Stepwise Gatekeeping Procedures in Clinical Trial Applications

Alex Dmitrienko[*,1], **Ajit C. Tamhane**[2], **Xin Wang**[2], and **Xun Chen**[3]

[1] Eli Lilly and Company, Indianapolis, IN 46285, USA
[2] Northwestern University, Evanston, IL 60208, USA
[3] Sanofi-Aventis, Bridgewater, NJ 08807, USA

Received 3 April 2006, revised 18 June 2006, accepted 14 July 2006

Summary

This paper discusses multiple testing problems in which families of null hypotheses are tested in a sequential manner and each family serves as a *gatekeeper* for the subsequent families. Gatekeeping testing strategies of this type arise frequently in clinical trials with multiple objectives, e.g., multiple endpoints and/or multiple dose-control comparisons. It is demonstrated in this paper that the parallel gatekeeping procedure of Dmitrienko, Offen and Westfall (2003) admits a simple stepwise representation (n null hypotheses can be tested in n steps rather than 2^n steps required in the closed procedure). The stepwise representation considerably simplifies the implementation of gatekeeping procedures in practice and provides an important insight into the nature of gatekeeping inferences. The derived stepwise gatekeeping procedure is illustrated using clinical trial examples.

Key words: Clinical trials; Multiple comparisons; Stepwise tests.

1 Introduction

Complex multiple testing strategies are becoming increasingly common in a clinical trial setting as clinical researchers attempt to improve the information/sample size ratio by pursuing multiple objectives representing multiple outcome variables, doses, analysis types (e.g., non-inferiority analysis versus superiority analysis), etc. (Chen, Luo and Capizzi, 2005; Chen et al., 2005; Dmitrienko et al., 2005). A gatekeeping testing approach introduced by Maurer, Hothorn and Lehmacher (1995) and Bauer et al. (1998) provides an efficient way of handling multiple testing problems of this kind. The gatekeeping approach relies on a sequential formulation of the problem, i.e., null hypotheses corresponding to multiple objectives are grouped into families which are then tested in a sequential manner.

Two types of gatekeeping testing procedures have been studied in the literature, *serial* and *parallel* gatekeeping procedures. Within the serial framework, one tests hypotheses within each family (gate) using a method that controls the familywise error rate (FWE) for each gate and proceeds to the next gate when *all* of the hypotheses in the current gate are rejected (Westfall and Krishen, 2001). A parallel gatekeeping strategy requires the rejection of *at least* one hypothesis in each gate (Dmitrienko et al., 2003).

It is interesting to note a connection between serial and parallel gatekeeping and intersection-union (IU) and union-intersection (UI) tests, respectively. Serial gatekeeping is analogous to the IU test (Berger, 1982) in which the null hypothesis, which is a union of several component hypotheses, is rejected iff all of them are rejected. On the other hand, parallel gatekeeping is analogous to the UI test in which the null hypothesis, which is an intersection of several component hypotheses, is rejected

[*] Corresponding author: e-mail: dmitrienko_alex@lilly.com, Phone: +1 317 277 1979, Fax: +1 317 277 3220

iff at least one of them is rejected. This connection may be useful in building a unified theory of gatekeeping procedures; however, that is not the purpose of the present paper.

An important difference between serial and parallel gatekeeping approaches is that the former is based on a straightforward sequential application of unadjusted tests and requires n steps to test n null hypotheses. By contrast, the parallel gatekeeping approach relies on complex procedures derived using the closed testing principle (Marcus, Peritz and Gabriel, 1976). In general, the number of computational steps grows exponentially with n (about 2^n calculations need to be performed to test n null hypotheses). Due to this property, parallel gatekeeping procedures are often considered computationally intractable in clinical applications involving a large number of null hypotheses (e.g., more than 10 null hypotheses).

The problem of finding "shortcuts" for closed testing procedures plays an important role in applications because it streamlines computational algorithms and leads to considerable savings. Shortcut procedures considered in the literature typically reduce the number of operations from order-2^n to order-n or order-n^2; see, for example, Grechanovsky and Hochberg (1999) and Westfall, Zaykin and Young (2001).

The goal of this paper is to prove that parallel gatekeeping procedures admit a shortcut. It is shown that the parallel Bonferroni gatekeeping procedure can be formulated as a stepwise procedure that requires order-n operations to test n null hypotheses. This simple stepwise representation enables clinical researchers to perform the gatekeeping procedures in a sequential manner by considering one gate at a time. The stepwise representation also facilitates understanding of the principles underlying the gatekeeping methodology because it explicitly demonstrates how the rejection of hypotheses in each gate affects inferences in subsequent gates.

The paper is organized as follows. Section 2 defines the parallel gatekeeping testing approach and outlines the algorithm for constructing gatekeeping procedures based on the closed testing principle. Section 3 introduces a stepwise version of the parallel Bonferroni gatekeeping procedure. Section 4 describes a clinical trial example to illustrate the stepwise gatekeeping procedure. Finally, the Appendix provides mathematical details.

2 Parallel Gatekeeping Procedure

In order to introduce a general framework of gatekeeping inferences, consider n null hypotheses tested in a clinical trial and assume that they are grouped into m families denoted by $F_1, \dots, F_m$. The n_i null hypotheses in ith family are denoted by $H_{i1}, \dots, H_{in_i}$ and $w_{i1}, \dots, w_{in_i}$ are the weights representing the importance of these null hypotheses within the family (note that $n_1 + \dots + n_m = n$, $0 < w_{ij} < 1$ and $w_{i1} + \dots + w_{in_i} = 1$). The common case of equal weights corresponds to $w_{ij} = 1/n_i$ ($j = 1, \dots, n_i$, $i = 1, \dots, m$).

The m families are tested in the following sequential manner. The null hypotheses in F_1 are examined first and tested with a suitable adjustment for multiplicity, e.g., using the Bonferroni test. Next, the null hypotheses in F_2 are tested with an appropriate adjustment for multiplicity provided the corresponding gatekeeper, F_1, is passed. Further, if the gatekeeper F_2 was successfully passed, one examines the null hypotheses in F_3 and so on. All of the gatekeepers are assumed parallel, i.e., at least one null hypothesis must be rejected in a gatekeeper to pass it.

Dmitrienko et al. (2003) introduced the parallel gatekeeping procedure based on the Bonferroni test for the case of two families of hypotheses (e.g., primary and secondary endpoints in a clinical trial) and Dmitrienko et al. (2005, Section 2.7) provided a general algorithm for constructing the parallel Bonferroni gatekeeping procedure for any number of families.

In general, gatekeeping testing procedures are constructed using a rather unwieldy algorithm based on the principle of closed testing proposed by Marcus, Peritz and Gabriel (1976). Let α be the prespecified familywise error rate. To define the parallel gatekeeping procedure for testing the null hypotheses in $F_1, \dots, F_m$ at the α level, consider the closed family consisting of all $2^n - 1$ nonempty intersections of these null hypotheses. Let $\mathcal{H}_{ij}$ denote the set of all intersection hypotheses in the closed family that imply H_{ij}, $i = 1, \dots, m$, $j = 1, \dots, n_i$, and, for any intersection hypothesis H, let $\delta_{ij}(H) = 1$ if $H \in \mathcal{H}_{ij}$ and $\delta_{ij}(H) = 0$ otherwise. For each H, one can define an n-dimensional vector of hypothesis weights, $v_{ij}(H)$, $i = 1, \dots, m$, $j = 1, \dots, n_i$, such that

$$0 \le v_{ij}(H) \le 1, \qquad v_{ij}(H) = 0 \quad \text{if} \quad \delta_{ij}(H) = 0, \qquad \sum_{i=1}^{m} \sum_{j=1}^{n_i} v_{ij}(H) \le 1 .$$

The algorithm for defining the hypothesis weights is given in the Appendix. The weighted Bonferroni p-value associated with H is given by $p_H = \min_{i,j} [p_{ij}/v_{ij}(H)]$. For any null hypothesis H_{ij} in $F_1, \dots, F_m$, the associated adjusted p-value is $\tilde{p}_{ij} = \max p_H$, where the maximum is computed over all $H \in \mathcal{H}_{ij}$, and H_{ij} is rejected if $\tilde{p}_{ij} \le \alpha$. The described parallel gatekeeping procedure controls the FWE in the strong sense at the pre-specified α level (Hochberg and Tamhane, 1987).

3 Stepwise Gatekeeping Procedure

Although the closed testing approach to the construction of gatekeeping procedures provides a means for testing any number of families of null hypotheses, it is generally quite complicated. In practice, it is highly desirable to have a simple set of decision rules that can help facilitate the implementation and interpretation of gatekeeping procedures. Consider, for example, a clinical trial with two families of null hypotheses denoted by $F_1 = \{H_{11}, H_{12}\}$ and $F_2 = \{H_{21}, H_{22}\}$. Assume that F_1 is a parallel gatekeeper and the hypotheses within each family are equally weighted, i.e., $w_{11} = w_{12} = 1/2$ and $w_{21} = w_{22} = 1/2$. These hypotheses may correspond to two co-primary and two secondary outcome variables as in the acute respiratory distress syndrome example discussed by Dmitrienko et al. (2003).

A close examination of the decision rule underlying the parallel Bonferroni gatekeeping procedure for testing the four null hypotheses reveals that the procedure has a simple stepwise structure. The parallel gatekeeping procedure rejects H_{11} or H_{12} whenever $p_{11} \le \alpha/2$ or $p_{12} \le \alpha/2$, respectively, and thus the null hypotheses in F_1 are tested using a Bonferroni rule. Further, it is easy to show that the null hypotheses in F_2 are tested by utilizing the Holm (1979) test. Specifically, let $p_{2(1)}$ and $p_{2(2)}$ be the ordered p-values in F_2 and $H_{2(1)}$ and $H_{2(2)}$ denote the null hypotheses corresponding to $p_{2(1)}$ and $p_{2(2)}$, respectively. Assume first that both H_{11} and H_{12} were rejected in F_1. In this case, $H_{2(1)}$ is rejected if $p_{2(1)} \le \alpha/2$ and $H_{2(2)}$ is rejected provided both $p_{2(1)} \le \alpha/2$ and $p_{2(2)} \le \alpha$. If, however, only one null hypothesis was rejected in F_1, it becomes more difficult to find significant outcomes in F_2. Specifically, the gatekeeping procedure rejects $H_{2(1)}$ if $p_{2(1)} \le \alpha/4$ and $H_{2(2)}$ if both $p_{2(1)} \le \alpha/4$ and $p_{2(2)} \le \alpha/2$.

Interestingly, this simple stepwise decision rule can be extended to the general case involving $m \ge 2$ gatekeepers. This decision rule uses penalized weighted Bonferroni tests for the first $m-1$ steps (for the first step, the penalized weighted Bonferroni test reduces to the ordinary weighted Bonferroni test) and a penalized weighted Holm test in the last step as described in the procedure given below. The tests are called "weighted" because they are based on weighted p-values, $q_{ij} = p_{ij}/w_{ij}$, $j = 1, \dots, n_i$. The tests are called "penalized" because the q_{ij} are compared with $\rho_i \alpha$ instead of α, where $0 \le \rho_i \le 1$ represents inverse of the penalty charged (the larger the ρ_i, the smaller the penalty). It is given by

$$\rho_1 = 1 \quad \text{and} \quad \rho_i = \prod_{k=1}^{i-1} \left[\sum_{j=1}^{n_k} r_{kj} w_{kj} \right], \quad i = 2, \dots, m ,$$

where $r_{kj} = 1$ if H_{kj} is rejected and 0 otherwise. Notice that if more rejections occur in earlier steps then ρ_i is larger resulting in a smaller penalty; for this reason we refer to ρ_i as the *rejection gain factor*. Note that the ρ_i must be calculated sequentially at each step after observing which hypotheses are rejected at that step.

3.1 Stepwise procedure in the general case

Step 1 For family F_1, use the weighted Bonferroni test to reject H_{1j} iff $q_{1j} \le \alpha$ for $j = 1, 2, \ldots, n_1$. If no H_{1j} is rejected then stop testing and retain all remaining hypotheses; otherwise go to Step 2.

Step *k* For family F_k, use the penalized weighted Bonferroni test to reject H_{kj} iff $q_{kj} \le \rho_k \alpha$ for $j = 1, 2, \ldots, n_k$. If no H_{kj} is rejected then stop testing and retain all remaining hypotheses. Otherwise let $k \leftarrow k + 1$. If $k < m$, return to Step k; otherwise go to Step m.

Step *m* For family F_m, first order the q_{mj} values as $q_{m(1)} \le \cdots \le q_{m(n_m)}$. Let $H_{m(1)}, \ldots, H_{m(n_m)}$ denote the corresponding hypotheses and $w_{m(1)}, \ldots, w_{m(n_m)}$ denote their weights. Use the penalized weighted Holm test to reject $H_{m(j)}$ iff

$$q_{m(i)} \le \frac{\rho_m \alpha}{w_{m(i)} + \ldots + w_{m(n_m)}} \quad \text{for all} \quad i = 1, \ldots, j\,.$$

This Holm test can be implemented in a stepwise manner by beginning with $H_{m(1)}$ and rejecting it iff $q_{m(1)} \le \rho_m \alpha$ and then proceeding to test $H_{m(2)}$, etc. Testing stops as soon as some hypothesis $H_{m(j)}$ cannot be rejected in which case all $H_{m(k)}$ for $k > j$ are retained automatically.

The critical values with which the raw p-values are compared, namely, $\alpha_{kj} = \alpha \rho_k w_{kj}$ in the penalized Bonferroni procedure for Steps $k = 1, \ldots, m-1$ and $\alpha_{m(j)} = \alpha \rho_m w_{m(j)} / [w_{m(j)} + \ldots + w_{m(n_m)}]$ in the penalized Holm procedure for Step m are referred to as *adjusted significance levels*.

Note that if $r_{ij} = 0$ for all j then $\rho_k = 0$ for all $k > i$. As a result, all $H_{kj} \in F_k$ would be retained automatically. Therefore F_k is tested iff all preceding gatekeepers are successfully passed, i.e., if at least one hypothesis was rejected in $F_1, \ldots, F_{k-1}$ $\left(\sum_{j=1}^{n_i} r_{ij} w_{ij} > 0, i = 1, \ldots, k-1 \right)$. The penalty one has to pay for performing multiple inferences in F_k depends on the number of the null hypotheses rejected in the previously examined gatekeepers. The rejection gain factor, ρ_k, equals 1 and thus no penalty is paid if all of the null hypotheses were rejected in $F_1, \ldots, F_{k-1}$. However, the rejection gain factor decreases with the number of hypotheses rejected in the preceding gatekeepers, which makes it more difficult to reject hypotheses later in the sequence.

Proposition 3.1 *The parallel Bonferroni gatekeeping procedure defined using the closed testing principle (Section 2) is equivalent to the stepwise gatekeeping procedure.*

The proof of the proposition is given in the Appendix.

4 Clinical Trial Example

To illustrate the utility of the stepwise version of the parallel Bonferroni gatekeeping procedure, we will use the clinical trial in patients with hypertension considered in Dmitrienko et al. (2005, Page 118). This clinical trial was conducted to test the efficacy and safety of four doses of an investigational drug versus placebo. The four doses will be labeled D1 (lowest dose) through D4 (highest dose) and placebo will be denoted by P. The primary efficacy endpoint was the reduction in diastolic blood pressure (measured in mm Hg).

The following hierarchical testing approach was defined prior to the beginning of the study. The four dose-placebo and four pairwise contrasts of interest were grouped into three families. Since Doses D3 and D4 were expected to be more efficacious than D1 and D2, the corresponding dose-placebo comparisons (D3 vs. P, D4 vs. P) were included in Family F_1. The other two dose-placebo comparisons (D1 vs. P, D2 vs. P) were included in Family F_2 and Family F_3 was comprised of the four pairwise contrasts (D4 vs. D1, D4 vs. D2, D3 vs. D1, D3 vs. D2). The null hypotheses in the first two families were to be tested in a parallel manner and null hypotheses within each family were

Table 1 Gatekeeping inferences in the hypertension trial example (P = Placebo, D1 through D4 denote the four doses of the experimental drug). The adjusted p-values are computed using the closed testing procedure and adjusted critical values are derived using the stepwise procedure. The familywise error rate is set at 0.05.

Family	Comparison	Raw p-value	Adjusted p-value	Adjusted significance level
F_1	D4-P	0.0008	0.0016	0.0250
	D3-P	0.0135	0.0269	0.0250
F_2	D2-P	0.0197	0.0394	0.0250
	D1-P	0.7237	1.0000	0.0250
F_3	D4-D1	0.0003	0.0394	0.0063
	D4-D2	0.2779	1.0000	0.0125
	D3-D1	0.0054	0.0394	0.0083
	D3-D2	0.8473	1.0000	0.0250

equally weighted ($w_{11} = w_{12} = 1/2$, $w_{21} = w_{22} = 1/2$ and $w_{31} = w_{32} = w_{33} = w_{34} = 1/4$). The FWE for the eight null hypotheses was set at $\alpha = 0.05$.

Hypothesis testing problems of this type arise in a variety of clinical trials with multiple endpoints when drug developers group these objectives into two or more categories, e.g., primary outcomes, more important secondary outcomes and tertiary outcomes. One encounters similar hierarchically ordered hypotheses in clinical trials designed to test several doses of an experimental drug versus placebo or an active control (Denne and Koch, 2002; Dmitrienko et al., 2006).

Table 1 displays the raw p-values for the eight null hypotheses in the hypertension clinical trial (computed using a two-sample t-test), adjusted p-values produced by the parallel Bonferroni gatekeeping approach based on a closed testing procedure and, finally, adjusted significance levels produced by the stepwise procedure.

It is instructive to compare the adjusted p-values and adjusted significance levels in Table 1. As indicated above, the adjusted p-values were obtained using a closed testing procedure by examining all $2^8 - 1 = 255$ intersection hypotheses in the closed family associated with F_1, F_2 and F_3. By contrast, the adjusted significance levels were computed using the stepwise algorithm of Section 3 in eight steps described below.

We will begin with the two hull hypotheses in Family F_1. The adjusted significance levels are computed using the weighted Bonferroni test ($\alpha_{11} = \alpha w_{11} = 0.025$ and $\alpha_{12} = \alpha w_{12} = 0.025$) and thus H_{11} (D4 vs. P) and H_{12} (D3 vs. P) are rejected. The rejection gain factor for Family F_2 is given by $\rho_2 = w_{11} + w_{12} = 1$. Given this rejection gain factor, the decision rule for H_{21} (D2 vs. P) and H_{22} (D1 vs. P) is based on comparing $p_{21} = 0.0197$ and $p_{22} = 0.7237$ to $\alpha_{21} = \alpha\rho_2 w_{21} = 0.025$ and $\alpha_{22} = \alpha\rho_2 w_{22} = 0.025$, respectively. Therefore H_{21} is rejected but H_{22} is retained. Since only one hypothesis was rejected in Family F_2, the rejection gain factor for the pairwise contrasts in Family F_3 is now less than 1, i.e., $\rho_3 = \rho_2 w_{21} = 0.5$. The null hypotheses H_{31} (D4 vs. D1), H_{32} (D4 vs. D2), H_{33} (D3 vs. D1) and H_{34} (D3 vs. D2) need to be tested using the penalized Holm test with $\rho_3 = 0.5$. Note that the adjusted significance level for $p_{3(j)}$, $j = 1, \dots, 4$, is given by

$$\alpha_{3(j)} = \alpha\rho_3 w_{3(j)} / \left(w_{3(j)} + \ldots + w_{3(4)}\right).$$

Comparing the raw p-values to the resulting adjusted significance level, it is easy to see that the stepwise procedure rejects H_{31} and H_{33}, whereas H_{32} and H_{34} are retained.

One can verify that the decisions based on the stepwise procedure are identical to those based on the original Bonferroni gatekeeping procedure (the adjusted p-value is no greater than $\alpha = 0.05$ iff the corresponding raw p-value is no greater than the adjusted significance level).

5 Conclusions

This paper introduces a streamlined algorithm for implementing the parallel gatekeeping procedure based on the Bonferroni test (Dmitrienko et al., 2003). It is shown that the parallel gatekeeping procedure, which was originally formulated as a general closed testing procedure, admits a useful shortcut – it is equivalent to a simple stepwise procedure. Using the stepwise procedure, one can test n null hypotheses in n steps rather than 2^n steps required in the general case. The stepwise procedure facilitates the implementation of gatekeeping procedures and provides an important insight into the nature of gatekeeping inferences. The only limitation of the stepwise procedure is that it cannot be used to compute adjusted p-values.

Appendix

Algorithm for defining hypothesis weights Select an arbitrary intersection hypothesis H from the closed family corresponding to $F_1, \dots, F_m$. The hypothesis weights, $v_{ij}(H)$, $i = 1, \dots, m$, $j = 1, \dots, n_i$, are defined using the following stepwise algorithm (Dmitrienko et al., 2005, Section 2.7):

$$v_{ij}(H) = \begin{cases} v_i^*(H)\,\delta_{ij}(H)\,w_{ij} & \text{if } i = 1, \dots, m-1\,, \\ v_i^*(H)\,\delta_{ij}(H)\,w_{ij} \Big/ \sum\limits_{\ell=1}^{n_m} \delta_{i\ell}(H)\,w_{i\ell} & \text{if } i = m\,, \end{cases}$$

where $v_1^*(H) = 1$ and $v_{i+1}^*(H) = v_i^*(H) - \sum\limits_{j=1}^{n_i} v_{ij}(H)$, $i = 1, \dots, m-1$.

Proof of Proposition 3.1 Denote by $r_i = \sum\limits_{j=1}^{n_i} r_{ij}$ the number of hypotheses rejected in F_i and, without loss of generality, assume that $H_{i1}, \dots, H_{ir_i}$ are the rejected hypotheses and $H_{ir_{i+1}}, \dots, H_{in_i}$ are the retained hypotheses in F_i. Unless otherwise specified, all of the references to rejection in the proof are to rejection by the closed testing procedure.

For F_1, the two procedures are equivalent because both reject H_{1j} iff $q_{1j} \le \alpha$. Next consider F_2. To derive the necessary and sufficient condition for rejecting a null hypothesis H_{2j}, $j = 1, \dots, r_1$, we need to obtain the corresponding condition for rejecting every intersection hypothesis $H \in \mathcal{H}_{2j}$. Note that, for any $H \in \mathcal{H}_{2j} \cap H_{1k}$, where $k = 1, \dots, r_1$, H is rejected automatically because $p_H \le q_{1k} \le \alpha$. Therefore we only need to consider H with $\delta_{1k}(H) = 0$, $k = 1, \dots, r_1$. Since $p_H \le p_{2j}/v_{2j}(H)$, a sufficient condition for rejecting H is

$$\frac{p_{2j}}{v_{2j}(H)} = \frac{p_{2j}}{w_{2j} v_2^*(H)} \le \frac{q_{2j}}{\min\limits_H v_2^*(H)} \le \alpha$$

and thus we need to find the intersection hypothesis $H \in \mathcal{H}_{2j}$ that minimizes $v_2^*(H)$. Now,

$$\min_H \; v_2^*(H) = 1 - \max_H \sum_{j=r_1+1}^{n_1} w_{1j}\delta_{1j}(H) = 1 - \sum_{j=r_1+1}^{n_1} w_{1j} = \sum_{j=1}^{n_1} r_{1j} w_{1j} = \rho_2$$

since $r_{1j} = 1$, $j = 1, \dots, r_1$ and $r_{1j} = 0$, $j = r_1 + 1, \dots, n_1$. Hence a sufficient condition for rejecting H is $q_{2j} \le \rho_2 \alpha$.

To show that this is also a necessary condition, consider $H^* = \bigcap\limits_{i=r_1+1}^{n_1} H_{1i} \cap H_{2j}$. For this hypothesis, as shown above, $v_2^*(H^*) = \min\limits_H v_1^*(H) = \rho_2$ and

$$p_{H^*} = \min \left\{ q_{1,r_1+1}, \dots, q_{1n_1}, \frac{q_{2j}}{\rho_2} \right\}.$$

By definition, $q_{1j} > \alpha$, $j = r_1 + 1, \dots, n_1$, and thus to reject H^* we must have

$$p_{H^*} \le \alpha \iff q_{2j} \le \rho_2 \alpha\,,$$

which is hence also the necessary condition for rejecting H^*. Since any $H \in \mathcal{H}_{2j}$ is rejected iff $q_{2j} \le \rho_2 \alpha$, H_{2j} is also rejected iff $q_{2j} \le \rho_2 \alpha$.

Next consider F_3. For any $H \in \mathcal{H}_{3j} \cap \mathcal{H}_{1k}$ or $H \in \mathcal{H}_{3j} \cap \mathcal{H}_{2\ell}$, where $k = 1, \ldots, r_1$, $\ell = 1, \ldots, r_2$, H is rejected automatically following a similar argument as F_2. Therefore consider H that is contained in H_{3j} but not in $\mathcal{H}_{1k}$ or $\mathcal{H}_{2\ell}$, $k = 1, \ldots, r_1$, $\ell = 1, \ldots, r_2$. Since $p_H \le p_{3j}/v_{3j}(H)$, a sufficient condition for rejecting $H \in \mathcal{H}_{3j}$ is

$$\frac{p_{3j}}{v_{3j}(H)} = \frac{p_{3j}}{w_{3j} v_3^*(H)} \le \frac{q_{3j}}{\min\limits_H \; v_3^*(H)} \le \alpha\,.$$

As before,

$$\min_H \; v_3^*(H) = \min_H \; v_2^*(H) \sum_{j=1}^{n_2} r_{2j} w_{2j} = \left[\sum_{j=1}^{n_1} r_{1j} w_{1j}\right] \left[\sum_{j=1}^{n_2} r_{2j} w_{2j}\right] = \prod_{i=1}^{2} \left[\sum_{j=1}^{n_i} r_{ij} w_{ij}\right] = \rho_3\,.$$

Therefore $q_{3j} \le \rho_3 \alpha$ is a sufficient condition for rejecting H. To show that this is also a necessary condition, consider $H^* = \bigcap\limits_{i=r_1+1}^{n_1} H_{1i} \bigcap\limits_{i=r_2+1}^{n_2} H_{2i} \cap H_{3j}$ and use the same argument as in the case of F_2. This proof extends to any $H_{ij} \in F_i$, $i = 4, \ldots, m-1$.

For F_m, the proof differs from the proof for the previous families because the weights are normalized at the last step and so the weights $v_{mj}(H)$ depend not only on $v_m^*(H)$ and w_{mj} but also on the w_{mk} values of other H_{mk} included in H. At the first step, all H_{mj} are eligible for rejection. To determine which particular H_{mj} is the first one eligible for rejection, note that, as shown for the previous families, the sufficient condition for rejecting every $H \in \mathcal{H}_{mj}$ is

$$\frac{p_{mj}}{\min\limits_H \; v_{mj}(H)} \le \alpha \Longleftrightarrow \frac{p_{mj}(w_{m1} + \cdots + w_{mn_m})}{w_{mj} \min\limits_H \; v_m^*(H)} \le \alpha \Longleftrightarrow q_{mj} \le \frac{\rho_m \alpha}{w_{m(1)} + \cdots + w_{m(n_m)}}\,,$$

where, as before, one can show that

$$\min_H \; v_m^*(H) = \prod_{i=1}^{m-1} \left[\sum_{j=1}^{n_i} r_{ij} w_{ij}\right] = \rho_m\,.$$

The above condition holds for at least one H_{mj} iff it holds for $H_{m(1)}$ since $q_{m(1)}$ is the smallest of the q_{mj} for $j = 1, \ldots, n_m$. The condition can be shown to be necessary by considering

$$H^* = \bigcap_{j=r_1+1}^{n_1} H_{1j} \cap \cdots \bigcap_{j=r_{m-1}+1}^{n_{m-1}} H_{m-1,j} \bigcap_{j=1}^{n_m} H_{m(j)}\,.$$

This is the first step of the penalized weighted Holm procedure for rejecting $H_{m(1)}$.

This argument can be extended to testing the next hypothesis. Since $H_{m(1)}$ was rejected, it is now excluded from H. Therefore the next H_{mj} to be rejected can be shown to be $H_{m(2)}$ associated with the smallest of the remaining q_{mj} and the sufficient condition for rejecting every $H \in \mathcal{H}_{m(2)}$ to be

$$q_{m(2)} \le \frac{\min\limits_H \; v_m^*(H)\, \alpha}{\sum\limits_{k=2}^{n_m} w_{m(k)}} = \frac{\rho_m \alpha}{\sum\limits_{k=2}^{n_m} w_{m(k)}}\,.$$

This can also be shown to be a necessary condition by considering

$$H^* = \bigcap_{j=r_1+1}^{n_1} H_{1j} \cap \cdots \bigcap_{j=r_{m-1}+1}^{n_{m-1}} H_{m-1,j} \bigcap_{j=2}^{n_m} H_{m(j)}\,.$$

Using the same argument it can be shown that $H_{m(j)}$ is rejected iff

$$q_{m(i)} \le \frac{\rho_m \alpha}{\sum\limits_{k=i}^{n_m} w_{m(k)}} \quad \text{for all} \quad i \le j, \quad j = 1, \ldots, m\,.$$

Hence the equivalence between the parallel gatekeeping procedure defined using the closed testing principle and the stepwise procedure is established.

References

Bauer, P., Röhmel, J., Maurer, W., and Hothorn, L. (1998). Testing strategies in multi-dose experiments including active control. *Statistics in Medicine* **17**, 2133–2146.

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics* **24**, 295–300.

Chen, X., Luo, X., and Capizzi, T. (2005). The application of enhanced parallel gatekeeping strategies. *Statistics in Medicine* **24**, 1385–1397.

Chen, X., Capizzi, T., Binkowitz, B., Quan, H., Wei, L., and Luo, X. (2005). Decision rule based multiplicity adjustment strategy. *Clinical Trials* **2**, 394–399.

Denne, J. S. and Koch, G. G. (2002). A sequential procedure for studies comparing multiple doses. *Pharmaceutical Statistics* **1**, 107–118.

Dmitrienko, A., Offen, W., and Westfall, P. (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.

Grechanovsky, E. and Hochberg, Y. (1999). Closed procedures are better and often admit a shortcut. *Journal of Statistical Planning and Inference* **76**, 79–91.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. Wiley, New York.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.

Marcus, R., Peritz, E., and Gabriel, K. R. (1976). On closed testing procedure 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. *Biometrie in der chemisch-pharmazeutischen Industrie*. Vollmar, J. (Editor). Stuttgart: Fischer Verlag. **6**, 3–18.

Westfall, P. H. and Krishen, A. (2001). Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures. *Journal of Statistical Planning and Inference* **99**, 25–40.

Westfall, P. H., Zaykin, D. V., and Young, S. S. (2001). Multiple tests for genetic effects in association studies. *Biostatistics Methods* (*Methods in Molecular Biology, Biotechnology and Medicine* series). Looney, S. (Editor). Humana Press, Inc.

