STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:1385–1397
Published online 29 November 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2005

The application of enhanced parallel gatekeeping strategies

Xun Chen*,†, Xiaohui Luo, and Tom Capizzi

Clinical Biostatistics, Merck Research Laboratories, RY34-A316, Rahway, NJ 07065, U.S.A.

SUMMARY

The parallel gatekeeping strategy proposed by Dmitrienko *et al.* (*Statist. Med.* 2003; **22**:2387–2400) provides a flexible framework for the pursuit of strong control on study wise type I error rate. This paper further explores the application of the weighted Simes parallel gatekeeping procedure recommended by Dmitrienko *et al.* and proposes some modifications to it to better incorporate the interrelationships of different hypotheses in actual clinical trials and to achieve better power performance. We first propose a simple method to quantitatively control the impact of secondary tests on the testing of primary hypotheses. We then introduce a matched gatekeeping procedure to exemplify how to address special relationships between individual primary and secondary tests following the parallel gatekeeping framework. Our simulation study demonstrates that the enhanced gatekeeping procedures generally result in more powerful tests than the parallel gatekeeping procedure in Dmitrienko *et al.* whenever applicable.
Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: closed testing; interrelationship; matching; minimum primary weight; strong control

1. INTRODUCTION

Multiple comparisons in clinical trials may arise in a variety of situations, for example, having multiple endpoints, comparing multiple treatments or multiple doses, analysing data in various predefined subgroups of interest, etc. Thorough evaluation of a novel therapeutic intervention and its impact on underlying disease usually requires tests on multiple endpoints. This may be due to the multifactorial nature of many diseases, the ability to measure many aspects of therapeutic benefit, the lack of consensus on appropriate measures of effectiveness, or regulatory considerations on standards of evidence, etc. It is also becoming more commonplace to evaluate several doses of a drug, a placebo control, and one or more active controls in one trial. Simon [1] stated that many of the uncertainties in the conclusions of clinical trials result from problems of multiplicity. It is well recognized that the control of study wise false positive rate is an important principle and is often of great value in the assessment of results

*Correspondence to: Xun Chen, Clinical Biostatistics, Merck Research Laboratories, RY34-A316, Rahway, NJ 07065, U.S.A.
†E-mail: xun_chen@merck.com

Received February 2004
Accepted July 2004
Copyright © 2004 John Wiley & Sons, Ltd.

from clinical trials. A testing procedure having strong control on study wise type I error rate provides the most rigorous multiplicity adjustment [2]. Although it is arguable regarding to the necessity of pursuing study wise strong control on type I error rate, i.e. eliminating the inflation of type I error under all possible null hypothesis configurations [3, 4], having more such kind of testing procedures is generally desirable to achieve more appropriate and more powerful strong control on the study wise type I error rate of various clinical trials whenever necessary.

The 'parallel' gatekeeping strategy proposed by Dmitrienko *et al.* [5] provides an alternative way to strongly control the study wise type I error rate of a clinical trial in contrast to the usual 'serial' gatekeeping procedure. Without loss of generality, consider two families of tests in a clinical trial—with primary hypotheses in the gatekeeper family and secondary hypotheses in the secondary family. The parallel gatekeeping procedure allows one to proceed to the secondary family when at least one of the tests in the gatekeeper family exhibits significance, whereas serial gatekeeping procedure allows one to proceed to the secondary family only if all of the hypotheses in the family are rejected. The parallel gatekeeping procedure is thus less restrictive and is more appropriate especially when the co-primary endpoints can lead to separate claims. Dmitrienko *et al*. presented the parallel gatekeeping procedure primarily in the format of weighted Bonferroni method and demonstrated two appealing properties of this procedure: (1) the adjusted p-values of the gatekeeper hypotheses do not depend on the significance of the p-values associated with the secondary hypotheses; (2) the adjusted p-values associated with the secondary hypotheses are always greater than the minimum of the adjusted p-values for the gatekeeper hypotheses, in other words, the secondary hypotheses can be tested if at least one gatekeeper hypothesis has been rejected. It was not clear, however, whether these two appealing properties held for the later recommended more powerful weighted Simes based parallel gatekeeping procedure. From the formulas of the weighted Simes procedure, where the adjusted primary test p-values are defined as functions of all p-values—of both primary tests and secondary tests—one has reason to ask whether the first property still holds for the weighted Simes parallel gatekeeping procedure. That is, in the weighted Simes procedure, the adjusted p-values of the gatekeeper hypotheses may be affected by the significance of the secondary hypotheses in contrast to the Bonferroni method. If this is true, it could be very worrisome as secondary hypotheses are often not sufficiently powered in many clinical trials. And it is very important for a statistician to become aware of the possible impacts of the non-significant secondary tests before applying the weighted Simes procedure to prevent possibly substantial loss of power in primary tests.

Another concern with the Dmitrienko *et al*.'s parallel gatekeeping procedure is that although this procedure gives us the ability to incorporate the secondary tests with less restriction (as compared to the serial gatekeeping procedure), it does not take into account the potential logical relationship between *individual* primary and secondary tests. For example, some secondary tests may not be meaningful or necessary after the failure of certain primary tests. If we can impose such kind of 'matching' relationships between individual primary and secondary tests into the gatekeeping procedure, more powerful tests may be expected.

In this paper, we focus our discussions on the application of the weighted Simes test following the recommendation by Dmitrienko *et al*. [5] (over the less powerful Bonferroni method). We propose modifications to the existing procedure to quantitatively control the impact of secondary tests on primary (gatekeeper) tests and to exemplify how to more appropriately address the special hierarchical relationships between individual primary and secondary tests

Copyright © 2004 John Wiley & Sons, Ltd.

through the gatekeeping procedure. The paper is organized as follows. Section 2 reviews the weighted Simes parallel gatekeeping procedure in contrast to the serial gatekeeping procedure. Section 3 proposes some enhancements to the weighted Simes procedure. The power of the various parallel gatekeeping procedures will be compared in Section 4 through a simulation study. An example is provided in Section 5.

2. SIMES TEST BASED GATEKEEPING PROCEDURE

In this paper, the multiplicity problems will be illustrated in a form of multiple endpoints with multiple doses of an active treatment compared to control. To simplify the presentation, we limit the number of families to two relating to the situation of the investigating two doses. The discussed methods can be easily generalized to other multiplicity problems and to problems with more than two families.

Suppose H_{ij} is the null hypothesis of comparing the ith dose to the control on the jth endpoint, and p_{ij} is the corresponding p-value, $i=1,2$ and $j=1,2,\ldots,k$. Assume no prioritization on the k endpoints. Dose 1 is known to be primary focus and Dose 2 is the secondary dose. The hypotheses can be grouped into two families, $F_1=\{H_{11},\ldots,H_{1k}\}$ and $F_2=\{H_{21},\ldots,H_{2k}\}$. The first family F_1 serves as a gatekeeper for F_2 such that F_2 is examined only if the gatekeeper has been successfully passed. We will refer to the hypotheses in F_1 as gatekeeper hypotheses or primary hypotheses and the hypotheses in F_2 as secondary family hypotheses in this paper hereafter.

The gatekeeping procedure in Reference [5] applies the closed testing principle [6] to achieve strong control on the study wise type I error rate with respect to families F_1 and F_2. That is, for an arbitrary intersection hypothesis H associated with t elemental hypotheses from $F_1\cup F_2=\{H_{11},\ldots,H_{2k}\}$, let p_H denote the p-value associated with H based on a test procedure when H is true; the gatekeeping procedure defines the adjusted p-value of hypothesis H_{ij} as

$$\tilde{p}_{ij}=\max_{H\in\aleph_{ij}} p_H \tag{1}$$

where $\aleph_{ij}$ denotes the set of all intersection hypotheses that contain H_{ij}. The procedure rejects H_{ij} when $\tilde{p}_{ij}\leqslant\alpha$. In other words, the gatekeeping procedure rejects hypothesis H_{ij} only if all of the intersection hypotheses that contain H_{ij} are rejected. The gatekeeping procedure thus controls the study wise type I error rate at α level with respect to hypotheses in $F_1\cup F_2$ in strong sense following the closed testing principle.

The weighted Simes test defines the p-value of an intersection hypothesis H as follows [7]. Let $p_{(1)H}\leqslant p_{(2)H}\leqslant\cdots\leqslant p_{(t)H}$ denote the ordered p-values for the t elemental hypotheses in H, and let these elemental hypotheses have weights (in H) equal to $v_{(1)}(H),\ldots,v_{(t)}(H)$, respectively, the p-value of the intersection hypothesis H is then defined as

$$p_H=\min_{1\leqslant l\leqslant t} p_{(l)H}\Big/\sum_{i=1}^{l} v_{(i)H} \tag{2}$$

As noted in Reference [5], the proof of type I error control for this procedure under positive regression dependency is available by Kling and Benjamini in an unpublished manuscript in 2002. The reader may also refer to Reference [8] for detailed discussions about the property

Copyright © 2004 John Wiley & Sons, Ltd.

of positive regression dependency. The serial gatekeeping procedure and the parallel gatekeeping procedure presented by Dmitrienko *et al.* define the weights, $v_{ij}(H)$, differently. The key difference is that when H contains partial gatekeeper hypotheses and partial or all secondary family hypotheses, the remaining weight of those partial gatekeeper hypotheses is assigned to the secondary family hypotheses in H according to their respective importance in F_2 in the parallel procedure, whereas in the serial procedure, the remaining weight of those partial gatekeeper hypotheses is re-assigned to the gatekeeper hypotheses in H according to their respective importance in F_1, and none to the secondary family hypotheses. And consequently, following the serial gatekeeping procedure, there will be no chance to test hypotheses in F_2 unless all F_1 tests exhibit significance; whereas following the parallel gatekeeping procedure, one is allowed to proceed to the secondary family even if some of the tests in F_1 are not significant. Specifically, let w_{ij} represent the relative importance of the corresponding hypothesis H_{ij} in family F_i with $w_{11}+\cdots+w_{1k}=1$ and $w_{21}+\cdots+w_{2k}=1$. The serial gatekeeping procedure defines $v_{ij}(H)$ as follows:

Case 1: If H contains at least one gatekeeper hypotheses, let

$$v_{1j}(H)=w_{1j}I(H_{1j}\in H)\Big/\sum_{i=1}^{k}w_{1i}I(H_{1i}\in H)\quad\text{and}\quad v_{2j}(H)=0\quad\text{for } j=1,\ldots,k$$

Case 2: If H does not contain any gatekeeper hypotheses, let $v_{1j}(H)=0$ and

$$v_{2j}(H)=w_{2j}I(H_{2j}\in H)\Big/\sum_{i=1}^{k}w_{2i}I(H_{2i}\in H)\quad\text{for } j=1,\ldots,k$$

In contrast, the parallel gatekeeping procedure defines $v_{ij}(H)$ as follows:

Case 1: If H contains all gatekeeper hypotheses, let $v_{1j}(H)=w_{1j}$ and

$$v_{2j}(H)=0\quad\text{for } j=1,\ldots,k$$

Case 2: If H only contains partial gatekeeper hypotheses but no secondary family hypotheses, let $v_{1j}(H)=w_{1j}I(H_{1j}\in H)\Big/\sum_{i=1}^{k}w_{1i}I(H_{1i}\in H)$ and $v_{2j}(H)=0$ for $j=1,\ldots,k$.

Case 3: If H contains partial gatekeeper hypotheses and partial or all secondary family hypotheses, let $v_{1j}(H)=w_{1j}I(H_{1j}\in H)$ and

$$v_{2j}(H)=\left(1-\sum_{i=1}^{k}w_{1i}I(H_{1i}\in H)\right)w_{2j}I(H_{2j}\in H)\Big/\sum_{i=1}^{k}w_{2i}I(H_{2i}\in H)\quad\text{for } j=1,\ldots,k$$

Case 4: If H does not contain any gatekeeper hypotheses, let $v_{1j}(H)=0$ and

$$v_{2j}(H)=w_{2j}I(H_{2j}\in H)\Big/\sum_{i=1}^{k}w_{2i}I(H_{2i}\in H)\quad\text{for } j=1,\ldots,k$$

Now that the parallel gatekeeping procedure allows to proceed to the secondary family F_2 despite of the non-significance of some tests in the gatekeeper family F_1, will the non-significance of the tests in F_2 affect the significance of the tests in F_1? When the weighted Bonferroni method is used to define p_H, Dmitrienko *et al.* showed that the adjusted p-values for the gatekeeper hypotheses do not depend on the significance of the secondary tests (where the adjusted p-values for the gatekeeper hypotheses are given by $\tilde{p}_{1j}=p_{1j}/w_{1j}$, independent of

Copyright © 2004 John Wiley & Sons, Ltd.

Table I. An illustration of the Simes parallel gatekeeping procedure in Reference [5] in different scenarios.

		Scenario 1		Scenario 2		Scenario 3	
Test	Weight	Raw p-value	Adjusted p-value (Simes)	Raw p-value	Adjusted p-value (Simes)	Raw p-value	Adjusted p-value (Simes)
Primary 1 (H_{11})	0.9	0.048	0.048	0.048	0.053	0.048	0.053
Primary 2 (H_{12})	0.1	0.003	0.030	0.003	0.030	0.030	0.056
Secondary 1 (H_{21})	0.5	0.026	0.048	0.060	0.060	0.060	0.060
Secondary 2 (H_{22})	0.5	0.002	0.040	0.002	0.040	0.002	0.048

the situations of tests in F_2). However when the weighted Simes method is used to define p_H, as shown in (2), the adjusted p-values are defined by functions involving all elementary test p-values—both primary and secondary. One has reason to suspect that the non-significance of tests in F_2 may affect the tests in F_1 when applying the weighted Simes parallel procedure.

To answer this question, we revisited one of the illustrative examples in Reference [5, p. 2395, Table III] and altered the raw p-values to check the performance of the weighted Simes procedure. The results of these illustration examples are all presented in Table I. Scenario 1 exactly copied Scenario 3 in Reference [5, Table III] and we note that given those raw p-values, the significance of the tests in F_1 is the same as they are tested alone (from the secondary tests). In Scenario 2, we altered one of the raw p-values of the secondary tests and made it greater than the significance level of 0.05. We find the significance of the primary tests is now obviously affected. And further in Scenario 3, we find that neither does the weighted Simes procedure hold the second property of the Bonferroni gatekeeping procedure, that is, the adjusted p-values of the secondary hypotheses could be smaller than the minimum of the adjusted p-values of the primary hypotheses in the weighted Simes procedure. In other words, the secondary tests could be significant even if none of the primary test is significant—the primary tests no longer gatekeep the secondary tests.

Hence the adjusted p-values of tests in F_1 could vary with the raw p-values of tests in F_2 when applying the weighted Simes parallel procedure. Consequently, one may substantially lose power in primary tests (as compared to the weighted Simes serial gatekeeping, for example) by 'lending' too many chances to the secondary tests. For example, assume there are five hypotheses in F_1 and five hypotheses in F_2. Let $w_{11} = \cdots = w_{15} = 0.2$, $w_{21} = \cdots = w_{25} = 0.2$. Following the weighted Simes parallel gatekeeping procedure in Dmitrienko *et al.*, for the intersection hypothesis $H = H_{11} \cap H_{21}$, the weight assigned to H_{11} will be 0.2 whereas the remaining weight 0.8 will be assigned to the secondary hypothesis H_{21}. Consequently, if $p_{21} > \alpha$, p_H will be greater than α only if $p_{11} \leqslant \alpha/5$. In other words, one will have no chance to reject H_{11} unless $p_{11} \leqslant \alpha/5$. The impact of the non-significant secondary tests (on the primary tests) seems too strong in this example. Proper modifications are needed to enhance the application of the Simes parallel gatekeeping procedure to avoid extensive loss of power for primary tests.

Also the parallel gatekeeping procedure discussed in Dmitrienko *et al*. does not take into account the potential special logical relationship between individual primary and secondary tests. For example, in the multiple endpoints multiple dose trial, the test on H_{11} is usually more closely related to the test on H_{21} (on the same endpoint) than to the test on H_{22} (on a different endpoint). In such a case, the weighting scheme of the parallel gatekeeping procedure needs

Copyright © 2004 John Wiley & Sons, Ltd.

to be modified to reflect the 'matching' gatekeeping pattern between F_1 and F_2 to achieve more appropriate and more efficient multiplicity adjustment.

3. MODIFICATIONS ON THE ORDINARY PARALLEL GATEKEEPING PROCEDURE

In this section, we discuss the different enhancements for the parallel gatekeeping procedure in Dmitrienko *et al.* (referred to as ordinary parallel gatekeeping procedure hereafter). At first, as we noted previously, when an intersection hypothesis H contains partial gatekeeper hypothesis and partial or all secondary family hypotheses, it may not be appropriate to assign w_{1j} to the gatekeeper hypotheses in H and distribute the remaining $1-\sum_{j=1}^{k} w_{1j}I(H_{1j}\in H)$ to the secondary family hypotheses in H. The reason is that w_{ij} only represents the relative importance of the gatekeeper hypothesis H_{1j} in the gatekeeper family F_1 not its relative importance to the hypotheses in the secondary family F_2. There is a need for the statistician to go further to explore the relative importance of the gatekeeper hypotheses to the secondary family hypotheses and determining the most appropriate weights for the individual hypothesis in an intersection H. Other than visiting any individualized weighting scheme, we discuss a general way to secure the power of tests in the gatekeeper family F_1 following the weighed Simes parallel gatekeeping framework.

We consider modifying the ordinary parallel gatekeeping procedure as follows:

Case 3: If H contains partial gatekeeper hypotheses and partial or all secondary family hypotheses, let

$$v_{1j}(H)=\max\left(\gamma,\sum_{i=1}^{k} w_{1i}I(H_{1i}\in H)\right)w_{1j}I(H_{1j}\in H)\Big/\sum_{i=1}^{k} w_{1i}I(H_{1i}\in H)$$

and

$$v_{2j}=\left(1-\max\left(\gamma,\sum_{i=1}^{k} w_{1i}I(H_{1i}\in H)\right)\right)w_{2j}I(H_{2j}\in H)\Big/\sum_{i=1}^{k} w_{2i}I(H_{2i}\in H)\quad \text{for } j=1,\ldots,k \tag{3}$$

where γ $(0\leqslant\gamma\leqslant 1)$ represents the 'minimum' relative importance of gatekeeper hypotheses as to secondary family hypotheses in an intersection. This modified procedure provides a simple way to impose the relative importance of gatekeeper hypotheses (beyond w_{ij}'s) into the weighting scheme. For example, by setting $\gamma=2/3$, in the earlier example with five gatekeeper hypotheses and five secondary family hypotheses, the intersection hypothesis $H=H_{12}\cap H_{21}$ will have weight $2/3$ for the primary hypothesis H_{11} and the remaining $1/3$ for the secondary hypothesis H_{21}, which is more reasonable than the original weight assignment of 0.2 vs 0.8 in general. Note when $\gamma=1$, the proposed gatekeeping procedure is converted to a serial gatekeeping procedure which provides maximum protection for the gatekeeper hypotheses (from the insignificance of secondary family tests), whereas when $\gamma=0$, the proposed gatekeeping procedure is converted to the ordinary parallel gatekeeping procedure in Dmitrienko *et al.* The value of the threshold γ should be pre-defined and should be selected to best reflect the

Copyright © 2004 John Wiley & Sons, Ltd.

Table II. Weights $v_t(H)$ for each intersection test in the parallel gatekeeping procedure with minimum primary weight $\frac{2}{3}$ and in the matched parallel gatekeeping procedure.

	Weights ($v_t(H)$)							
	Minimum primary weight $\gamma = 2/3$				Matched parallel gatekeeping procedure			
Intersection hypothesis H	H_{11}	H_{12}	H_{21}	H_{22}	H_{11}	H_{12}	H_{21}	H_{22}
$H_{11} \cap H_{12} \cap H_{21} \cap H_{22}$	0.5	0.5	0	0	0.5	0.5	0	0
$H_{11} \cap H_{12} \cap H_{21}$	0.5	0.5	0	0	0.5	0.5	0	0
$H_{11} \cap H_{12} \cap H_{22}$	0.5	0.5	0	0	0.5	0.5	0	0
$H_{11} \cap H_{12}$	0.5	0.5	0	0	0.5	0.5	0	0
$H_{11} \cap H_{21} \cap H_{22}$	2/3	0	1/6	1/6	0.5	0	0	0.5
$H_{11} \cap H_{21}$	2/3	0	1/3	0	1.0	0	0	0
$H_{11} \cap H_{22}$	2/3	0	0	1/3	0.5	0	0	0.5
H_{11}	1	0	0	0	1	0	0	0
$H_{12} \cap H_{21} \cap H_{22}$	0	2/3	1/6	1/6	0	0.5	0.5	0
$H_{12} \cap H_{21}$	0	2/3	1/3	0	0	0.5	0.5	0
$H_{12} \cap H_{22}$	0	2/3	0	1/3	0	1	0	0
H_{12}	0	1	0	0	0	1	0	0
$H_{21} \cap H_{22}$	0	0	0.5	0.5	0	0	0.5	0.5
H_{21}	0	0	1	0	0	0	1	0
H_{22}	0	0	0	1	0	0	0	1

relative importance between gatekeeper hypotheses and secondary family hypotheses and to balance the power for primary tests and secondary tests.

We can also enhance the ordinary parallel gatekeeping procedure by imposing pre-specified special logical relationships between individual primary and secondary tests into the weighting scheme. For example, in the multiple endpoints multiple doses clinical trial, it may be felt *a priori* that it is not meaningful to exhibit the significance of test for H_{21} (endpoint 1 in lower dose) unless the significance of test for H_{11} (endpoint 1 in higher dose) is achieved; and the same for H_{12} and H_{22}. To reflect such kind of 'matching' parallel gatekeeping relationship, we may modify the ordinary parallel gatekeeping procedure as follows (without loss of generality, assume one to one match between H_{1j} and H_{2j}, $j = 1, \ldots, k$):

Case 3: If H contains partial gatekeeper hypotheses and partial or all secondary family hypotheses, let $v_{1j}(H) = w_{1j} I(H_{1j} \in H)$ and

$$v_{2j} = w_{2j} I(H_{2j} \in H) I(H_{1j} \notin H) \left(1 - \sum_{i=1}^{k} v_{1i}(H)\right) \Big/ \sum_{i=1}^{k} w_{2i} I(H_{2i} \in H)\, I(H_{1i} \notin H)$$
$$\text{for } j = 1, \ldots, k \tag{4}$$

The modified weighting schemes are illustrated in Table II assume $k = 2$ and $w_{11} = w_{12} = w_{21} = w_{22} = 0.5$. Note that when H contains partial gatekeeper hypotheses and partial or all secondary family hypotheses, in the matched gatekeeping procedure, 0 weights are assigned to those secondary family hypotheses if their matched gatekeepers are in the intersection. Consequently, larger weights will be assigned to the remaining secondary tests in the intersection, and more powerful secondary tests may be expected. For example, in the case of H

Copyright © 2004 John Wiley & Sons, Ltd.

equals $H_{11} \cap H_{21} \cap H_{22}$, the weights assigned to the two secondary hypotheses in the ordinary parallel gatekeeping procedure is (0.25, 0.25) whereas in the 'matched' parallel gatekeeping procedure the weights are (0,0.5) as H_{21} is tested only if H_{11} is rejected.

The two modifications proposed above may be applied jointly whenever applicable and the weight may be defined as

Case 3: If H contains partial gatekeeper hypotheses and partial or all secondary family hypotheses, let

$$v_{1j}(H) = \max\left(\gamma, \sum_{i=1}^{k} w_{1i} I(H_{1i} \in H)\right) w_{1j} I(H_{1j} \in H) \Big/ \sum_{i=1}^{k} w_{1i} I(H_{1i} \in H)$$

and

$$v_{2j} = \left(1 - \max\left(\gamma, \sum_{i=1}^{k} w_{1i} I(H_{1i} \in H)\right)\right) w_{2j} I(H_{2j} \in H) I(H_{1j} \notin H) \Big/ \sum_{i=1}^{k} (w_{2i} I(H_{2i} \in H) I(H_{1i} \notin H)) \quad \text{for } j = 1, \ldots, k$$

4. POWER COMPARISON

To achieve better understanding to the performance of the proposed procedures, a simulation study similar to that in Reference [5] was conducted to assess the power of the various testing procedures discussed in this paper. We considered the case of four null hypotheses grouped into two families, say, H_{11} and H_{12} in the primary (gatekeeper) family, and H_{21} and H_{22} in the secondary family. The four individual null hypotheses are of the form $\mu_{ij} = 0$ $(i = 1, 2,\ j = 1, 2)$ where $\mu_{11}, \ldots, \mu_{22}$ represent the means of four normally distributed random variables $X_{11}, \ldots, X_{22}$ with standard deviation 1 and a common correlation coefficient ρ. Assuming that the null hypotheses are equally weighted within each family, that is, $w_{11} = \cdots = w_{22} = 0.5$, and the significance of the test for the secondary hypothesis H_{2j} is meaningful only if the corresponding primary hypothesis H_{1j} is rejected. Note that under these settings, the minimum primary weight with value $0 \leqslant \gamma \leqslant 1/2$ in (3) all results in the ordinary parallel gatekeeping procedure. The ordinary parallel procedure is compared with the parallel procedure with a different minimum primary weight, say, $\gamma = 2/3$ and two different versions of the matched parallel gatekeeping procedure, one with the minimum primary weight $0 \leqslant \gamma \leqslant 1/2$, the other with the minimum primary weight $\gamma = 2/3$. The performance of the serial gatekeeping procedure, which corresponds to a maximum 'minimum primary weight' $\gamma = 1$, is also evaluated and compared with the different parallel gatekeeping procedures. The simulation results are summarized from $1,000,000$ replications. The tested nominal level is the conventional $\alpha = 5$ per cent. That is, the different gatekeeping procedures are to control the study wise type I error at 5 per cent level.

Table III summarizes the results of the simulation study. Note when applying the 'non-matched' parallel gatekeeping procedure, it is possible that a test for H_{2j} has an adjusted p-value less than the significance level yet the test for H_{1j} has an adjusted p-value greater than the significance level. Such kind of significance of H_{2j} should not be counted based on the priorly agreed hierarchic relationship between H_{1j} and H_{2j}. However, to achieve better

Copyright © 2004 John Wiley & Sons, Ltd.

Table III. Estimated power (per cent) of the different gatekeeping procedure (nominal level $\alpha = 5$ per cent).

	H_{11}					H_{21}				
Parameters $(\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}; \rho)$	Non-Matched parallel $0 \leqslant \gamma \leqslant 1/2^*$	Non-Matched parallel $\gamma = 2/3$	Matched parallel $0 \leqslant \gamma \leqslant 1/2$	Matched parallel $\gamma = 1$	Serial $\gamma = 1$	Non-matched parallel $0 \leqslant \gamma \leqslant 1/2$ Pass H_{11}/ Pass F_1	Non-matched parallel $\gamma = 2/3$ Pass H_{11}/ Pass F_1	Matched parallel $0 \leqslant \gamma \leqslant 1/2$	Matched parallel $\gamma = 2/3$	Serial $\gamma = 1$
(0, 0, 0, 0; 0)	2.5	2.5	2.5	2.5	2.6	0.0/0.1	0.0/0.0	0.1	0.0	0.0
(3, 0, 0, 0; 0)	77.6	77.7	77.6	77.8	77.9	1.0/1.1	0.7/0.7	2.0	1.3	0.1
(1, 3, 0, 0; 0)	10.8	12.7	11.1	12.9	16.1	0.3/1.2	0.3/0.9	0.3	0.3	0.4
(3, 3, 0, 0; 0)	77.6	80.4	77.9	80.5	84.0	1.9/2.1	1.9/2.0	2.0	2.0	1.9
(1, 0, 3, 0; 0)	10.8	10.9	10.8	10.9	11.1	7.5/9.2	7.1/8.6	8.4	8.0	0.7
(3, 0, 3, 0; 0)	77.6	77.8	77.6	77.8	78.0	54.1/54.6	50.4/50.8	60.2	56.8	3.3
(3, 1, 3, 0; 0)	77.6	78.2	77.6	78.2	78.9	54.8/56.8	51.9/53.4	60.3	57.6	11.3
(1, 3, 3, 0; 0)	11.0	12.8	11.0	12.9	16.0	8.5/57.0	9.8/53.5	8.6	10.0	11.2
(3, 3, 3, 0; 0)	77.9	80.6	78.0	80.6	84.0	59.6/72.1	61.0/70.9	60.8	62.2	56.5
(3, 3, 3, 1; 0)	78.5	80.8	78.6	80.9	83.9	60.9/73.0	62.1/71.8	62.0	63.2	57.1
(3, 3, 0, 3; 0)	77.9	80.4	83.0	83.4	83.9	3.6/3.9	3.5/3.7	3.6	3.5	3.4
(1, 3, 1, 3; 0)	11.5	13.2	15.3	15.5	16.1	2.4/8.8	2.4/7.2	2.4	2.4	2.3
(3, 3, 1, 3; 0)	78.6	81.0	83.1	83.5	84.1	12.7/13.9	12.4/13.3	12.8	12.5	11.6
(1, 1, 3, 3; 0)	11.6	11.7	11.7	11.8	11.9	9.1/16.0	8.7/15.0	9.3	8.9	2.4
(3, 3, 3, 3; 0)	82.2	83.0	83.0	83.5	84.0	68.8/78.2	68.5/77.0	69.0	68.8	60.8

*Ordinary parallel gatekeeping procedure.

Copyright © 2004 John Wiley & Sons, Ltd.

understanding to the performance of these methods, we presented the significance of H_{2j} both with and without counting the significance of test on H_{1j}, which was denoted as (significance conditional on) 'passing H_{1j}' and 'passing F_1', respectively, in Table III.

In summary, from Table III, we find that the serial gatekeeping procedure generates the highest power for the primary test but the lowest power for the secondary test—the loss of power for the secondary test is overwhelming in the serial gatekeeping procedure when there is large chance that one of the primary tests will be not significant. The ordinary parallel gatekeeping procedure provides more opportunities for the secondary tests as compared to the serial gatekeeping procedure but as a tradeoff, it could suffer up to 10 per cent loss of power or more for the primary test in the simulated examples (and could be worse with more tests involved). Both the matched and non-matched gatekeeping procedure with larger pre-specified minimum primary weight $\frac{2}{3}$ generates higher power for the primary test but slightly lower power for the secondary test in contrast to the matched and non-matched procedure with smaller pre-specified minimum primary weight (say, $0 \leqslant \gamma \leqslant 1/2$ in the simulation), respectively. The two matched parallel gatekeeping procedures outperform their corresponding non-matched parallel gatekeeping procedures in both primary and secondary tests (if the significance of a secondary test is counted only if its 'matched' primary test is rejected). Specifically, the matched gatekeeping protects a secondary test from its unmatched gatekeeper (e.g. (3,0,3,0;0) and (3,1,3,0;0)); and it protects a primary test from its matched secondary test (e.g. (3,3,0,3;0), (1,3,1,3;0), and (3,3,1,3;0)). On the other hand, we should note that when there is no 'natural' or 'pre-defined' matching hierarchical relationship between individual primary tests and secondary tests, the matched gatekeeping procedure should not be used although it may potentially lead to larger powers for primary tests—because as shown in Table III, such kind of unnecessarily forced matching relationship between primary and secondary tests could lead to substantial power loss for secondary tests.

Due to limited space, only powers of tests for H_{11} and H_{21} are presented in Table III. We should note that powers of tests for H_{12} and H_{22} follow the similar patterns. We also only reported the simulation results with correlation coefficient $\rho = 0$ in Table III as similar results were seen with other common correlation coefficients. This finding actually coincides with that in Dmitrienko *et al.* which suggests the relative robustness of the proposed gatekeeping procedures to different correlation structures and therefore modification to accommodate correlation (e.g. using the resampling method) may generally not be needed. The probability of passing the front gate (F_1) (that is, the probability of showing at least one significance in the testing of primary hypotheses) were also evaluated for each tested parallel gatekeeping procedure. We note that, as expected, the procedure with larger minimum primary weight (γ) always has slightly larger chance to pass the front gate across different scenarios, and the matched procedure always has slightly larger chance to pass the front gate than the non-matched procedure under the same γ. The differences for the probabilities of passing front gate are generally marginal for the different parallel procedures tested in this simulation especially when the correlation coefficient ρ is small.

5. EXAMPLE

A randomized trial was conducted to compare the effect of mitoxantrone 5 mg/m^2 ($n = 64$) or 12 mg/m^2 ($n = 64$) after 24 months of treatment to placebo ($n = 60$) on patients with

Copyright © 2004 John Wiley & Sons, Ltd.

Table IV. Adjusted p-values for different gatekeeping procedures in the mitoxantrone trial ($\alpha = 0.05$).

	Endpoint	Raw p-value	Ordinary parallel $\gamma = 0$	Parallel with $\gamma = 2/3$	Matched parallel $\gamma = 0$	Matched parallel $\gamma = 2/3$	Serial $\gamma = 1$
12 mg/m^2							
H_{11}	EDSS	0.0194	0.06	0.0459	0.06	0.0402	0.0306
H_{12}	Ambulation	0.0306	0.06	0.0459	0.0306	0.0306	0.0306
H_{13}	# of relapse	0.0002	0.001	0.001	0.001	0.001	0.001
H_{14}	Time to relapse	0.0004	0.002	0.002	0.002	0.0018	0.0016
H_{15}	Neuro. status	0.0268	0.06	0.0459	0.06	0.0402	0.0306
5 mg/m^2							
H_{21}	EDSS	0.0100	0.0382	0.0367	0.06	0.0402	0.0306
H_{22}	Ambulation	0.06*	0.06	0.06	0.06	0.06	0.06
H_{23}	# of relapse	0.0002*	0.0025	0.003	0.001	0.0024	0.0306
H_{24}	Time of relapse	0.0004*	0.004	0.0048	0.002	0.0036	0.0306
H_{25}	Neuro. status	0.0268*	0.0536	0.0536	0.06	0.0536	0.0536

*These presumed p-values are not in the original paper of Hartung *et al.* [9].

Copyright © 2004 John Wiley & Sons, Ltd.

worsening relapsing-remitting or secondary progressive multiple sclerosis [9]. The primary efficacy variables for the trial consisted of five clinical measures: change from baseline in EDSS (expanded disability status scale), change from baseline in ambulation index, number of relapses treated with corticosteroids, time to first treated relapse, and change from baseline in standardized neurological status. The $12\,mg/m^2$ dose was the primary dose for the trial and the $5\,mg/m^2$ treatment group was included in the trial only for exploratory purpose. All p-values including those presumed p-values for the other three variables for the $5\,mg/m^2$ group are presented in Table IV. We see in this example, the ordinary parallel gatekeeping procedure performed poorly—the insignificance of two of the five secondary tests heavily affected the primary tests. Specifying a minimum primary weight ($\frac{2}{3}$, e.g.) would have helped protect the primary tests yet preserve reasonable power for the secondary tests. One may further match the primary and secondary tests by endpoint in the testing procedure if dose response is assumed. In this specific example, matching or not matching the primary and secondary tests would not make a difference.

6. CONCLUSIONS AND DISCUSSIONS

The parallel gatekeeping strategy proposed by Dmitrienko *et al.* [5] provides a flexible framework for the pursuit of strong control on study wise type I error rate. This paper further explores the application of the weighted Simes parallel gatekeeping procedure recommended by Dmitrienko *et al*. and proposes some enhancements for the procedure to better incorporate the interrelationships between different hypotheses in a clinical trial and to achieve better power performance. At first, we find unlike the less powerful weighted Bonferroni procedure, the weighted Simes procedure cannot guarantee to protect the tests of primary hypotheses from being affected by some non-significant secondary tests—which could lead to substantial loss of power for primary tests. We proposed a simple method to quantitatively control the impact of secondary tests when applying the weighted Simes procedure. We also introduced a matched gatekeeping procedure to exemplify how to address the special relationships between individual primary and secondary tests in a gatekeeping procedure. Our simulation study shows that whenever applicable, these enhanced gatekeeping procedures generally result in more powerful tests than the ordinary gatekeeping procedure.

In this paper, we defined the relative importance of gatekeeper hypotheses to secondary family hypotheses in a simple and general way. A more specific and individualized weighting scheme might be used to most appropriately reflect the relative importance of gatekeeper hypotheses to secondary family hypotheses and to best balance the powers for primary and secondary tests. In such a situation, it is quite straightforward to apply it to the parallel gatekeeping procedure as we exemplified in this paper. However, it is important to prespecify the weighting scheme (say, the w_{ij}'s and γ) and to ensure that such a scheme adequately addresses the objectives of the study. Specifically to identify the most appropriate value of γ, a simulation study could be conducted based on available prior knowledge at design stage to help the trialist to evaluate the power performance of the gatekeeping procedure in different scenarios with different γ values. And a γ may be selected to achieve the 'best' overall power performance around the most possible scenarios.

Copyright © 2004 John Wiley & Sons, Ltd.

Although we illustrated the application of the ordinary and modified gatekeeping procedures in the case of two families, these procedures can be easily generalized to trials with more than two families of hypotheses and with more complicated logical relationships between families. We should note, however, one major disadvantage of the discussed gatekeeping procedures is that to derive the adjusted elemental p-values, one needs to calculate the p-values of all hypotheses intersections—for a study of K hypotheses, the number of intersection hypotheses equals $2^K - 1$, say 255 when $K = 8$. The 'final' significance of each elemental hypothesis is not straightforwardly predictable (based on the respective raw p-value) and the assistance of computer software is generally needed to implement such kind of closed test based gatekeeping procedure. A general computer algorithm of the proposed parallel gatekeeping procedures which is applicable for different hypothesis structure is desirable but not straightforward. Pros and cons of alternative testing procedures that provide strong control on study wise type I error rate yet is easier to implement, for example, the procedure proposed by Quan *et al.* [10], which compares the unadjusted individual p-values with some pre-determined adjusted significance levels, are worth further exploring. Further research on these open questions is needed.

REFERENCES

1. Simon R. Problems of multiplicity in clinical trials. *Journal of Statistical Planning and Inference* 1994; **42**: 209–221.
2. Hochberg Y, Westfall P. On some multiplicity problems and multiple comparison procedures in biostatistics. *Handbook of Statistics*, vol. 18. Elsevier: Amsterdam. 2000; 75–113.
3. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *Series B*, *Methodological* 1995; **57**:289–300.
4. Proschan M, Waclawiw M. Practical guidelines for multiplicity adjustment in clinical trials. *Controlled Clinical Trials* 2001; **21**:527–539.
5. Dmitrienko A, Offen W, Westfall P. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine* 2003; **22**:2387–2400.
6. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
7. Benjamini Y, Hochberg Y. Multiple hypothesis testing and weights. *Scandinavian Journal of Statistics* 1997; **24**:407–418.
8. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 2001; **29**:1165–1188.
9. Hartung HP, Gonsette R, Konig N, Kwiecinski H, Guseo A, Morrissey SP, Krapf H, Zwingers T, Mitoxantrone in Multiple Sclerosis Study Group. Mitoxantrone in progressive multiple sclerosis: a placebo-controlled, double-blind, randomized, multicentre trial. *Lancet* 2002; **360**:2018–2025.
10. Quan H, Luo E, Capizzi T, Chen X, Wei L, Binkowitz B. Multiplicity adjustments for multiple endpoints in clinical trials with multiple doses of an active treatment. *BARDS Technical Report #094*, Merck Research Lab, 2003.

Copyright © 2004 John Wiley & Sons, Ltd.