STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:2465–2478
Published online 19 October 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2716

Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives

Alex Dmitrienko[1,*,†], Brian L. Wiens[2], Ajit C. Tamhane[3] and Xin Wang[4]

[1]*Eli Lilly and Company, Indianapolis, IN 46285, U.S.A.*
[2]*Myogen Inc., Westminster, CO 80021, U.S.A.*
[3]*Northwestern University, Evanston, IL 60208, U.S.A.*
[4]*Sanofi-Aventis, Bridgewater, NJ 08807, U.S.A.*

SUMMARY

This paper discusses a new class of multiple testing procedures, tree-structured gatekeeping procedures, with clinical trial applications. These procedures arise in clinical trials with hierarchically ordered multiple objectives, for example, in the context of multiple dose–control tests with logical restrictions or analysis of multiple endpoints. The proposed approach is based on the principle of closed testing and generalizes the serial and parallel gatekeeping approaches developed by Westfall and Krishen (*J. Statist. Planning Infer.* 2001; **99**:25–41) and Dmitrienko *et al.* (*Statist. Med.* 2003; **22**:2387–2400). The proposed testing methodology is illustrated using a clinical trial with multiple endpoints (primary, secondary and tertiary) and multiple objectives (superiority and non-inferiority testing) as well as a dose-finding trial with multiple endpoints. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: multiple tests; multiple endpoints; dose-finding; clinical trials

1. INTRODUCTION

Hypothesis testing problems encountered in clinical applications are often based on testing families of null hypotheses in a sequential manner. For example, drug developers can consider a multistage testing strategy in the analysis of a clinical trial with multiple objectives. In this trial the first family of null hypotheses describes the trial's primary outcomes, the second family includes more important secondary outcomes and, finally, the third family is formulated in terms of the less

*Correspondence to: Alex Dmitrienko, Lilly Research Laboratories, Eli Lilly and Company, Lilly Corporate Centre, Drop Code 2233, Indianapolis, IN 46285, U.S.A.
†E-mail: mitrienko_alex@lilly.com

Contract/grant sponsor: National Heart, Lung and Blood Institute; contract/grant number: 1 R01 HL082725-01A1 (Prof. Ajit Tamhane)

Received 19 February 2006
Accepted 28 August 2006

important secondary outcomes such as efficacy findings at intermediate time points and subgroup analyses. To account for the ordering of the endpoints, inferences in each family depend on the acceptance or rejection of null hypotheses in all previously examined families. Since each family serves as a gatekeeper for the families later in the sequence, this testing approach is commonly known as the *gatekeeping testing approach*. Several types of gatekeeping procedures for clinical trials with ordered objectives have been proposed in the literature.

Westfall and Krishen [1], Maurer *et al.* [2] and Bauer *et al.* [3] considered a multistage testing procedure that has the following form. Consider m families of null hypotheses (gates) corresponding to multiple endpoints, objectives or dose–control comparisons. The hypotheses in the first family are tested using any multiple test that controls the familywise error rate (FWER) at a pre-specified α level within that family. The testing is stopped at first failure to reject. Otherwise, one passes the first gate and the next family is tested using any α-level multiple test. Testing continues in this manner until either all hypotheses are rejected or at least one of them is accepted (retained).

An important feature of the multiple testing framework considered in [1–3] is that each family of null hypotheses is examined only if all of the tests carried out before are significant. Dmitrienko *et al.* [4] discussed an alternative scenario in which one sequentially tests families of hypotheses provided at least one hypothesis in each previously examined family is rejected. There is an interesting analogy between the two described gatekeeping procedures and reliability theory problems. In reliability theory, the setting discussed in [1–3] corresponds to a system in which basic elements are connected in series and the strength of the system depends on each individual element. As a result, this method is termed the *serial gatekeeping method*. In contrast, the scenario studied in [4] is conceptually similar to a system with elements connected in parallel and is thus termed the *parallel gatekeeping approach*.

A review of multiple testing problems arising in clinical trials reveals that they typically extend well beyond the simple serial or parallel gatekeeping frameworks [5–7]. As trial designs are becoming increasingly more complex, clinical researchers commonly encounter situations when some null hypotheses are tested serially and the others are examined in a parallel fashion with additional logical restrictions. Examples include dose-finding trials with multiple endpoints in which secondary tests are restricted to the doses at which the primary endpoint was significant or clinical trials conducted to study multiple endpoints and pursue multiple objectives (non-inferiority *versus* superiority).

The goal of this paper is to develop a general framework for setting up hybrid multistage testing procedures in clinical trials with hierarchically ordered objectives. The proposed approach is termed the *tree-structured gatekeeping approach* (abbreviated as *tree gatekeeping approach* hereafter) to emphasize that the decision-making process no longer exhibits a simple sequential structure but rather relies on a decision tree with multiple branches corresponding to individual objectives. The proposed testing strategies extend the serial and parallel gatekeeping methods and can be used to effectively manage multiple tests in a wide variety of applications. The tree gatekeeping tests are constructed based upon the principle of closed testing [8] and protect the FWER in the strong sense at a pre-specified α level [9].

This paper is organized as follows. Section 2 introduces the tree gatekeeping framework. Section 3 reviews the principles used in the construction of tree gatekeeping procedures based on the Bonferroni test in the simple case of two families of hypotheses and Section 4 demonstrates how one can improve the power of Bonferroni-based procedures *via* resampling. Sections 5 and 6 apply the proposed tree gatekeeping procedures to two clinical trials with hierarchically ordered

DOI: 10.1002/sim

multiple objectives and compare the tree gatekeeping and regular gatekeeping procedures. Finally, Section 7 summarizes the conclusions and the Appendix defines the algorithm for constructing tree gatekeeping procedures for the general case involving an arbitrary number of families.

2. GENERAL FRAMEWORK

To introduce the general framework, consider n null hypotheses grouped into m families $F_1, \ldots, F_m$ as shown in Table I. The hypotheses in F_1 may be related to a set of primary analyses in a clinical trial whereas the hypotheses in the other families may represent sequentially ordered secondary analyses (see Sections 5 and 6 for clinical trial examples). As indicated in Table I, $w_{i1}, \ldots, w_{ik_i}$ are the weights representing the importance of the k_i hypotheses within the ith family (note that $w_{i1} + \cdots + w_{ik_i} = 1$) and $p_{i1}, \ldots, p_{ik_i}$ are the associated raw p-values. Finally, let α be the pre-specified FWER.

In the regular gatekeeping framework, each family serves as a gatekeeper for all subsequent families. As was explained in the Introduction, the serial gatekeeping procedure must reject all null hypotheses in a gate (e.g. Family F_i) in order to proceed to the next gates (Families $F_{i+1}, \ldots, F_m$). Likewise, with the parallel gatekeeping approach, at least one test must be significant in a family to pass the gate.

In the more general case considered in this paper, multiple testing is performed in stages and a decision to test a particular null hypothesis at the next stage depends on the rejection of selected (rather than all or at least one) null hypotheses at some or all previous stages. For each individual hypothesis in Family F_i, $i = 2, \ldots, m$, say, H_{ij}, we define two sets of hypotheses denoted by R_{ij}^S and R_{ij}^P. The selected hypothesis is tested only if all hypotheses in the first set, termed the *serial rejection set*, are rejected and at least one hypothesis in the other set, known as the *parallel rejection set*, is found false. Note that, for each null hypothesis H_{ij}, the rejection sets R_{ij}^S and R_{ij}^P can include null hypotheses from $F_1, \ldots, F_{i-1}$ and at least one of them must be non-empty.

Once the two sets have been defined for each null hypothesis in $F_2, \ldots, F_m$, multiple testing is carried out in the following manner. First, all hypotheses in F_1 are tested independently. When testing is complete, the second family, F_2, is considered. For each hypothesis in F_2, say, H_{2j}, one first needs to determine whether or not it is 'testable.' The null hypothesis H_{2j} will be tested only if all hypotheses in R_{2j}^S and at least one hypothesis in R_{2j}^P are rejected. Otherwise, H_{2j} is automatically accepted and the next hypothesis in F_2 is considered. Hypotheses in the other families are tested in a similar fashion.

Table I. Notation used in the paper.

Family	Null hypotheses	Hypothesis weights	Raw p-values
F_1	$H_{11}, \ldots, H_{1k_1}$	$w_{11}, \ldots, w_{1k_1}$	$p_{11}, \ldots, p_{1k_1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
F_i	$H_{i1}, \ldots, H_{ik_i}$	$w_{i1}, \ldots, w_{ik_i}$	$p_{i1}, \ldots, p_{ik_i}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
F_m	$H_{m1}, \ldots, H_{mk_m}$	$w_{m1}, \ldots, w_{mk_m}$	$p_{m1}, \ldots, p_{mk_m}$

DOI: 10.1002/sim

It is easy to verify that the introduced tree gatekeeping approach simplifies to serial gatekeeping testing if $R_{ij}^S = F_{i-1}$ and R_{ij}^P is empty for every H_{ij}, $i>1$. On the other hand, if $R_{ij}^S = \emptyset$ and $R_{ij}^P = F_{i-1}$, the tree gatekeeping strategy turns into a parallel gatekeeping strategy.

3. BONFERRONI-BASED TREE GATEKEEPING PROCEDURE IN THE TWO-FAMILY CASE

Tree gatekeeping procedures can be constructed using the powerful principle of closed testing which guarantees strong FWER control. Before we define the general rules for setting up tree gatekeeping procedures, we will consider a special case of two families of null hypotheses, the primary family $F_1 = \{H_{11}, \ldots, H_{1k}\}$ and secondary family $F_2 = \{H_{21}, \ldots, H_{2k}\}$. In what follows we will introduce a tree gatekeeping procedure for testing the $2k$ null hypotheses based on the basic Bonferroni test. This procedure will be constructed using the *decision matrix approach* [10].

The closed testing family associated with the $2k$ null hypotheses includes $2^{2k} - 1$ intersection hypotheses. For each intersection hypothesis H, let $\delta_{ij}(H) = 1$ if H contains H_{ij} and 0 otherwise. Also, let $\xi_{2j}(H)$ be an indicator that reflects the logical restrictions in F_2. Specifically, $\xi_{2j}(H) = 0$ if H contains at least one null hypothesis from R_{2j}^S or all null hypotheses from R_{2j}^P. Otherwise, $\xi_{2j}(H) = 1$.

To define the tree gatekeeping procedure, we will need to construct a $2k$-dimensional vector for the selected H:

$$v(H) = (v_{11}(H), \ldots, v_{1k}(H), v_{21}(H), \ldots, v_{2k}(H))$$

The following algorithm defines $v(H)$ for each individual intersection hypothesis H:

Case 1: If H contains all null hypotheses from F_1, for any $j = 1, \ldots, k$,

$$v_{1j}(H) = w_{1j}\delta_{1j}(H)$$

$$v_{2j}(H) = 0$$

Case 2: If H contains at least one null hypothesis (but not all null hypotheses) from F_1, for any $j = 1, \ldots, k$,

$$v_{1j}(H) = w_{1j}\delta_{1j}(H)$$

$$v_{2j}(H) = v_1^* w_{2j}\delta_{2j}(H)\xi_{2j}(H) \Big/ \sum_{l=1}^{k} w_{2l}\xi_{2l}(H)$$

where

$$v_1^* = 1 - \sum_{l=1}^{k} w_{1l}\delta_{1l}(H)$$

and $0/0$ is set to 0.

DOI: 10.1002/sim

Case 3: If H does not contain any null hypotheses from F_1, for any $j = 1, \ldots, k$,

$$v_{1j}(H) = 0$$

$$v_{2j}(H) = v_1^* w_{2j} \delta_{2j}(H) \xi_{2j}(H) \Big/ \sum_{l=1}^{k} w_{2l} \delta_{2l}(H) \xi_{2l}(H)$$

Again, 0/0 is set to 0.

The Bonferroni p-value for testing H is given by

$$p_H = \min_{i,j} \{p_{ij} / v_{ij}(H)\}$$

Once the Bonferroni p-values have been computed for each intersection hypothesis in the closed family, a multiplicity-adjusted p-value for a null hypothesis H_{ij} (denoted by $\widetilde{p}_{ij}$) is defined as the largest Bonferroni p-value among all intersection hypotheses containing H_{ij}. The hypothesis H_{ij} is rejected if $\widetilde{p}_{ij} \leqslant \alpha$. By the principle of closed testing, the constructed multiple test controls FWER in the strong sense at the α level.

The weights defined in this algorithm were chosen using the following simple rules:

1. If an intersection hypothesis H contains a null hypothesis from the primary family, say, H_{1j}, the weight assigned to this hypothesis, $v_{1j}(H)$, is equal to the weight reflecting its importance, w_{1j}.
2. For any H containing null hypotheses from F_1 and F_2, one first computes the weight remaining after testing the primary hypotheses, v_1^*. If the weight is positive, it is distributed among the testable secondary hypotheses according to their importance. Also, the weights assigned to the null hypotheses from F_2, i.e. $v_{21}(H), \ldots, v_{2k}(H)$, are normalized using the total weight of the testable hypotheses,

$$\sum_{l=1}^{k} w_{2l} \xi_{2l}(H)$$

   This rule is an extension of a similar rule utilized in the Bonferroni parallel gatekeeping procedure [4].
3. When H contains a null hypothesis from F_2, say, H_{2j}, and at least one hypothesis from its serial rejection set, R_{2j}^S, the indicator $\xi_{2j}(H)$ is set to 0 and thus H_{2j} is given a zero weight. This is done to ensure that H_{2j} cannot be rejected if some hypotheses in R_{2j}^S fail to be rejected. A similar principle is applied when H contains H_{2j} and all hypotheses from the associated parallel rejection set, R_{2j}^P. In this case, $\xi_{2j}(H)$ is set to 0 to prevent the rejection of H_{2j} when no hypotheses in R_{2j}^P are found false.

It is important to note that the sum of the weights assigned to an intersection hypothesis H is no greater than one, i.e. $\sum_{j=1}^{k} v_{ij}(H) \leqslant 1$. This sum can be less than one for some intersection hypotheses, e.g. $H = \{H_{1j}\}$, $j = 1, \ldots, k$. As shown in the Appendix, this property of the tree gatekeeping procedure ensures that the following *independence condition* is satisfied: the adjusted p-values for the null hypotheses in F_i do not depend on the raw p-values for the null hypotheses in $F_{i+1}, \ldots, F_m$, which implies that a decision to reject a null hypothesis in F_i is independent

DOI: 10.1002/sim

of decisions made in $F_{i+1}, \ldots, F_m$. This condition plays an important role in clinical trials with multiple primary and secondary endpoints because it guarantees that inferences with respect to primary hypotheses are unaffected by which and how many secondary hypotheses are rejected. See Dmitrienko *et al*. [10] (Section 2.7) for a detailed discussion of this condition in a clinical trial setting.

The described algorithm for the Bonferroni-based tree gatekeeping procedure is easy to extend to the case of $m>2$ families. The general algorithm for defining tree gatekeeping procedures is given in the Appendix. This algorithm is based on the principles described above to ensure that the procedures meet the serial and parallel gatekeeping criteria, i.e.

$$\widetilde{p}_{ij} \geqslant \max_{H_{kl} \in R_{ij}^S} \widetilde{p}_{kl}, \quad \widetilde{p}_{ij} \geqslant \min_{H_{kl} \in R_{ij}^P} \widetilde{p}_{kl}$$

The first condition states that a null hypothesis, H_{ij}, cannot be rejected unless all of the hypotheses in its serial rejection set, R_{ij}^S, are rejected. Similarly, the second condition states that H_{ij} can be rejected only after at least one of the hypotheses in its parallel rejection set, R_{ij}^P, is found false.

4. RESAMPLING-BASED TREE GATEKEEPING PROCEDURES

The tree gatekeeping procedure introduced in Section 3, as well as its general version described in the Appendix, rely on the Bonferroni test. Specifically, the Bonferroni test is carried out for each intersection hypothesis in the closed family and thus the testing procedure depends entirely on the marginal distributions of the p-values for testing the null hypotheses in $F_1, \ldots, F_m$. It is well known that multiple testing procedures of this kind tend to become conservative in clinical trials applications when the test statistics are strongly positively correlated which, in turn, leads to loss of power.

In order to improve the power of the Bonferroni-based procedures, one can account for the underlying correlation structure *via* resampling [11]. The following is an example of a non-parametric resampling algorithm described in Westfall and Young [11]. Begin with the original sample and generate N bootstrap or permutation samples from an estimated complete null distribution (e.g. from pooled residuals that are obtained by mean centring the observations in each treatment group). Consider the closed family associated with the null hypotheses in $F_1, \ldots, F_m$. For each intersection hypothesis H, compute the Bonferroni p-value from the original sample (denoted by p_H) as well as N Bonferroni p-values from the bootstrap or permutation samples (the p-value from the kth sample is denoted by $p_H(k)$). The resampling p-value for testing H is defined as follows:

$$p_H^* = \frac{1}{N} \sum_{k=1}^{N} I\{p_H(k) < p_H\}$$

where $I\{\}$ is an indicator function. After that, as in Section 3, the adjusted p-value for H_{ij} is defined as the maximum p_H^* over all intersection hypotheses containing H_{ij}.

When the joint distribution of the test statistics is normal or nearly normal, one can define a tree gatekeeping procedure based on parametric resampling. The process of setting up this procedure is similar to the process of constructing a parallel gatekeeping procedure that incorporates correlation *via* parametric resampling [4].

DOI: 10.1002/sim

When implementing these or other resampling algorithms, it is critical to ensure that they retain the correlation structure of the test statistics. For example, in clinical applications involving multiple endpoints resampling must be performed at the patient level so that each patient's measurements are kept together.

Resampling-based procedures control the FWER in the strong sense at a pre-specified α level under the subset pivotality condition [11]. This condition is met in a wide variety of multiple testing problems arising in clinical trials, including multiple testing in general ANOVA models.

5. HYPERTENSION TRIAL EXAMPLE

Consider a clinical trial in patients with hypertension which is conducted to compare an experimental drug to an active control with respect to four endpoints:

- Primary endpoint (P): Mean reduction in systolic blood pressure.
- Two secondary endpoints (S1 and S2): Mean reduction in diastolic blood pressure and proportion of patients with controlled systolic/diastolic blood pressure.
- Tertiary endpoint (T): Average blood pressure based on ambulatory blood pressure monitoring.

The primary comparison in this trial is non-inferiority with a desire to test for superiority for each endpoint conditional on showing non-inferiority for that endpoint. There are eight null hypotheses of interest, a non-inferiority and a superiority hypothesis for each of the four endpoints.

Decision rules are shown in Figure 1: if non-inferiority is shown for an endpoint, superiority will be considered for that endpoint. For example, if non-inferiority is shown for Endpoint P, a superiority analysis for P as well as non-inferiority analysis for S1 and S2 will be performed. If non-inferiority is shown for either S1 or S2 (or both S1 and S2), superiority can be tested for that endpoint and non-inferiority can be tested for Endpoint T. Lastly, if non-inferiority is shown for Endpoint T, superiority can be tested for that endpoint.

It is clear from Figure 1 that the decision tree in this clinical trial example does not have a simple stepwise structure required by gatekeeping tests. While most of the analyses have a single

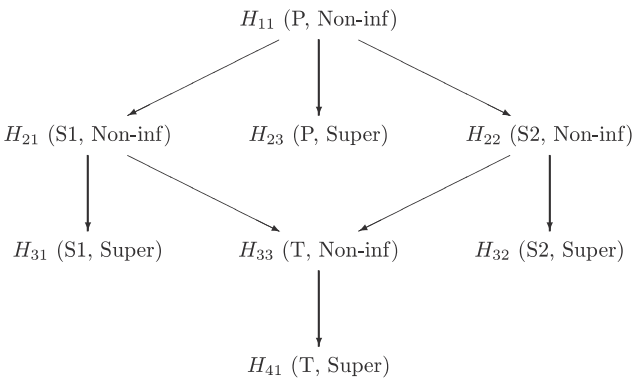

Figure 1. Decision rule in the hypertension clinical trial. The diagram includes references to the primary (P), two secondary (S1 and S2) and tertiary (T) endpoints as well as non-inferiority (Non-inf) and superiority (Super) analyses.

DOI: 10.1002/sim

Table II. Bonferroni-based tree gatekeeping procedure in the hypertension clinical trial.

Family	Null hypothesis	Parallel rejection set	Hypothesis weight	Raw p-value	Adjusted p-value
F_1	H_{11}	NA	1	0.001	0.001*
F_2	H_{21}	$\{H_{11}\}$	1/3	0.008	0.024*
	H_{22}	$\{H_{11}\}$	1/3	0.026	0.078
	H_{23}	$\{H_{11}\}$	1/3	0.003	0.009*
F_3	H_{31}	$\{H_{21}\}$	1/3	0.208	0.624
	H_{32}	$\{H_{22}\}$	1/3	0.302	0.906
	H_{33}	$\{H_{21}, H_{22}\}$	1/3	0.010	0.045*
F_4	H_{41}	$\{H_{31}\}$	1	0.578	0.906

The serial rejection sets for the null hypotheses in F_2–F_4 are empty. The asterisk identifies the adjusted p-values that are significant at the 0.05 level.

predecessor, the non-inferiority test for Endpoint T has two predecessors and depends on them in a 'parallel manner.'

To set up a tree gatekeeping procedure, one first needs to group the described eight null hypotheses into families. Family F_1 consists of all null hypotheses that do not depend on other null hypotheses. Family F_2 consists of all null hypotheses that depend on null hypotheses in F_1, etc. Using this algorithm, the null hypotheses displayed in Figure 1 can be grouped into four families:

Family F_1. Non-inferiority analysis for Endpoint P (H_{11}).

Family F_2. Non-inferiority analysis for Endpoints S1 and S2 (H_{21} and H_{22}) and superiority analysis for Endpoint P (H_{23}).

Family F_3. Superiority analysis for Endpoints S1 and S2 (H_{31} and H_{32}) and non-inferiority analysis for Endpoint T (H_{33}).

Family F_4. Superiority analysis for Endpoint T (H_{41}).

Secondly, based on the decision rule depicted in Figure 1, it is easy to set up rejection sets for the null hypotheses in Families F_2, F_3 and F_4 (Table II). Note that a serial rejection set is equivalent to a parallel rejection set when it includes a single null hypothesis. For the sake of simplicity, Table II focuses on the parallel rejection sets and the serial rejection sets are defined as empty sets. As mentioned above, most null hypotheses in this clinical trial depend on a single null hypothesis from the previous family, for example, $R^P_{21} = \{H_{11}\}$ and $R^P_{32} = \{H_{22}\}$. The only exception is the non-inferiority analysis of the tertiary endpoint (H_{33}) for which $R^P_{33} = \{H_{21}, H_{22}\}$.

In order to compute multiplicity-adjusted p-values for the eight null hypotheses in this example, one first needs to obtain the Bonferroni p-values for the $2^8 - 1 = 255$ intersection hypotheses as shown in the Appendix. As an illustration, we will consider the computation of the Bonferroni p-value for $H = H_{23} \cap H_{41}$. Note that $w_{23} = \frac{1}{3}$ and $w_{41} = 1$. Therefore, $v_{23}(H) = \frac{1}{3}$ and the remaining weight, $\frac{2}{3}$, will be carried over to the last family. Since H includes only one null hypothesis in the last family, i.e. H_{41}, all of the remaining weight will be spent on this null hypothesis. In

DOI: 10.1002/sim

other words, $v_{41}(H) = \frac{2}{3}$ and thus

$$p_H = \min(3p_{23}, 3p_{41}/2) = \min(0.009, 0.867) = 0.009$$

After the Bonferroni p-values have been calculated for the 255 intersection hypotheses in the closed family, an adjusted p-value for a null hypothesis, say, H_{ij}, is found as the maximum Bonferroni p-value among the intersection hypotheses containing H_{ij}.

Table II displays the raw p-values for the eight null hypotheses as well as adjusted p-values produced by the Bonferroni-based tree gatekeeping procedure. The procedure rejects the null hypothesis of lack of non-inferiority for Endpoint P (H_{11}) and proceeds to test the null hypotheses in the second family. The superiority test for Endpoint P (H_{23}) is significant and so is the non-inferiority test for Endpoint S1 (H_{21}). Since H_{21} is in the parallel rejection set for H_{31} and H_{33}, these two null hypotheses are tested next. The testing procedure fails to reject the former but rejects the latter. Note also that, since H_{22} is accepted, one can 'cut off' the associated branch which includes the superiority test for Endpoint S2 (H_{32}). By doing this, one increases the power of other significance tests in Families F_3 and F_4. Reviewing the results, we conclude that the experimental drug is superior to the active control with respect to Endpoint P and non-inferior for Endpoints S1 and T.

6. DOSE-FINDING DIABETES TRIAL EXAMPLE

The second example deals with a clinical trial in patients with Type II diabetes. The trial compares three doses of an experimental drug (Doses L, M and H) *versus* placebo (Plac). The efficacy profile of the drug will be studied using three outcome variables:

- Primary endpoint (P): Haemoglobin A1c.
- Secondary endpoint (S1): Fasting serum glucose.
- Secondary endpoint (S2): HDL cholesterol.

The endpoints will be examined at each of the three doses. To build a testing procedure, we need to define nine null hypotheses and group them into three families. Family F_1 consists of the H-Plac (H_{11}), M-Plac (H_{12}) and L-Plac (H_{13}) comparisons for Endpoint P. Families F_2 and F_3 include the dose–placebo comparisons for Endpoints S1 and S2, respectively.

The three families will be tested sequentially with the following caveat. The tests for Endpoint S1 (Family F_2) will be restricted to the doses at which Endpoint P demonstrated a significant effect. Likewise, a dose–placebo test for Endpoint S2 (Family F_3) will be carried out only if the corresponding tests for Endpoints P and S1 produced significant results. This logical restriction arises in a large number of registration trials and helps the trial sponsors streamline drug labelling.

As in the hypertension trial example, it is easy to 'quantify' the described decision rule by defining rejection sets for the null hypotheses in F_2 and F_3. Table III displays the serial rejection sets for the six null hypotheses (the corresponding parallel rejection sets are empty). Using these rejection sets, we can now carry out the Bonferroni- and resampling-based tree gatekeeping tests and compare their performance in this clinical trial.

Beginning with the tree gatekeeping procedure derived from the Bonferroni test, we can see from Table III that four adjusted p-values are significant at the 0.05 level. First, the two higher doses demonstrate a significant effect on Endpoint P compared to placebo (H_{11} and H_{12}). No

DOI: 10.1002/sim

Table III. Bonferroni- and resampling-based tree gatekeeping procedures in the Type II diabetes clinical trial.

Family	Null hypothesis set	Serial rejection	Hypothesis weight	Raw p-value	Adjusted p-value: Bonferroni procedure	Adjusted p-value: Resampling procedure
F_1	H_{11}	NA	1/3	0.005	0.015*	0.015*
	H_{12}	NA	1/3	0.011	0.033*	0.033*
	H_{13}	NA	1/3	0.018	0.054	0.035*
F_2	H_{21}	$\{H_{11}\}$	1/3	0.009	0.027*	0.026*
	H_{22}	$\{H_{12}\}$	1/3	0.026	0.078	0.076
	H_{23}	$\{H_{13}\}$	1/3	0.013	0.054	0.035*
F_3	H_{31}	$\{H_{11}, H_{21}\}$	1/3	0.010	0.030*	0.029*
	H_{32}	$\{H_{12}, H_{22}\}$	1/3	0.006	0.078	0.076
	H_{33}	$\{H_{13}, H_{23}\}$	1/3	0.051	0.076	0.076

The parallel rejection sets for the null hypotheses in F_2 and F_3 are empty. The asterisk identifies the adjusted p-values that are significant at the 0.05 level.

significant effect is detected at the lowest dose. Therefore, we can focus on the H-Plac and M-Plac comparisons in the second family. Next, the H-Plac contrast is significant for Endpoint S1 (H_{21}) which means that only the highest dose will be compared to placebo in the third family. The H-Plac contrast also turns out to be significant for Endpoint S2 (H_{31}) and thus we conclude that the experimental drug separates from placebo for all three endpoints at Dose H and exhibits a significant effect for Endpoint P at Dose M.

Table III also shows the adjusted p-values generated by the tree gatekeeping procedure that accounts for the correlation among the dose–placebo contrasts and endpoints. This procedure is based on the non-parametric resampling algorithm described in Section 4. The resampling-based adjusted p-values were computed from $N = 50\,000$ bootstrap samples and are virtually uniformly smaller than the Bonferroni-adjusted p-values. As a result, the resampling-based procedure rejects more null hypotheses. For example, since the resampling-based procedure detects a significant difference between L and Plac for Endpoint P (H_{13}), we can follow the branch and test the L-Plac contrast for Endpoint S1 (H_{23}). This test also produces a significant outcome.

It is instructive to compare the described tree gatekeeping strategy with a parallel gatekeeping approach that treats F_1 and F_2 as parallel gatekeepers and thus ignores the logical restrictions. Table IV shows that the parallel gatekeeping procedure can be written as a tree gatekeeping procedure by appropriately defining parallel rejection sets for the null hypotheses in F_2 and F_3. This table also displays adjusted p-values produced by the Bonferroni-based parallel gatekeeping procedure. It is easy to verify that this testing procedure finds only three significant dose–placebo contrasts, the H-Plac contrast for Endpoints P and S1 (H_{11} and H_{21}) and M-Plac contrast for Endpoint P (H_{12}). The H-Plac comparison for Endpoint S2 (H_{31}) turns out to be non-significant because the parallel gatekeeping test does not take the logical restrictions into account and thus does not 'cut off' branches when it encounters a non-significant result. This leads to power loss for the tests that are placed later in the sequence (in this case, dose–placebo comparisons for Endpoint S2).

DOI: 10.1002/sim

Table IV. Parallel gatekeeping procedure in the Type II diabetes clinical trial.

Family	Null hypothesis	Raw p-value	Parallel rejection set	Adjusted p-value
F_1	H_{11}	0.005	NA	0.015*
	H_{12}	0.011	NA	0.033*
	H_{13}	0.018	NA	0.054
F_2	H_{21}	0.009	$\{H_{11}, H_{12}, H_{13}\}$	0.041*
	H_{22}	0.026	$\{H_{11}, H_{12}, H_{13}\}$	0.078
	H_{23}	0.013	$\{H_{11}, H_{12}, H_{13}\}$	0.054
F_3	H_{31}	0.010	$\{H_{21}, H_{22}, H_{23}\}$	0.054
	H_{32}	0.006	$\{H_{21}, H_{22}, H_{23}\}$	0.054
	H_{33}	0.051	$\{H_{21}, H_{22}, H_{23}\}$	0.076

The serial rejection sets for the null hypotheses in F_2 and F_3 are empty. The adjusted p-values are computed using the regular Bonferroni approach. The asterisk identifies the adjusted p-values that are significant at the 0.05 level.

Table V. Tree gatekeeping procedure with equally important secondary endpoints in the Type II diabetes clinical trial.

Family	Null hypothesis	Raw p-value	Serial rejection set	Adjusted p-value
F_1	H_{11}	0.005	NA	0.015*
	H_{12}	0.011	NA	0.033*
	H_{13}	0.018	NA	0.054
F_2	H_{21}	0.009	$\{H_{11}\}$	0.045*
	H_{22}	0.026	$\{H_{12}\}$	0.052
	H_{23}	0.013	$\{H_{13}\}$	0.054
F_3	H_{31}	0.010	$\{H_{11}\}$	0.045*
	H_{32}	0.006	$\{H_{12}\}$	0.036*
	H_{33}	0.051	$\{H_{13}\}$	0.054

The parallel rejection sets for the null hypotheses in F_2 and F_3 are empty. The adjusted p-values are computed using the regular Bonferroni approach. The asterisk identifies the adjusted p-values that are significant at the 0.05 level.

To gain another insight into the nature of tree gatekeeping procedures and appreciate the flexibility of the proposed testing framework, one can compare the scenario reviewed above (Endpoints S1 and S2 are ordered) to the case of equally important secondary endpoints. To construct a tree gatekeeping procedure for this case, all we need to do is to re-define the serial rejection sets for the null hypotheses in F_3 and then re-run the algorithm to compute a new set of adjusted p-values. The modified rejection sets and resulting adjusted p-values are shown in Table V. Since the dose–placebo tests in F_3 no longer depend on the outcome of the corresponding tests in F_2, it might be possible to find more significant results for Endpoint S2. Table V demonstrates that this is actually the case. Comparing the Bonferroni-adjusted p-values displayed in Table V to

DOI: 10.1002/sim

those shown in Table III, we discover that the four dose–placebo comparisons that were significant in the ordered case retain significance and, in addition, the M-Plac contrast for Endpoint S2 (H_{32}) turns out to be significant.

7. CONCLUSIONS

This paper introduces a general family of flexible testing procedures (tree gatekeeping procedures) for multiple testing problems encountered in clinical trials with hierarchically ordered objectives. The described framework helps unify and extend a variety of multiple testing procedures proposed in the literature, including serial gatekeeping tests [1–3], parallel gatekeeping tests [4] and gatekeeping tests with logical restrictions [5].

The paper defines tree gatekeeping procedures derived from the basic Bonferroni test and also shows how these procedures can be extended to account for the correlation among the individual test statistics. Clinical trial examples illustrate the power and flexibility of the proposed testing approach and demonstrate that it can be used to effectively manage multiplicity issues arising in a wide variety of applications.

APPENDIX

Weight assignment algorithm. Consider the general case of testing null hypotheses in m families. The ideas presented in Section 3 can be extended to define Bonferroni-based tree gatekeeping procedures using the principle of closed testing and decision matrix approach.

Consider the closed family associated with the n null hypotheses in Families $F_1, \ldots, F_m$. For each intersection hypothesis H, define the indicator functions $\delta_{ij}(H)$ and $\xi_{ij}(H)$ as in Section 3. To define the general tree gatekeeping procedure, we need to construct an n-dimensional weight vector for H. In order to facilitate this process, the following algorithm sequentially defines m subvectors, one for each family of null hypotheses (it is assumed in the algorithm that $0/0 = 0$):

Step 1: Family F_1. Let $v_{1j}(H) = w_{1j}\delta_{1j}(H)$, $j = 1, \ldots, k_1$, and let v_1^* be the remaining weight that can be used in Families $F_2, \ldots, F_m$, i.e. $v_1^* = 1 - \sum_{j=1}^{k_1} v_{1j}(H)$.

$\vdots$

Step l: Family F_l. Let

$$v_{lj}(H) = v_1^* w_{lj}\delta_{lj}(H)\xi_{lj}(H) \Bigg/ \sum_{s=1}^{k_l} w_{ls}\xi_{ls}(H)$$

where $j = 1, \ldots, k_l$. The remaining weight at this step is given by $v_l^* = v_{l-1}^* - \sum_{j=1}^{k_l} v_{lj}(H)$.

$\vdots$

Step m: Family F_m. Let

$$v_{mj}(H) = v_{m-1}^* w_{mj}\delta_{mj}(H)\xi_{mj}(H) \Bigg/ \sum_{l=1}^{k_m} w_{ml}\delta_{ml}(H)\xi_{ml}(H)$$

where $j = 1, \ldots, k_m$.

DOI: 10.1002/sim

As in Section 3, the Bonferroni p-value for the selected intersection hypothesis H is given by

$$p_H = \min_{i,j} \{p_{ij}/v_{ij}(H)\}$$

and the adjusted p-value for a null hypothesis, say, H_{ij}, is found by computing the maximum p_H over all intersection hypotheses containing H_{ij}.

As shown in Dmitrienko *et al.* [10] (Section 2.7), this algorithm is easy to implement by constructing a $2^{n-1} \times n$ matrix in which each row corresponds to an intersection hypothesis in the closed family and each column corresponds to a null hypothesis. The matrix is populated by the weights, $v_{ij}(H)$, and the Bonferroni p-values are computed for each row. After that, adjusted p-values for the original null hypotheses are obtained by finding the maximum in each column. This approach is implemented in the SAS macro (%TreeGatekeeper) that can be downloaded from the BioPharmNet web site (www.biopharmnet.com/code).

Independence condition. In order to understand why the proposed tree gatekeeping procedure satisfies the independence condition given in Section 3, note that, for any intersection hypothesis H, the weights, $v_{ij}(H)$, are defined sequentially and determined solely by the higher ranked hypotheses contained in H. The presence or absence of lower ranked hypotheses in H does not affect the weights assigned to the higher ranked ones and thus, intuitively, the adjusted p-value for H_{ij} will not depend on the p-values for the null hypotheses in $F_{i+1}, \ldots, F_m$.

To provide a more formal proof, consider a null hypothesis H_{ij}, where $i = 1, \ldots, m-1$ (note that the independence condition is relevant only for null hypotheses in $F_1, \ldots, F_{m-1}$). Let $\mathscr{H}_{ij}$ be the set of all intersection hypotheses containing H_{ij} and let $\mathscr{H}_{ij}^{-}$ denote the set of all intersection hypotheses that contain null hypotheses in $F_1, \ldots, F_{i-1}$ and H_{ij}.

First, let H^* be a non-empty intersection formed by any hypotheses, other than H_{ij}, from $F_i, \ldots, F_m$. For any $H \in \mathscr{H}_{ij}^{-}$,

$$p_{H \cap H^*} \leqslant p_H$$

since, by adding H^* to H, no previously assigned weights in H are changed. Next, the adjusted p-value $\widetilde{p}_{ij}$ is the largest p_H over all intersection hypotheses in $\mathscr{H}_{ij}$. By the obtained inequality, the maximum is achieved at an intersection hypothesis $H \in \mathscr{H}_{ij}^{-}$, i.e.

$$\widetilde{p}_{ij} = \max_{H \in \mathscr{H}_{ij}^{-}} p_H$$

It immediately follows from this representation of the adjusted p-value that $\widetilde{p}_{ij}$ is independent of the p-values for the null hypotheses in $F_{i+1}, \ldots, F_m$ and the independence condition holds.

ACKNOWLEDGEMENT

This research was partially supported by grant award 1 R01 HL082725-01A1 from National Heart, Lung and Blood Institute to Prof. Ajit Tamhane at Northwestern University.

REFERENCES

1. Westfall PH, Krishen A. Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *Journal of Statistical Planning and Inference* 2001; **99**:25–41.

DOI: 10.1002/sim

2. Maurer W, Hothorn L, Lehmacher W. Multiple comparisons in drug clinical trials and preclinical assays: a priori ordered hypotheses. In *Biometrie in der chemisch-pharmazeutischen Industrie*, Vollmar J (ed.). Fischer: Stuttgart, 1995; 3–18.
3. Bauer P, Röhmel J, Maurer W, Hothorn L. Testing strategies in multi-dose experiments including active control. *Statistics in Medicine* 1998; **17**:2133–2146.
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. Chen X, Luo X, Capizzi T. The application of enhanced parallel gatekeeping strategies. *Statistics in Medicine* 2005; **24**:1385–1397.
6. Chen X, Capizzi T, Binkowitz B, Quan H, Wei L, Luo X. Decision rule based multiplicity adjustment strategy. *Clinical Trials* 2005; **2**:394–399.
7. 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.
8. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
9. Hochberg Y, Tamhane AC. *Multiple Comparison Procedures*. Wiley: New York, 1987.
10. Dmitrienko A, Molenberghs G, Chuang-Stein C, Offen W. *Analysis of Clinical Trials Using SAS*: *A Practical Guide*. SAS Press: Cary, NC, 2005.
11. Westfall PH, Young SS. *Resampling-Based Multiple Testing*: *Examples and Methods for P-Value Adjustment*. Wiley: New York, 1993.

DOI: 10.1002/sim