Research Article

Statistics in Medicine

Received 29 January 2010, Accepted 26 May 2010 Published online 18 April 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4008

Mixtures of multiple testing procedures for gatekeeping applications in clinical trials

Alex Dmitrienko[a*†] **and Ajit C. Tamhane**[b]

This paper proposes a general framework for constructing gatekeeping procedures for clinical trials with hierarchical objectives. Such problems frequently exhibit complex structures including multiple families of hypotheses and logical restrictions. The proposed framework is based on combining multiple procedures across families. It enables the construction of powerful and flexible gatekeeping procedures that account for general logical restrictions among the hypotheses of interest. A clinical trial in patients with schizophrenia is used to illustrate the approach for parallel gatekeeping, whereas another clinical trial in patients with hypertension is used to illustrate the approach for gatekeeping with general logical restrictions. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: multiple comparisons; familywise error rate; closed procedures; gatekeeping procedures; Bonferroni procedure; Dunnett procedure; logical restrictions; mixing function; error rate function; separable procedures

1. Introduction

Gatekeeping procedures address the problems of testing hierarchically ordered and logically related null hypotheses that arise in clinical trials involving multiple endpoints, multiple doses, noninferiority–superiority tests, subgroup analyses, etc. Because of the practical importance of these problems, gatekeeping procedures have become an active area of research since the past decade; see Dmitrienko and Tamhane [1, 2] for recent reviews. Much of this work deals with the so-called serial and parallel gatekeeping procedures, and their generalization to the so-called tree-structured gatekeeping procedures. The goal of the present paper is to provide a powerful method for constructing gatekeeping procedures, especially in the case where there are general logical restrictions among the null hypotheses that are outside the scope of tree-structured gatekeeping restrictions.

Consider a clinical trial with multiple objectives that can be classified into primary objectives (e.g. primary endpoints), secondary objectives (e.g. secondary endpoints) and possibly other, less important objectives. Suppose there are $k \geqslant 2$ hypotheses associated with all these objectives. To account for the hierarchical structure of these objectives, the hypotheses are grouped into $m \geqslant 2$ ordered families, $F_1, \ldots, F_m$, with $k_1, \ldots, k_m$ hypotheses, respectively, such that $\sum_{i=1}^{m} k_i = k$. In serial gatekeeping, the hypotheses in F_{i+1} are tested if and only if (iff) all the hypotheses in F_i are rejected, whereas in parallel gatekeeping the hypotheses in F_{i+1} are tested iff at least one hypothesis in F_i is rejected. If the appropriate condition is not met then all hypotheses in F_j for $j > i$ are accepted without testing. Thus F_i serves as a gatekeeper for families F_j for $j > i$. Serial gatekeeping procedures were studied by Maurer *et al.* [3], Bauer *et al.* [4] and Westfall and Krishen [5], whereas Dmitrienko *et al.* [6] were the first to study parallel gatekeeping procedures.

[a]*Eli Lilly and Company, Indianapolis, IN 46285, U.S.A.*
[b]*Northwestern University, Evanston, IL 60208, U.S.A.*
[*]*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, alex.dmitrienko@gmail.com*

Further research in this area was conducted along two directions. In the first direction, Dmitrienko *et al.* [7, 8] generalized serial and parallel gatekeeping procedures, which assume a simple set of logical restrictions among the hypotheses, to the tree-structured gatekeeping procedures. In these procedures, the decision to test any particular null hypothesis in F_i for $i>1$ or to accept it without a test is conditional on rejection of all hypotheses from one subset of hypotheses (called a serial rejection set) and rejection of at least one hypothesis from another subset of hypotheses (called a parallel rejection set) from $F_1, \ldots, F_{i-1}$ (one of these rejection sets may be empty). Dmitrienko *et al.* [7] derived procedures to deal with these tree-structured gatekeeping restrictions using the closure method of Marcus *et al.* [9] based on the Bonferroni procedure as the local test for each intersection hypothesis. Bretz *et al.* [10] used a graphical approach and Burman *et al.* [11] used a closely related recycling approach to construct Bonferroni-based gatekeeping procedures with general logical relationships among the hypotheses.

In the second direction, Dmitrienko *et al.* [12] considered the problem of building multistage parallel gatekeeping procedures based on the procedures that are more powerful than the Bonferroni procedure. They introduced a general method for this purpose derived from a broad class of the so-called separable procedures (explained in the sequel) using the notion of the error rate function. This general method eliminates the need to use the closure method thus simplifying the computations.

These previous works have certain limitations, however. First, the tree-structured gatekeeping approach is designed to use only the Bonferroni procedures. Second, although the framework introduced in [12] uses more powerful separable procedures than the Bonferroni procedures used in [13, 14], it can only deal with parallel gatekeeping restrictions. In the present paper, we propose a new approach for constructing gatekeeping procedures that not only overcome these limitations, thus yielding more powerful procedures, but also extend their applicability to general logical restrictions which are not representable in the tree-structured gatekeeping framework.

The new approach is based on mixtures of multiple testing procedures. The term *mixture* is used here to make an analogy with a mixture of distributions [15]. To specify a mixture distribution, one needs to specify component distributions and a mixing distribution. Similarly, in the case of a mixture procedure, one needs to specify component procedures and a mixing function. Mixtures of multiple testing procedures were implicitly used as a tool for building gatekeeping procedures in other papers, e.g. mixtures of the Dunnett procedures were considered in [16, 17], the tree-structured gatekeeping approach in [7, 8] was based on mixtures of the Bonferroni procedures and the algorithm for constructing multistage parallel gatekeeping procedures [12] employed mixtures of general multiple testing procedures.

To introduce mixture procedures, we have chosen to restrict to the simple setting of two families, F_1 and F_2, in the present paper. We will refer to F_1 and F_2 as the primary and secondary families and their corresponding hypotheses as the primary and secondary null hypotheses, respectively. Multiple families pose no new conceptual problems but involve more complicated notation and arguments. In the present paper, we will focus on the procedures and not on their properties. The theory of mixture procedures for multiple families and their properties will be studied in a separate paper.

All procedures considered in this paper will be assumed to satisfy the following strong familywise error rate (FWER) control requirement [18, 19]:

$$\text{FWER} = P(\text{Reject at least one true null hypothesis}) \leqslant \alpha$$

for a specified α for any combination of the true and false null hypotheses. In addition, gatekeeping procedures are generally required to satisfy the desirable property of *independence* [6] which states that inferences on the hypotheses in any family F_i must not depend on the test statistics or the p-values of the hypotheses in family F_j for $j>i$. In the case of two families considered in this paper, the independence condition implies that inferences on the hypotheses in family F_1 do not depend on the inferences on the hypotheses in family F_2.

The outline of this paper is as follows. Section 2 introduces the mixture procedures in the simple setting of parallel gatekeeping. Section 3 extends the mixture framework to general monotone gatekeeping restrictions through the so-called restriction functions. Clinical trial examples are given in Sections 2 and 3 to illustrate mixture procedures for parallel and general gatekeeping restrictions. Section 4 gives the concluding remarks.

2. Mixture procedures for parallel gatekeeping restrictions

We will use the following example involving subgroup analyses to illustrate the procedures for parallel gatekeeping. Consider a parallel-group clinical trial in patients with schizophrenia which is conducted to

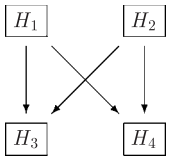

Figure 1. Decision tree in the schizophrenia trial example with parallel gatekeeping restrictions (the secondary null hypotheses H_3 and H_4 are tested iff at least one primary null hypothesis, i.e. H_1 or H_2, is rejected).

evaluate the efficacy profiles of two doses of a treatment (Dose 1: low dose; Dose 2: high dose) compared with that of a placebo in a general population of patients as well as in a prespecified subpopulation (subgroup) defined by a genotypic classifier. Suppose that the primary efficacy endpoint is continuous and the treatment effect is defined to be the mean difference between each dose and placebo. Let δ_i and δ_i' denote the true mean differences for the comparison between the ith dose and placebo in the general population and subpopulation of classifier-positive patients, respectively ($i=1,2$). The primary family F_1 consists of the null hypotheses of no treatment effect in the general population, i.e.

$$H_1:\delta_1=0, \quad H_2:\delta_2=0, \tag{1}$$

and the secondary family F_2 consists of the null hypotheses of no treatment effect in the subpopulation, i.e.

$$H_3:\delta_1'=0, \quad H_4:\delta_2'=0. \tag{2}$$

The decision tree used in this clinical trial is displayed in Figure 1. The secondary null hypotheses are tested iff at least one primary null hypothesis is rejected, i.e. there is evidence of a beneficial treatment effect in at least one dose group in the general population. Thus the primary family serves as a parallel gatekeeper for the secondary family.

We now discuss how to set up mixture procedures for parallel gatekeeping with arbitrary numbers of hypotheses in each of two families.

2.1. Mixture procedures

Consider the general problem of testing $k\geqslant 2$ null hypotheses grouped into two families, F_1 with k_1 primary null hypotheses and F_2 with k_2 secondary null hypotheses, where F_1 serves as a parallel gatekeeper for F_2 and $k_1+k_2=k$. The procedures used in the primary and secondary families will be termed *component procedures* and denoted by $\mathscr{P}_1$ and $\mathscr{P}_2$, respectively. We assume that each $\mathscr{P}_i$ is a closed testing procedure which controls the FWER within F_i at a prespecified α_i level ($i=1,2$) and $\mathscr{P}_1$ is a *separable* procedure [12]. The class of separable procedures includes the Bonferroni and single-step Dunnett [20] procedures. Popular stepwise procedures such as the Holm [21], Hochberg [22], Wiens' fallback [23, 24], step-up Dunnett [25] and step-down Dunnett [26] are not separable. However, they can be made separable by applying them using convex combinations of their critical values with those of one of the separable procedures, e.g. Bonferroni for Holm, Hochberg, and fallback, and single-step Dunnett for step-up and step-down Dunnett. The resulting procedures are referred to as *truncated procedures* and their properties are discussed in [12].

We are interested in constructing a *mixture procedure* $\mathscr{P}$ from $\mathscr{P}_1$ and $\mathscr{P}_2$ for testing the null hypotheses in the combined family $F=F_1\cup F_2$ such that $\mathscr{P}$ controls the FWER with respect to all k null hypotheses. We use the closure method to set up $\mathscr{P}$.

Let

$$K_1=\{1,\ldots,k_1\}, \quad K_2=\{k_1+1,\ldots,k_1+k_2\}, \quad K=K_1\cup K_2=\{1,\ldots,k\}$$

be the index sets of the null hypotheses in F_1, F_2 and F, respectively. Consider closed families associated with the primary and secondary families of null hypotheses. A closed family for F_i ($i=1,2$) consists of all nonempty intersections of the null hypotheses in F_i. Let

$$H(I_i)=\cap_{j\in I_i}H_j$$

denote an intersection hypothesis, where $I_i\subseteq K_i$ is the index set of the null hypotheses included in $H(I_i)$. Let $p_i(I_i)$ denote the local p-value for testing the intersection hypothesis $H(I_i)$. This p-value

defines a *local test* for this intersection hypothesis. Recall that the component procedure $\mathscr{P}_i$ is a closed testing procedure and thus each local test is an α-level test. For example, if $\mathscr{P}_i$ is the Holm procedure then the local test of $H(I_i)$ is the Bonferroni test and

$$p_i(I_i)=|I_i|\min_{j\in I_i}(p_j), \tag{3}$$

where $|I_i|$ is the cardinality of the index set I_i and p_j is the raw p-value for H_j.

For given α, the error rate function of $\mathscr{P}_1$ is defined as

$$e_1(I_1|\alpha)=P\{p_1(I_1)\leqslant\alpha|H(I_1)\}.$$

If this probability depends upon the false null hypotheses H_i, $i\notin I_1$ then we take the supremum of the probability over all configurations such that H_i, $i\in I_1$ are true and H_i, $i\notin I_1$ are false. As $\mathscr{P}_1$ is a separable procedure, we have $e_1(I_1|\alpha)\leqslant\alpha$ for all $I_1\subseteq K_1$ with equality holding iff $I_1=K_1$. Note that $e_1(I_1|\alpha)$ is a monotone function, i.e. $e_1(I_1|\alpha)\leqslant e_1(I_1'|\alpha)$ if $I_1\subseteq I_1'$. Generally, an exact expression for the error rate function of any procedure is difficult to obtain and hence we use an easily computable upper bound, which we treat as the actual error rate function of that procedure. For example, if $\mathscr{P}_1$ is the Bonferroni procedure, then we use $e_1(I_1|\alpha)=|I_1|\alpha/k_1$. Error rate functions of some popular procedures were derived in [12].

To define the mixture procedure $\mathscr{P}$ for all null hypotheses in the combined family F, consider the associated closed family. Let I be a nonempty subset of K such that $I=I_1\cup I_2$, where $I_1\subseteq K_1$ and $I_2\subseteq K_2$ are the primary and secondary index sets at least one of which is nonempty. Let $H(I)=\cap_{i\in I}H_i$ denote an intersection hypothesis. $\mathscr{P}$ is a closed procedure with a local test of level α of each intersection hypothesis $H(I)$ defined as follows:‡

$$\text{Reject } H(I) \text{ if } \begin{cases} p_i(I_i)\leqslant\alpha & \text{if } I=I_i\,(i=1,2),\\ \phi_I(p_1(I_1),p_2(I_2))\leqslant\alpha & \text{if } I=I_1\cup I_2,\ I_1 \text{ and } I_2 \text{ are nonempty.}\end{cases} \tag{4}$$

In other words, if the intersection hypothesis $H(I)$ contains only primary or secondary null hypotheses ($I=I_1$ or $I=I_2$), the decision rule for $H(I)$ is directly based on the local test associated with the corresponding component procedure. Further, if both primary or secondary null hypotheses are included in $H(I)$ ($I=I_1\cup I_2$), the decision rule is constructed by combining the local tests associated with the primary and secondary component procedures $\mathscr{P}_1$ and $\mathscr{P}_2$ based on the *mixing function* $\phi_I(p_1(I_1),p_2(I_2))$. The mixing function is defined over the interval $(0,1)$ such that the local test of $H(I)$ is an α-level test, i.e.

$$P\{\phi_I(p_1(I_1),p_2(I_2))\leqslant\alpha|H(I)\}\leqslant\alpha. \tag{5}$$

In addition, we require that

$$\phi_I(p_1(I_1),p_2(I_2))\leqslant p_1(I_1). \tag{6}$$

This property guarantees that the mixture procedure $\mathscr{P}$ is equivalent to $\mathscr{P}_1$ within F_1 and thus it satisfies the independence condition, i.e. the inferences in the primary family will not be affected by the inferences in the secondary family. Using the closure principle, $\mathscr{P}$ rejects any individual null hypothesis H_i at multiple level α iff all $H(I)$ such that $i\in I$ are rejected using their local tests (4).

We consider a class of mixing functions of the following general form:

$$\phi_I(p_1(I_1),p_2(I_2))=\min\left(p_1(I_1),\frac{p_2(I_2)}{c(I_1,I_2|\alpha)}\right), \tag{7}$$

where $c(I_1,I_2|\alpha)$ is a coefficient that, in general, depends on the subsets I_1,I_2 and on α. It satisfies $0\leqslant c(I_1,I_2|\alpha)\leqslant 1$. The last inequality ensures that the secondary hypotheses are less important than the primary hypotheses, i.e. $p_2(I_2)$ receives a weight less than or equal to the weight of 1 on $p_1(I_1)$. Different mixing functions differ in their choice of $c(I_1,I_2|\alpha)$ which is chosen to satisfy (5). We will consider two mixing functions: Bonferroni and parametric. They are defined in Sections 2.2 and 2.4, respectively. The Bonferroni mixing function ignores correlations among the test statistics whereas

‡*Note that 'reject $H(I)$' in (4) refers to the local test of $H(I)$ taken in isolation. As $\mathscr{P}$ is a closed procedure, $H(I)$ can be rejected iff all $H(J)$ for $J\supseteq I$ are rejected.*

the parametric mixing function takes them into account by assuming a joint distribution for the test statistics.

Computation of local p-values for the intersection hypotheses in the closed family depends on the property of α-*consistency* [27, 28], which states that the rejection region for each local test must be monotonically nondecreasing in $\alpha \in (0, 1)$. It can be seen from (7) that α-consistency is guaranteed if $\alpha c(I_1, I_2|\alpha)$ is nondecreasing in α. We will explain why the procedures used in the examples discussed in the sequel are α-consistent. The general problem of α-consistency will be addressed in a follow-up paper.

When the α-consistency condition is satisfied, local p-values are computed as follows. Let $p(I)$ denote the p-value for $H(I)$. Noting that $p(I)$ is the smallest α at which $H(I)$ can be rejected, we see that if $I = I_i$ then $p(I) = p_i(I_i)$ for $i = 1, 2$ regardless of the type of mixing function used. If $I = I_1 \cup I_2$, where I_1 and I_2 are both nonempty subsets, one must numerically find the smallest α such that $\phi_I(p_1(I_1), p_2(I_2)) \leqslant \alpha$. Note that since $0 < \phi_I(p_1(I_1), p_2(I_2)) < 1$, the inequality (5) is satisfied for $\alpha = 1$ but not for $\alpha = 0$. Therefore a smallest α that satisfies (5) always exists. In the special case when $c(I_1, I_2|\alpha)$ does not depend on α we can simply set

$$p(I) = \phi_I(p_1(I_1), p_2(I_2)).$$

Once the local p-values for all the intersection hypotheses are computed, the adjusted p-value of any H_i can be computed from

$$\widetilde{p}_i = \max_{I: i \in I} p(I),$$

where the maximum is taken over all index sets I that include i, $i \in K$. The mixture procedure $\mathscr{P}$ rejects the null hypothesis H_i, $i \in I$, iff $\widetilde{p}_i \leqslant \alpha$.

2.2. *Bonferroni mixing function*

Consider an intersection hypothesis $H(I)$, where $I = I_1 \cup I_2$ and I_1 and I_2 are nonempty index sets corresponding to the null hypotheses from F_1 and F_2 included in $H(I)$. The Bonferroni mixing function uses $c(I_1, I_2|\alpha) = 1 - e_1(I_1|\alpha)/\alpha$, which is independent of I_2. We thus have

$$\phi_I(p_1(I_1), p_2(I_2)) = \min\left(p_1(I_1), \frac{p_2(I_2)}{1 - e_1(I_1|\alpha)/\alpha}\right). \tag{8}$$

By the Bonferroni inequality and the definition of the error rate function

$$\begin{aligned} P\{\phi_I(p_1(I_1), p_2(I_2)) \leqslant \alpha | H(I)\} &= P\{p_1(I_1) \leqslant \alpha \text{ or } p_2(I_2) \leqslant \alpha c(I_1, I_2|\alpha) | H(I)\} \\ &\leqslant P\{p_1(I_1) \leqslant \alpha | H(I_1)\} + P\{p_2(I_2) \leqslant \alpha[1 - e_1(I_1|\alpha)/\alpha] | H(I_2)\} \\ &\leqslant e_1(I_1|\alpha) + \alpha[1 - e_1(I_1|\alpha)/\alpha] \\ &= \alpha \end{aligned}$$

and thus (5) is satisfied. Note that if $\mathscr{P}_1$ is the Bonferroni procedure then $e_1(I_1|\alpha) = |I_1|\alpha/n_1$ and hence $c(I_1, I_2|\alpha) = 1 - |I_1|/n_1$, which is independent of both I_2 and α; we denote it by $c(I_1)$.

Note that since the error rate function is a monotone function of its argument, as the index set I_1 gets bigger, $e_1(I_1|\alpha)$ increases. In other words, with the increasing number of primary null hypotheses included in $H(I)$, the primary component procedure $\mathscr{P}_1$ consumes a larger fraction of α. As a result, a smaller fraction of α can be allocated to the secondary null hypotheses. This α-allocation scheme is conceptually similar to that in [12] for calculating α_2 from $\alpha_1 = \alpha$. Furthermore, if $I_1 = K_1$, we have $e_1(K_1|\alpha) = \alpha$ and thus $\phi_I(p_1(K_1), p_2(I_2)) = p_1(K_1)$ for any index set $I_2 \subseteq K_2$. This is an important property which translates into the parallel gatekeeping restriction, i.e. all secondary null hypotheses are automatically accepted whenever all primary null hypotheses are accepted. A detailed treatment of the conditions that guarantee the parallel gatekeeping restriction will be provided in a follow-up paper.

2.3. *Nonparametric mixture procedure*

We now illustrate the Bonferroni mixing function in the special case of the schizophrenia trial example with parallel gatekeeping restrictions for which $k_1 = k_2 = 2$. There are four null hypotheses in this

Table I. Index sets and local p-values for the nonparametric mixture procedure (mixture of the Bonferroni and Holm procedures based on the Bonferroni mixing function) in the schizophrenia trial example with parallel gatekeeping restrictions.

Index set			Coefficient	Local
I	I_1	I_2	$c(I_1)$*	p-value ($p(I)$)
$\{1,2,3,4\}$	$\{1,2\}$	$\{3,4\}$	—	$2\min(p_1, p_2)$
$\{1,2,3\}$	$\{1,2\}$	$\{3\}$	—	$2\min(p_1, p_2)$
$\{1,2,4\}$	$\{1,2\}$	$\{4\}$	—	$2\min(p_1, p_2)$
$\{1,2\}$	$\{1,2\}$	$\emptyset$	—	$2\min(p_1, p_2)$
$\{1,3,4\}$	$\{1\}$	$\{3,4\}$	0.5	$2\min(p_1, 2\min(p_3, p_4))$
$\{1,3\}$	$\{1\}$	$\{3\}$	0.5	$2\min(p_1, p_3)$
$\{1,4\}$	$\{1\}$	$\{4\}$	0.5	$2\min(p_1, p_4)$
$\{1\}$	$\{1\}$	$\emptyset$	—	$2p_1$
$\{2,3,4\}$	$\{2\}$	$\{3,4\}$	0.5	$2\min(p_2, 2\min(p_3, p_4))$
$\{2,3\}$	$\{2\}$	$\{3\}$	0.5	$2\min(p_2, p_3)$
$\{2,4\}$	$\{2\}$	$\{4\}$	0.5	$2\min(p_2, p_4)$
$\{2\}$	$\{2\}$	$\emptyset$	—	$2p_2$
$\{3,4\}$	$\emptyset$	$\{3,4\}$	—	$2\min(p_3, p_4)$
$\{3\}$	$\emptyset$	$\{3\}$	—	p_3
$\{4\}$	$\emptyset$	$\{4\}$	—	p_4

*In cases where $p(I)=p_1(I_1)$ or $p(I)=p_2(I_2)$ (e.g. when $I=I_1$ or $I=I_2$ or $I_1=K_1$) we do not need to specify $c(I_1)$ since the mixing function (7) is not used to compute $p(I)$.

example grouped into two families: $F_1=\{H_1, H_2\}$ and $F_2=\{H_3, H_4\}$. The 15 intersection hypotheses in the associated closed family and the index sets I_1 and I_2 for each intersection hypothesis are shown in Table I. Define the index sets $K_1=\{1,2\}$, $K_2=\{3,4\}$ and $K=\{1,2,3,4\}$. Let p_i denote the raw p-value for testing the null hypothesis H_i $(1\leqslant i\leqslant 4)$.

We will define a nonparametric mixture procedure with $\mathscr{P}_1$ being the Bonferroni procedure and $\mathscr{P}_2$ being the Holm procedure. Numerical illustration for calculations of the adjusted p-values for this nonparametric mixture procedure is given in Example 1 (Section 2.6). To set up this nonparametric mixture procedure, we first need to compute the local p-values for the two-component procedures, i.e. $p_1(I_1)$ for the Bonferroni procedure and $p_2(I_2)$ for the Holm procedure. The local p-values for the Bonferroni procedure are given by

$$p_1(I_1)=\begin{cases} 2\min(p_1, p_2) & \text{if } I_1=\{1,2\},\\ 2p_1 & \text{if } I_1=\{1\},\\ 2p_2 & \text{if } I_1=\{2\}.\end{cases}$$

Using (3), the local p-values for the Holm procedure are given by

$$p_2(I_2)=\begin{cases} 2\min(p_3, p_4) & \text{if } I_2=\{3,4\},\\ p_3 & \text{if } I_2=\{3\},\\ p_4 & \text{if } I_2=\{4\}.\end{cases}$$

Next we show how to compute the local p-values for the nonparametric mixture procedure. Choose an arbitrary nonempty subset I of K. If $I=I_1$ or I_2 then $p(I)=p_1(I_1)$ or $p_2(I_2)$, respectively, as noted before. If $I=I_1\cup I_2$ and both I_1 and I_2 are nonempty then we use the Bonferroni mixing function (8) to calculate the local p-value $p(I)$. Here, we have

$$c(I_1)=1-\frac{e_1(I_1|\alpha)}{\alpha}=\begin{cases} 1 & \text{if } I_1=\emptyset,\\ \frac{1}{2} & \text{if } I_1=\{1\},\{2\},\\ 0 & \text{if } I_1=\{1,2\}.\end{cases}$$

As $c(I_1)$ is independent of α, we can set $p(I)=\phi_I(p_1(I_1), p_2(I_2))$, thus yielding

$$p(I)=\begin{cases} 2\min(p_1, p_2) & \text{if } I_1=\{1,2\}, \quad I_2=\{3,4\},\{3\},\{4\}, \\ 2\min(p_i, 2\min(p_3, p_4)) & \text{if } I_1=\{i\}, \quad i=1,2, \quad I_2=\{3,4\}, \\ 2\min(p_i, p_j) & \text{if } I_1=\{i\}, \quad i=1,2, \quad I_2=\{j\}, \quad j=3,4. \end{cases}$$

These calculations are summarized in Table I, which displays all 15 intersection hypotheses with the associated index sets I, I_1, I_2 and the $p(I)$-values.

It can be shown that the nonparametric mixture procedure is equivalent to the general parallel gatekeeping procedure proposed in [12] in which the primary family is tested using the Bonferroni procedure at level $\alpha_1=\alpha$ and the secondary family is tested using the Holm procedure at level

$$\alpha_2=\alpha_1-e_1(I_1|\alpha_1).$$

Here I_1 is the set of accepted hypotheses from F_1 and $e_1(I_1|\alpha)=|I_1|\alpha/2$ is the error rate function of the Bonferroni procedure.

2.4. Parametric mixing function

To specify a parametric mixing function, we need to specify a joint distribution for the test statistics. We assume that t-statistics, denoted by t_i, are used to test the null hypotheses H_i $(1\leqslant i\leqslant k)$. Further suppose that under the overall null hypothesis $H(K)=\bigcap_{i=1}^{k} H_i$, $(t_1, \ldots, t_k)$ has a k-variate t-distribution with ν degrees of freedom (d.f.) and known correlation matrix $R=\{\rho_{ij}\}$, where ρ_{ij} is the correlation coefficient between the numerators of t_i and t_j.

Suppose that $\mathscr{P}_1$ is the Dunnett procedure and $\mathscr{P}_2$ is the step-down Dunnett procedure. The Dunnett procedure tests and rejects any intersection hypothesis $H(I_1)$ at level α if

$$t_1(I_1)=\max_{i\in I_1} t_i>d_1(\alpha),$$

where $d_1(\alpha)$ is the upper α critical point of the null distribution of $t_1(K_1)$. The step-down Dunnett procedure, on the other hand, being a shortcut to the closed procedure that uses the Dunnett procedure at level α to test each intersection hypothesis, rejects any intersection hypothesis $H(I_2)$ at level α if

$$t_2(I_2)=\max_{i\in I_2} t_i>d_2(\alpha),$$

where $d_2(\alpha)$ is the upper α critical point of the null distribution of $t_2(I_2)$ (and thus varies with I_2). To make the local test (5) of $H(I)$ of level α using the mixing function (7), $c(I_1, I_2|\alpha)$ must satisfy the following equation:

$$\begin{aligned} P\{\phi_I(p_1(I_1), p_2(I_2))\leqslant\alpha|H(I)\} &= P\{p_1(I_1)\leqslant\alpha \text{ or } p_2(I_2)\leqslant c(I_1, I_2|\alpha)\alpha|H(I)\} \\ &= P\{t_1(I_1)\geqslant d_1(\alpha) \text{ or } t_2(I_2)\geqslant d_2(c(I_1, I_2|\alpha)\alpha)|H(I)\} \\ &= \alpha. \end{aligned} \tag{9}$$

As in the nonparametric case considered in Section 2.2, it is easy to verify that the parametric mixing function incorporates the parallel gatekeeping restriction. Indeed, $P\{p_1(K_1)\leqslant\alpha\}=\alpha$ due to the assumptions on the error rate function of $\mathscr{P}_1$ and thus $c(K_1, I_2|\alpha)=0$ for any α. This implies that $\phi_I(p_1(K_1), p_2(I_2))=p_1(K_1)$ and thus no secondary null hypotheses can be rejected if all primary null hypotheses are accepted.

2.5. Parametric mixture procedure

Parametric mixing function will be illustrated for the same special case, $k_1=k_2=2$, considered in Section 2.3 for the nonparametric mixture procedure. To define the test statistics in this case, let n_0, n_1 and n_2 denote the total sample sizes for the placebo and the two dose groups, respectively, in the general population and let n_0', n_1' and n_2' denote the corresponding subsample sizes restricted to the classifier-positive subpopulation. We assume that all sample and subsample sizes are fixed by design and the patient responses are independent normal with a common variance. Let $\widehat{\delta}_i$ and $\widehat{\delta}_i'$ be the estimates of the true dose effects δ_i and δ_i' $(i=1,2)$, in the general and classifier-positive population, respectively.

Further let s^2 denote the pooled estimate of the common variance based on $\nu=n_0+n_1+n_2-3$ d.f. Let $v_0=1/\sqrt{n_0}$, $v_0'=1/\sqrt{n_0'}$, $v_i=\sqrt{1/n_0+1/n_i}$ and $v_i'=\sqrt{1/n_0'+1/n_i'}$ $(i=1,2)$. Then the t-statistics for testing the null hypotheses H_i $(1\leqslant i\leqslant 4)$ are given by

$$t_1=\frac{\widehat{\delta}_1}{sv_1},\quad t_2=\frac{\widehat{\delta}_2}{sv_2},\quad t_3=\frac{\widehat{\delta}_1'}{sv_1'},\quad t_4=\frac{\widehat{\delta}_2'}{sv_2'}.$$

It is straightforward to show that, when all four null hypotheses are true, (t_1,t_2,t_3,t_4) has a central 4-variate t-distribution with ν d.f. and correlation matrix $R=\{\rho_{ij}\}$ whose elements are

$$\begin{aligned}\rho_{12}&=\frac{v_0^2}{v_1v_2},\quad \rho_{13}=\frac{v_1^2}{v_1v_1'},\quad \rho_{14}=\frac{v_0^2}{v_1v_2'},\\ \rho_{23}&=\frac{v_0^2}{v_2v_1'},\quad \rho_{24}=\frac{v_2^2}{v_1v_2'},\quad \rho_{34}=\frac{v_0'^2}{v_1'v_2'}.\end{aligned}\tag{10}$$

Thus all the correlations are known, being simple functions of the sample and subsample sizes.

The test statistics for testing $H(I_1)$ is t_i if $I_1=\{i\}$ for $i=1,2$ and $\max(t_1,t_2)$ if $I_1=\{1,2\}$. Similarly, the test statistics for testing $H(I_2)$ is t_j if $I_2=\{j\}$ for $j=3,4$ and $\max(t_3,t_4)$ if $I_2=\{3,4\}$. Denoting the c.d.f. of the univariate t-distribution with ν d.f. by $F_1(t|\nu)$ and that of $\max(t_i,t_j)$ under $H_i\cap H_j$ (where (t_i,t_j) has a central bivariate t-distribution with ν d.f. and correlation coefficient ρ_{ij}) by $F_2(t|\nu,\rho_{ij})$, the local p-values for $H(I_1)$ and $H(I_2)$ are given by

$$p_1(I_1)=\begin{cases}1-F_2(t_i|\nu,\rho_{12}) & \text{if } I_1=\{i\},\quad i=1,2,\\ 1-F_2(\max(t_1,t_2)|\nu,\rho_{12}) & \text{if } I_1=\{1,2\}\end{cases}$$

and

$$p_2(I_2)=\begin{cases}1-F_1(t_j|\nu) & \text{if } I_1=\{j\},\quad j=3,4,\\ 1-F_2(\max(t_3,t_4)|\nu,\rho_{34}) & \text{if } I_1=\{3,4\}.\end{cases}$$

The common critical value used to test all intersection hypotheses $H(I_1)$ is $d_1(\alpha)=t_2^*(\alpha|\nu,\rho_{12})=F_2^{-1}(1-\alpha|\nu,\rho_{12})$. The critical value used to test $H(I_2)$ depends on the index set I_2. If $I_2=\{j\}$ $(j=3,4)$ then $d_2(\alpha)=t_1^*(\alpha|\nu)=F_1^{-1}(1-\alpha|\nu)$ and if $I_2=\{3,4\}$ then $d_2(\alpha)=t_2^*(\alpha|\nu,\rho_{34})$.

As before, $p(I)=p_1(I_1)$ or $p_2(I_2)$ if $I=I_1$ or I_2, respectively. If the intersection hypothesis $H(I)$ contains both primary and secondary null hypotheses then we need to find $c(I_1,I_2|\alpha)$ in order to calculate $p(I)$. The coefficients $c(I_1,I_2|\alpha)$ are obtained by solving the following equations:

- If $I_1=\{i\}$, $i=1,2$, $I_2=\{j\}$, $j=3,4$ then

$$P\{t_i\geqslant t_2^*(\alpha|\nu,\rho_{12})\quad\text{or}\quad t_j\geqslant t_1^*(c(I_1,I_2|\alpha)\alpha|\nu)\}=\alpha.\tag{11}$$

- If $I_1=\{i\}$, $i=1,2$, $I_2=\{3,4\}$ then

$$P\{t_i\geqslant t_2^*(\alpha|\nu,\rho_{12})\quad\text{or}\quad \max(t_3,t_4)\geqslant t_2^*(c(I_1,I_2|\alpha)\alpha|\nu,\rho_{34})\}=\alpha.\tag{12}$$

We have numerically checked that $\alpha c(I_1,I_2|\alpha)$ is nondecreasing in α in the cases displayed above, which immediately implies that the α-consistency condition is met for this mixture procedure.

2.6. *Examples*

Example 1 (*Nonparametric and parametric mixture procedures for the schizophrenia trial*)

Assume that the sample size per dose group (placebo, low dose and high dose) is 300 patients and the size of the classifier-positive subpopulation is 100 patients per dose group. Further assume that the t-statistics for testing the null hypotheses of no treatment effect in the general population and classifier-positive subpopulation are given by $t_1=2.04$, $t_2=2.46$, $t_3=2.22$ and $t_4=2.66$ with 897 d.f. The raw one-sided p-values for the four null hypotheses computed from these t-statistics are $p_1=0.021$,

$p_2=0.007$, $p_3=0.013$ and $p_4=0.004$. The correlation coefficients of the 4-variate t-distribution can be calculated using (10) as follows:

$$\rho_{12}=\frac{1}{2},\quad \rho_{13}=\frac{1}{\sqrt{3}},\quad \rho_{14}=\frac{1}{2\sqrt{3}},\quad \rho_{23}=\frac{1}{2\sqrt{3}},\quad \rho_{24}=\frac{1}{\sqrt{3}},\quad \rho_{34}=\frac{1}{2}.$$

Beginning with the nonparametric mixture procedure introduced in Section 2.3, we first calculate the local $p(I)$-values for all intersection hypotheses by substitution in Table I. The $p(I)$-values are displayed in Table II. Recall that the adjusted p-values for the individual H_i's are set equal to the maximum of the $p(I)$-values for which $i\in I$ following the closure principle. The adjusted p-values for the four null hypotheses are listed in Table III. For example, the adjusted p-value for H_2 is given by

$$\widetilde{p}_2=\max(p(\{1,2,3,4\}),p(\{1,2,3\}),p(\{1,2,4\}),p(\{1,2\}),p(\{2,3,4\}),p(\{2,3\}),p(\{2,4\}),p(\{2\}))$$
$$=0.014.$$

Using the one-sided $\alpha=0.025$, the mixture procedure rejects the null hypothesis H_2 and thus establishes superiority of the high dose versus placebo in the general population. Note, however, that there is no evidence of efficacy at the low dose. As the primary family serves a parallel gatekeeper for the secondary family, we can pass the gatekeeper and proceed toward testing the null hypotheses of no treatment effect in the subpopulation, i.e. H_3 and H_4. The mixture procedure barely fails to reject H_3 but rejects H_4 and thus we conclude that the high dose is also significantly different from the placebo in the subpopulation of classifier-positive patients.

Next we turn to the parametric mixture procedure defined in Section 2.5. The $c(I_1,I_2)$-values are calculated from (11) and (12). As the α-consistency is satisfied in this case, the $p(I)$-values were

Table II. Local $p(I)$-values* for the nonparametric mixture procedure (mixture of the Bonferroni and Holm procedures based on the Bonferroni mixing function) and parametric mixture procedure (mixture of the single-step Dunnett and step-down Dunnett procedures based on the parametric mixing function) in the schizophrenia trial example with parallel gatekeeping restrictions.

Index set I	Nonparametric mixture procedure	Parametric mixture procedure
{1, 2, 3, 4}	0.014	0.013
{1, 2, 3}	0.014	0.013
{1, 2, 4}	0.014	0.013
{1, 2}	0.014	0.013
{1, 3, 4}	0.016	0.015
{1, 3}	0.027	0.025
{1, 4}	0.008	0.008
{1}	0.042	0.038
{2, 3, 4}	0.014	0.013
{2, 3}	0.014	0.013
{2, 4}	0.008	0.008
{2}	0.014	0.013
{3, 4}	0.008	0.008
{3}	0.013	0.013
{4}	0.004	0.004

*The raw one-sided p-values for the four null hypotheses are $p_1=0.021$, $p_2=0.007$, $p_3=0.013$, $p_4=0.004$.

Table III. Raw and adjusted p-values produced by the nonparametric mixture procedure (mixture of the Bonferroni and Holm procedures based on the Bonferroni mixing function) and parametric mixture procedure (mixture of the single-step Dunnett and step-down Dunnett procedures based on the parametric mixing function) in the schizophrenia trial example with parallel gatekeeping restrictions.

			Adjusted p-value	
Family	Null hypothesis	Raw p-value	Nonparametric mixture procedure	Parametric mixture procedure
F_1	H_1	0.021	0.042	0.038
	H_2	0.007	0.014*	0.013*
F_2	H_3	0.013	0.027	0.025*
	H_4	0.004	0.016*	0.015*

*Identifies the adjusted p-values that are significant at the one-sided 0.025 level.

computed using the method defined at the end of Section 2.1. These $p(I)$-values are shown in Table II. The resulting adjusted p-values (given in Table III) are no larger than those for the nonparametric mixture procedure. As in the nonparametric case, the mixture procedure passes the parallel gatekeeper due to a significant outcome for the null hypothesis H_2 and, in the secondary family, it barely rejects H_3 in addition to H_4. Thus, the parametric procedure is able to declare a significant difference also for the low dose in the subpopulation. In general, the parametric mixture procedure will be uniformly more powerful than the nonparametric mixture procedure due to the fact that it makes use of the joint distribution of the test statistics within each family and also across the families.

The local and adjusted p-values in this example were computed using R programs that can be downloaded from `www.multxpert.com`.

Example 2 (*A parametric mixture procedure in a trial with multiple primary and secondary endpoints*)

Consider a parallel-group clinical trial for evaluating the efficacy of a single dose of a treatment compared to a placebo with k_1 primary and k_2 secondary endpoints with $k=k_1+k_2$. F_1 and F_2 consist of the null hypotheses on the primary and secondary endpoints, respectively. The primary family serves as a parallel gatekeeper for the secondary family.

Assume that n_0 patients are enrolled in the placebo group and n_1 patients in the treatment group and let $v=\sqrt{1/n_0+1/n_1}$. Assume further that the responses in the trial are normally distributed and let $\widehat{\delta}_i$ denote the sample estimate of the true treatment difference δ_i on the ith endpoint. This estimate is distributed as $\mathrm{N}(\delta_i, v^2\sigma_i^2)$, where σ_i^2 is the variance of the individual observations on the ith endpoint. Let ρ_{ij} be the correlation coefficient between the ith and jth endpoints ($1\leqslant i<j\leqslant k$), which is also the correlation between $\widehat{\delta}_i$ and $\widehat{\delta}_j$. Denoting by s_i^2 the sample estimate of σ_i^2 with $\nu=n_0+n_1-2$ d.f., the test statistics for the ith endpoint is given by

$$t_i=\frac{\widehat{\delta}_i}{s_i v}(1\leqslant i\leqslant k).$$

To define a mixture procedure based on the parametric mixing function in this example, we need to define local p-values for all intersection hypotheses in the closed family associated with the k null hypotheses. To test the intersection hypotheses $H(I_1)$, $I_1\subseteq K_1$, and $H(I_2)$, $I_2\subseteq K_2$, we may use the union-intersection statistics

$$t_1(I_1)=\max_{i\in I_1} t_i, \quad t_2(I_2)=\max_{i\in I_2} t_i.$$

Note that under H_i the marginal distribution of t_i follows the t-distribution with ν d.f., but the joint distribution of the t_i's ($1\leqslant i\leqslant k$) is not multivariate t because different denominators s_i, which are correlated, are used for the t_i. This distribution may be called the generalized multivariate t, which was studied for the bivariate case by Siddiqui [29]. In general, this distribution is difficult to evaluate. The difficulty is further compounded by the fact that the correlations ρ_{ij} are unknown. Therefore the probability in (9) needs to be evaluated via resampling. We do not pursue further discussion of the details of resampling as it is not germane to the topic of this paper.

3. Mixture procedures for general logical restrictions

To motivate general logical restrictions among the null hypotheses consider the following example. A parallel-group clinical trial for an anti-hypertensive treatment is designed to evaluate the efficacy of two doses (Dose 1: low dose; Dose 2: high dose) versus an active control. Assume that the primary outcome variable in the trial is normally distributed and δ_1 and δ_2 denote the true differences between the two dose means and the active control mean. Each dose is tested for noninferiority with respect to the active control first and the superiority test is carried out after noninferiority is established. Thus the primary family F_1 consists of the noninferiority null hypotheses:

$$H_1:\delta_1\leqslant-\gamma, \quad H_2:\delta_2\leqslant-\gamma, \tag{13}$$

where $\gamma>0$ is a prespecified noninferiority margin, and the secondary family F_2 consists of the superiority null hypotheses:

$$H_3:\delta_1\leqslant 0, \quad H_4:\delta_2\leqslant 0. \tag{14}$$

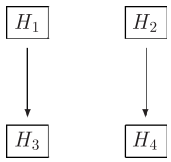

Figure 2. Decision tree in the hypertension trial example with general gatekeeping restrictions (the secondary null hypothesis H_3 is tested iff the corresponding primary null hypothesis H_1 is rejected and, similarly, H_4 is tested iff H_2 is rejected).

The decision rules in this multiple testing problem are displayed in Figure 2. Note that the null hypotheses in this problem do not have a simple serial or parallel logical relationship. For example, the superiority null hypothesis for the first dose, i.e. H_3 is tested only if H_1 is rejected. This is an example of tree-structured gatekeeping. We propose to handle such and more general logical relationships by using what we call as *restriction functions*.

3.1. Restriction functions

In this section we will introduce a general method for defining logical relationships among the primary and secondary null hypotheses which extends the class of tree-structured restrictions. As in Section 2, define two families, F_1 and F_2 with index sets K_1 and K_2 and let $K=K_1\cup K_2$. General logical restrictions mean that whether each secondary null hypothesis is *nontestable* (i.e. whether it is automatically accepted without testing) or *testable* (i.e. whether it is to be tested and either accepted or rejected) depends upon whether a specified condition on rejection of primary null hypotheses is met. (In the general case of $m>2$ families, the testability of each hypothesis in F_i for $i>1$ depends upon whether a specified condition on rejection of hypotheses in $F_1,\ldots,F_{i-1}$ is met.) Such a condition on rejection of primary null hypotheses can be specified by an indicator function, $L_j(I_1)$, that is defined on the collection of all subsets I_1 of K_1 and for each secondary null hypothesis H_j, $j\in K_2$. Assume that the primary null hypotheses H_i, $i\in I_1$, are accepted and H_i, $i\notin I_1$, are rejected. Then the restriction function $L_j(I_1)$ equals 0 if H_j is nontestable and $L_j(I_1)=1$ if H_j is testable.

We require that logical restrictions (and hence the corresponding restriction functions) satisfy the following natural and important conditions:

- Monotonicity condition: If a secondary null hypothesis is not testable given a set of accepted primary null hypotheses, it remains nontestable if more primary null hypotheses are accepted. In other words, for any $j\in K_2$, if $L_j(I_1)=0$ then $L_j(I_1')=0$ for $I_1\subseteq I_1'\subseteq K_1$.
- Parallel gatekeeping condition: All secondary null hypotheses are nontestable if all primary null hypotheses are accepted, i.e. $L_j(K_1)=0$ for all $j\in K_2$.

To demonstrate that the above method for defining logical relationships among the null hypotheses is more general than the tree gatekeeping method, consider the following example. Suppose $F_1=\{H_1,H_2,H_3\}$ and $F_2=\{H_4\}$ and H_4 is testable only if at least two of the three null hypotheses in F_1 are rejected. For instance, F_1 may refer to three primary endpoints and F_2 may refer to a secondary endpoint and the secondary endpoint will be tested only if the treatment shows a significant effect on at least two out of the three primary endpoints. This logical restriction cannot be modeled in the tree-structured gatekeeping framework; however, it is easily defined using the restriction functions introduced above. Specifically, the restriction function for H_4 is given by: $L_4(\{1,2\})=L_4(\{1,3\})=L_4(\{2,3\})=L_4(\{1,2,3\})=0$ and $L_4(\emptyset)=L_4(\{1\})=L_4(\{2\})=L_4(\{3\})=1$.

To define a mixture procedure for multiple testing problems with general logical restrictions, we will use a method very similar to that described in Section 2. Consider the closed family associated with the combined family F and let I be a nonempty subset of K such that $I=I_1\cup I_2$, where $I_1\subseteq K_1$ and $I_2\subseteq K_2$. Furthermore, let I_2^* be the restricted secondary index set of testable secondary null hypotheses from I_2 under the assumption that H_i, $i\in I_1$, are accepted. In other words, $I_2^*=\{j\in I_2\colon L_j(I_1)=1\}$. The mixture procedure $\mathscr{P}$ is defined as a closed procedure based on the following local tests of level α:

$$\text{Reject } H(I) \text{ if } \begin{cases} p_i(I_i)\leqslant\alpha & \text{if } I=I_i\ (i=1,2) \quad \text{or} \quad I_2^* \text{ is empty,} \\ \phi_I(p_1(I_1),p_2(I_2^*))\leqslant\alpha & \text{if } I=I_1\cup I_2, I_1 \quad \text{and} \quad I_2^* \text{ are nonempty.} \end{cases} \tag{15}$$

Table IV. Index sets, restriction functions and local p-values for the nonparametric mixture procedure (mixture of the Bonferroni and Holm procedures based on the Bonferroni mixing function) in the hypertension trial example with general gatekeeping restrictions.

Index set				Restriction function		
I	I_1	I_2	I_2^*	$L_3(I_1)^*$	$L_4(I_1)^*$	Local p-value ($p(I)$)
$\{1,2,3,4\}$	$\{1,2\}$	$\{3,4\}$	$\emptyset$	0	0	$2\min(p_1, p_2)$
$\{1,2,3\}$	$\{1,2\}$	$\{3\}$	$\emptyset$	0	—	$2\min(p_1, p_2)$
$\{1,2,4\}$	$\{1,2\}$	$\{4\}$	$\emptyset$	—	0	$2\min(p_1, p_2)$
$\{1,2\}$	$\{1,2\}$	$\emptyset$	$\emptyset$	—	—	$2\min(p_1, p_2)$
$\{1,3,4\}$	$\{1\}$	$\{3,4\}$	$\{4\}$	0	1	$2\min(p_1, p_4)$
$\{1,3\}$	$\{1\}$	$\{3\}$	$\emptyset$	0	—	$2p_1$
$\{1,4\}$	$\{1\}$	$\{4\}$	$\{4\}$	—	1	$2\min(p_1, p_4)$
$\{1\}$	$\{1\}$	$\emptyset$	$\emptyset$	—	—	$2p_1$
$\{2,3,4\}$	$\{2\}$	$\{3,4\}$	$\{3\}$	1	0	$2\min(p_2, p_3)$
$\{2,3\}$	$\{2\}$	$\{3\}$	$\{3\}$	1	—	$2\min(p_2, p_3)$
$\{2,4\}$	$\{2\}$	$\{4\}$	$\emptyset$	—	0	$2p_2$
$\{2\}$	$\{2\}$	$\emptyset$	$\emptyset$	—	—	$2p_2$
$\{3,4\}$	$\emptyset$	$\{3,4\}$	$\{3,4\}$	1	1	$2\min(p_3, p_4)$
$\{3\}$	$\emptyset$	$\{3\}$	$\{3\}$	1	—	p_3
$\{4\}$	$\emptyset$	$\{4\}$	$\{4\}$	—	1	p_4

*The restriction functions $L_3(I_1)$ and $L_4(I_1)$ are defined only for the intersection hypotheses containing H_3 and H_4, respectively.

Given the local tests, the local p-values for the intersection hypotheses in the closed family and adjusted p-values for the null hypotheses are computed using the algorithms defined in Section 2.1.

3.2. Nonparametric mixture procedure

To illustrate the process of applying logical restrictions, consider the hypertension trial example. In this example there are four null hypotheses. $F_1=\{H_1, H_2\}$ is the family of noninferiority null hypotheses and $F_2=\{H_3, H_4\}$ is the family of superiority null hypotheses. The 15 intersection hypotheses and the index sets I_1 and I_2 are listed in Table IV. This table also shows the restriction functions $L_3(I_1)$ and $L_4(I_1)$ and restricted secondary index sets I_2^*.

As in Section 2.3, consider a nonparametric mixture procedure based on the Bonferroni mixing function that uses the Bonferroni procedure as $\mathscr{P}_1$ and the Holm procedure as $\mathscr{P}_2$. Let p_i denote the raw p-value for testing the null hypothesis H_i ($1\leqslant i\leqslant 4$). The local p-values, $p_1(I_1)$ and $p_2(I_2)$ which equal $p(I)$ if $I=I_1$ or I_2, respectively, are the same as those given in Section 2.3; the $c(I_1)$-values are also the same as given there.

To calculate the $p(I)$-values when null hypotheses from both families are part of an intersection hypothesis, we need to specify the restriction functions. The first four intersection hypotheses in Table IV contain both primary null hypotheses and thus both $L_3(I_1)$ and $L_4(I_1)$ are equal to 0. This implies that I_2^*, the set of testable secondary null hypotheses, is empty. As a result, the local p-values for these intersection hypotheses are based only on the primary p-value, i.e. $p(I)=p_1(\{1,2\})=2\min(p_1, p_2)$ for all four $H(I)$. This definition of the local p-value ensures that the secondary null hypotheses cannot be tested if both primary null hypotheses are accepted.

Furthermore, consider the intersection hypotheses containing one primary null hypothesis and at least one secondary null hypothesis. In this case, if an intersection hypothesis includes H_1, we exclude H_3 from the secondary index set (i.e. set $L_3(I_1)=0$); similarly, if an intersection hypothesis includes H_2 then we exclude H_4. Consider, for example, the intersection hypothesis $H(I)$ with $I=\{1,3,4\}$. Even though $I_2=\{3,4\}$, the restricted secondary index set I_2^* is set to $\{4\}$. As $\mathscr{P}_1$ is the Bonferroni procedure, $c(\{1\})=1/2$ and the associated local p-value equals

$$p(I)=p(\{1,3,4\})=\phi_I(p_1(\{1\}), p_2(\{4\}))=\min(2p_1, p_4/(1/2))=2\min(p_1, p_4),$$

which does not depend on p_3. The local p-values calculated in this way for all intersection hypotheses are given in Table IV. The resulting decision rule is consistent with the logical restriction defined above, e.g. the mixture procedure cannot reject H_3 if H_1 is accepted.

3.3. Parametric mixture procedure

Given that the data are normally distributed and homoscedastic in the hypertension trial example, we can also construct a parametric mixture procedure which takes into account the joint distribution of the four test statistics. This mixture procedure is similar to the one introduced in Section 2.5, i.e. it is also based on the parametric mixing function and uses the single-step Dunnett procedure in the primary family and the step-down Dunnett procedure in the secondary family.

Let n_0, n_1 and n_2 denote the total sample sizes in the placebo and the two dose groups, respectively, and let $\widehat{\delta}_i$ denote the sample estimate of the true mean difference between the ith dose and the control. This sample estimate is normally distributed with variance $v_i^2\sigma^2$, where σ^2 is the variance of the individual observations and $v_i=\sqrt{1/n_0+1/n_i}$ $(i=1,2)$. If s^2 denotes the sample estimate of σ^2 with $\nu=n_0+n_1+n_2-3$ d.f., the test statistics for the noninferiority and superiority null hypotheses is given by

$$t_1=\frac{\widehat{\delta}_1+\gamma}{sv_1},\quad t_2=\frac{\widehat{\delta}_2+\gamma}{sv_2},\quad t_3=\frac{\widehat{\delta}_1}{sv_1},\quad t_4=\frac{\widehat{\delta}_2}{sv_2}.$$

Both (t_1,t_2) and (t_3,t_4) have bivariate t-distributions with ν d.f. and correlation coefficient

$$\rho=\sqrt{\frac{n_1n_2}{(n_0+n_1)(n_0+n_2)}}.$$

Furthermore, the pairs (t_1,t_3) and (t_2,t_4) have correlation 1 and the pairs (t_1,t_4) and (t_2,t_3) have correlation ρ. Thus this is a fully specified 4-variate t-distribution.

As $\mathscr{P}_1$ and $\mathscr{P}_2$ are the same as those in the schizophrenia trial example, the local p-values for the intersection hypotheses that contain only the primary or secondary null hypotheses are also the same as those given in Section 2.5. Next, consider the intersection hypotheses containing both primary and secondary null hypotheses. The computation of local p-values for these intersection hypotheses is based on the method introduced in Section 2.5 with minor modifications to account for the logical restrictions. For example, consider the intersection hypothesis $H(I)$ with $I=\{1,3,4\}$. In this case, $I_1=\{1\}$ and $I_2^*=\{4\}$. The local p-values for $H(I_1)$ and $H(I_2^*)$ are given by

$$p_1(I_1)=1-F_2(t_1|\nu,\rho),\quad p_2(I_2^*)=1-F_1(t_4|\nu).$$

Analogous to (11), the coefficient $c(I_1,I_2^*|\alpha)$ can be found from the equation:

$$P\{t_1\geqslant t_2^*(\alpha|\nu,\rho)\text{ or }t_4\geqslant t_1^*(c(I_1,I_2^*|\alpha)\alpha|\nu)\}=\alpha,$$

where t_1 and t_4 follow a bivariate t-distribution with ν d.f. and correlation coefficient ρ. The α-consistency is satisfied in this case and the local p-value for the selected intersection hypothesis is given by the smallest α for which

$$\min\left(p_1(I_1),\frac{p_2(I_2^*)}{c(I_1,I_2^*|\alpha)}\right)\leqslant\alpha.$$

3.4. Example 3 (*Nonparametric and parametric mixture procedures for the hypertension trial*)

Suppose that the hypertension trial is conducted with 200 patients on each of the three arms (active control, low dose and high dose). Assume that the t-statistics for the four null hypotheses are given by $t_1=1.90$, $t_2=2.26$, $t_3=1.87$ and $t_4=2.23$ with 597 d.f. The corresponding raw p-values equal $p_1=0.029$, $p_2=0.012$, $p_3=0.031$ and $p_4=0.013$. The correlation coefficient ρ equals 1/2.

The local $p(I)$-values for all intersection hypotheses for the nonparametric mixture procedure introduced in Section 3.2 are obtained by simple substitutions in Table IV. The results are listed in Table V from which the adjusted p-values for the four null hypotheses are obtained by the closure method as illustrated in Example 1 (Section 2.6). These adjusted p-values are listed in Table VI. If the one-sided α is set to 0.025 then the nonparametric mixture procedure rejects only H_2 in the primary family, i.e. if noninferiority is established at the high dose but cannot be established at the low dose. As H_1 is not rejected, the null hypothesis H_3 becomes non-testable due to the logical restrictions defined in Table IV (there is no sense in carrying out a superiority test if there is no evidence of noninferiority) and we focus on testing H_4 in the secondary family. However, the mixture procedure cannot reject this null hypothesis and thus superiority cannot be shown at either dose.

Table V. Local p-values* for the nonparametric mixture procedure (mixture of the Bonferroni and Holm procedures based on the Bonferroni mixing function) and parametric mixture procedure (mixture of the single-step Dunnett and step-down Dunnett procedures based on the parametric mixing function) in the hypertension trial example with general gatekeeping restrictions.

Index set I	Nonparametric mixture procedure	Parametric mixture procedure
{1, 2, 3, 4}	0.024	0.022
{1, 2, 3}	0.024	0.022
{1, 2, 4}	0.024	0.022
{1, 2}	0.024	0.022
{1, 3, 4}	0.026	0.025
{1, 3}	0.058	0.052
{1, 4}	0.026	0.025
{1}	0.058	0.052
{2, 3, 4}	0.024	0.022
{2, 3}	0.024	0.022
{2, 4}	0.024	0.022
{2}	0.024	0.022
{3, 4}	0.026	0.025
{3}	0.031	0.031
{4}	0.013	0.013

*The raw one-sided p-values for the four null hypotheses are $p_1 = 0.029$, $p_2 = 0.012$, $p_3 = 0.031$, $p_4 = 0.013$.

Table VI. Raw and adjusted p-values produced by the nonparametric mixture procedure (mixture of the Bonferroni and Holm procedures based on the Bonferroni mixing function) and parametric mixture procedure (mixture of the single-step Dunnett and step-down Dunnett procedures based on the parametric mixing function) in the hypertension trial example with general gatekeeping restrictions.

			Adjusted p-value	
Family	Null hypothesis	Raw p-value	Nonparametric mixture procedure	Parametric mixture procedure
F_1	H_1	0.029	0.058	0.052
	H_2	0.012	0.024*	0.023*
F_2	H_3	0.031	0.058	0.052
	H_4	0.013	0.026	0.025*

*Identifies the adjusted p-values that are significant at the one-sided 0.025 level.

Turning to the parametric mixture procedure from Section 3.3, the $c(I_1, I_2^*|\alpha)$-values are calculated in the same way as in Example 1. The $p(I)$-values for all intersection hypotheses are listed in Table V and the adjusted p-values for the four null hypotheses computed from these $p(I)$-values are given in Table VI. As in Example 1, the adjusted p-values generated by the parametric mixture procedure are smaller than those for the nonparametric mixture procedure, which again illustrates the higher power of the parametric approach. The parametric mixture procedure rejects only one null hypothesis in the primary family (H_2) and, given that H_3 is non-testable, proceeds to testing H_4 in the secondary family. This null hypothesis is rejected and thus the parametric mixture procedure establishes not only the noninferiority but also the superiority of the high dose. The local and adjusted p-values in the hypertension trial example were computed using R programs that can be downloaded from `www.multxpert.com`.

4. Concluding remarks

This paper introduces a unified approach to constructing multiple testing procedures for addressing multiplicity issues arising in clinical trials with multiple families of null hypotheses. The objectives considered in trials of thess kinds are often interrelated and we consider a very general framework of hierarchical objectives based on a new method for defining logical restrictions among individual null

hypotheses. In addition, this paper introduces a novel method for setting up gatekeeping procedures for the overall family of null hypotheses based on a mixture of procedures defined within each individual family. As shown in the paper, this mixture method serves as a versatile tool for constructing powerful gatekeeping procedures. This paper gives several examples of gatekeeping procedures, including parametric mixture procedures for trials with parallel gatekeeping restrictions (Examples 1 and 2 in Section 2) and mixture procedures for trials with general gatekeeping restrictions (Example 3 in Section 3). These procedures enable trial sponsors to enrich product labels by including the results of multiple clinically relevant analyses while controlling the familywise error rate in the strong sense.

Acknowledgements

Professor Ajit Tamhane's research was supported by NHLBI Award 1 RO1 HLO 82725-01A1. The authors are grateful to two referees for their insightful comments which helped improve this paper.

References

1. Dmitrienko A, Tamhane AC. Gatekeeping procedures with clinical trial applications. *Pharmaceutical Statistics* 2007; **6**:171–180.
2. Dmitrienko A, Tamhane AC. Gatekeeping procedures in clinical trials. In *Multiple Testing Problems in Pharmaceutical Statistics*, Dmitrienko A, Tamhane AC, Bretz F (eds), Chapter 5. Chapman & Hall/CRC Press: New York, 2009.
3. 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 Verlag: Stuttgart, 1995; **6**:3–18.
4. Bauer P, Röhmel J, Maurer W, Hothorn L. Testing strategies in multi-dose experiments including active control. *Statistics in Medicine* 1998; **17**:2133–2146.
5. Westfall PH, Krishen A. Optimally weighted, fixed-sequence, and gatekeeping multiple testing procedures. *Journal of Statistical Planning and Inference* 2001; **99**:25–40.
6. 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.
7. 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.
8. Dmitrienko A, Tamhane AC, Liu L, Wiens BL. A note on tree gatekeeping procedures in clinical trials. *Statistics in Medicine* 2008; **27**:3446–3451.
9. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
10. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* 2009; **28**:586–604.
11. Burman CF, Sonesson C, Guilbaud O. A recycling framework for the construction of Bonferroni-based multiple tests. *Statistics in Medicine* 2009; **28**:739–761.
12. Dmitrienko A, Tamhane AC, Wiens B. General multistage gatekeeping procedures. *Biometrical Journal* 2008; **50**:667--677.
13. Dmitrienko A, Tamhane AC, Wang X, Chen X. Stepwise gatekeeping procedures in clinical trial applications. *Biometrical Journal* 2006; **48**:984–991.
14. Guilbaud O. Bonferroni parallel gatekeeping—generalizations, adjusted p-values, and short direct proofs. *Biometrical Journal* 2007; **49**:917–927.
15. Everitt BS, Hand DJ. *Finite Mixture Distributions*. Chapman & Hall: New York, 1981.
16. 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.
17. Xu H, Nuamah I, Liu J, Lim P, Sampson A. A Dunnett–Bonferroni-based parallel gatekeeping procedure for dose–response clinical trials with multiple endpoints. *Pharmaceutical Statistics* 2009; **8**:301–316.
18. Hochberg Y, Tamhane AC. *Multiple Comparison Procedures*. Wiley: New York, 1987.
19. Dmitrienko A, Bretz F, Westfall PH, Troendle J, Wiens BL, Tamhane AC, Hsu JC. Multiple testing methodology. In *Multiple Testing Problems in Pharmaceutical Statistics*. Dmitrienko A, Tamhane AC, Bretz F (eds), Chapter 2. Chapman & Hall/CRC Press: New York, 2009.
20. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* 1955; **50**:1096–1121.
21. Holm S. A simple sequentially rejective multiple test procedure. *Scandanavian Journal of Statistics* 1979; **6**:65–70.
22. Hochberg Y. A sharper Bonferroni procedure for multiple significance testing. *Biometrika* 1988; **75**:800–802.
23. Wiens B. A fixed-sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics* 2003; **2**:211–215.
24. Wiens B, Dmitrienko A. The fallback procedure for evaluating a single family of hypotheses. *Journal of Biopharmaceutical Statistics* 2005; **15**:929–942.
25. Dunnett CW, Tamhane AC. A step-up multiple test procedure. *Journal of the American Statistical Association* 1992; **87**:162–170.

26. Naik UD. Some selection rules for comparing p processes to a standard. *Communications in Statistics. Series A* 1975; **4**:519–535.
27. Roth AJ. Multiple comparison procedures for discrete test statistics. *Journal of Statistical Planning and Inference* 1999; **82**:101–117.
28. Lehmann EL, Romano JP. Generalizations of the familywise error rate. *Annals of Statistics* 2005; **33**:1138–1154.
29. Siddiqui MM. A bivariate t-distribution. *Annals of Mathematical Statistics* 1967; **38**:162–166.