STATISTICS IN MEDICINE
Statist. Med. 2003; **22**:2387–2400 (DOI: 10.1002/sim.1526)

Gatekeeping strategies for clinical trials that do not require all primary effects to be significant

Alexei Dmitrienko[1,*,†], Walter W. Offen[1] and Peter H. Westfall[2]

[1]*Lilly Research Laboratories, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285, U.S.A.*
[2]*Department of Information Systems and Quantitative Sciences, Texas Tech University, Lubbock, TX 79409, U.S.A.*

SUMMARY

In this paper we describe methods for addressing multiplicity issues arising in the analysis of clinical trials with multiple endpoints and/or multiple dose levels. Efficient 'gatekeeping strategies' for multiplicity problems of this kind are developed. One family of hypotheses (comprising the primary objectives) is treated as a 'gatekeeper', and the other family or families (comprising secondary and tertiary objectives) are tested only if one or more gatekeeper hypotheses have been rejected. We discuss methods for constructing gatekeeping testing procedures using weighted Bonferroni tests, weighted Simes tests, and weighted resampling-based tests, all within the closed testing framework. The new strategies are illustrated using an example from a clinical trial with co-primary endpoints, and using an example from a dose-finding study with multiple endpoints. Power comparisons with competing methods show the gatekeeping methods are more powerful when the primary objective of the trial must be met.

KEY WORDS: multiple tests; closed testing; clinical trial; *P*-value; resampling

1. INTRODUCTION

Hypotheses tested in clinical trials are commonly divided into primary and secondary. The primary hypothesis is related to the primary trial endpoint which describes the most important features of the disease under study. O'Neill [1] defines the primary endpoint as 'a clinical endpoint that provides evidence sufficient to fully characterize clinically the effect of a treatment in a manner that would support a regulatory claim for the treatment'. In many cases, the primary hypothesis test determines the overall conclusion from the trial. Secondary hypotheses also play an important role in characterizing the effects of the study drug. However, a significant improvement in a secondary endpoint in isolation is not generally considered as substantial evidence of therapeutic benefit.

*Correspondence to: Alex Dmitrienko, Lilly Research Laboratories, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285, U.S.A.
†E-mail: dmitrienko_alex@lilly.com

Received June 2002
Accepted February 2003

The interpretation of a positive finding with respect to a secondary outcome variable depends heavily on its clinical importance. The following general types of secondary outcome variables are based on D'Agostino [2]:

Type I: *Separate components of the primary trial objective***.** These may be very important secondary endpoints that are difficult to incorporate into the formal power calculation because the expected improvement is relatively small. All-cause mortality plays this role in a large number of trials including cardiovascular and critical care studies.

Type II: *Endpoints that help interpret the primary findings***.** These endpoints are very helpful for understanding the big picture, for example, understanding the benefits with respect to various aspects of the disease. The effect of osteoarthritis drugs is typically measured using pain and physical function indices; however, there are other important outcome variables such as patient global assessment and quality-of-life measures.

A similar classification of secondary endpoints is proposed in the guideline entitled 'Points to consider on multiplicity issues in clinical trials' published by the Committee for Proprietary Medicinal Products (CPMP) [3].

With this classification in mind, it is critically important to prospectively define not only the hypotheses of interest but also a decision rule that will be used to guide the decision-making process at study completion (Chi [4]). The decision rule can have a flexible hierarchical structure with several data-driven clinical decision paths. The clinical statistician 'translates' this decision rule into a statistical procedure that incorporates the interrelationships among the primary and secondary trial hypotheses.

Multiplicity problems arising in the context of primary and secondary trial hypotheses can be effectively dealt with using hierarchical testing procedures, also known as 'gatekeeping procedures'. Consider two families of outcome variables, for example, primary variables that can lead to new regulatory claims (gatekeeper) and secondary variables that may become the basis for additional claims. The gatekeeper family is tested without an adjustment for the other family, and the second family is examined only if the gatekeeper has been successfully passed; see Bauer *et al.* [5], Westfall and Krishen [6], and Gong *et al.* [7].

Gatekeeping procedures proposed in the literature have been designed for the case when the gatekeeper family is passed only if all of the hypotheses in the family have been rejected, which we refer to as 'serial' gatekeeping procedures. Serial procedures can be too restrictive; for example, the requirement to reject all primary trial hypotheses before performing the secondary analyses may be inappropriate when the co-primary endpoints can lead to separate regulatory claims. Similarly, the analysis of individual doses of an experimental drug in a dose-finding setting can be enhanced by using a hierarchical testing strategy, for example, one examines the higher doses first and studies the lower doses if at least one of the higher doses (but not necessarily both) has shown a significant difference from the control.

Thus, while it is well known that one can proceed in serial fashion, it is apparently not known that gatekeeping tests also can be performed in 'parallel' fashion, where one may proceed to the secondary family when *at least one* of the primary tests exhibits significance. The main contribution of this paper is the development of such parallel gatekeeping procedures.

An alternative to gatekeeping strategies is the prospective alpha allocation scheme (PAAS) proposed by Moyé [8], in which the primary and secondary endpoints are tested simultaneously at levels less than the common 0.05 threshold, ensuring that the total threshold is still 0.05. This approach has been designed for the cases when the secondary endpoints may potentially

provide the basis for a new regulatory claim. If the primary effects are truly null, the PAAS strategy is more powerful than a gatekeeping strategy. On the other hand, when some of the primary hypotheses are false, the power gains of gatekeeping strategies can be substantial, as we demonstrate in Section 6.

In this paper, we develop parallel gatekeeping methods using the powerful closed testing principle of Marcus *et al*. [9]. Section 2 introduces closed gatekeeping procedures based on the weighted version of the Bonferroni test. Section 3 discusses extensions based on the weighted Simes test and parametric resampling. The described gatekeeping strategies are illustrated in Sections 4 and 5 using two clinical trial examples, one with two co-primary endpoints and the other a dose-finding study. Section 6 compares the power of the various testing procedures.

2. BONFERRONI GATEKEEPING PROCEDURES

Consider a family of null hypotheses $H_1,\ldots,H_m$ and assume that the hypotheses are grouped into two families, $F_1=\{H_1,\ldots,H_k\}$ and $F_2=\{H_{k+1},\ldots,H_m\}$. The first family serves as a gatekeeper in the sense that F_1 is tested without an adjustment for F_2 and $H_{k+1},\ldots,H_m$ are examined only if the gatekeeper has been successfully passed. It will be assumed throughout the paper that F_1 is a parallel gatekeeper, that is, $H_{k+1},\ldots,H_m$ can be tested if at least one hypothesis in F_1 has been rejected. In the clinical trial context, F_1 and F_2 can represent sets of the primary and secondary trial hypotheses, respectively. We are interested in constructing a testing strategy that controls the familywise error rate (FWE) with respect to both families of hypotheses in the strong sense (Hochberg and Tamhane [10]).

To apply the closed testing principle to the problem of testing F_1 and F_2, consider an arbitrary intersection hypothesis H in the closed family associated with $H_1,\ldots,H_m$. Let $p_1,\ldots,p_m$ denote the raw p-values for $H_1,\ldots,H_m$. Further, let p_H denote the p-value associated with H, obtained from a test procedure whose size is no more than α when H is true. The closed testing principle states that an original hypothesis is rejected provided all of the p-values associated with the intersection hypotheses containing it are significant. Therefore, the adjusted p-value associated with the hypothesis H_i equals $\tilde{p}_i=\max_{H\in\mathscr{H}_i} p_H$, where $\mathscr{H}_i$ denotes the set of all intersection hypotheses that contain H_i. The closed testing procedure rejects H_i when $\tilde{p}_i\leqslant\alpha$, and strongly controls the FWE at the α level, over the combined family $(H_1,\ldots,H_m)$. Note that the adjusted p-value depends on the test procedure chosen to test H; the key to our gatekeeping procedures is the particular choice of tests for each H, as we now describe.

Parallel gatekeeping procedures can be defined using weighted Bonferroni tests for the intersection hypotheses. Select an intersection hypothesis H and consider a set of m weights $v_1(H),\ldots,v_m(H)$ such that

$$0\leqslant v_i(H)\leqslant 1,\quad v_i(H)=0\quad\text{if }\delta_i(H)=0,\quad \sum_{i=1}^{m} v_i(H)\leqslant 1$$

Here $\delta_i(H)=1$ if $H\in\mathscr{H}_i$ and 0 otherwise. The weighted Bonferroni p-value associated with H is given by

$$p_H=\min_{1\leqslant i\leqslant m}\left(\delta_i(H)\,p_i/v_i(H)\right)$$

where $\delta_i(H)\,p_i/v_i(H)=1$ if $v_i(H)=0$, and H is rejected if $p_H\leqslant\alpha$. By the Bonferroni inequality, the size of this test is no greater than α. Therefore, the resulting closed testing

procedure for the original hypotheses controls the FWE in the strong sense for any set of weight vectors.

The PAAS procedure fits with the current framework as follows. Suppose the weights are fixed for all hypotheses ($v_i(H) \equiv v_i$) and that H is tested using $p_H^{(\mathrm{PAAS})} = \min_{1 \leqslant i \leqslant m} \delta_i(H) p_i / v_i$. Then 'reject all H_i for which $p_i \leqslant v_i \alpha$' is equivalent to the closed testing procedure using p-values $p_H^{(\mathrm{PAAS})}$ for the intersection hypotheses. Viewed as a closed testing procedure, one can see that (i) the weights for the PAAS procedure are identical for all intersections, and (ii) the PAAS method uses extremely conservative tests for some of the intersection hypotheses; for example, the levels of the singletons are $v_i \alpha$, potentially much less than α. More power can be obtained using more powerful tests for the intersections; this is the essential contribution of Holm [18].

In what follows we describe an FWE-controlling closed procedure for testing F_1 and F_2 that meets the following parallel gatekeeping criteria:

Condition 1. The adjusted p-values for the gatekeeper hypotheses $H_1, \ldots, H_k$ do not depend on the significance of the p-values associated with $H_{k+1}, \ldots, H_m$.

Condition 2. The adjusted p-values associated with the secondary hypotheses $H_{k+1}, \ldots, H_m$ are greater than the minimum of $\tilde{p}_1, \ldots, \tilde{p}_k$. This means that the hypotheses in F_2 can be tested if at least one hypothesis in F_1 has been rejected.

The following algorithm shows how to choose the weight vectors to satisfy conditions 1 and 2. Suppose that $w_1, \ldots, w_m$ represent the relative importance of the null hypotheses in F_1 and F_2 with $w_1 + \cdots + w_k = 1$ and $w_{k+1} + \cdots + w_m = 1$. For example, w_1 may be set to 0.8 if H_1 corresponds to the most important primary trial endpoint and 0.2 may be distributed evenly across the remaining gatekeeper hypotheses associated with the less important outcome variables. Using the w_i, we define weights $v_i(H)$ to use in a Bonferroni test of hypothesis H. Every hypothesis H will be tested using a potentially different set of weights $(v_1(H), \ldots, v_m(H))$.

2.1. *Algorithm 1*: *Selecting the weights for the parallel Bonferroni gatekeeping procedure*

Select a hypothesis in the closed family and denote it by H. Consider the following three mutually exclusive cases:

Case 1. If H contains all k gatekeeper hypotheses, that is, $H \in \bigcap_{i=1}^{k} \mathscr{H}_i$, let $v_i(H) = w_i$, $i = 1 \ldots, k$, and $v_i(H) = 0$, $i = k+1, \ldots, m$.

Case 2. If H contains r gatekeeper hypotheses $(1 \leqslant r \leqslant k-1)$, that is, $H \in (\bigcup_{i=1}^{k} \mathscr{H}_i) \cap (\bigcap_{i=1}^{k} \mathscr{H}_i)^c$, let $v_i(H) = w_i \delta_i(H)$, $i = 1, \ldots, k$, and

$$v_i(H) = w_i \delta_i(H) \left(1 - \sum_{j=1}^{k} w_j \delta_j(H)\right) \Bigg/ \sum_{j=k+1}^{m} w_j \delta_j(H), \quad i = k+1, \ldots, m$$

Case 3. If H does not contain any gatekeeper hypotheses, that is, $H \in (\bigcup_{i=1}^{k} \mathscr{H}_i)^c$, let $v_i(H) = 0, i = 1, \ldots, k$, and

$$v_i(H) = w_i \delta_i(H) \Bigg/ \sum_{j=k+1}^{m} w_j \delta_j(H), \quad i = k+1, \ldots, m$$

Table I. Weights assigned to the intersection hypothesis tests.

Intersection hypothesis	Weights			
	H_1	H_2	H_3	H_4
$H_1 \cap H_2 \cap H_3 \cap H_4$	0.5	0.5	0.0	0.0
$H_1 \cap H_2 \cap H_3$	0.5	0.5	0.0	0.0
$H_1 \cap H_2 \cap H_4$	0.5	0.5	0.0	0.0
$H_1 \cap H_2$	0.5	0.5	0.0	0.0
$H_1 \cap H_3 \cap H_4$	0.5	0.0	0.25	0.25
$H_1 \cap H_3$	0.5	0.0	0.5	0.0
$H_1 \cap H_4$	0.5	0.0	0.0	0.5
H_1	0.5	0.0	0.0	0.0
$H_2 \cap H_3 \cap H_4$	0.0	0.5	0.25	0.25
$H_2 \cap H_3$	0.0	0.5	0.5	0.0
$H_2 \cap H_4$	0.0	0.5	0.0	0.5
H_2	0.0	0.5	0.0	0.0
$H_3 \cap H_4$	0.0	0.0	0.5	0.5
H_3	0.0	0.0	1.0	0.0
H_4	0.0	0.0	0.0	1.0

For example, suppose there are two primary and two secondary hypotheses, with equal weights assumed for all tests. There are $2^4 - 1 = 15$ intersection hypotheses H. Table I shows the weights $v_i(H)$ that are used for each test. The parallel gatekeeping procedure simply uses weighted Bonferroni tests for every intersection, with weights as shown.

It is also instructive to compare the proposed weighting scheme with the weighting scheme underlying serial gatekeeping procedures discussed by Westfall and Krishen [6]. Westfall and Krishen showed that serial gatekeeping strategies can be set up by sequentially carrying out two weighted Holm tests [18], that is, by using the following algorithm for defining weight vectors in the closed test for $H_1, \ldots, H_m$:

2.2. Algorithm 2: The weighting scheme for the serial Bonferroni gatekeeping procedure

Select a hypothesis in the closed family and denote it by H. Consider the following two mutually exclusive cases:

Case 1. If H contains at least one gatekeeper hypothesis, that is, $H \in \bigcup_{i=1}^{k} \mathscr{H}_i$, let $v_i(H) = w_i\delta_i(H)/\sum_{j=1}^{k} w_j\delta_j(H)$, $i = 1, \ldots, k$, and $v_i(H) = 0$, $i = k+1, \ldots, m$.

Case 2. If H does not contain any gatekeeper hypotheses, that is, $H \in (\bigcup_{i=1}^{k} \mathscr{H}_i)^c$, let $v_i(H) = 0$, $i = 1, \ldots, k$, and $v_i(H) = w_i\delta_i(H)/\sum_{j=k+1}^{m} w_j\delta_j(H)$, $i = k+1, \ldots, m$.

One can verify that this choice of the weight vectors ensures that the adjusted p-values associated with $H_{k+1}, \ldots, H_m$ are greater than the maximum of $\tilde{p}_1, \ldots, \tilde{p}_k$. In other words, one can test the hypotheses in F_2 only if all hypotheses in F_1 have been rejected. To construct a parallel gatekeeping procedure, one needs to modify the serial weighting scheme by assigning smaller weights to $H_1, \ldots, H_k$ when H contains some but not all gatekeeper hypotheses. For example, assume that H is the intersection of $H_1, \ldots, H_{k-1}$ and $H_{k+1}, \ldots, H_m$ but does not contain H_k. The weight vector associated with the serial gatekeeping approach

(see algorithm 2) is

$$\left(w_1 \Big/ \sum_{j=1}^{k-1} w_j, \ldots, w_{k-1} \Big/ \sum_{j=1}^{k-1} w_j, 0, \ldots, 0 \right)$$

It is important to note that weights for $H_1, \ldots, H_{k-1}$ are defined in such a way that they add up to 1. This immediately implies that the weights assigned to the hypotheses in F_2 are set to zero. To modify the serial weighting scheme, note that the weights associated with the gatekeeper hypotheses contained in H do not add up to 1 in this scenario, that is, $w_1 + \cdots + w_{k-1} < 1$. This gives us an ability to incorporate the secondary p-values into the decision rule. Specifically, the gatekeeper hypotheses will receive the prespecified weights, that is, $w_1, \ldots, w_{k-1}$, and the remainder (that is, $1 - \sum_{j=1}^{k-1} w_j$) will be distributed among the secondary hypotheses according to their importance, that is, the following weight vector will be constructed:

$$\left(w_1, \ldots, w_{k-1}, 0, w_{k+1} \left(1 - \sum_{j=1}^{k-1} w_j \right), \ldots, w_m \left(1 - \sum_{j=1}^{k-1} w_j \right) \right)$$

An important property of the parallel weighting scheme defined in algorithm 1 is that the adjusted p-values associated with the gatekeeper hypotheses are given by $\tilde{p}_i = p_i / w_i$, $i = 1, \ldots, k$. This implies that the hypotheses in F_1 are tested using the weighted Bonferroni tests that take into account the relative importance of the hypotheses. As a result, the gatekeeper hypotheses will be rejected whenever their Bonferroni-adjusted p-values are significant regardless of the values of $p_{k+1}, \ldots, p_m$ and thus condition 1 is satisfied. Further, it is easy to demonstrate that the adjusted p-values $\tilde{p}_{k+1}, \ldots, \tilde{p}_m$ are greater than the minimum of $\tilde{p}_1, \ldots, \tilde{p}_k$. The closed testing procedure based on the weighting scheme in algorithm 1 satisfies condition 2 and therefore it presents a valid parallel gatekeeping procedure.

3. GATEKEEPING PROCEDURES BASED ON SIMES AND RESAMPLING TESTS

The gatekeeping procedure introduced in the previous section is based on the Bonferroni test and thus ignores the correlation among the individual test statistics. It is troubling that (i) Bonferroni-based tests are conservative for large correlation and (ii) Bonferroni-based tests may all be insignificant even when all unadjusted p-values are significant. Resampling-based procedures [11] can alleviate problem (i), while use of the Simes [12] test can alleviate problem (ii).

Simes proposed the following test for an intersection hypothesis H in the closed family. Let t denote the number of elemental hypotheses contained in H and $p_{(1)H} \leqslant p_{(2)H} \leqslant \cdots \leqslant p_{(t)H}$ denote the ordered p-values for the elemental hypotheses. The Simes p-value is then given by $p_H = t \min_{1 \leqslant j \leqslant t} p_{(j)H}/j$. Exact type I error control was proven under independence by Simes [12] and conservative type I error control was established under positive dependency among p-values by Sarkar [13]. This test always produces p-values as small or smaller than the Bonferroni test, for which $p_H = t p_{(1)H}$. The Simes test is popular because of its improved power and because of its close connection with procedures that control the false discovery rate (Benjamini and Hochberg [14]).

Since our methodology uses weighted Bonferroni tests, we use the weighted Simes test proposed by Benjamini and Hochberg [15]. Consider an intersection hypothesis H and a weight vector $v_i(H)$, $i=1,\ldots,m$. Let t and $p_{(1)H},\ldots,p_{(t)H}$ be defined as above and let $v_{(1)H},\ldots,v_{(t)H}$ denote the weights corresponding to the ordered p-values. The weighted Simes p-value is equal to

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

Proof of type I error control for this procedure under positive regression dependency is given by Kling and Benjamini (unpublished manuscript, 2002), thus any closed testing procedure that uses such tests for the intersection hypotheses controls the FWE strongly under the same conditions.

Resampling-based p-values for $H_1,\ldots,H_m$ can be obtained by using parametric resampling. The resulting adjusted p-values are useful (i) to assess robustness of the Simes-based and Bonferroni-based tests to correlation that may be typical for clinical trials, and (ii) to provide alternative large-sample multiplicity-adjusted p-values that directly incorporate correlation. To obtain the parametric resampling-based p-values, consider an intersection hypothesis H and let p_H denote the observed p-value (weighted Bonferroni or weighted Simes) for H. Assuming the usual multivariate normal MANOVA assumptions for our data, with true correlation matrix ρ, the 'true p-value' is $p_H(\rho)=P(P_H\leqslant p_H|\rho)$. An approximate p-value is obtained as the plug-in estimator $p_H(\hat{\rho})$, which can easily be simulated as follows (see Westfall and Young (reference [11], pp. 122–125) and Westfall *et al.* (reference [16], pp. 130–131) for details):

1. Given a consistent estimate of ρ (denoted by $\hat{\rho}$), generate B sets of n independent identically distributed $N(0,\hat{\rho})$ vectors, where n is the combined sample size and B is the number of simulations.
2. Compute a vector of raw p-values for $H_1,\ldots,H_m$ from each simulated data set. Calculate the combined Bonferroni or Simes p-value for each intersection hypothesis in the closed family using the weight vectors defined by algorithm 1. The obtained p-value for H from the ith simulated data set will be denoted by $p_H^*(i)$.
3. The resampling-based p-value for H is equal to

$$\hat{p}_H(\hat{\rho}) = \frac{1}{B}\sum_{i=1}^{B} \delta(p_H^*(i)\leqslant p_H)$$

where $\delta(\cdot)$ is the indicator function.

4. A CLINICAL TRIAL WITH CO-PRIMARY ENDPOINTS

Consider a clinical trial in patients with acute respiratory distress syndrome (ARDS). The trial is conducted to compare one dose of a new drug to placebo. The therapeutic benefits of experimental treatments in ARDS trials are commonly measured using the number of days alive and off mechanical ventilation during a 28-day study period and 28-day all-cause mortality rate (see reference [17] for a detailed description of a recent trial in ARDS patients). Let H_1 and H_2 denote the null hypotheses of no treatment effect with respect to the number of ventilator-free days and 28-day all-cause mortality. It typically takes fewer patients to detect

Table II. Decision matrix for the parallel Bonferroni gatekeeping procedure.

Intersection hypothesis	P-values for intersection hypotheses	Original hypotheses			
		H_1	H_2	H_3	H_4
H_{1111}	$p_{1111} = \min(p_1/0.9, p_2/0.1)$	p_{1111}	p_{1111}	p_{1111}	p_{1111}
H_{1110}	$p_{1110} = \min(p_1/0.9, p_2/0.1)$	p_{1110}	p_{1110}	p_{1110}	0
H_{1101}	$p_{1101} = \min(p_1/0.9, p_2/0.1)$	p_{1101}	p_{1101}	0	p_{1101}
H_{1100}	$p_{1100} = \min(p_1/0.9, p_2/0.1)$	p_{1100}	p_{1100}	0	0
H_{1011}	$p_{1011} = \min(p_1/0.9, p_3/0.05, p_4/0.05)$	p_{1011}	0	p_{1011}	p_{1011}
H_{1010}	$p_{1010} = \min(p_1/0.9, p_3/0.1)$	p_{1010}	0	p_{1010}	0
H_{1001}	$p_{1001} = \min(p_1/0.9, p_4/0.1)$	p_{1001}	0	0	p_{1001}
H_{1000}	$p_{1000} = p_1$	p_{1000}	0	0	0
H_{0111}	$p_{0111} = \min(p_2/0.1, p_3/0.45, p_4/0.45)$	0	p_{0111}	p_{0111}	p_{0111}
H_{0110}	$p_{0110} = \min(p_2/0.1, p_3/0.9)$	0	p_{0110}	p_{0110}	0
H_{0101}	$p_{0101} = \min(p_2/0.1, p_4/0.9)$	0	p_{0101}	0	p_{0101}
H_{0100}	$p_{0100} = p_2$	0	p_{0100}	0	0
H_{0011}	$p_{0011} = \min(p_3/0.5, p_4/0.5)$	0	0	p_{0011}	p_{0011}
H_{0010}	$p_{0010} = p_3$	0	0	p_{0010}	0
H_{0001}	$p_{0001} = p_4$	0	0	0	p_{0001}

Note: The table shows p-values associated with the intersection hypotheses. The adjusted p-values for the original hypotheses H_1, H_2, H_3 and H_4 are defined as the largest p-value in the corresponding column in the right-hand panel of the table (see equation (1)).

a clinically relevant improvement in the number of ventilator-free days compared to 28-day mortality. For this reason, the number of ventilator-free days often serves as the primary endpoint in ARDS trials. However, either of these two endpoints can be used to make regulatory claims. Additionally, there is interest in including information about the drug effects on the number of days the patients were out of the intensive care unit (ICU-free days) and general quality of life in the product label. Denote the secondary hypotheses associated with these secondary endpoints by H_3 and H_4. We wish to develop a parallel gatekeeping procedure that will test the secondary hypotheses only if at least one primary hypothesis has been rejected.

Suppose that the weights for the two gatekeeper hypotheses are given by $w_1 = 0.9$ and $w_2 = 0.1$ and the secondary hypotheses are equally weighted, that is, $w_3 = w_4 = 0.5$. To define the adjusted p-values for $H_1, \ldots, H_4$, consider the closed family associated with the four hypotheses of interest. The closed family includes 15 intersection hypotheses. It is convenient to adopt the following binary representation of the intersection hypotheses. If an intersection hypothesis equals H_1, it will be denoted by H_{1000}. Similarly

$$H_{1100} = H_1 \cap H_2, \quad H_{1010} = H_1 \cap H_3, \quad H_{1001} = H_1 \cap H_4 \text{ etc}$$

The decision matrix in Table II serves as a useful tool that facilitates the computation of p-values for each intersection hypothesis in the closed family and also the adjusted p-values associated with the four original hypotheses. The p-values shown in Table II are based on the weighted Bonferroni rule. To implement the Simes gatekeeping procedure, one needs to test the individual intersection hypotheses using the weighted Simes test introduced in Section 3.

Each row in Table II corresponds to an intersection hypothesis. The p-values associated with the intersection hypotheses are defined using the weighting scheme defined in algorithm 1.

Table III. Bonferroni and Simes gatekeeping procedures in the acute respiratory distress syndrome trial.

Family	Endpoint	Weight	Raw p-value	Adjusted p-value	
				Bonferroni	Simes
Scenario 1					
Primary	Vent-free days	0.9	0.024	0.0267	0.0260
Primary	Mortality	0.1	0.003	0.0300	0.0260
Secondary	ICU-free days	0.5	0.026	0.0289	0.0260
Secondary	Quality of life	0.5	0.002	0.0267	0.0253
Scenario 2					
Primary	Vent-free days	0.9	0.084	0.0933	0.0840
Primary	Mortality	0.1	0.003	0.0300	0.0300
Secondary	ICU-free days	0.5	0.026	0.0933	0.0840
Secondary	Quality of life	0.5	0.002	0.0400	0.0400
Scenario 3					
Primary	Vent-free days	0.9	0.048	0.0533	0.0480
Primary	Mortality	0.1	0.003	0.0300	0.0300
Secondary	ICU-free days	0.5	0.026	0.0533	0.0480
Secondary	Quality of life	0.5	0.002	0.0400	0.0400

The adjusted p-value for H_1, H_2, H_3 and H_4 equals the largest p-value in the corresponding column, that is,

$$\begin{aligned}
\tilde{p}_1 &= \max[p_{1111}, p_{1110}, p_{1101}, p_{1100}, p_{1011}, p_{1010}, p_{1001}, p_{1000}] \\
\tilde{p}_2 &= \max[p_{1111}, p_{1110}, p_{1101}, p_{1100}, p_{0111}, p_{0110}, p_{0101}, p_{0100}] \\
\tilde{p}_3 &= \max[p_{1111}, p_{1110}, p_{1011}, p_{1010}, p_{0111}, p_{0110}, p_{0011}, p_{0010}] \\
\tilde{p}_4 &= \max[p_{1111}, p_{1101}, p_{1011}, p_{1001}, p_{0111}, p_{0101}, p_{0011}, p_{0001}]
\end{aligned} \tag{1}$$

Inferences with respect to H_1, H_2, H_3 and H_4 are performed by comparing these adjusted p-values with the prespecified α.

As an illustration, Table III presents the raw and adjusted p-values produced by the Bonferroni and Simes gatekeeping procedures under three scenarios. Table III shows that all four Bonferroni- and Simes-adjusted p-values are significant at the 0.05 level under the assumptions of scenario 1. This means that the gatekeeping procedures have rejected both gatekeeper hypotheses and continued to test the secondary hypotheses, both of which were also rejected. It is worth noting that the adjusted p-values for the mortality endpoint are considerably larger than the corresponding raw p-value. This is caused by the fact that this endpoint was considered less important than the number of ventilator-free days and was assigned a small weight ($w_2 = 0.1$). Further, the adjusted p-values for the quality of life assessment are also larger than the corresponding raw p-value. This happened because the gatekeeping procedures adjusted the raw p-value upward to make it consistent with the p-value associated with the more important gatekeeper hypothesis. In general, the amount by which secondary p-values are adjusted upward is determined largely by the magnitude of raw p-values associated with the gatekeeper hypotheses. Note also that the weighted Simes test always produces p-values that are as small or smaller than those of the weighted Bonferroni test. As a result, the

Simes gatekeeping procedure produced adjusted p-values that are uniformly smaller than the corresponding Bonferroni-adjusted p-values, without sacrificing type I error control.

Considering scenario 2, we see that an increase in the raw p-value for the number of ventilator-free days did not affect the magnitude of the Bonferroni-adjusted p-value for mortality. This highlights the fact that the gatekeeper hypotheses are tested independently of each other when the Bonferroni gatekeeping procedure is used. Since both gatekeeping procedures have rejected the mortality hypothesis, the secondary analyses were undertaken and, as a result, the secondary hypothesis with the highly significant p-value was rejected. Again, the adjusted p-value for the quality of life assessment is substantially larger than the raw one. The magnitude of the upward adjustment reflects the fact that the more important gatekeeper (vent-free days) is less significant, thus the Bonferroni and Simes procedures need more evidence to reject the secondary (quality of life) hypothesis.

Scenario 3 illustrates an important property of the Simes gatekeeping procedure. This procedure is known to reject all null hypotheses whenever all raw p-values are significant. Thus, the Simes gatekeeping procedure rejected all four null hypotheses in scenario 3, whereas the Bonferroni procedures failed to detect significance with respect to the number of ventilator- and ICU-free days.

5. A CLINICAL TRIAL WITH MULTIPLE ENDPOINTS AND MULTIPLE DOSES

Table IV summarizes the results of a dose-finding study in patients with hypertension. The study was conducted to evaluate effects of low, medium and high doses of an investigational drug compared to placebo. The effects were measured by computing the reduction in systolic and diastolic blood pressure (SBP and DBP) measurements.

The design of the testing strategy is of course done prior to data collection, but having the data visible helps to clarify what is done. In this study, it was felt *a priori* that (i) SBP is more indicative of true effect than DBP, and hence was placed higher in the hierarchy, and (ii) both the medium and high doses were considered equally important, and potentially equally powerful, while the lower dose was considered less likely to exhibit significance. Accordingly, the following four families of null hypotheses were considered: F_1 consisted of the null hypotheses related to the high versus placebo and medium versus placebo comparisons for SBP; F_2 consisted of the null hypotheses related to the high versus placebo and medium versus placebo comparisons for DBP; F_3 contained the null hypothesis for the low versus placebo comparison for SBP; and F_4 contained the null hypothesis for the low versus placebo

Table IV. Results of a dose-finding study in patients with hypertension.

Treatment group	n	Change in SBP (mmHg)		Change in DBP (mmHg)	
		Mean	Standard Deviation	Mean	Standard Deviation
Placebo	29	4.02	7.85	2.31	5.88
Low dose	23	−2.16	8.71	−0.67	5.51
Medium dose	24	−5.03	11.01	−3.18	6.89
High dose	24	−2.60	8.96	−1.44	6.15

Table V. Basic and resampling-based Bonferroni and Simes gatekeeping procedures in the hypertension trial.

Family	Endpoint (comparison)	Weight	Raw p-value	Adjusted p-value			
				Bonferroni	Simes	Resampling Bonferroni	Resampling Simes
F_1	SBP (H versus P)	0.5	0.0101	0.0203	0.0203	0.0196	0.0199
F_1	SBP (M versus P)	0.5	0.0005	0.0011	0.0011	0.0011	0.0011
F_2	DBP (H versus P)	0.5	0.0286	0.0573	0.0573	0.0536	0.0553
F_2	DBP (M versus P)	0.5	0.0016	0.0064	0.0064	0.0062	0.0063
F_3	SBP (L versus P)	1	0.0174	0.0348	0.0286	0.0333	0.0281
F_4	DBP (L versus P)	1	0.0848	0.0848	0.0848	0.0848	0.0848

P = placebo, L = low dose, M = medium dose and H = high dose of the investigational drug.

comparison for DBP. The null hypotheses in F_1 and F_2 were tested in parallel fashion and were equally weighted within each family, reflecting equal importance of the high and medium doses.

Table V displays the raw and adjusted p-values for the individual tests. The raw p-values were computed using the pooled-variance ANOVA two-sided contrast tests. The hierarchical strategy is better in this application than is the usual Bonferroni–Holm [18] or Simes–Hommel procedures [19]; the adjusted p-values for the Simes–Hommel procedure are (in the vertical order of the table) 0.0348, 0.0032, 0.0573, 0.0080, 0.0430 and 0.0848, showing generally less significance. On the other hand, this example shows a case where the serial gatekeeping strategy would have worked better, since both hypotheses are significant in the first gate F_1. However, had the p-values for the first gate not both been significant, the serial strategy would not have allowed continuation, as is allowed with our parallel testing procedure.

The resampling-based calculations shown in Table V used $N = 50\,000\,000$, so that the Monte Carlo error is very small. We can see that, while resampling generally makes the adjusted p-values smaller, as is guaranteed for Bonferroni tests and has been proven recently for weighted Simes tests with positively dependent tests statistics (Benjamini, personal communication), the results are not much affected by correlation. Thus, as correlation has little effect on the adjusted p-values, we can recommend general use of the procedures without parametric resampling, except perhaps in borderline cases. (On the other hand, if non-normality is a great concern, then one should consider non-parametric resampling.)

6. POWER COMPARISONS

A study was conducted to compare the performance of the Bonferroni, Simes and PAAS testing procedures in the case of four null hypotheses grouped into two families (for example,

two primary and two secondary hypotheses). It was assumed that the four individual null hypotheses are of the form

$$H_i = \{\mu_i = 0\}, \quad i = 1, 2, 3, 4$$

where $\mu_1, \ldots, \mu_4$ represent the means of four normally distributed random variables $X_1, \ldots, X_4$ with standard deviation 1 and common correlation coefficient ρ. The unadjusted two-sided p-values were defined as $p_i = 2\min[\Phi(X_i), 1 - \Phi(X_i)]$, $i = 1, \ldots, 4$, where Φ denotes the cumulative distribution function of the standard normal distribution.

The Bonferroni and Simes gatekeeping procedures were performed under the assumption that the null hypotheses were equally weighted within each family, that is, $w_1 = \cdots = w_4 = 0.5$. The adjusted p-values associated with these two gatekeeping procedures were computed as outlined in Sections 2 and 3. Two versions of the PAAS method were used. First, the four null hypotheses were weighted equally and the adjusted p-values were defined as $\tilde{p}_i = 4p_i$, $i = 1, \ldots, 4$. Second, the two primary hypotheses received more weight and the PAAS-adjusted p-values were defined as follows:

$$\tilde{p}_1 = 2.5p_1, \quad \tilde{p}_2 = 2.5p_2, \quad \tilde{p}_3 = 10p_3, \quad \tilde{p}_4 = 10p_4$$

Results for Bonferroni and Simes were obtained using simulation with 1 000 000 samples; all PAAS results were calculated analytically.

Table VI summarizes the results of the study. It shows that the PAAS procedures are more powerful than the gatekeeping procedures with respect to the secondary analyses when the primary null hypotheses are both true ($\mu_1 = \mu_2 = 0$), but that there are substantial power gains from using the Bonferroni and Simes gatekeeping procedures for testing the primary variables. In particular, if the regulatory mandate is to show a significance in the primary analysis, then the gatekeeping procedures show uniformly higher probability of meeting the regulatory mandate, as shown in the 'F_1' column.

It is important to emphasize that the PAAS procedures treat the four hypotheses in the two families as co-primary and test them simultaneously. This approach is justified when secondary endpoints may provide the basis for a new regulatory claim (type I secondary endpoints by the D'Agostino classification [2]). Under the gatekeeping strategies considered here, one does not test the secondary hypotheses unless at least one primary null hypothesis has been rejected. In other words, the gatekeeping approach assumes that the secondary findings will not lead to separate regulatory claims but can only provide supportive evidence for the claims based on the primary endpoints (type II secondary endpoints by the D'Agostino classification).

As suggested by Westfall and Krishen [6], the greatest gains in efficiency for the gatekeeping approach occur when the primary hypotheses have higher power; in our simulation we see that the gatekeeping procedures beat the PAAS procedures even for the secondary endpoints in these cases. An example of such a case is in dose-finding, where the higher doses are expected to exhibit higher power than the lower doses.

However, in fairness, it is important to note that for many disorders the secondary endpoints have greater power than the primary endpoints. Examples include oncology, where regulatory agencies may insist that sponsors demonstrate a benefit with respect to survival (time to death), although a key secondary endpoint, time to progressive disease, generally has greater power. Another example is in depression, where regulators have insisted on the 17-item Hamilton Depression Scale total score as the primary outcome measure. The literature contains several

Table VI. Estimated power of the Bonferroni (B), Simes (S), PAAS with equal weights (PE) and PAAS with unequal weights (PU) testing procedures in the case of two primary and two secondary hypotheses.

Parameters $(\mu_1, \mu_2, \mu_3, \mu_4; \rho)$	H_1				H_3				F_1			
	B	S	PE	PU	B	S	PE	PU	B	S	PE	PU
(0, 0, 0, 0; 0)	2.4	2.4	1.3	2.0	0.2	0.2	1.3	0.5	4.8	4.8	2.5	4.0
(0, 0, 3, 3; 0)	2.4	2.4	1.3	2.0	3.6	3.7	69.2	57.7	4.8	4.8	2.5	4.0
(3, 3, 3, 3; 0)	77.8	82.3	69.2	75.0	76.2	78.2	69.2	57.7	94.9	95.4	90.5	93.7
(3, 3, 0, 0; 0)	77.8	77.8	69.2	75.0	2.1	2.2	1.3	0.5	94.9	94.9	90.5	93.7
(3, 3, 2, 2; 0)	77.8	79.5	69.2	75.0	39.0	41.5	30.9	21.0	94.9	95.1	90.5	93.7
(4, 4, 3, 3; 0)	96.1	97.4	93.3	95.3	82.6	83.6	69.2	57.7	99.9	99.9	99.6	99.8
(4, 4, 2, 2; 0)	96.1	96.6	93.3	95.3	44.2	45.8	30.9	21.0	99.9	99.9	99.6	99.8
(2, 2, 3, 3; 0)	40.6	44.6	30.9	37.2	49.8	52.0	69.2	57.7	64.5	65.4	52.3	60.6
(3, 3, 4, 4; 0)	77.8	83.8	69.2	75.0	92.2	93.0	93.3	88.4	94.9	95.5	90.5	93.7
(2, 2, 4, 4; 0)	40.5	45.9	30.9	37.2	62.3	63.7	93.3	88.4	64.5	65.7	52.3	60.6
(0, 0, 0, 0; .5)	2.4	2.4	1.3	2.0	0.4	0.5	1.3	0.5	4.5	4.5	2.4	3.7
(0, 0, 3, 3; .5)	2.4	2.6	1.3	2.0	2.7	2.9	69.2	57.7	4.5	4.6	2.4	3.7
(3, 3, 3, 3; .5)	77.8	81.4	69.2	75.0	74.8	77.0	69.2	57.7	89.7	90.2	83.7	88.0
(3, 3, 0, 0; .5)	77.8	77.8	69.2	75.0	1.7	1.9	1.3	0.5	89.7	89.7	83.7	88.0
(3, 3, 2, 2; .5)	77.8	78.8	69.2	75.0	42.0	44.2	30.9	21.0	89.7	89.8	83.7	88.0
(4, 4, 3, 3; .5)	96.1	96.8	93.3	95.3	81.5	82.7	69.2	57.7	99.2	99.2	98.2	98.9
(4, 4, 2, 2; .5)	96.1	96.2	93.3	95.3	44.8	46.4	30.9	21.0	99.2	99.2	98.2	98.9
(2, 2, 3, 3; .5)	40.5	45.8	30.9	37.2	50.5	52.5	69.2	57.7	56.6	57.8	45.5	52.9
(3, 3, 4, 4; .5)	77.8	83.1	69.2	75.0	87.8	88.9	93.3	88.4	89.7	90.5	83.7	88.0
(2, 2, 4, 4; .5)	40.5	46.4	30.9	37.2	56.1	57.6	93.3	88.4	56.6	58.0	45.5	52.9

F_1 denotes the probability of passing the front gate.

subscales that have been shown to have greater power in discriminating between active drug and placebo [20]. Thus it may be true that under such scenarios the PAAS method would have greater power at detecting significant differences from a set of two primary and two secondary endpoints, for example. However, regulatory agencies generally would not consider the study as supporting efficacy unless the primary null hypothesis was rejected.

7. DISCUSSION

This paper presents a framework for performing parallel gatekeeping inferences and outlines applications of gatekeeping procedures in clinical trials. Using the closed testing principle, we show how to construct powerful testing procedures that allow one to proceed to lower levels of the hierarchy when not all of the co-primary tests are significant. The benefits of the parallel gatekeeping approach are (i) regulatory acceptability in cases where a front gate of co-primary endpoints must be passed with at least one significance, and (ii) good power. The gatekeeping approach is most appropriate when the secondary analyses do not lead to separate regulatory claims but play a supportive role.

While the method is easiest to motivate and present in terms of Bonferroni-based gatekeeping procedure, we see power gains with the weighted Simes-based procedure and can therefore recommend it over the Bonferroni method. Furthermore, we find that the Simes gatekeeping

procedure is relatively robust to correlation structure in a clinical example when correlations induced by multiple dose group comparisons and multiple endpoints, and therefore using that modification to accommodate correlation may not generally be needed.

REFERENCES

1. O'Neill RT. Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance. *Controlled Clinical Trials* 1997; **18**:550–556.
2. D'Agostino RB. Controlling alpha in clinical trials: the case for secondary endpoints. *Statistics in Medicine* 2000; **19**:763–766.
3. Committee for Proprietary Medicinal Products. Points to consider on multiplicity issues in clinical trials (CPMP/EWP/908/99). London 19 September 2002.
4. Chi GYH. Multiple testings: multiple comparisons and multiple endpoints. *Drug Information Journal* 1998; **32**:1347S–1362S.
5. Bauer P, Röhmel J, Maurer W, Hothorn L. Testing strategies in multi-dose experiments including active control. *Statistics in Medicine* 1998; **17**:2133–2146.
6. Westfall PH, Krishen A. Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures. *Journal of Statistical Planning and Inference* 2001; **99**:25–40.
7. Gong J, Pinheiro JC, DeMets DL. Estimating significance level and power comparisons for testing multiple endpoints in clinical trials. *Controlled Clinical Trials* 2000; **21**:313–329.
8. Moyé LA. Alpha calculus in clinical trials: considerations and commentary for the new millennium. *Statistics in Medicine* 2000; **19**:767–779.
9. Marcus R, Peritz E, Gabriel KR. On closed testing procedure with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
10. Hochberg Y, Tamhane AC. *Multiple Comparison Procedures*. Wiley: New York, 1987.
11. Westfall PH, Young SS. *Resampling-Based Multiple Testing*: *Examples and Methods for P-Value Adjustment*. Wiley: New York, 1993.
12. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **63**:655–660.
13. Sarkar S. Some probability inequalities for ordered MTP2 random variables: a proof of the Simes conjecture. *Annals of Statistics* 1998; **26**:494–504.
14. 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* 1995; **57**:289–300.
15. Benjamini Y, Hochberg Y. Multiple hypothesis testing and weights. *Scandinavian Journal of Statistics* 1997; **24**:407–418.
16. Westfall PH, Ho SY, Prillaman BA. Properties of multiple intersection-union tests for multiple endpoints in combination therapy trials. *Journal of Biopharmaceutical Statistics* 2001; **11**:125–138.
17. ARDS Network. Ventilation with lower tidal volumes for acute lung injury and the acute respiratory distress syndrome. *New England Journal of Medicine* 2000; **342**:1301–1308.
18. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979; **6**: 65–70.
19. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 1988; **75**:383–386.
20. Faries D, Herrera J, Rayamajhi J, DeBrota D, Demitrack M, Potter WZ. The responsiveness of the Hamilton depression rating scale. *Journal of Psychiatric Research* 2000; **34**:3–10.

