General Multistage Gatekeeping Procedures

Alex Dmitrienko[*, 1], **Ajit C. Tamhane**[2], and **Brian L. Wiens**[3]

[1] Eli Lilly and Company, Indianapolis, IN 46285, USA
[2] Northwestern University, Evanston, IL 60208, USA
[3] Gilead Colorado, Westminster, CO 80021, USA

Received 29 October 2007, revised 21 February 2008, accepted 9 July 2008

Summary

A general multistage (stepwise) procedure is proposed for dealing with arbitrary gatekeeping problems including parallel and serial gatekeeping. The procedure is very simple to implement since it does not require the application of the closed testing principle and the consequent need to test all nonempty intersections of hypotheses. It is based on the idea of carrying forward the Type I error rate for any rejected hypotheses to test hypotheses in the next ordered family. This requires the use of a so-called separable multiple test procedure (MTP) in the earlier family. The Bonferroni MTP is separable, but other standard MTPs such as Holm, Hochberg, Fallback and Dunnett are not. Their truncated versions are proposed which are separable and more powerful than the Bonferroni MTP. The proposed procedure is illustrated by a clinical trial example.

Key words: Bonferroni test; Closed procedure; Fallback test; Hochberg test; Holm test; Multiple comparisons; Stepwise procedure; Truncated tests.

1 Introduction

Gatekeeping procedures have become popular in recent years as they provide a convenient way to handle logical relationships between multiple hierarchical objectives that clinical trials are often required to address. Questions concerning different objectives are formulated as hypotheses and the hierarchy of relationships is modeled by dividing them into ordered families. Suppose there are $n \geq 2$ hypotheses divided into $m \geq 2$ ordered families, F_i $(1 \leq i \leq m)$. Let $F_i = \{H_{i1}, \ldots, H_{in_i}\}$, where $\sum_{i=1}^{m} n_i = n$. Generally, familywise error rate (FWER) control at a designated level α is desired for the family $F = \bigcup_{i=1}^{m} F_i$ of all n hypotheses.

Westfall and Krishen (2001) proposed procedures for the serial gatekeeping problem in which the hypotheses in F_{i+1} are tested if and only if (iff) all hypotheses in F_i are rejected $(1 \leq i \leq m-1)$. Dmitrienko, Offen and Westfall (2003) proposed procedures for the parallel gatekeeping problem in which the hypotheses in F_{i+1} are tested iff at least one hypothesis in F_i is rejected $(1 \leq i \leq m-1)$. In both cases, if the specified rejection criterion for F_i is not satisfied then all hypotheses in F_j for $j > i$ are automatically accepted. Dmitrienko, Wiens, Tamhane and Wang (2007) proposed tree gatekeeping procedures which unify and generalize these two types of procedures (see also Dmitrienko, Tamhane, Liu and Wiens, 2008); the present paper does not cover general tree gatekeeping procedures. Dmitrienko and Tamhane (2007) have given a review of these recent developments.

A serial gatekeeping procedure tests the families F_i $(1 \leq i \leq m)$ sequentially, each at an α-level, using any suitable MTP (Maurer, Hothorn and Lehmacher, 1995). Testing stops as soon as at least one hypothesis in a family under test is not rejected. Thus these procedures are naturally stepwise and do not require further elucidation.

[*] Corresponding author: e-mail: dmitrienko alex@lilly.com, Phone: 317 277 1979, Fax: 317 277 3220.

To construct a parallel gatekeeping procedure, Dmitrienko et al. (2003) employed the closure principle of Marcus, Peritz and Gabriel (1976). This involves testing up to $2^n - 1$ intersection hypotheses. Each intersection hypothesis is tested using a weighted test where weights are assigned to the hypotheses in the intersection to take into account their a priori order. Dmitrienko et al. (2003) used the weighted Bonferroni MTP for testing intersection hypotheses. They gave a decision matrix approach to systematize the calculation of multiplicity adjusted p-values for all hypotheses from which acceptance/rejection decisions on them can be readily made. The main difficulty of this algorithm is its lack of transparency – not so much its computational complexity – which makes its decisions not easily interpretable to clinicians.

Dmitrienko, Tamhane, Wang and Chen (2006) showed that the closed parallel gatekeeping procedure described above simplifies to a stepwise procedure. In this procedure the first $m-1$ families are tested using the Bonferroni MTP which tests F_i at level $\varrho_i\alpha$ $(1 \le i \le m-1)$. The family F_m is tested at level $\varrho_m\alpha$ using the Holm (1979) MTP. Here ϱ_i is the so-called *rejection gain factor* for family F_i. If the hypotheses are equally weighted in each family then ϱ_i is given by

$$\varrho_1 = 1, \qquad \varrho_i = \prod_{j=1}^{i-1} \left(\frac{r_j}{n_j}\right), \quad (2 \le i \le m), \tag{1}$$

where r_j is the number of rejected hypotheses in F_j; thus ϱ_i is the product of the proportions of rejected hypotheses in F_1 through F_{i-1}. Note that rejection criterion becomes more stringent in later families if fewer hypotheses are rejected in earlier families. If no hypotheses are rejected in some family F_i, $\varrho_j = 0$ for all $j > i$. Therefore, all hypotheses in F_j for $j > i$ are automatically accepted, thus satisfying the parallel gatekeeping condition. On the other hand, if all hypotheses are rejected in F_1 through F_{i-1}, then $\varrho_i = 1$ and thus full α level is used to test F_i. In general, if the procedure does not use the fraction of α assigned to a particular hypothesis (i.e., this hypothesis is not rejected), this fraction is lost (i.e., it cannot be carried over to the next family). We refer to this as the "use it or lose it" principle.

Guilbaud (2007) showed that a stepwise parallel gatekeeping procedure can be directly constructed without appeal to the closure principle. For $m = 2$ families, this procedure uses the Bonferroni MTP for F_1 and any FWER-controlling MTP (not necessarily the Holm MTP) for F_2. For more than two families, he applied this two-stage procedure recursively.

In this paper we capture the essence of the Guilbaud procedure which allows us to generalize it to develop more powerful procedures. The key property of the Bonferroni MTP used to test F_1 is that it does not exhaust the designated Type I error rate, α, unless all hypotheses are true, and thus one can carry over the remaining α to test F_2. In fact, the actual FWER of the Bonferroni MTP is less than or equal to α when all hypotheses are true, but the upper bound α is used to construct the procedure. We show that any MTP that does not exhaust α unless all hypotheses are true can be used in place of the Bonferroni MTP. Such an MTP is said to be *separable*.

We introduce the separability condition in Section 2. A general class of MTPs that satisfies the separability condition and are more powerful than the Bonferroni MTP is introduced in Section 3. Next we present a general multistage gatekeeping procedure in Section 4 in which any separable MTP can be used. Calculation of the adjusted p-values for the proposed procedure is discussed in Section 5. The paper concludes with an illustrative clinical trial example in Section 6. The proof of the main result is given in the Appendix.

2 Separability Condition

To introduce the separability condition, it is convenient to consider the case of a single family of hypotheses, $F = \{H_1, \ldots, H_n\}$. For any $I \subseteq N = \{1, 2, \ldots, n\}$, the *error rate function* of an MTP is defined as

$$e(I) = \sup_{H_I} P\left\{\bigcup_{i \in I} (\text{Reject } H_i) \,\middle|\, H_I\right\},$$

where "Reject H_i" represents the event (the subset of the sample space) that corresponds to rejection of H_i and the supremum of the probability is taken over the entire null space defined by

$H_I = \bigcap_{i \in I} H_i$, including any false hypotheses H_j, $j \notin I$. Thus $e(I)$ is the maximum probability of making at least one Type I error in the subfamily $\{H_i,\ i \in I\}$.

Generally, a closed form and readily computable upper bound $e^*(I)$ on $e(I)$ is used to determine the critical values of an MTP since an exact expression for $e(I)$ is difficult to derive if the test statistics are correlated; if the correlations are unknown, which is often the case, then $e(I)$ cannot be computed at all. For example, for the Bonferroni MTP, which rejects any H_i if $p_i \leq \alpha/n$, the upper bound used is

$$e^*(I) = \alpha|I|/n\,, \tag{2}$$

where $|I|$ is the cardinality of set I. For any MTP, if an exact computable expression for $e(I)$ is available then we set $e^*(I) = e(I)$; otherwise we will treat $e^*(I)$ itself as the error rate function and state all the formulas in terms of $e^*(I)$.

The $e^*(I)$ function of an MTP will be used in the construction of the multistage parallel gatekeeping procedure defined in Section 4 as follows. Consider an MTP operating at an α-level and let A denote the index set of accepted hypotheses. The part of α that is "unused" and hence can be carried over to test the hypotheses in the next family is $\alpha - e^*(A)$ or the corresponding fraction is $1 - e^*(A)/\alpha$. Obviously, this fraction equals 1 if $A = \emptyset$ (i.e., if all hypotheses are rejected). To satisfy the *parallel gatekeeping condition* we require this fraction to be 0 if $A = N$ (i.e., if all hypotheses are accepted). In addition, this fraction must be positive, and hence $e^*(I)$ must be strictly less than α, if only a subset of hypotheses are accepted. We refer to this condition as the *separability condition*, and state it as

$$e^*(I) < \alpha \quad \text{for} \quad \text{all } I \subset N\,. \tag{3}$$

An α-level MTP that satisfies the separability condition is said to be *separable*. The Bonferroni MTP can be readily seen to be separable since $e^*(I) = \alpha|I|/n < \alpha$ if $I \subset N$.

Commonly, we want the fraction of α carried over to the next stage, namely $1 - e^*(A)/\alpha$, to be a monotone increasing function of the index set $R = N \backslash A$ of rejected hypotheses. Therefore we require

$$e^*(\emptyset) = 0, \quad e^*(I) \leq e^*(J) \text{ if } I \subseteq J \quad \text{and} \quad e^*(N) = \alpha\,. \tag{4}$$

It is readily seen that the Bonferroni MTP satisfies these conditions. Also, it is worth noting that, if $e^*(I)$ does not satisfy the monotonicity condition, this condition can always be enforced. For any separable MTP, one can define the upper bound as $e^*(I) = \max_{I' \subseteq I} e^*(I')$; then it follows that $e^*(I)$ will meet the separability and monotonicity conditions.

3 Truncated MTPs

Many standard MTPs such as the Holm (1979) and Hochberg (1988) do not satisfy the separability condition (3). In order that they do, while maintaining their power advantage over the Bonferroni MTP, we modify them by taking a convex combination of their critical constants with the Bonferroni critical constants. We refer to the resulting modified MTPs as truncated MTPs.

Consider again a single family of hypotheses, $F = \{H_1, \ldots, H_n\}$, with p-values, $p_1, \ldots, p_n$. For convenience, we will assume that the hypotheses are equally weighted. The same principle can be applied to construct truncated MTPs for weighted hypotheses. All MTPs are assumed to be of nominal α level.

3.1 Holm step-down MTP

In the Holm (1979) MTP the p-values are first ordered, $p_{(1)} \leq \cdots \leq p_{(n)}$. Let $H_{(1)}, \ldots, H_{(n)}$ be the corresponding hypotheses. At the first stage, $H_{(1)}$ is tested by comparing $p_{(1)}$ with α/n. If $p_{(1)} > \alpha/n$, then all hypotheses are accepted and testing stops. Otherwise $H_{(1)}$ is rejected and one proceeds to test $H_{(2)}$ by comparing $p_{(2)}$ with $\alpha/(n-1)$. In general, $H_{(i)}, \ldots, H_{(n)}$ are accepted and testing stops if

$$p_{(i)} > \frac{\alpha}{n - i + 1}\,; \tag{5}$$

otherwise $H_{(i)}$ is rejected and testing continues with $H_{(i+1)}$.

The Holm MTP incorrectly rejects any true hypothesis with probability α and hence is not separable. To see this, consider the problem of testing $H_i : \mu_i = 0$ versus $H_i' : \mu_i > 0$, $(1 \le i \le n)$. Suppose that $\mu_j = 0$ for some j and $\mu_i \to \infty$, $i \ne j$. Then $p_i \to 0$ for $i \ne j$ and p_j will be the largest p-value. Therefore, p_j will be compared with α and H_j will be rejected with probability α.

To make the Holm MTP separable, we truncate its critical constants by taking their convex combination with the Bonferroni MTP constants as follows. In (5), for specified γ $(0 \le \gamma < 1)$, called the *truncation fraction*, we replace the critical constant for comparing $p_{(i)}$ with

$$w_i\alpha = \left[\frac{\gamma}{n-i+1} + \frac{1-\gamma}{n}\right]\alpha\,. \tag{6}$$

We refer to this procedure as the *truncated Holm MTP*. The power of this MTP is strictly increasing in γ. For $\gamma = 0$ and $\gamma = 1$, this MTP simplifies to the Bonferroni MTP and the Holm MTP, respectively.

Note that the truncated Holm MTP is a step-down shortcut to a closed procedure that tests any intersection hypothesis $H_I = \bigcap_{i \in I} H_i$ using the weighted Bonferroni MTP with weights $w_i(I)$ and finds it significant if $p_i \le w_i(I)\alpha$ for at least one $i \in I$, where

$$w_i(I) = \frac{\gamma}{|I|} + \frac{1-\gamma}{n}\,.$$

Recalling that a closed procedure rejects H_I iff all H_J for $J \supseteq I$ are significant, an upper bound on the error rate function of the truncated Holm MTP is given by

$$e^*(I) = \begin{cases} \sum\limits_{i \in I} w_i(I)\alpha = [\gamma + (1-\gamma)\,|I|/n]\,\alpha & \text{if } |I| > 0, \\ 0 & \text{if } |I| = 0. \end{cases} \tag{7}$$

Therefore for any $I \subset N$ and $\gamma \in [0, 1)$,

$$e^*(I) < [\gamma + (1-\gamma)]\,\alpha = \alpha\,.$$

Hence the truncated Holm MTP is separable. Note that the $e^*(I)$ function of the truncated Holm MTP satisfies (4).

3.2 Hochberg step-up MTP

The Hochberg MTP uses the same Holm critical constants (5) but tests the hypotheses in a step-up manner (it begins with the hypotheses corresponding to the least significant p-value). The Hochberg MTP is more powerful than the Holm MTP. However, the Hochberg MTP (as well as the Hommel MTP mentioned in the sequel) requires independence among the p-values since it is based on the Simes (1986) test, which assumesindependence; Sarkar and Chang (1997) have shown that the independence assumption can be relaxed to the positive dependence assumption. The error rate function of the Hochberg MTP also equals α under the same configuration for which the error rate function of the Holm MTP equals α, namely one hypothesis is true and the others are infinitely false. Hence the Hochberg MTP is not separable.

A *truncated Hochberg MTP* uses the same critical constants (6) as does the truncated Holm MTP, but is more powerful than the latter. At the first stage, this MTP rejects all hypotheses and stops testing if

$$p_{(n)} \le w_n\alpha = \left[\gamma + \frac{(1-\gamma)}{n}\right]\alpha\,;$$

otherwise it accepts $H_{(n)}$ and goes on to test $H_{(n-1)}$. In general, having accepted $H_{(n)}, \ldots, H_{(i+1)}$, it rejects $H_{(i)}, \ldots, H_{(1)}$ and stops testing if $p_{(i)} \le w_i\alpha$ where w_i is defined in (6); otherwise, it accepts $H_{(i)}$ and goes on to test $H_{(i-1)}$.

As is well-known, the Hochberg (1988) MTP is a conservative shortcut to the closed procedure in which each intersection hypothesis $H_I = \bigcap_{i \in I} H_i$ is tested using the Simes (1986) test. Similarly, the truncated Hochberg MTP is a conservative shortcut to the closed procedure based on the truncated Simes test in which any intersection hypothesis $H_I = \bigcap_{i \in I} H_i$ is rejected if

$$p_{(i)}(I) \leq \left[\frac{\gamma}{|I| - i + 1} + \frac{1 - \gamma}{n}\right]\alpha \quad \text{for at least one} \quad i \in I\,,$$

where $p_{(i)}(I)$ is the ith ordered p-value in the index set I and $\gamma \in [0, 1)$. Therefore an upper bound on the error rate function of the truncated Hochberg MTP is given by

$$e^*(I) = 1 - P\left\{p_{(i)}(I) > \left[\frac{\gamma}{|I| - i + 1} + \frac{1 - \gamma}{n}\right]\alpha \text{ for all } i \in I\right\}$$

if $|I| > 0$ and $e^*(I) = 0$ if $|I| = 0$. Using the Simes (1986) identity, it is readily seen that $e^*(I) < \alpha$ for $I \subset N$ and $\gamma \in [0, 1)$. Therefore, the truncated Hochberg MTP is separable. In general, $e^*(I)$ above does not satisfy the monotonicity condition (4); therefore the latter may need to be enforced as explained following its statement. For independent p-values, $e^*(I)$ can be computed using the recursive formula given in the following result due to Sen (1999).

Präposition 3.1 *Let $U_{(1)} < \ldots < U_{(k)}$ denote the order statistics of $k \geq 1$ i.i.d. observations from a uniform $(0, 1)$ distribution. For any $0 < a_1 < \ldots < a_k < 1$,*

$$P(a_1, \ldots, a_k) = P(U_{(i)} > a_i \text{ for all } i = 1, \ldots, k) = k! H_k(1)\,,$$

where

$$H_i(u) = \int_{a_i}^{u} H_{i-1}(v)\,\mathrm{d}v, \quad i = 1, \ldots, k \quad \text{and} \quad H_0(u) = I(u \geq a_1)\,,$$

and $I(\cdot)$ is an indicator function.

3.3 Fallback MTP

Wiens (2003) proposed a step-down MTP in which the hypotheses are *a priori* ordered (in contrast to the Holm MTP which orders the hypotheses according to their observed p-values). The total α is allocated to the n ordered hypotheses as $\alpha_1, \ldots, \alpha_n$ such that $\sum_{i=1}^{n} \alpha_i = \alpha$. For simplicity, we shall restrict to the equal allocation case: $\alpha_i = \alpha/n$ $(1 \leq i \leq n)$. The MTP begins by testing H_1 at level α/n; more generally, it tests a hypothesis H_i at level $(i - t)\,\alpha/n$, where t is the index of the last accepted hypothesis ($t = 0$ if none of the previous hypotheses is accepted). This MTP also follows the "use it or lose it" principle so that the α_i's for the rejected hypotheses in the sequence are carried forward to test the later hypotheses.

This fallback MTP is not separable. Suppose, for example, that $H_1, \ldots, H_{n-1}$ are infinitely false and H_n is true, so that $p_1, \ldots, p_{n-1} \to 0$ and p_n is compared with α. Then the probability of rejecting H_n is α.

The *truncated fallback MTP* tests H_i at level

$$w_i(t)\alpha = \left(\frac{\gamma(i - t)}{n} + \frac{1 - \gamma}{n}\right)\alpha\,,$$

where $0 \leq \gamma < 1$ and t is the index of the last accepted hypothesis ($t = 0$ if none of the previous hypotheses is accepted).

Extending the arguments in the proof of Theorem 1 of Wiens and Dmitrienko (2005), it can be shown that this MTP is a shortcut to a closed procedure which rejects any intersection hypothesis $H_I = \bigcap_{i \in I} H_i$ if $p_i \leq w_i(t_I)\alpha$ for at least one $i \in I$, where t_I is the largest index in I that is smaller

than i if i is not the smallest index in I and $t_I = 0$ if i is the smallest index in I. Therefore, using the closure principle and the Bonferroni inequality, an upper bound on the error rate function of the truncated fallback MTP is given by

$$e^*(I) = \begin{cases} \sum_{i \in I} w_i(t_I)\alpha & \text{if} \quad |I| > 0 \\ 0 & \text{if} \quad |I| = 0 \end{cases}.$$

Note that $e^*(I) < \alpha$ for any $I \subset N$, and hence the truncated fallback MTP is separable if $\gamma \in [0, 1)$. Also, $e^*(I)$ satisfies the conditions (4).

3.4 Dunnett step-down MTP

The Dunnett (1955) MTP can be thought of as a parametric version of the Bonferroni MTP and the step-down Dunnett MTP (Marcus, Peritz and Gabriel, 1976) is analogous to the Holm MTP. The step-down Dunnett MTP does not satisfy the separability condition because it incorrectly rejects any true hypothesis with probability α. The *truncated Dunnett MTP* is defined as a convex combination of the regular and step-down Dunnett MTPs with $0 \leq \gamma < 1$. Let $t_1, \ldots, t_n$ be the test statistics associated with $H_1, \ldots, H_n$. Let $t_{(1)} > \ldots > t_{(n)}$ be the ordered test statistics and $H_{(1)}, \ldots, H_{(n)}$ denote the corresponding null hypotheses. Further, let $T_1, \ldots, T_n$ denote the random variables corresponding to the observed statistics $t_1, \ldots, t_n$ and assume that they follow a multivariate t-distribution under the global null hypothesis.

For any $I \subseteq N$, let $c(I)$ be the critical value for the maximum test statistic associated with H_i, $i \in I$, such that

$$P\left(\max_{i \in I} T_i > c(I) \mid H_I = \bigcap_{i \in I} H_i\right) = \alpha\,.$$

The computation of these critical values can be performed using the algorithm for calculating multivariate t probabilities due to Genz and Bretz (2002).

For any $i = 1, \ldots, n$, let $I_{(i)} = \{(i), \ldots, (n)\}$. The truncated Dunnett MTP begins with the hypothesis, $H_{(1)}$, corresponding to the most significant t-statistic, $t_{(1)}$. This hypothesis is rejected if $t_{(1)} > c(I_{(1)})$ and is accepted otherwise. If $H_{(1)}$ is rejected, the next hypothesis in the sequence, $H_{(2)}$, is tested. In general, the MTP rejects $H_{(j)}$ if

$$t_{(i)} > (1 - \gamma)c(I_{(1)}) + \gamma c(I_{(i)}) \text{ for all } i = 1, \ldots, j\,.$$

Otherwise, $H_{(j)}, \ldots, H_{(n)}$ are accepted and testing stops. The truncated Dunnett MTP simplifies to the regular Dunnett MTP if $\gamma = 0$ and to the step-down Dunnett MTP if $\gamma = 1$.

The computation of the error rate function for the truncated Dunnett MTP can be performed by using its closed representation. This MTP is equivalent to a closed testing procedure that rejects the intersection hypothesis $\bigcap_{i \in I} H_i$, $I \subseteq N$, if

$$\max_{i \in I}\ T_i > (1 - \gamma)\, c(N) + \gamma c(I)\,.$$

Therefore, using the same argument as used for the Holm MTP, an upper bound on the error rate function of the truncated Dunnett MTP is given by

$$e^*(I) = \begin{cases} P(\max_{i \in I} T_i > (1 - \gamma)\, c(N) + \gamma c(I)) & \text{if} \quad |I| > 0 \\ 0 & \text{if} \quad |I| = 0 \end{cases}.$$

It is easy to see that the truncated Dunnett MTP satisfies the separability condition for any $I \subset N$ if $0 \leq \gamma < 1$. However, the upper bound $e^*(I)$ on its error rate function may not satisfy the monotonicity condition (4), in which case the latter may need to be enforced as explained following its statement.

4 General Multistage Gatekeeping Procedure

To explain the key principles underlying the general multistage gatekeeping procedure, we will begin with a simple case of two families of hypotheses:

$$F_1 = \{H_{11}, \ldots, H_{1n_1}\} \quad \text{and} \quad F_2 = \{H_{21}, \ldots, H_{2n_2}\}.$$

The hypotheses in F_1 and F_2 are tested using a two-stage gatekeeping procedure described below. Let α denote the FWER for this procedure. Further Let $N_1 = \{1, \ldots, n_1\}$ and $A_1 \subseteq N_1$ be the index set corresponding to the accepted hypotheses in F_1. The hypotheses in F_1 are tested at the $\alpha_1 = \alpha$ level using an MTP that controls the FWER within F_1 and meets the separability condition introduced in Section 2. Next, F_2 is tested using an MTP that controls the FWER within F_2 at level

$$\alpha_2 = \alpha_1 - \mathrm{e}_1^*(A_1)$$

and $\mathrm{e}_1^*(I)$ is an appropriate upper bound on the error rate function of the MTP used at the first stage of this procedure. The second-stage MTP is assumed to be α-consistent (Roth, 1999), i.e., if it rejects H_{2j}, $j = 1, \ldots, n_2$, at the α level, it will also reject it at the α' level, where $\alpha < \alpha'$. Note that all popular MTPs are α-consistent.

Proposition 4.1 *The two-stage gatekeeping procedure controls the FWER at the α level.*

Proof Given in the Appendix.

The simple two-stage procedure provides useful insights into the nature of gatekeeping inferences. Since $\mathrm{e}_1^*(\emptyset) = 0$, the second-stage MTP is carried out at the α level if all hypotheses are rejected in F_1. Secondly, due to the monotonicity of the $\mathrm{e}_1^*(I)$ function (4), a greater fraction of α will be propagated to F_2 (and consequently more hypotheses could be rejected in F_2) if more hypotheses are rejected in F_1. Finally, since $\alpha_2 = 0$ if $A_1 = N_1$ according to (4), this procedure satisfies the parallel gatekeeping condition.

It is important to note that any FWER-controlling MTP can be used at the second stage of the two-stage gatekeeping procedure. Therefore one can construct gatekeeping procedures with an arbitrary number of stages by a recursive application of the two-stage procedure. This approach is conceptually similar to the recursive algorithm for constructing combination tests proposed by Brannath, Posch and Bauer (2002) in the context of multistage adaptive clinical trials. Since a serial gatekeeper can be expressed as a series of single-hypothesis families, multistage gatekeeping procedures obtained via the recursive algorithm can have a very flexible structure that combines serial gatekeepers and parallel gatekeepers.

To define the multistage gatekeeping procedure, consider $m \geq 2$ families, $F_i = \{H_{i1}, \ldots, H_{in_i}\}$ $(1 \leq i \leq m)$. Let $N_i = \{1, \ldots, n_i\}$ and $A_i \subseteq N_i$ be the index set corresponding to the accepted hypotheses in F_i. The algorithm for applying the procedure is as follows.

- **Start** Initialize $\alpha_1 = \alpha$.
- **Stages 1** through $m - 1$ Test F_i at an α_i level using any separable MTP with a suitable upper bound on the error rate function $e_i^*(I)$. Set

$$\alpha_{i+1} = \alpha_i - \mathrm{e}_i^*(A_i).$$

- If $A_i = N_i$, i.e., no hypotheses are rejected in F_i, then $\mathrm{e}_i^*(A_i) = \alpha_i$ and hence $\alpha_{i+1} = 0$. In that case, stop testing and accept all hypotheses in $F_{i+1}, \ldots, F_m$; otherwise go to the next stage.
- **Stage** m Use any FWER-controlling MTP to test F_m at an α_m level.

The following remarks may be noted regarding this procedure.

1. If all hypotheses are rejected at the i-th stage $(1 \leq i \leq m - 1)$, then $A_i = \emptyset$ and $\alpha_{i+1} = \alpha_i$. Thus full α_i is carried over to the next stage.
2. At the final stage, any FWER controlling MTP may be used, but a truncated MTP should not be used since it is less powerful than its untruncated version.

If the Bonferroni MTP is used at the first $m-1$ stages, it follows from (2) that

$$\alpha_i = \alpha_{i-1} - e^*_{i-1}(A_{i-1}) = \left[1 - \left(\frac{a_{i-1}}{n_{i-1}}\right)\right]\alpha_{i-1} = \left(\frac{r_{i-1}}{n_{i-1}}\right)\alpha_{i-1}\,,$$

where a_{i-1} and r_{i-1} are the numbers of accepted and rejected hypotheses, respectively, in F_{i-1}. Applying this formula recursively we get the formula (1) for the rejection gain factor ϱ_i.

5 Adjusted *p*-values

Multiple tests are commonly performed by computing adjusted p-values. Given the stepwise structure of multistage gatekeeping procedures, it seems natural to derive a stepwise algorithm for computing adjusted p-values for the hypotheses in $F_i = \{H_{i1}, \ldots, H_{in_i}\}$ $(1 \le i \le m)$. Theoretically, this algorithm can be constructed along the lines of Guilbaud (2007). However, the algorithm quickly becomes quite complex due to a large number of minimization/maximization steps.

An alternative approach can be based on the general definition of adjusted p-values: The multiplicity adjusted p-value for a given null hypothesis and an MTP is defined as the significance level at which the procedure rejects the hypothesis (Westfall and Young, 1993). Using this definition, it is easy to compute adjusted p-values associated with the gatekeeping procedure by looping through a discrete grid of significance levels. Thus let $\alpha = k/K$, $(0 < k < K)$ for some sufficiently large value of K. The adjusted p-value, $\tilde{p}_{ij}$, for hypotheses H_{ij} is the smallest α (corresponding to the smallest k) for which H_{ij} is rejected. Since multistage gatekeeping procedures have a simple stepwise form, this direct-calculation algorithm is quite fast even when the number of hypotheses is large.

An SAS code for computing multiplicity-adjusted p-values can be downloaded from the BioPharmNet web site: http://www.biopharmnet.com/code. A special case of this code is used in the following clinical trial example.

6 Example

To illustrate the implementation of multistage gatekeeping procedures, consider the Type II diabetes clinical trial example from Dmitrienko, Wiens, Tamhane and Wang (2007, Section 6). The trial compares three doses (Low (L), Medium (M) and High (H)) of an experimental drug versus placebo (Plac) with respect to a primary endpoint (Endpoint P, Hemoglobin A1c) and two key secondary endpoints (Endpoint S1, Fasting serum glucose; Endpoint S2, HDL cholesterol). Nine null hypotheses in this trial are grouped into three families:

- F_1 includes the H-Plac (H_{11}), M-Plac (H_{12}) and L-Plac (H_{13}) comparisons for Endpoint P.
- F_2 includes the H-Plac (H_{21}), M-Plac (H_{22}) and L-Plac (H_{23}) comparisons for Endpoint S1.
- F_3 includes the H-Plac (H_{31}), M-Plac (H_{32}) and L-Plac (H_{33}) comparisons for Endpoint S2.

The hypotheses are equally weighted within each family and the FWER is set at 0.05. Raw p-values for the nine hypotheses computed from two-sample t tests are displayed in Table 1.

It is worth noting that the original example in Dmitrienko, Wiens, Tamhane and Wang (2007) included logical restrictions, i.e., the dose-placebo tests for Endpoint S1 were restricted to the doses at which there was a significant treatment effect for Endpoint P and, similarly, the dose-placebo tests for Endpoint S2 were restricted to the doses at which Endpoints P and S1 demonstrated a significant effect. For the sake of simplicity, we will not impose these restrictions in this example and F_1 will be assumed to serve as a parallel gatekeeper for F_2, which, in turn, will serve as a parallel gatekeeper for F_3.

The nine hypotheses in the Type II diabetes clinical trial will be tested using the three-stage gatekeeping procedure defined below. The procedure uses the truncated Holm MTP with $\gamma = 0$, 0.25 and 0.5 at the first two stages (F_1 and F_2) and the Holm MTP at the last stage (F_3).

Table 1 Three-stage gatekeeping procedure in the Type II diabetes clinical trial example. The hypotheses in F_1 and F_2 are tested using the truncated Holm MTP with the truncation fraction γ and the hypotheses in F_3 are tested using the regular Holm MTP. The asterisk identifies the adjusted p-values that are significant at the 0.05 level

Family	Null hypothesis	Raw p-value	Adjusted p-value		
			$\gamma = 0$	$\gamma = 0.25$	$\gamma = 0.5$
F_1	H_{11}	0.005	0.015^*	0.015^*	0.015^*
	H_{12}	0.011	0.033^*	0.029^*	0.027^*
	H_{13}	0.018	0.054	0.036^*	0.027^*
F_2	H_{21}	0.009	0.041^*	0.036^*	0.027^*
	H_{22}	0.026	0.078	0.052	0.039^*
	H_{23}	0.013	0.054	0.036^*	0.031^*
F_3	H_{31}	0.010	0.054	0.040^*	0.039^*
	H_{32}	0.006	0.054	0.036^*	0.039^*
	H_{33}	0.051	0.076	0.052	0.051

To illustrate the process of applying a multistage gatekeeping procedure, consider the case when $\gamma = 0.25$. Using the algorithm defined in Section 4, initialize $\alpha_1 = 0.05$. The hypotheses in F_1 are tested using the truncated Holm MTP at the overall α_1 level. All hypotheses are rejected at this level and, by the definition of the error rate function for this truncated MTP,

$$\alpha_2 = \alpha_1 = 0.05\,.$$

The three tests in F_2 are also carried out using the truncated Holm MTP at the overall α_2 level. Two hypotheses (H_{21} and H_{23}) are rejected in F_2 and thus the significance level in F_3 is given by

$$\alpha_3 = \alpha_2 - \left[\gamma + (1-\gamma)\frac{|A_2|}{n}\right]\alpha_2 = \alpha_2 - \left[\gamma + \frac{(1-\gamma)}{3}\right]\alpha_2 = 0.025\,,$$

where $|A_2| = 1$ and $n = 3$. Two hypotheses (H_{31} and H_{32}) are rejected by the regular Holm MTP at this level in F_3.

The multiplicity-adjusted p-values associated with the three-stage gatekeeping procedure are computed using the direct-calculation algorithm defined in Section 5 with $K = 10\,000$. The adjusted p-values are given in Table 1. In general, the power of the truncated Holm MTP in F_1 is an increasing function of γ but the power of the MTPs in F_2 and F_3 may not be a monotone function of the truncation fraction. In this particular case, the number of rejected hypotheses in these two families increases with γ because more hypotheses are found false in F_1 and, as a consequence, a larger fraction of α is carried over to F_2 and F_3. To see this, compare the columns for $\gamma = 0$ and $\gamma = 0.25$ andthe columns for $\gamma = 0.25$ and $\gamma = 0.5$ in Table 1.

As an aside note, it is instructive to compare the unrestricted gatekeeping procedure defined above to the procedure with logical restrictions considered in Dmitrienko, Wiens, Tamhane and Wang (2007). In this example, the logical restriction condition is not met for $\gamma = 0.25$ since H_{22} is not rejected but H_{32} is rejected.

Lastly, note that the truncated Holm MTP with $\gamma = 0$ is equivalent to the Bonferroni MTP and thus the three-stage procedure simplifies in this case to the Bonferroni-based parallel gatekeeping procedure. The adjusted p-values for the three-stage procedure with $\gamma = 0$ are equal to those presented by Dmitrienko, Wiens, Tamhane and Wang (2007) in Table IV.

7 Conclusions

This paper introduced a general approach to constructing multistage gatekeeping procedures. The resulting procedures are fairly flexible in the sense that one can choose from a broad class of MTPs to define a test for each individual stage. Among the p-value-based procedures, we recommend using the truncated Hochberg or Hommel MTPs if the statistics are positively dependent. Otherwise, the truncated Holm procedure can be used. The truncated fallback MTP can be used if the hypotheses within a family can be naturally ordered. If the normality assumption holds then the truncated Dunnett MTP should be used.

If a truncated MTP is used at the i-th stage of a multistage gatekeeping procedure, the truncation fraction γ can be thought of as the "weight" of Family F_i relative to the subsequent families. The choice of γ for the truncated Holm MTP was briefly described in Dmitrienko and Tamhane (2007, Section 4) and similar arguments apply to other truncated MTPs. In particular, the power of the tests in F_i is an increasing function of γ but the relationship between the power of the tests in $F_{i+1}, \ldots, F_m$ and the truncation fraction used in F_i is more complicated. The tests in $F_{i+1}, \ldots, F_m$ can gain or lose power with increasing γ depending on the number of true hypotheses in $F_{i+1}, \ldots, F_m$, their weights and effect sizes for false hypotheses. It is worth noting that for $m > 2$ families different γ's can be used at different stages.

Appendix

Proof of Proposition 4.1 Define the following events:

$$B_1 = \{\text{One or more true null hypotheses are rejected in } F_1\},$$
$$B_2(x) = \{\text{One or more true null hypotheses are rejected at level } x \text{ in } F_2\}.$$

Note that, due to α-consistency of the second-stage MTP, $B_2(x) \subseteq B_2(y)$ if $x \leq y$. Further, since the MTP controls the FWER within F_2, $P(B_2(x)) \leq x$. Also, let $e_1^*(I)$ be an upper bound on the error rate function of the first-stage MTP, $I \subseteq N_1$, α_2 be the random level at which the second-stage MTP is carried out within the two-stage procedure and $\bar{E}$ be the complement of the event E.

The FWER of the two-stage gatekeeping procedure can be written as

$$P(B_1 \cup B_2(\alpha_2)) = P(B_1) + P(\bar{B}_1 \cap B_2(\alpha_2)).$$

Let $T_1 \subseteq N_1$ denote the set of indices corresponding to the true null hypotheses in F_1. By the definition of the error rate function, $P(B_1) \leq \mathrm{e}_1^*(T_1)$.

Next consider $\bar{B}_1 \cap B_2(\alpha_2)$. Since

$$\bar{B}_1 = \{\text{No true null hypotheses are rejected in } F_1\},$$

we have $T_1 \subseteq A_1$ and thus, due to the monotonicity condition (4),

$$\alpha_2 = \alpha - \mathrm{e}_1^*(A_1) \leq \alpha - \mathrm{e}_1^*(T_1)$$

when $\bar{B}_1$ is true. Therefore,

$$\bar{B}_1 \cap B_2(\alpha_2) \subseteq \bar{B}_1 \cap B_2(\alpha - \mathrm{e}_1^*(T_1))$$

and

$$P(\bar{B}_1 \cap B_2(\alpha_2)) \leq P(\bar{B}_1 \cap B_2(\alpha - \mathrm{e}_1^*(T_1)) \leq P(B_2(\alpha - \mathrm{e}_1^*(T_1)) \leq \alpha - \mathrm{e}_1^*(T_1).$$

Therefore $P(B_1 \cup B_2(\alpha_2)) \leq \mathrm{e}_1^*(T_1) + \alpha - \mathrm{e}_1^*(T_1) = \alpha$ and thus the two-stage procedure controls the FWER at the α level. The proof of Proposition 4 is complete.

The asterisk identifies the adjusted p-values that are significant at the 0.05 level.

Acknowledgements *The authors would like to thank two anonymous referees for their valuable comments. This research was supported by grants from the National Institutes of Health and National Security Agency to Prof. Ajit Tamhane at Northwestern University.*

Conflict of Interests Statement
The authors have declared no conflict of interest.

References

Brannath, W., Posch, M., and Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association* **97**, 236–244.

Dmitrienko, A., Offen, W. W., and Westfall, P. H. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine* **22**, 2387–2400.

Dmitrienko, A., Molenberghs, G., Chuang-Stein, C., and Offen, W. (2005). *Analysis of Clinical Trials Using SAS: A Practical Guide*. SAS Press: Cary, NC.

Dmitrienko, A., Offen, W., Wang, O., and Xiao, D. (2006). Gatekeeping procedures in dose-response clinical trials based on the Dunnett test. *Pharmaceutical Statistics* **5**, 19–28.

Dmitrienko, A., Tamhane, A. C., Wang, X., and Chen, X. (2006). Stepwise gatekeeping procedures in clinical trial applications. *Biometrical Journal* **48**, 984–991.

Dmitrienko, A., Wiens, B. L., Tamhane, A. C., and Wang, X. (2007). Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Statistics in Medicine* **26**, 2465–2478.

Dmitrienko, A. and Tamhane, A. C. (2007). Gatekeeping procedures with clinical trial applications. *Pharmaceutical Statistics* **6**, 171–180.

Dmitrienko, A., Tamhane, A. C., Liu, L., and Wiens, B. L. (2008). A note on tree gatekeeping procedures in clinical trials. *Statistics in Medicine*. To appear.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* **50**, 1096–1121.

Genz, A. and Bretz, F. (2002). Methods for the computation of multivariate *t*-probabilities. *Journal of Computational and Graphical Statistics* **11**, 950–971.

Guilbaud, O. (2007). Bonferroni parallel gatekeeping – Transparent generalizations, adjusted *p*-values, and short direct proofs. *Biometrical Journal* **49**, 917–927.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple significance testing. *Biometrika* **75**, 800–802.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75**, 383–386.

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

Maurer, W., Hothorn, L., and Lehmacher, W. (1995). Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. In *Biometrie in der chemisch-pharmazeutischen Industrie*. Vollmar J. (ed.). Stuttgart: Fischer Verlag. **6**, 3–18.

Roth, A. J. (1999). Multiple comparison procedures for discrete test statistics. *Journal of Statistical Planning and Inference* **82**, 101–117.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **63**, 655–660.

Sarkar, S. and Chang, C. K. (1997). Simes' method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* **92**, 1601–1608.

Sen, P. K. (1999). Some remarks on Simes-type multiple tests of significance. *Journal of Statistical Planning and Inference* **82**, 139–145.

Wiens, B. (2003). A fixed-sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics* **2**, 211–215.

Wiens, B. and Dmitrienko, A. (2005). The fallback procedure for evaluating a single family of hypotheses. *Journal of Biopharmaceutical Statistics* **15**, 929–942.

