null and alternative parameters.

Final stage 2. If the second group is sampled, the trial stops with the following conclusion,

$$\begin{cases} \text{Stop and reject } H_0 & \text{if } \Lambda_2 \ge c, \\ \text{Stop and accept } H_0 & \text{if } \Lambda_2 < c. \end{cases}$$

Table 3: Comparison with the Pollock-Golhar test under H_0.

(α, β)	δ	Pollock-Golhar test			Zhao-Baron test			
		m	$E_{\theta_0}(mT)$	$P(T=K)$	(α^*, β^*)	m	$E_{\theta_0}(mT)$	$P(T=K)$
(0.01, 0.01)	0.5	55	69.85	0.270	(0, 0.005)	49	57.03	0.164
(0.02, 0.02)	0.5	40	54.40	0.360	(0, 0.009)	38	46.80	0.232
(0.05, 0.05)	0.5	24	35.52	0.480	(0, 0.020)	25	32.98	0.319
(0.01, 0.02)	0.5	48	60.96	0.270	(0, 0.009)	43	50.49	0.174
(0.05, 0.10)	0.5	22	31.02	0.410	(0, 0.036)	21	26.68	0.270
(0.01, 0.01)	0.2	275	382.25	0.390	(0, 0.005)	301	356.44	0.184
(0.02, 0.02)	0.2	215	309.60	0.440	(0, 0.009)	236	292.52	0.239
(0.05, 0.05)	0.2	139	211.28	0.520	(0, 0.020)	156	206.69	0.325
(0.01, 0.02)	0.2	274	361.68	0.320	(0, 0.009)	267	315.54	0.182
(0.05, 0.10)	0.2	137	193.17	0.410	(0, 0.036)	126	166.72	0.323

The expected stopping time for the two-stage procedure is $\mathbf{E}(T) = 1 + \mathbf{P}\{T > 1\}$, leading to the expected sample size

$$\begin{aligned}&\mathbf{E}_0(mT)\\ &= m + m\left(1 - \mathbf{P}_0\{\Lambda_1 \geq a\} - \mathbf{P}_0\{\Lambda_1 \leq b\}\right)\\ &= m\left(2 - \alpha^* - \Phi((m\delta^2 + b)/(\delta\sqrt{m}))\right) \quad (15)\end{aligned}$$

under $H_0 : \theta = \theta_0$ and

$$\begin{aligned}&\mathbf{E}_A(mT)\\ &= m + m\left(1 - \mathbf{P}_1\{\Lambda_1 \leq b\} - \mathbf{P}_1\{\Lambda_1 \geq a\}\right)\\ &= m\left(2 - \beta^* - \Phi((m\delta^2 - a)/(\delta\sqrt{m}))\right) \quad (16)\end{aligned}$$

under $H_A : \theta = \theta_1$.

In Tables 3-4, we compare expected sample sizes (15)-(16) with those of a similar truncated group sequential test of Pollock and Golhar [13] who elaborate a recursive method of finding the earliest truncation point to control the given error probabilities α and β.

Table 3 shows performance of both tests under H_0 for $\delta = 0.5$, $\delta = 0.2$, and various combinations of α and β. Table 4 shows performance characteristics under H_A. The columns containing $\mathbf{P}(T = K) = \mathbf{P}(\Lambda_1 \in (b, a))$ for the two methods show the probabilities of continuing sampling after the first group and requiring the maximum sample size.

Results imply that the TSPRT proposed in this section requires a smaller expected sample size than the Pollock-Golhar method under both H_0 and H_A. This is achieved by the optimal choice of the decision boundary c, the group size m, and the error spending α^* and β^*.

4 Stepwise Procedure for Multiple Testing

Now we use the test developed in Section 3 to derive a truncated sequential procedure for stepwise testing of multiple hypotheses (1) that controls both Type I and Type II familywise error rates as in (3) requiring a smaller sample size than the Bonferroni procedure.

4.1 Construction of a stepwise truncated group sequential test

Again, d tests about parameters $\theta_1, \ldots, \theta_d$ are conducted simultaneously based on sequentially collected data.

Derivation of the stepwise truncated sequential method for multiple testing is based on the Holm [9] approach that is adapted for purely sequential open-ended experiments in [6]. The test will be based on log-likelihood ratio statistics

$$\Lambda_k^{(j)} = \log \frac{f(X_1, \ldots, X_{mk} \mid \theta_1^{(j)})}{f(X_1, \ldots, X_{mk} \mid \theta_0^{(j)})},$$

for parameter $\theta^{(j)}$ at the interim point $k = 1, \ldots, K$, where m is the group size.

Table 4: Comparison with the Pollock-Golhar test under H_A.

(α, β)	δ	Pollock-Golhar test			Zhao-Baron test			
		m	$E_{\theta_1}(mT)$	$P(T=K)$	(α^*, β^*)	m	$E_{\theta_1}(mT)$	$P(T=K)$
(0.01, 0.01)	0.5	55	69.85	0.270	(0.005, 0)	49	57.03	0.164
(0.02, 0.02)	0.5	40	54.4	0.360	(0.009, 0)	38	46.80	0.232
(0.05, 0.05)	0.5	24	35.52	0.480	(0.020, 0)	25	33.07	0.323
(0.01, 0.02)	0.5	48	64.32	0.430	(0.005, 0)	43	53.27	0.239
(0.05, 0.10)	0.5	22	33.22	0.510	(0.018, 0)	20	28.53	0.427
(0.01, 0.01)	0.2	275	382.25	0.390	(0.005, 0)	301	356.44	0.184
(0.02, 0.02)	0.2	215	309.60	0.440	(0.009, 0)	236	292.52	0.239
(0.05, 0.05)	0.2	139	211.28	0.520	(0.020, 0)	156	206.69	0.325
(0.01, 0.02)	0.2	274	380.86	0.390	(0.005, 0)	268	332.92	0.242
(0.05, 0.10)	0.2	137	206.87	0.510	(0.018, 0)	123	178.32	0.450

Following the Holm [9] stepwise approach for controlling the Type I familywise error rate and its sequential adaptation in [6], at each interim point k, we order test statistics $\Lambda_k^{[1]} \geq ... \geq \Lambda_k^{[d]}$ and let $H_0^{[j]}$ be the tested null hypotheses that corresponds to $\Lambda_k^{[j]}$. Consider the *stopping boundaries*

$$a_j = -\log \frac{\alpha^*}{d-j+1}, \quad b_j = \log \frac{\beta^*}{j}, \qquad (17)$$

$a_1 \geq \ldots \geq a_d \geq b_1 \geq \ldots \geq b_d$, for $k < K$, $j = 1, \ldots, d$, $\alpha^* \in (0, \alpha)$, and $\beta^* \in (0, \beta)$, and *decision boundaries* at $k = K$,

$$c_1^{[1]} = -\log \frac{\alpha - \alpha^*}{d-j+1}, \quad c_2^{[j]} = \log \frac{\beta - \beta^*}{j}. \quad (18)$$

Comparing with the Bonferroni method, increasing the significance levels for the marginal Type I and Type II error probabilities, as in (17)–(18), causes the familywise error rates to increase. On the other hand, all the stopping boundaries in (17) become tighter than they are under the Bonferroni procedure, leading to widened acceptance and rejection regions, and ultimately, to a higher probability of early stopping. Thus, the stepwise method requires a smaller sample size $\mathbf{E}(T)$ under any combination of true and false hypotheses. In this sense, the resulting stepwise procedure is an improvement of Bonferroni schemes under the given constraints on familywise error rates.

In order to control both FWER_I and FWER_II, m is chosen to be the smallest group size satisfying

$$\bigcap_{j=1}^{d} [c_1^{(j)},\ c_2^{(j)}] \neq \varnothing,$$

and in this case, the decision boundary at the truncation point $k = K$ is any number c that belongs to the intersection

$$c \in \bigcap_{j=1}^{d} [c_1^{(j)},\ c_2^{(j)}]. \qquad (19)$$

Since the lower bound for $c_1^{(j)}$ tends to $-\infty$ and the upper bound for $c_2^{(j)}$ tends to ∞, as $m \to \infty$, the required group size m exists.

Sampling continues until the stopping time

$$T = \min\left\{K,\ \min\left(k :\ \bigcap_{j=1}^{d} \Lambda_k^{[j]} \in (b_j, a_j)\right)\right\}, \qquad (20)$$

where $\Lambda_k^{[j]}$ is the j-th largest log-likelihood ratio at time k. Then, if $T < K$, acceptance or rejection of each $H_0^{(j)}$ is decided according to whether $\Lambda_k^{[j]} \leq b_j$ or $\Lambda_k^{[j]} \geq a_j$. At $T = K$, all the null hypotheses $H_0^{(j)}$ corresponding to $\Lambda_k^{(j)} \geq c$ are rejected, and all the others are accepted.

4.2 Simultaneous control of error rates

Theorem 5. *The introduced stepwise truncated sequential multiple testing procedure with stopping*

boundaries (17), decision boundary (19), and stopping time (20) guarantees simultaneous control of the Type I familywise error rate at level α and the Type II familywise error rate at level β and stopping no later than at $T = K$.

Proof. Let us first establish the control of FWER_{I}. The proof of FWER_{II}control is similar. The main steps follow the ideas of [9] and [6], adapting them to our truncated sequential scheme.

Consider the null hypotheses $H_0^{[1]}, H_0^{[2]}, ..., H_0^{[d]}$, ordered according to the non-increasing order of the corresponding log-likelihood ratios $\Lambda_k^{(j)}$ at the interim point k. Let $J_0 \in \{1, ..., d\}$ be the index set of true null hypothesis, and let $j_0 = \min\{j : H_0^{[j]} \text{ is true}\}$ be the index of the largest and the most significant log-likelihood ratio at $k = T$, among all true null hypotheses. Then the number of false hypotheses is $|J_A| = d - |J_0| \geq j_0 - 1$. Thus, there are at least $(j_0 - 1)$ false hypotheses, which leads to the inequality $|J_0| \leq d - j_0 + 1$.

The first $(j_0 - 1)$ tests (in the non-increasing order of $\Lambda_T^{[j]}$) cannot result in a Type I error because the corresponding null hypotheses are false. Additionally, if there is no Type I error on the j_0-th test and $H_0^{[j_0]}$ is not rejected, then no $H_0^{[j]}$ is rejected for $j > j_0$, and therefore, there is no Type I error at all. Indeed, acceptance of $H_0^{[j_0]}$ implies that $\Lambda_T^{[j]} \leq \Lambda_T^{[j_0]} \leq b_{j_0} \leq a_j$ for $T < K$, and $\Lambda_T^{[j]} \leq \Lambda_T^{[j_0]} \leq c$ for $T = K$, and thus, $H_0^{[j]}$ is accepted for any $j > j_0$.

Hence, we have Type I errors if and only if there is a Type I error on $H_0^{[j_0]}$, so that

$$
\begin{aligned}
FWER_I &= \mathbf{P}\left\{H_0^{[j_0]} \text{ is rejected }\right\} \\
&= \mathbf{P}\left\{T < K \cap \Lambda_T^{[j_0]} \geq a_{j_0}\right\} \\
&\quad + \mathbf{P}\left\{T = K \cap \Lambda_K^{[j_0])} \geq c_1^{[j_0]}\right\} \\
&\leq \mathbf{P}\left\{T < K \cap \Lambda_T^{[j_0]} \geq -\log\frac{\alpha^*}{|J_0|}\right\} \\
&\quad + \mathbf{P}\left\{T = K \cap \Lambda_K^{[j_0]} \geq -\log\frac{\alpha - \alpha^*}{|J_0|}\right\} \quad (21)
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbf{P}\left\{T < K \cap \max_{j\in J_0} \Lambda_T^{[j]} \geq -\log\frac{\alpha^*}{|J_0|}\right\} \\
&\quad + \mathbf{P}\left\{T = K \cap \max_{j\in J_0} \Lambda_K^{(j)} \geq -\log\frac{\alpha - \alpha^*}{|J_0|}\right\} \\
&= \mathbf{P}\left\{T < K \cap \bigcup_{j\in J_0} \Lambda_T^{[j]} \geq -\log\frac{\alpha^*}{|J_0|}\right\} \\
&\quad + \mathbf{P}\left\{T = K \cap \bigcup_{j\in J_0} \Lambda_K^{[j]} \geq -\log\frac{\alpha - \alpha^*}{|J_0|}\right\} \\
&\leq \sum_{j\in J_0} \mathbf{P}\left\{T < K \cap \Lambda_T^{[j]} \geq -\log\frac{\alpha^*}{|J_0|}\right\} \\
&\quad + \sum_{j\in J_0} \mathbf{P}\left\{T = K \cap \Lambda_K^{[j]} \geq -\log\frac{\alpha - \alpha^*}{|J_0|}\right\} \\
&\leq \sum_{j\in J_0} \exp\left(\log\frac{\alpha^*}{|J_0|}\right) + \sum_{j\in J_0} \exp\left(\log\frac{\alpha - \alpha^*}{|J_0|}\right) \\
&= \alpha^* + (\alpha - \alpha^*) = \alpha. \quad (22)
\end{aligned}
$$

Here, formulas (17) and (18) are applied for a_{j_0} and $c_1^{[j_0]}$ in (21), as well as the inequality $d - j_0 + 1 \geq |J_0|$, established earlier. On the next line, we used the definition of index j_0. Finally, inequalities in (22) follow from Lemma 2 of [6] which states that $\mathbf{P}\{\Lambda_T^{(j)} \geq a \mid \theta_0^{(j)}\} \leq e^{-a}$ and $\mathbf{P}\{\Lambda_T^{(j)} \leq b \mid \theta_1^{(j)}\} \leq e^b$ for any stopping time T where $\Lambda_T^{(j)} \notin (b, a)$ with probability one, and the null hypothesis $H_0^{(j)}$ is accepted or rejected depending on whether $\Lambda_T^{(j)} \leq b$ or $\Lambda_T^{(j)} \geq a$.

Control of FWER_{II} is proven similarly, and truncation $\mathbf{P}\{T \leq K\} = 1$ is evident from (20). □

4.3 Performance evaluation and comparison

To evaluate the actual performance of the new stepwise truncated sequential procedure for multiple testing, we consider the same testing problem as in Section 2. That is, we conduct $d = 4$ tests of Normal means simultaneously, under the desired $FWER_I = 0.05$ and $FWER_{II} = 0.10$, with the maximum of $K = 6$ allowed groups of sampling units, where the difference between the null and alternative parameter values is $\theta_1^{(j)} - \theta_0^{(j)} = 0.5$ with a standard deviation of $\sigma = 1.2$.

Number of true null hypotheses	Expected number of groups	Expected sample size	Type I familywise error rate	Type II familywise error rate
Stepwise truncated group sequential procedure				
0	4.779	81.243	0	0.093
1	5.101	86.717	0.015	0.079
2	5.153	87.601	0.021	0.044
3	5.028	85.476	0.038	0.024
4	4.469	75.973	0.047	0
Bonferroni truncated group sequential procedure				
0	5.411	91.987	0	0.089
1	5.322	90.474	0.011	0.074
2	5.340	90.780	0.020	0.039
3	5.290	89.930	0.032	0.022
4	5.182	88.094	0.044	0

Table 5: Performance characteristics of the stepwise truncated group sequential multiple testing procedure with $d = 4$ tests. Nominal familywise error rates are $\alpha = 0.05$ and $\beta = 0.10$.

The optimal (rounded) group size in this case is $m = 17$, with error spending $\alpha = \alpha^* + (\alpha - \alpha^*) = 0.038 + 0.012$ and $\beta = \beta^* + (\beta - \beta^*) = 0.069 + 0.031$ and decision boundary $c = 0.72$.

Simulation results are summarized in Table 5. Both familywise error rates are controlled, as it is already proven in Theorem 5. At the same time, in terms of the expected sample size $\mathbf{E}(mT)$, there is a substantial improvement over the Bonferroni procedure based on the Pocock test and the O'Brien-Fleming test for the same testing problem (Table 1). The expected required sample size of our stepwise procedure is on the average reduced by 21 patients, which is an approximately 19% reduction comparing with the O'Brien-Fleming test with Bonferroni adjustment for multiple tests. Comparing with the Pocock test, the presented stepwise procedure results in an average saving of 30 patients, which is a 25% reduction.

An insightful comparison is with the Bonferroni procedure that is also based on the TSPRT, the truncated sequential probability ratio test elaborated in Chapter 3. Results for this test are given in the second portion of Table 5. The use of Bonferroni inequality, without the Holm's adjustment for the ordered statistics resulted in lower error rates than the required $\alpha = 0.05$ and $\beta = 0.10$. Our stepwise procedure also uses the Bonferroni inequality in the proof of Theorem 5, and its actual familywise error rates are also below α and β, however, it uses tighter boundaries. This is the consequence of dividing α and β levels by factors of $(d - j + 1)$ and j, respectively, instead of the constant Bonferroni factor of d. As a result, the difference between the actual and nominal error rates is not as large as in the Bonferroni case, the tighter boundaries are in general attained earlier, and for this reason, the expected sample size $\mathbf{E}(mT)$ is reduced.

Comparing with the Bonferroni procedure, our stepwise approach requires 6.8 patients fewer, on the average, which is approximately a 7.6% reduction.

5 Summary and Conclusions

Tests of multiple hypotheses in group sequential clinical trials are elaborated to control Type I and Type II familywise error rates while requiring at most K groups of patients. Open-ended sequential procedures such as SPRT are not practical in clinical trials. Instead, truncation techniques based on stepwise sequential procedures for multiple testing coupled with

optimal error spending are proposed.

Bonferroni approach satisfies the error rates and truncation requirements. However, it is not optimal due to the use of Bonferroni inequality which is not sharp for moderate to a large number of tests. There exist tests that require smaller samples while still controlling the error rates at the given levels.

Advanced methods based on Holm's stepwise approach are developed to enhance the performance of sequential multiple testing procedures. Among them, the truncated sequential probability

Ratio test is constructed with the optimal choice of stopping boundaries, a decision boundary at the truncation point, and a group size that controls $FWER_{I}$ and $FWER_{II}$.

As a by-product of optimal error spending between the truncation point and early stopping, the one-dimensional version of this test outperforms the existing Pollock-Golhar truncated SPRT in terms of requiring a smaller overall expected sample size.

This effort focused on optimizing the expected sample size results in lower costs of sequential clinical trials, and subsequently, in lower costs of medical treatments.

Acknowledgement

Research of both authors at the University of Texas at Dallas was funded by the National Science Foundation. This support is greatly appreciated.

References

1. Turk DC, Dworkin RH, Allen RR, Bellamy N, Brandenburg N et al. Core outcome domains for chronic pain clinical trials: IMMPACT recommendations. Pain. 2003, 106(3): 337–345.

2. Neuhauser M, Steindans VW, Bretz F. The evaluation of multiple clinical endpoints, with application to asthma. Drug Information Journal. 1999, 33(2): 471–477.

3. Castagno D, Jhund PS, McMurray JJ, Lewsey JD, Erdmann E et al. Improved survival with bisoprolol in patients with heart failure and renal impairment: an analysis of the cardiac insufficiency bisoprolol study II (CIBIS-II) trial. European Journal of Heart Failure. 2010, 12(6): 607–616.

4. Claggett B, Tian L, Castagno D, Wei LJ. Treatment selections using risk-benefit profiles based on data from comparative randomized clinical trials with multiple endpoints. Biostatistics. 2015, 16(1): 60–72.

5. Urschel HC, Hanselka LL, Baron M. A controlled trial of flumazenil and gabapentin for the initial treatment of methylamphetamine dependence. J of Psychopharmacology. 2011, 25(2): 254–262.

6. Urschel HC, Hanselka LL, Gromov I, White L, Baron M. Open-label study of a proprietary treatment program targeting type a γ-aminobutyric acid receptor dysregulation in methamphetamine dependence. Mayo Clinic Proceedings. 2007, 82(10): 1170–1178.

7. O'Brien PC, Fleming TR. A multiple testing procedures for clinical trials. Biometrika. 1979, 35(3): 549–556.

8. Edwards DG, Hsu JC. Multiple comparisons with the best treatment. J Amer Stat Assoc. 1983, 78(384): 965–971.

9. Jennison C, Turnbull BW. Group sequential methods with applications to clinical trials. Chapman & Hall, Boca Raton, FL, 2000.

10. Zacks S. Stage-wise Adaptive Designs. Wiley, Hoboken, NJ, 2009.

11. Rhodes D, Tomlins S, Williams P, Sadis S, Wyngaard P et al. P1-07-04: Gene expression module biomarkers to stratify multiple clinical and therapeutic endpoints for universal breast cancer companion diagnostic. Cancer Research. 2011, 71(24 Supplement): P1–07.

12. Wald A. Sequential Analysis. Wiley, New York, 1947.

13. Govindarajulu Z. Sequential Statistics. World Scientific Publishing Co, Singapore, 2004.

14. Tartakovsky AG, Nikiforov IV, Basseville M. Sequential Analysis: Hypothesis Testing and Change-Point Detection. Chapman & Hall/CRC, 2014.

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

16. De S, Baron M. Step-up and step-down methods for testing multiple hypotheses in sequential experiments. J Statist Plann Inference. 2012, 142(7): 2059–2070.

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

18. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979, 6(2): 65–70.

19. Pollock SM, Golhar DY. Efficient recursions for truncation of the SPRT. Sequential Analysis. 1986, 5(3): 253–262.

20. Casella G, Berger RL. Statistical Inference. Duxbury Press, Belmont, CA, 2002.

21. Wald A, Wolfowitz J. Optimal character of the sequential probability ratio test. Ann Math Statist. 1948, 19(3): 326–339.