PHARMACEUTICAL STATISTICS
Pharmaceut. Statist. 2007; **6**: 171–180
Published online 21 June 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/pst.291

WILEY InterScience® DISCOVER SOMETHING GREAT

Gatekeeping procedures with clinical trial applications

MAIN PAPER

Alex Dmitrienko[1,*,†] and Ajit C. Tamhane[2]
[1] *Eli Lilly and Company, Indianapolis, IN, USA*
[2] *Northwestern University, Evanston, IL, USA*

The objective of this paper is to give an overview of a relatively new area of multiplicity research that deals with the analysis of hierarchically ordered multiple objectives. Testing procedures for this problem are known as gatekeeping procedures and have found a variety of applications in clinical trials. This paper reviews main classes of these procedures, including serial and parallel gatekeeping procedures, and tree gatekeeping procedures that account for logical restrictions among multiple objectives. We focus on procedures based on marginal p-values; extensions to procedures that exploit the joint distribution of the p-values are also noted. Clinical trial examples are used to illustrate the procedures and their important properties. Copyright © 2007 John Wiley & Sons, Ltd.

Keywords: ***multiple comparisons; multiple endpoints; clinical trials***

1. INTRODUCTION

It is becoming increasingly common to consider designs with multiple endpoints, analyses and objectives in registration studies because additional information on the efficacy and safety profiles of an experimental drug helps patients, prescribing physicians and payers better understand its properties. More complicated study designs give rise to more sophisticated analysis methods. In the area of multiple comparisons, these problems motivated research on novel testing strategies for hierarchically ordered objectives [1–4].

This paper gives an overview of recent developments in this area with emphasis on testing strategies for multiple families of analyses. The analyses are often related to multiple endpoints (both primary and secondary) but can also represent dose–control comparisons, noninferiority and superiority tests or inferences at several time points. Testing strategies considered here are commonly referred to as *gatekeeping strategies*. This terminology highlights the fact that the families of analyses are examined sequentially and each one serves as a gatekeeper for the subsequent families. The sequential testing approach reflects the hierarchical nature of the problem and improves the power of the

*Correspondence to: Alex Dmitrienko, Lilly Research Laboratories, Eli Lilly and Company, Lilly Corporate Center, Drop Code 2233, Indianapolis, IN 46285, USA.
† E-mail: dmitrienko_alex@lilly.com

Copyright © 2007 John Wiley & Sons, Ltd.

Pharmaceutical STATISTICS

more important analyses placed early in the sequence.

The paper is organized as follows. Section 2 introduces the basic concepts about gatekeeping and establishes notation. Section 3 describes *serial gatekeeping procedures* and Section 4 describes *parallel gatekeeping procedures*. A generalization of the two approaches, termed the *tree gatekeeping approach*, is considered in Section 5. Tree gatekeeping procedures enable clinical trial researchers to construct testing strategies that combine serial and parallel gatekeepers and also account for logical restrictions among multiple analyses conducted in a clinical trial. The procedures discussed in Sections 3–5 are based on marginal p-values. In Section 6 we briefly discuss methods that exploit the joint distribution of the p-values; these include resampling and normal theory methods. Section 7 gives a summary and references for downloading the SAS macros for applying the above procedures. Clinical trial examples are provided to illustrate key properties of gatekeeping procedures.

2. BASIC CONCEPTS AND NOTATION

This paper assumes that the reader is familiar with the key concepts in the theory of multiple comparisons. For more information about multiple comparison procedures, see Hochberg and Tamhane [5]. To introduce the concepts underlying gatekeeping testing strategies, consider a clinical trial with multiple objectives. Each objective is associated with a null hypothesis of no treatment effect and each hypothesis is tested using an appropriate significance test. The objectives are hierarchically ordered, for example, primary, secondary and tertiary objectives are defined. To account for the hierarchical structure of the testing problem, the hypotheses are grouped into families. Consider m families denoted by $F_1, \ldots, F_m$ and let $H_{i1}, \ldots, H_{in_i}$ denote the hypotheses included in F_i, $i = 1, \ldots, m$. Further, let $n = n_1 + \cdots + n_m$ denote the total number of hypotheses. The families are examined sequentially beginning with F_1 that corresponds to the most important objectives. Inferences in this family are performed without adjusting for tests of hypotheses in the other families. However, when significance tests are carried out in the subsequent families, one needs to introduce a multiplicity adjustment to account for the previously examined families.

Each of the first $m-1$ families serves as a gatekeeper for the families placed later in the sequence. A family is termed a *serial gatekeeper* if and only if (iff) one must reject all hypotheses in this family to test subsequent families. Serial gatekeeping procedures were considered by Maurer *et al*. [1], Bauer *et al*. [2] and Westfall and Krishen [3]. As an example, in clinical trials for Alzheimer's disease, two primary endpoints are generally required: Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) and Clinical Global Impression of Change (CGIC). The trial is declared successful only if the treatment effect on both endpoints is significant. Therefore, the primary gatekeeper can be passed only if the null hypotheses for both the endpoints are rejected.

The concept of a *parallel gatekeeper* was introduced in Dmitrienko *et al*. [4]. To pass a parallel gatekeeper, one needs to reject at least one hypothesis in the family. The acute respiratory distress syndrome (ARDS) trial in [4] provides an example of a parallel gatekeeper. It had two primary endpoints (mortality and lung function endpoints). A significant treatment effect with respect to either of these two endpoints allowed the researcher to test for efficacy with respect to secondary endpoints.

As pointed out in [6], the serial gatekeeping approach is analogous to intersection–union testing [7] in which the union of several component hypotheses is rejected iff all of them are rejected. Likewise, the parallel gatekeeping approach is similar to union–intersection testing [8] in which the intersection of several component hypotheses is rejected iff at least one of them is rejected.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Gatekeeping procedures are defined as multiple testing procedures that meet the following conditions:

- *Type I error rate control*: The familywise error rate (FWER) associated with the null hypotheses in $F_1, \ldots, F_m$ is controlled in the strong sense at a prespecified α level [5].
- *Serial and parallel gatekeeping conditions*: Consider F_i, $i = 1, \ldots, m-1$. If F_i is a serial gatekeeper, then hypotheses in F_{i+1} are tested iff all hypotheses in F_i are rejected. In other words, if $\widetilde{p}_{i1}, \ldots, \widetilde{p}_{in_i}$ are multiplicity adjusted p-values in F_i, then the hypotheses in F_{i+1} are tested iff

  $$\max(\widetilde{p}_{i1}, \ldots, \widetilde{p}_{in_i}) \leqslant \alpha$$

  If F_i is a parallel gatekeeper, then hypotheses in F_{i+1} are tested iff one or more hypotheses in F_i are rejected, i.e.

  $$\min(\widetilde{p}_{i1}, \ldots, \widetilde{p}_{in_i}) \leqslant \alpha$$

  The untested hypotheses are automatically accepted.
- *Independence condition*: Inferences in F_i, $i = 1, \ldots, m-1$, are independent of the p-values for the hypotheses in $F_{i+1}, \ldots, F_m$.

The independence condition plays a key role in clinical applications [9, Section 2.7]. It ensures that more important analyses (e.g. analysis of primary endpoints) will not depend on the results of less important analyses (e.g. analysis of secondary endpoints). However, relaxing this condition, when it is consistent with the objectives of a clinical trial, can result in some power gains, see Chen *et al*. [10], Dmitrienko *et al*. [9, Section 2.7.3] and Hommel *et al*. [11].

3. SERIAL GATEKEEPING PROCEDURES

Serial gatekeeping procedures have a straightforward stepwise form that facilitates their use in multiplicity problems arising in clinical studies. As an example, consider a trial with two families of hypotheses, F_1 and F_2, the first of which is a serial gatekeeper, and suppose the overall FWER is to be controlled at the α level. Since it is required that all hypotheses in F_1 must be rejected to test the hypotheses in F_2, a powerful procedure to use would be the intersection–union test of Berger [7]. This test tests each hypothesis in F_1 at the α level. Hypotheses in F_2 can be tested using any multiple test that controls the FWER for that family at the α level, e.g. the Holm test [12] or the Hochberg test [13] (assuming that conditions under which the Hochberg test controls the FWER are satisfied, see Sarkar and Chang [14]).

This simple setting is easily extended to the general case of m families in which the first $m-1$ families are serial gatekeepers. Since any coherent gatekeeping procedure can be expressed as a closed testing procedure [15], this serial gatekeeping procedure protects the FWER in the strong sense.

As an illustration, consider a clinical trial in patients with Alzheimer's disease that was conducted to evaluate the efficacy and safety of a single dose of an experimental drug compared to placebo. The primary objective of the trial was to assess the effect of the experimental drug on two endpoints, P1 (ADAS-Cog) and P2 (CGIC). The null hypotheses associated with the primary endpoints were included in F_1. This family served as a serial gatekeeper for F_2 which contained two hypotheses related to the secondary endpoints, S1 (a biochemical endpoint) and S2 (an imaging endpoint). The hypotheses in both families were equally weighted and the FWER was set at $\alpha = 0.05$. The raw p-values produced by the primary and secondary tests are given in Table I.

F_1 is a serial gatekeeper and the hypotheses corresponding to P1 and P2 are tested by using an intersection–union test that does not require an adjustment for multiplicity. The p-values in F_1 are significant at the 0.05 level and thus the primary objective of the trial is met. Since the primary tests are both significant, the testing procedure passes the serial gatekeeper and can now examine the hypotheses in F_2. The secondary tests are carried out using the Holm test. Comparing the Holm-adjusted p-values in F_2 to 0.05, it is easy to see that

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Pharmaceutical STATISTICS

Table I. Serial gatekeeping procedure in the Alzheimer's disease trial.

Family	Endpoint	Weight	Raw p-value	Multiple test	Adjusted p-value	Test outcome
F_1	P1	0.5	0.023	IU	0.023	S
F_1	P2	0.5	0.018	IU	0.018	S
F_2	S1	0.5	0.014	Holm	0.028	S
F_2	S2	0.5	0.106	Holm	0.106	NS

Primary endpoints, P1 (ADAS-Cog) and P2 (CGIC). Secondary endpoints, S1 (a biochemical endpoint) and S2 (an imaging endpoint). Multiple test, IU (intersection–union test) and Holm (Holm test). The adjusted p-values are identical to the raw p-values in F_1 and are produced by the Holm test in F_2. Test outcome, S (significant at the 0.05 level) and NS (not significant at the 0.05 level).

only S1 is significant. The overall conclusion is that the experimental drug is significantly different from placebo with respect to P1, P2 and S1 at the 0.05 level.

4. PARALLEL GATEKEEPING PROCEDURES

The most basic parallel gatekeeping procedure, derived from the Bonferroni test, was proposed in Dmitrienko *et al.* [4]. This procedure was formulated as a closed testing procedure and guaranteed strong control of the FWER due to the closed testing principle [16]. To satisfy the parallel gatekeeping and independence conditions, a weighted Bonferroni test was defined for each individual intersection hypothesis in the closed family induced by the null hypotheses in $F_1, \ldots, F_m$. Since the closed family contains $2^n - 1$ intersection hypotheses, this process is generally computationally intensive; the *decision matrix algorithm* [9, Section 2.7] systematizes these computations.

A detailed examination of the underlying decision rule reveals that the Bonferroni parallel gatekeeping procedure has, in fact, a simple stepwise structure that provides important insights into the nature of gatekeeping inferences. This stepwise procedure, proposed in Dmitrienko *et al.* [6], is described below. The procedure is built around the concept of a *rejection gain factor*. At the first stage of the procedure, inferences are performed at the α level, where α is the FWER. At each subsequent stage, significance tests are carried out at the $\rho_k\alpha$ level, $k = 2, \ldots, m$. The rejection gain factor, $0 \leqslant \rho_k \leqslant 1$, depends on the number and importance of hypotheses rejected at the earlier stages.

In mathematical terms, let $w_{i1}, \ldots, w_{in_i}$ be the weights representing the importance of null hypotheses in F_i, $i = 1, \ldots, m$ (it is assumed that $0 < w_{ij} < 1$ and $w_{i1} + \cdots + w_{in_i} = 1$). The stepwise parallel gatekeeping procedure for testing the null hypotheses in $F_1, \ldots, F_m$ is as follows:

- *Family* F_k, $k = 1, \ldots, m-1$: Test the null hypotheses using the Bonferroni test at the $\rho_k\alpha$ level.
- *Family* F_m: Test the null hypotheses using the weighted Holm test [12] at the $\rho_m\alpha$ level.

The rejection gain factors $\rho_1, \ldots, \rho_m$ are given by

$$\rho_1 = 1, \quad \rho_k = \prod_{i=1}^{k-1} \left(\sum_{j=1}^{n_i} r_{ij} w_{ij} \right), \quad k = 2, \ldots, m$$

where $r_{ij} = 1$ if H_{ij} is rejected and 0 otherwise. For equally weighted hypotheses ($w_{ij} = 1/n_i$), the formula for ρ_k simplifies to

$$\rho_k = \prod_{i=1}^{k-1} \left(\frac{r_i}{n_i} \right), \quad k = 2, \ldots, m$$

where $r_i = \sum_j r_{ij}$ is the number of rejected hypotheses in F_i. Thus, ρ_k is the product of the proportions of rejected hypotheses in F_1 through F_{k-1}.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Table II. Stepwise parallel gatekeeping procedure based on the Bonferroni test in the ARDS trial.

Family	Endpoint	Weight	Raw p-value	Rejection gain factor	Multiple test	Adjusted p-value	Test outcome
F_1	P1	0.9	0.048	1	Bonf	0.053	NS
F_1	P2	0.1	0.003	1	Bonf	0.030	S
F_2	S1	0.5	0.026	0.1	Holm	0.260	NS
F_2	S2	0.5	0.002	0.1	Holm	0.040	S

Primary endpoints, P1 (Lung function) and P2 (Mortality). Secondary endpoints, S1 (ICU-free days) and S2 (Quality of life). Multiple test, Bonf (weighted Bonferroni test) and Holm (Holm test). The adjusted p-values are produced by the weighted Bonferroni test (for F_1) and Holm test (for F_2). Test outcome, S (significant at the 0.05 level) and NS (not significant at the 0.05 level).

In order to incorporate the rejection gain factors into the decision rule, it is convenient to re-define the adjusted p-values for the hypotheses in the last $m-1$ families. The modified adjusted p-value for H_{ij}, $i=2,\ldots,m$, is given by $\widetilde{p}_{ij}/\rho_i$, where $\widetilde{p}_{ij}$ is the usual adjusted p-value produced by the multiple test in F_i. After this modification, inferences in $F_2, \ldots, F_m$ can be performed by comparing adjusted p-values to the prespecified FWER, α.

It follows from the formula for ρ_k that F_k is tested iff the procedure passed the preceding gatekeepers, i.e. iff at least one hypothesis is rejected in $F_1,\ldots,F_{k-1}$ and thus ρ_k is positive. Further, the combined weight of the null hypotheses rejected at the earlier stages determines the penalty one has to pay for performing multiple inferences in F_k. No penalty is involved, i.e. $\rho_k = 1$, if the procedure rejects all hypotheses in $F_1,\ldots,F_{k-1}$. Otherwise, ρ_k is strictly less than 1 and therefore the significance level for F_k is adjusted downward.

To illustrate the utility of the stepwise gatekeeping procedure, consider the ARDS trial example given in [4, Section 4]. The trial was designed to compare a single dose of an experimental drug to placebo. Two families of endpoints were considered in this trial. F_1 consisted of two hypotheses related to the primary endpoints, P1 (lung function) and P2 (mortality), and F_2 consisted of two hypotheses related to the secondary endpoints, S1 (ICU-free days) and S2 (quality of life). F_1 was a parallel gatekeeper. P1 was deemed more important than P2 in F_1 ($w_{11}=0.9$, $w_{12}=0.1$), but S1 and S2 were equally weighted ($w_{21}=0.5$, $w_{22}=0.5$). The raw p-values for the treatment comparisons are given in Table II. The FWER is to be controlled at $\alpha=0.05$.

To apply the stepwise parallel gatekeeping procedure, one first considers the adjusted p-values produced by the weighted Bonferroni test and Holm test for the null hypotheses in F_1 and F_2, respectively. Since $\rho_1=1$, the primary hypotheses are tested at the full $\alpha=0.05$ level. The P2 comparison is significant at this level, whereas the P1 comparison is not. Therefore, the rejection gain factor for the secondary family is $\rho_2=w_{12}=0.1$ and the adjusted p-values for S1 and S2 are $0.026/\rho_2=0.260$ and $0.004/\rho_2=0.040$, respectively. It is clear that only the hypothesis concerning S2 is rejected. These conclusions are identical to those based on the parallel gatekeeping procedure that was derived using the closed testing principle (compare to Table III, Scenario 3 of [4]).

Generalizations of the Bonferroni-based parallel gatekeeping procedure were studied by Dmitrienko *et al.* [17] and Hommel *et al.* [11]. Dmitrienko *et al.* discussed procedures with an extended parallel gatekeeping property. These procedures are derived from the fallback test [18] and enable researchers to carry over a predetermined fraction of the Type I error rate to the next family even if no hypotheses are rejected in the previous family. Dmitrienko *et al.* demonstrated how this testing approach can be used in dose-finding studies with multiple endpoints. Hommel *et al.* described a general family of Bonferroni-based stepwise testing procedures and applications to clinical trials with several dose–control comparisons and outcome variables.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Pharmaceutical STATISTICS

Table III. Parallel gatekeeping procedure based on the truncated Holm test in the ARDS trial.

				Test outcome		
Family	Endpoint	Weight	Raw p-value	$\gamma_1 = 0$	$\gamma_1 = 0.5$	$\gamma_1 = 0.9$
F_1	P1	0.9	0.048	NS	NS	S
F_1	P2	0.1	0.003	S	S	S
F_2	S1	0.5	0.026	NS	NS	S
F_2	S2	0.5	0.002	S	NS	S

Primary endpoints, P1 (Lung function) and P2 (Mortality). Secondary endpoints, S1 (ICU-free days) and S2 (Quality of life). Test outcome, S (significant at the 0.05 level) and NS (nonsignificant at the 0.05 level).

The stepwise testing framework described above relies on the basic Bonferroni test in the first $m-1$ families and it is natural to ask whether alternative tests can be utilized to improve the power for the more important objectives. One can consider more powerful Bonferroni-based tests (e.g. the truncated Holm test defined below) and Simes-based tests (e.g. the Hochberg test). These alternative approaches are described below. Parallel gatekeeping procedures based on resampling and parametric tests (e.g. the Dunnett test [19]) are described in Section 6.

The most straightforward extension of the Bonferroni gatekeeping procedure relies on replacing the Bonferroni test at the first $m-1$ stages with a more powerful test. The Holm test cannot be used for this purpose because it 'spends' all of the Type I error rate at each stage [9, Section 2.7.5]. As an alternative, one can consider the truncated Holm test based on a convex combination of the Bonferroni and Holm tests. Consider, for the sake of simplicity, the case of equally weighted hypotheses ($w_{ij} = 1/n_i$). Let $p_{i(1)} < \cdots < p_{i(n_i)}$ denote the ordered p-values in F_i. The truncated Holm test rejects the hypothesis $H_{i(k)}$ corresponding to $p_{i(k)}$, $k = 1, \ldots, n_i$, if

$$p_{i(j)} \leqslant \alpha \left[\frac{1-\gamma_i}{n_i} + \frac{\gamma_i}{n_i - j + 1} \right], \quad j = 1, \ldots, k$$

where $0 \leqslant \gamma_i < 1$ is the *truncation fraction* for Stage i between the Bonferroni and Holm tests (it is analogous to the *relative importance factor* defined in [9, Section 2.7.5]). When $\gamma_i = 0$, this test simplifies to the Bonferroni test and, when $\gamma_i = 1$, it is equivalent to the regular Holm test. The power of the truncated Holm test is an increasing function of γ_i.

A parallel gatekeeping procedure based on the truncated Holm test was constructed in [9, Section 2.7.5] and can be illustrated using the ARDS trial example. Table III gives the results produced by the gatekeeping procedure based on the truncated Holm test with $\gamma_1 = 0$, 0.5 and 0.9. When $\gamma_1 = 0$, this procedure is equivalent to the Bonferroni-based procedure and thus the conclusions are identical to those presented in Table II. Setting $\gamma_1 = 0.5$ leads to a nonsignificant outcome for the S2 endpoint. However, with $\gamma_1 = 0.9$, the treatment differences become significant for all primary and secondary endpoints.

In general, γ_1 can be thought of as a leverage factor that determines the power of the significance tests in F_1 relative to the power of the remaining tests. The power of the tests in F_1 is an increasing function of γ_1. The power of the tests in $F_2, \ldots, F_m$ can increase or decrease with increasing γ_1 depending on the number of true hypotheses in all families as well as the effect sizes for false hypotheses.

Parallel gatekeeping procedures based on the Simes test [20] were studied by several authors (this work was done under the assumption that conditions under which the Simes and related tests control the FWER are met [14]). The first attempt to construct a Simes-based parallel gatekeeping procedure was made in [4]. It was based on a straightforward extension of the principles underlying the Bonferroni-based procedure and was subsequently shown to have certain undesirable properties such as the violation of the

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

independence condition (see [9, 10, Section 2.7.3]). Quan *et al*. [21] proposed parallel gatekeeping procedures that relied on a Bonferroni-type modification of the Hochberg test to achieve control of the Type I error rate. Another Simes-based parallel gatekeeping procedure was introduced by Chen *et al*. [10]; however, this procedure does not satisfy the independence condition. More recently, Wang [22] developed a gatekeeping method based on a combination of the Bonferroni and Simes tests that satisfies the three conditions given in Section 2. Further research is required to improve the power advantage of this procedure over the Bonferroni parallel gatekeeping procedure.

5. TREE GATEKEEPING PROCEDURES

The gatekeeping procedures described in Sections 3 and 4 are based on the assumption that ordered objectives have a simple 'one-dimensional' form defined by a sequence of serial or parallel gatekeepers. Testing problems encountered in clinical trials with multiple objectives often exhibit a more complicated 'multi-dimensional' structure with one dimension corresponding to multiple outcome variables, another to multiple doses and yet another to multiple analysis objectives (e.g. noninferiority and superiority tests). In addition, one may need to account for logical relationships among the multiple comparisons, for example, require that secondary tests in dose-finding studies with multiple endpoints be restricted to the doses at which the primary endpoints are significant. In order to develop testing procedures for complex problems of this kind, the standard gatekeeping framework needs to be extended. Dmitrienko *et al*. [23] proposed a testing approach, termed the *tree gatekeeping approach*, that supports decision trees with multiple branches.

The tree gatekeeping approach assumes the setting described in Section 2 that involves m families of hypotheses, $F_1, \ldots, F_m$. The families are tested sequentially as described below. The algorithm begins with the hypotheses in F_1, which are tested using an appropriate test with local (for F_1) level α. When the other families are examined, one first determines whether each particular hypothesis is *testable*. Consider, for instance, F_i, $i = 2, \ldots, m$, and select a hypothesis, say, H_{ij}, $j = 1, \ldots, n_i$. This hypothesis is tested by the tree gatekeeping procedure iff it meets the following two conditions:

- All hypotheses from a prespecified subset of hypotheses in $F_1, \ldots, F_{i-1}$, denoted by R_{ij}^S, are rejected. This subset is referred to as the *serial rejection set* for H_{ij}.
- One or more hypothesis from a prespecified subset of hypotheses in $F_1, \ldots, F_{i-1}$, denoted by R_{ij}^P, are rejected. This subset is referred to as the *parallel rejection set* for H_{ij}.

If either condition is not satisfied, H_{ij} is automatically accepted. Otherwise, it is tested with an appropriate adjustment for multiplicity. The other hypotheses are tested in a similar manner.

Using the principle of closed testing, Dmitrienko *et al*. [23] developed Bonferroni-based tree gatekeeping procedures. These procedures can be applied to a wide variety of testing problems encountered in clinical trial applications. Examples considered in [23] include clinical trials with (1) ordered primary/secondary endpoints and noninferiority/superiority assessments and (2) ordered primary/secondary endpoints and multiple dose levels. Another important application of the tree gatekeeping approach involves testing problems with several treatment groups and noninferiority/superiority assessments. This application is described below.

Suppose a parallel-group trial is conducted to compare a new formulation of an insulin therapy (Formulation A) to a standard formulation (Formulation B) in patients with Type 2 diabetes. Patients are allocated to three treatment groups (A, B and A + B) and the efficacy analysis is based on the mean change in hemoglobin A1c from baseline to a 6-month endpoint. The three pairwise comparisons among the treatment groups are ordered according to their clinical relevance. The primary objective of the study is to compare the new formulation to the standard one (A versus B).

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

Pharmaceutical STATISTICS

After that, the combination is compared to the standard formulation (A + B versus B) and to the new formulation (A + B versus A). Each comparison begins with a noninferiority test followed by a superiority test if noninferiority is established.

According to this strategy, the six null hypotheses are grouped into four families:

- Family $F_1 = \{H_1\}$, where H_1 states that A is inferior to B.
- Family $F_2 = \{H_2, H_3\}$, where H_2 states that A is not superior to B and H_3 states that A + B is inferior to B.
- Family $F_3 = \{H_4, H_5\}$, where H_4 states that A +B is not superior to B and H_5 states that A + B is inferior to A.
- Family $F_4 = \{H_6\}$, where H_6 states that A + B is not superior to A.

The four families are tested as shown in the decision tree (Figure 1). To set up a tree gatekeeping procedure, one needs to define serial and parallel rejection sets for the hypotheses in F_2, F_3 and F_4. In this case, the parallel rejection sets can be set to be empty and serial rejection sets defined as given in Table IV. This table also gives the raw p-values for each analysis and multiplicity adjusted p-values produced by the tree gatekeeping procedure based on the Bonferroni test.

The procedure begins with the single hypothesis in F_1 and rejects it at the 0.05 level. Due to this rejection, the procedure passes the first gatekeeper and proceeds to testing the hypotheses H_2 and H_3 in F_2. The two hypotheses are tested using the Bonferroni test and are both rejected at the 0.05 level. Note that F_3 depends on H_3 and thus the next step is to examine the hypotheses H_4 and H_5. These hypotheses are again tested using the Bonferroni test, H_4 is found false but H_5 is accepted. Since H_5 is in the serial rejection set of H_6, testing stops and H_6 is accepted without testing. The overall conclusion is that Formulation A is superior to Formulation B and the combination of A and B is superior to B.

The computation of adjusted p-values in this example is performed using the closed testing procedure proposed in Dmitrienko *et al.* [23]. The computational algorithm relies on a complete

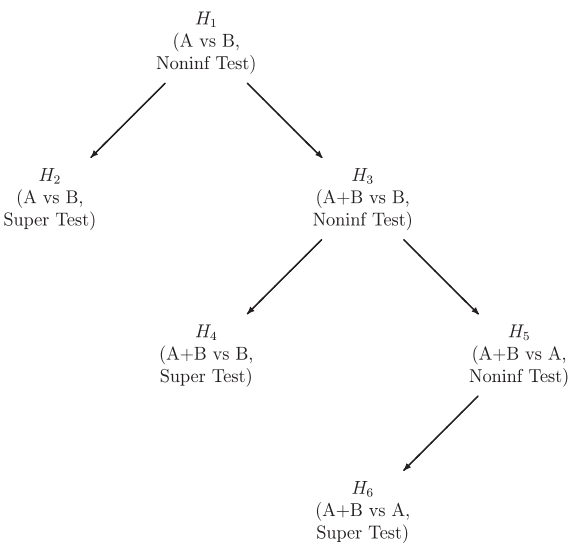

Figure 1. Decision tree in the Type 2 diabetes trial. Noninferiority (Noninf) and superiority (Super) tests.

Table IV. Tree gatekeeping procedure based on the Bonferroni test in the Type 2 diabetes trial.

Family	Hypothesis	Serial rejection set	Raw p-value	Adjusted p-value	Test outcome
F_1	H_1	NA	0.011	0.011	S
F_2	H_2	H_1	0.023	0.046	S
F_2	H_3	H_1	0.006	0.012	S
F_3	H_4	H_3	0.018	0.046	S
F_3	H_5	H_3	0.042	0.084	NS
F_4	H_6	H_5	0.088	0.088	NS

The parallel rejection sets are empty. The adjusted p-values are produced by the tree gatekeeping procedure. Test outcome, S (significant at the 0.05 level) and NS (nonsignificant at the 0.05 level).

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst

enumeration of all intersections in the closed family but in some cases it assumes a simple sequential form. For example, the adjusted p-value for H_4 is the maximum of the Bonferroni p-value for H_4 ($2p_4 = 0.036$) and the largest adjusted p-value associated with the hypotheses in F_1 and F_2 (0.046). The maximum is taken in this calculation to account for the sequential nature of this procedure.

6. PARALLEL GATEKEEPING PROCEDURES BASED ON JOINT DISTRIBUTIONS OF TEST STATISTICS

Further improvements of the Bonferroni-based gatekeeping procedure discussed in Section 4 can be achieved by considering tests that account for the joint distribution of the test statistics associated with the null hypotheses in $F_1, \ldots, F_m$. A resampling-based approach to construct parallel gatekeeping procedures was proposed in [4] (see also [9, Section 2.7.4]). This approach relies on the closed testing principle and replaces the potentially conservative Bonferroni test for each intersection hypothesis in the closed family with parametric or nonparametric resampling tests described in Westfall and Young [24] (assuming that resampling-based procedures preserve the FWER, e.g. the subset pivotality condition is met). The resulting procedure takes into account the correlations among the test statistics within each family and across families. For an application of a resampling-based gatekeeping procedure to the analysis of multiple dose–placebo comparisons, see [4, Section 5]. In this example, the use of a parametric resampling procedure led to uniformly smaller adjusted p-values compared to the Bonferroni-based gatekeeping procedure.

In dose-finding studies, instead of using the Bonferroni test for comparing doses to a placebo control, one can use the more powerful Dunnett test [19] if the normality assumption is satisfied. This was done in Dmitrienko *et al.* [25]. The Dunnett test used there accounted for not only the correlations among the dose–placebo contrasts, but also between the endpoints. It was shown via simulations that the Dunnett-based gatekeeping procedure is more powerful than the Bonferroni-based procedure. The power advantage of the parametric procedure increased with increasing correlations among the endpoints, especially in the case when all primary dose–control comparisons were significant.

7. SUMMARY

This paper reviewed developments in the growing area of multiple comparison research, namely, multiple testing procedures for hierarchically ordered objectives. The gatekeeping framework described in the paper provides clinical trial researchers with useful tools for managing multiplicity in clinical trials that guarantee strong control of the FWER. We described two basic approaches to gatekeeping, namely serial and parallel gatekeeping, and a unified approach of tree gatekeeping.

The gatekeeping procedures discussed in the paper can be carried out using a number of SAS macros freely available on the Internet. The Bonferroni-based gatekeeping procedure described in Section 4 can be carried out using the %Gatekeeper macro available at `http://biopharmnet.com/books/book40005.html`. The tree gatekeeping approach (Section 5) is implemented in the %TreeGatekeeper macro that can be downloaded from `http://www.biopharmnet.com/code`.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the useful comments from three referees. This research was supported by grants from the National Institute of Heart, Lung and Blood Institute and National Security Agency to Professor Ajit Tamhane.

REFERENCES

1. Maurer W, Hothorn L, Lehmacher W. Multiple comparisons in drug clinical trials and preclinical

Copyright © 2007 John Wiley & Sons, Ltd.

assays: a-priori ordered hypotheses. In *Biometrie in der Chemisch-pharmazeutischen Industrie*, Vollmar J (ed.), vol. 6. Fischer Verlag: Stuttgart, 1995; 3–18.
2. Bauer P, Röhmel J, Maurer W, Hothorn L. Testing strategies in multi-dose experiments including active control. *Statistics in Medicine* 1998; **17**:2133–2146.
3. Westfall PH, Krishen A. Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *Journal of Statistical Planning and Inference* 2001; **99**:25–41.
4. Dmitrienko A, Offen WW, Westfall PH. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine* 2003; **22**:2387–2400.
5. Hochberg Y, Tamhane AC. *Multiple Comparison Procedures*. Wiley: New York, 1987.
6. Dmitrienko A, Tamhane AC, Wang X, Chen X. Stepwise gatekeeping procedures in clinical trial applications. *Biometrical Journal* 2006; **48**:984–991.
7. Berger RL. Multiparameter hypothesis testing and acceptance sampling. *Technometrics* 1982; **24**: 295–300.
8. Roy SN. On a heuristic method for test construction and its use in multivariate analysis. *The Annals of Statistics* 1953; **24**:220–238.
9. Dmitrienko A, Molenberghs G, Chuang-Stein C, Offen W. *Analysis of Clinical Trials Using SAS*: *A Practical Guide*. SAS Press: Cary, NC, 2005.
10. Chen X, Luo X, Capizzi T. The application of enhanced parallel gatekeeping strategies. *Statistics in Medicine* 2005; **24**:1385–1397.
11. Hommel G, Bretz F, Maurer W. Powerful shortcuts for multiple testing procedures with special reference to gatekeeping strategies. *Statistics in Medicine* 2007, in press.
12. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979; **6**:65–70.
13. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988; **75**:800–802.
14. Sarkar SK, Chang CK. Simes' method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* 1997; **92**:1601–1608.
15. Grechanovsky, E, Hochberg Y. Closed procedures are better and often admit a shortcut. *Journal of Statistical Planning and Inference* 1999; **76**:79–91.
16. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
17. Dmitrienko A, Wiens BL, Westfall PH. Fallback tests in dose–response clinical trials. *Journal of Biopharmaceutical Statistics* 2006; **16**:745–755.
18. Wiens BL. A fixed-sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics* 2003; **2**:211–215.
19. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* 1955; **50**:1096–1121.
20. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **63**:655–660.
21. Quan H, Luo X, Capizzi T. Multiplicity adjustment for multiple endpoints in clinical trials with multiple doses of an active control. *Statistics in Medicine* 2005; **24**:2151–2170.
22. Wang X. Gatekeeping procedures for multiple endpoints. *Doctoral Dissertation*. Department of Statistics, Northwestern University, Evanston, IL, 2006.
23. Dmitrienko A, Wiens BL, Tamhane AC, Wang X. Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Statistics in Medicine* 2007; **26**:2465–2478.
24. Westfall PH, Young SS. *Resampling-Based Multiple Testing*: *Examples and Methods for P-Value Adjustment*. Wiley: New York, 1993.
25. Dmitrienko A, Offen W, Wang O, Xiao D. Gatekeeping procedures in dose–response clinical trials based on the Dunnett test. *Pharmaceutical Statistics* 2006; **5**:19–28.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/pst