PHARMACEUTICAL STATISTICS
Pharmaceut. Statist. 2009; **8**: 301–316
Published online 8 December 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/pst.358

A Dunnett–Bonferroni-based parallel gatekeeping procedure for dose–response clinical trials with multiple endpoints

MAIN PAPER

Haiyan Xu[1,*,†], Isaac Nuamah[1], Jingyi Liu[2], Pilar Lim[1], and Allan Sampson[3]

[1]*Clinical Biostatistics, Johnson & Johnson Pharmaceutical Research & Development, L.L.C., Titusville, NJ, USA*
[2]*Department of Statistics, University of California, Davis, CA, USA*
[3]*Department of Statistics, University of Pittsburgh, Pittsburgh, PA, USA*

This paper proposes a Dunnett–Bonferroni-based parallel gatekeeping procedure in a setting of dose–response clinical trials with multiple endpoints. It follows the Dunnett-based parallel gatekeeping strategy of Dmitrienko et al. (Pharm. Stat. 2006; 5:19–28), but differs in the calculation of the critical values. The implementation of the Dunnett-based parallel gatekeeping procedure relies on assumptions difficult to justify in typical clinical trials, namely (a) that the joint distribution of the test statistics from different endpoints can be approximated by a multivariate-t distribution and (b) that the true correlation between multiple endpoints can be well estimated using observed data. The proposed Dunnett–Bonferroni-based parallel gatekeeping procedure relaxes the preceding assumptions by splitting type I error rate among families using the Bonferroni inequality. While it is potentially less powerful than a Dunnett-based procedure when both procedures are applicable, the power loss is very minimal. Our proposed method avoids assumptions that might be challenged by regulatory agencies and does so with virtually no cost. Moreover, in most cases this method is easier to implement compared with the Dunnett-based procedure. Copyright © 2008 John Wiley & Sons, Ltd.

Keywords: ***Dunnett test; Bonferroni inequality; parallel gatekeeping procedures; dose–response; clinical trials; multiple endpoints***

*Correspondence to: Haiyan Xu, Clinical Biostatistics, E22504, Johnson & Johnson Pharmaceutical Research & Development, L.L.C., 1125 Trenton-Harbourton Road, Titusville, NJ 08560, USA.
[†]E-mail: hxu22@its.jnj.com

Pharmaceutical
STATISTICS

1. INTRODUCTION

The problem of multiplicity adjustment arises commonly in clinical studies. Strong control of the study-wise false-positive rate is often required in many multiple testing scenarios. For example, in a clinical study with multiple dose-to-control comparisons and multiple endpoints, it is becoming more common to control the overall false-positive rate across both endpoints and dose-to-control comparisons.

Gatekeeping testing procedures are increasingly implemented in clinical studies with hierarchically ordered multiple objectives. An example of clinical studies with hierarchically ordered objectives is a dose–response study with multiple endpoints. Gatekeeping procedures assure that the less important goals (e.g. secondary endpoints) are not studied until the more important goals (e.g. primary endpoint) have been reached. This is achieved by testing multiple families of analyses sequentially and each family serves as a gatekeeper for subsequent families. As pointed out by Dmitrienko and Tamhane [1], multiple families of analyses are often related to multiple endpoints but can also represent dose-to-control comparisons, non-inferiority and superiority tests or inferences at several time points in a longitudinal study. In this paper, we focus on the problem of dose-to-control comparisons with multiple endpoints.

Two basic types of gatekeeping procedures have been studied in the literature: serial gatekeeping (see [1–4]) and parallel gatekeeping (see [1,5–7]). More recent tree-structured gatekeeping procedures [8] are a generalization of serial and parallel gatekeeping procedures. Our focus is on parallel gatekeeping procedures in this paper.

Most of the recent developments of parallel gatekeeping procedures do not account for the joint distribution of test statistics. Dmitrienko *et al.* [7,9] proposed a Dunnett-based parallel gatekeeping method in a dose–response clinical study setting, which does take into account distributional assumptions concerning test statistics.

The Dunnett-based parallel gatekeeping procedure relies on some assumptions that could be difficult to justify in typical clinical trials, namely (a) that the joint distribution of the test statistics from different endpoints can be approximated by a multivariate-t distribution and (b) that the true correlation between multiple endpoints can be well estimated using observed data. In this paper, we introduce a Dunnett–Bonferroni-based parallel gatekeeping procedure in a dose–response clinical study setting, which relaxes the preceding assumptions by splitting type I error rate α among families using the Bonferroni inequality. The α-splitting/Bonferronization is applied in testing all intersection hypotheses in a closed testing [10] framework. While it is potentially less powerful than a Dunnett-based parallel gatekeeping procedure when both procedures are applicable, we show that the power loss is very minimal. Moreover, in most cases the newly proposed method is easier to implement compared with the Dunnett-based procedure.

Section 2 provides the background about parallel gatekeeping procedures and Section 3 gives an example of the dose–response study with multiple endpoints. Section 4 gives an overview of the Dunnett-based parallel gatekeeping procedure. Section 5 introduces our proposed Dunnett–Bonferroni-based parallel gatekeeping procedure. A clinical application of the Dunnett–Bonferroni-based procedure is illustrated in Section 6. Different parallel gatekeeping procedures are compared in Section 7.

2. NOTATION AND BACKGROUND ABOUT PARALLEL GATEKEEPING PROCEDURES

A gatekeeping approach usually groups null hypotheses into families with hierarchical orders. Each family becomes a gatekeeper for subsequent families. For example, assume that there are s endpoints with m active doses to be compared with placebo in a dose–response study. Then for each endpoint, there are m null hypotheses stating that the treatment effect of an active dose is not superior to placebo: $H_1^i, H_2^i, \ldots, H_m^i$, where

DOI: 10.1002/pst

$i = 1, 2, \ldots, s$. The m null hypotheses corresponding to each endpoint i form a family of hypotheses. Assume that these s families $F_1, F_2, \ldots, F_s$ are hierarchically ordered with descending importance; then according to the gatekeeping strategy, F_i serves as a gatekeeper for $F_{i'}$, for $i < i'$.

A parallel gatekeeping procedure, firstly introduced by Dmitrienko *et al*. [5], requires at least one hypothesis in a gatekeeper family to be declared significant in order to proceed to its subsequent families. The originally proposed parallel gatekeeping procedure [5] (referred to as *the original parallel gatekeeping procedure* hereafter) was formulated based on the closed testing principle [10]. A testing procedure constructed based on the closed testing principle tests all possible non-empty intersections of null hypotheses. An individual null hypothesis can be rejected only when all the intersection hypotheses containing it are rejected.

A common condition to be followed when applying a parallel gatekeeping procedure is the independence condition [11, Section 2.7]. The independence condition means that the inference made in the primary family should be independent of, or in other words, not affected by, the inference made in the subsequent families. It is worth noting that relaxing the independence condition may result in some power gain when doing so is in line with the objectives of clinical studies (see [6; 11, Section 2.7.3; 12]).

There have been many recent developments of the original parallel gatekeeping procedure. The original parallel gatekeeping procedure was based on a weighted Bonferroni adjustment. A straightforward extension of it was to replace the Bonferroni test in the first $m-1$ stages with a more powerful test: truncated Holm test [1; 11, Section 2.7.5], which is based on a convex combination of the Bonferroni and Holm tests. It has also been proposed to replace the weighted Bonferroni test with another more powerful test: Simes test [5,13]. The Simes test controls type I error if some distributional assumptions are met, for example, positive regression dependence or independence [14]. Chen *et al*. [6] further enhanced this Simes-test-based procedure so that its weighting scheme can take care of special logical relationships between individual primary and secondary/tertiary tests.

The recent developments of parallel gatekeeping procedures mainly include extensions to the original Bonferroni-based parallel gatekeeping procedure, which does not account for the joint distribution of test statistics. Dmitrienko *et al*. [7,9] proposed a Dunnett-based parallel gatekeeping procedure in a dose–response clinical study setting. This procedure was developed in a closed testing [10] framework similar to the original parallel gatekeeping procedure [5]. Unlike the original procedure and its extensions, it takes into account the joint distribution of test statistics by carrying out multiplicity adjustment through Dunnett's method. In addition to the independence condition [11, Section 2.7], which states that the inference made in the primary endpoint should not be affected by the inference made in the subsequent endpoints, it also follows a special logical relationship between endpoints, that is, the efficacy for a dose level in the subsequent endpoints cannot be claimed unless the efficacy for that dose level was supported in the primary endpoint analysis.

3. CLINICAL STUDY WITH THREE DOSE-TO-CONTROL COMPARISONS AND TWO ENDPOINTS

In this section, we consider a dose–response study with three treatment-to-placebo comparisons and two endpoints. This study setting will be used to illustrate the Dunnett-based and Dunnett–Bonferroni-based parallel gatekeeping procedures in Sections 4 and 5.

Suppose that the primary and secondary endpoints are indexed by a superscript $i = P, S$, respectively, and the treatment groups (placebo, low dose, middle dose and high dose) are indexed by $j = 0, 1, 2, 3$. Let μ_j^i denote the mean response of the jth treatment group for the ith endpoint, and let n_j^i denote the number of observations for the ith endpoint and the jth treatment group. Let $Y_{j1}^i, \ldots, Y_{jn_j^i}^i$ be the measurements of the ith endpoint for the jth treatment

DOI: 10.1002/pst

Pharmaceutical STATISTICS

group. Then we suppose that, for $i = P, S$, $j = 0, 1, 2, 3$ and $k = 1, \ldots, n_j^i$,

$$\begin{pmatrix} Y_{jk}^P \\ Y_{jk}^S \end{pmatrix} \overset{\text{i.i.d.}}{\sim} N\left(\begin{pmatrix} \mu_j^P \\ \mu_j^S \end{pmatrix}, \begin{pmatrix} (\sigma^P)^2 & \rho\sigma^P\sigma^S \\ \rho\sigma^P\sigma^S & (\sigma^S)^2 \end{pmatrix} \right) \quad (1)$$

Furthermore, measurements of subjects from different treatment groups are independent.

Define $\theta_j^i = \mu_j^i - \mu_0^i$ to be the difference between the means of the jth treatment group and the placebo for the ith endpoint with $i = P, S$ and $j = 1, 2, 3$. Let H_1^P, H_2^P and H_3^P denote the null hypotheses that low, middle and high doses are not superior to placebo for the primary endpoint. Let H_1^S, H_2^S and H_3^S denote the corresponding null hypotheses for the secondary endpoint. We then have the following six original null hypotheses:

$$H_1^P : \theta_1^P \le 0, \quad H_2^P : \theta_2^P \le 0, \quad H_3^P : \theta_3^P \le 0$$

$$H_1^S : \theta_1^S \le 0, \quad H_2^S : \theta_2^S \le 0, \quad H_3^S : \theta_3^S \le 0$$

We refer to H_1^P, H_2^P and H_3^P as the primary (family of) hypotheses, and H_1^S, H_2^S and H_3^S as the secondary (family of) hypotheses.

We also assume that the test statistics T_j^i corresponding to each individual hypothesis H_j^i is

$$T_j^i = \frac{\bar{Y}_{j.}^i - \bar{Y}_{0.}^i}{\hat{\sigma}^i \sqrt{1/n_0^i + 1/n_j^i}}, \quad i = P, S \text{ and } j = 1, 2, 3 \quad (2)$$

where $\bar{Y}_{j.}^i = (1/n_j^i) \sum_k Y_{jk}^i$,

$$\hat{\sigma}^i = \sqrt{\frac{\sum_j \sum_k (Y_{jk}^i - \bar{Y}_{j.}^i)^2}{\sum_j (n_j^i - 1)}}$$
$$i = P, S, \ \ j = 0, 1, 2, 3, \ \ k = 1, \ldots, n_j^i \quad (3)$$

For the purpose of simplicity, we assume that n_j^i, the number of observations for the jth dose level and the ith endpoint, is equal for any combination of i and j, and we denote it as n. Let N denote the error degrees of freedom (df) related to the estimation of σ^P and σ^S; thus N is equal to $\sum_{j=0,1,2,3} (n_j^i - 1) = \sum_{j=0,1,2,3} (n - 1) = 4*(n - 1)$.

4. DUNNETT-BASED PARALLEL GATEKEEPING PROCEDURE AND ITS REQUIRED ASSUMPTIONS

In this section, we illustrate the Dunnett-based parallel gatekeeping procedure [7] using the Section 3 example and explain the assumptions required for implementing this procedure.

4.1. Dunnett-based parallel gatekeeping procedure

A parallel gatekeeping procedure is formulated as a closed testing procedure [5,7,10]; therefore, we consider $2^6 - 1 = 63$ intersection hypotheses from the 6 original null hypotheses. We adopt the binary representation of the intersection hypotheses from [5,7]. For example, $H_{100011} = H_1^P \cap H_2^S \cap H_3^S$, $H_{100001} = H_1^P \cap H_3^S$ and $H_{100000} = H_1^P$. An original null hypothesis H_j^i can be rejected only if all the intersection hypotheses containing H_j^i are rejected.

There are three rules for setting up tests for intersection hypotheses H as explained by Dmitrienko *et al.* [7]:

- If H includes all three primary hypotheses H_1^P, H_2^P and H_3^P, the decision rule for H should not depend on T_1^S, T_2^S or T_3^S. This is to ensure that a secondary hypothesis cannot be rejected unless at least one primary hypothesis was rejected.
- The same critical values should be used for testing the three primary hypotheses. This is to ensure that the inference made in the primary family is not affected by the inference made in the secondary family.
- If H includes a primary hypothesis H_j^P and its corresponding secondary hypothesis H_j^S, the decision rule for H should not depend on T_j^S, the test statistic for H_j^S. This is to ensure that H_j^S cannot be rejected unless H_j^P was rejected.

Table AI (Appendix A) presents a set of decision rules for testing H_1^P, H_2^P, H_3^P, and H_1^S, H_2^S, H_3^S, when n_j^i is equal for any combination of i and j. This set of decision rules follows the above three rules and involves the critical values $c_1, c_2, \ldots, c_7$.

DOI: 10.1002/pst

The calculation of c_1 involves test statistics from the primary family only, and the calculation of c_5, c_6 and c_7 involves test statistics from the secondary family only. Thus, they can be easily obtained through either a Dunnett-t distribution or a univariate-t distribution [7]. The calculation of c_2, c_3 and c_4 involves test statistics from different families. According to Dmitrienko *et al.* [7], under the global null hypothesis, the joint distribution of the six test statistics T_j^i ($i = P, S, j = 1, 2, 3$) follows a central multivariate-t distribution, and the critical values c_2, c_3 and c_4 can be obtained by applying the Genz and Bretz method [15].

4.2. Assumptions required in implementing the Dunnett-based procedure

In the Dunnett-based parallel gatekeeping procedure [7], the critical values c_2, c_3 and c_4 were obtained using the Genz and Bretz method [15], which is a method for finding quantiles of the central multivariate-t distribution when the correlation structure for the T-statistics is *known*. Hence, the accuracy of the critical values found through this method or similar algorithms relies on an assumption that the true correlation between multiple endpoints can be well estimated using the observed data. Considering that the strong control of familywise error rate (FWER) is commonly defined as the supremum of type I error made across the whole parameter space, the use of either a point estimate or a 95% lower bound of the point estimate of correlation for finding the critical values can be easily challenged by regulatory agencies regarding the control of FWER. Others have considered less restrictive approaches to define the FWER. For example, Pollard and van der Laan [16], in a bioinformatics setting, considered a modified FWER for which the nuisance parameters are fixed and equal to their true values. Dmitrienko *et al.*'s method [7] controls this modified FWER when the true parameter values are estimated and the multivariate-t critical values are accurately obtained.

Furthermore, the Genz and Bretz method [15] deals with a specific form of the multivariate-t distribution, which requires the denominators of each T-statistic to be identical random variables or, equivalently, requires the estimators of σ^P and σ^S to be the same. For a typical dose–response clinical study with multiple endpoints, the assumption of $\hat{\sigma}^P = \hat{\sigma}^S$ does not hold, since the estimators of σ^P and σ^S are based on data from different endpoints. When the assumption of $\hat{\sigma}^P = \hat{\sigma}^S$ does not hold, Liu *et al.* [17] actually showed through simulation based on the correct generalized multivariate-t distribution (see [17, Table VI]) that, in terms of type I error control, it could be liberal to apply the Genz and Bretz algorithm [15] for finding the critical values in the Dunnett-based procedure.

In summary, the Dunnett-based procedure could inflate type I error because of (1) using estimated correlation to find the critical values and (2) using the Genz and Bretz method [15] for finding the critical values in a way that is not intended to be used, such as in the case of unequal variance estimates.

To avoid making assumptions on the correlation structure between endpoints or assuming $\hat{\sigma}^P = \hat{\sigma}^S$, Liu *et al.* [17] proposed to assume the independence of test statistics from different families for finding the critical values. They showed that, when the correlation between two families is *non-negative*, assuming the independence of test statistics from two families is conservative in terms of type I error control. In mathematical terms,

$$\Pr_{\rho \geq 0}\left\{\left(\max_{k\in K} T_k^P \leq c_1\right) \cap \left(\max_{l\in L} T_l^S \leq c^S\right)\right\} \geq \Pr\left\{\max_{k\in K} T_k^P \leq c_1\right\} \Pr\left\{\max_{l\in L} T_l^S \leq c^S\right\}$$

where K is the index set for test statistics from the primary family and L is the index set for test statistics from the secondary family. Note that, to be conservative in controlling type I error, this method requires that $\rho \geq 0$, which in some ways is a more relaxed assumption than a known ρ as required in the Genz and Bretz method [15].

The reliance on the assumptions as mentioned above could be challenged by regulatory agencies, which is our primary motivation in proposing the

DOI: 10.1002/pst

Dunnett–Bonferroni-based parallel gatekeeping procedure.

5. DUNNETT–BONFERRONI-BASED PARALLEL GATEKEEPING PROCEDURE

5.1. Methodology

An intuitive solution for finding the critical values for decision rules that involve test statistics from different families while avoiding hard-to-justify assumptions is that, for any such decision rule, the Bonferroni inequality can be adopted to 'split' α among families, without worrying about the joint distribution of test statistics from different families.

Our proposed Dunnett–Bonferroni-based parallel gatekeeping procedure follows the gatekeeping strategy of the Dunnett-based procedure [7], but differs in the way the critical values are calculated. Therefore, its decision rules have the same form as that of the decision rules for the Dunnett-based procedure (as shown in Table AI), while having different values of c.

There are three basic steps for finding the critical values in our Dunnett–Bonferroni-based procedure. Assume that $K \subseteq \{1, 2, \ldots, m\}$ is the index set for test statistics from the primary family F_1 involved in a decision rule and $L \subseteq \{1, 2, \ldots, m\}$ is the index set for test statistics from the secondary family F_2 involved in a decision rule. For example, a decision rule $T_1^P > c_1$ or $T_2^P > c_1$ or $T_3^S > c_2$ has $K = \{1, 2\}$ and $L = \{3\}$. K and L must be disjoint to ensure that H_j^S cannot be rejected unless H_j^P was rejected, for $j = 1, 2, \ldots, m$, according to the third rule of setting up intersection hypotheses tests as described in Section 4.1. Let $\Theta_0 = \{\theta_j^P \leq 0, \theta_j^S \leq 0, j = 1, 2, \ldots, m, -1 \leq \rho \leq 1\}$. The three steps of finding the critical values for each decision rule are:

(1) The critical value c_1 corresponding to test statistics from F_1 is calculated so that the probability equality $\sup_{\theta \in \Theta_0} \Pr_\theta \{\max_{k \in \{1,2,\ldots,m\}} T_k^P > c_1\} = \alpha$ holds. Here c_1 is fixed for all test statistics from F_1, regardless of how many test statistics from F_1 or F_2 are involved in a decision rule. This ensures the independence condition [11, Section 2.7] so that the inference made in F_1 is not affected by the inference made in F_2.

(2) The type I error spent in testing the primary family F_1 hypotheses corresponding to the index set K is then given by $\alpha' = \sup_{\theta \in \Theta_0} \Pr_\theta \{\max_{k \in K} T_k^P > c_1\}$ so that α' is the chance of making any false positive when testing these hypotheses from F_1 and $\alpha' \leq \alpha$.

(3) Based on the α' for the tests about the index set K, we calculate the critical value c^S corresponding to the hypotheses defined by the index set L from F_2 so that the following holds: $\sup_{\theta \in \Theta_0} \Pr_\theta \{\max_{l \in L} T_l^S > c^S\} = \alpha - \alpha'$.

Since the Dunnett–Bonferroni-based parallel gatekeeping method is constructed as a closed testing procedure [10], it controls type I error strongly as long as type I error is strongly controlled in each intersection hypothesis. With the critical values obtained from the above three steps, the Dunnett–Bonferroni-based procedure controls type I error rate at level α for any intersection hypothesis. The reason is as follows. For each intersection hypothesis:

$$\begin{aligned}
&\sup_{\theta \in \Theta_0} \Pr_\theta\{\text{Make any false rejection}\} \\
&= \sup_{\theta \in \Theta_0} \Pr_\theta\left\{\left(\max_{k \in K} T_k^P > c_1\right) \text{ or } \left(\max_{l \in L} T_l^S > c^S\right)\right\} \\
&\leq \sup_{\theta \in \Theta_0} \Pr_\theta\left\{\max_{k \in K} T_k^P > c_1\right\} + \sup_{\theta \in \Theta_0} \Pr_\theta\left\{\max_{l \in L} T_l^S > c^S\right\} \\
&= \alpha' + (\alpha - \alpha') \\
&= \alpha
\end{aligned} \tag{4}$$

The inequality in (4) holds from the Bonferroni inequality.

It is worthy to note that our Dunnett–Bonferroni-based procedure is equivalent to the Dunnett-based procedure [7], when the correlation ρ between test statistics from two families is equal to -1 (see proof in Appendix B). If one applies the Dunnett-based procedure [7] and truly wants to control the type I error by taking the supremum over the nuisance parameter space, rather than

DOI: 10.1002/pst

just estimating ρ, Slepian's inequality (see [18, Appendix A.3]) shows that the maximum type I error for the Dunnett-based procedure occurs at $\rho = -1$. Thus, in order to strongly control type I error by taking the supremum over the parameter space, the Dunnett-based procedure is indeed reduced to the Dunnett–Bonferroni-based procedure.

5.2. Illustration

In this subsection, we use the dose–response clinical study example in Section 3 to illustrate how to obtain the critical values for our proposed Dunnett–Bonferroni-based procedure.

Case 1: The critical value c_1 for any hypothesis from F_1 can be obtained from a Dunnett-t distribution with 3 and N df so that $\Pr\{\max(T_1^P, T_2^P, T_3^P) > c_1\} = \alpha$. This is the same as the way of calculating c_1 for the Dunnett-based parallel gatekeeping procedure (Section 4.1).

Case 2: For a decision rule involving test statistics from both F_1 and F_2, it is easy to see how to compute the critical values c_2, c_3 and c_4 by considering examples as listed below.

(1) For a decision rule involving two test statistics from F_1 and one from F_2, for example, the decision rule for H_{110001}, α' is first calculated for F_1 as $\alpha' = \Pr\{T_1^P > c_1 \text{ or } T_2^P > c_1\}$, where (T_1^P, T_2^P) follows a Dunnett-t distribution with 2 and N df. Here c_2 can then be calculated so that $\Pr\{T_3^S > c_2\} = \alpha - \alpha'$, where T_3^S follows a univariate-t distribution with N df.

(2) For a decision rule involving one test statistic from F_1 and two from F_2, for example, the decision rule for H_{100011}, $\alpha' = \Pr\{T_1^P > c_1\}$, where T_1^P follows a univariate-t distribution with N df. c_3 can then be calculated so that $\Pr\{T_2^S > c_3 \text{ or } T_3^S > c_3\} = \alpha - \alpha'$, where (T_2^S, T_3^S) follows a Dunnett-t distribution with 2 and N df.

(3) For a decision rule involving only two test statistics, one from F_1 and the other from F_2, for example, the decision rule for H_{100110}, c_4 can be obtained so that $\Pr\{T_2^S > c_4\} = \alpha - \alpha'$, where $\alpha' = \Pr\{T_1^P > c_1\}$. Both T_1^P and T_2^S follow a univariate-t distribution with N df.

Case 3: For decision rules not involving test statistics from F_1, the critical values can be easily obtained through either a Dunnett-t or a univariate-t distribution, which is the same as the way of calculating c_5, c_6 and c_7 for the Dunnett-based parallel gatekeeping procedure (Section 4.1).

One extreme scenario involves Case 1 when $\alpha' = \alpha$. In this case, all the hypotheses from F_1 are tested and no testing on hypotheses from F_2 will be conducted. In contrast, when $\alpha' = 0$ (Case 3), the hypotheses testing in F_2 is carried out at a full α level. In the middle ground, if some null hypotheses from F_1 are tested (Case 2), one has to pay the price for performing multiple tests in F_1, which makes it more difficult to reject hypotheses in F_2. In addition, since c_1 is fixed for all T_j^P's (test statistics from F_1), the more the T_j^P's involved in a decision rule, the smaller a 'split' of the pre-specified type I error rate α that can be passed to F_2. As a result, generally $c_2 \geq c_3$ and $c_3 \geq c_4$.

5.3. Unbalanced case

Until now, we have only discussed the implementation of our Dunnett–Bonferroni-based parallel gatekeeping procedure when the number of observations is assumed to be the same for any dose level or endpoint. In practice, there could be different numbers of observations across endpoints for the same dose level or there could be different numbers of observations across dose levels for the same endpoint. In the unbalanced case, the values of c_2, c_3 and c_6 (as in Table AI) may differ depending on which intersection hypothesis H they correspond to.

For example, for the intersection hypotheses $H_{110101}, H_{110011}, H_{110001}$ and H_{110111}, c_2 is calculated so that $\Pr\{T_3^S > c_2\} = \alpha - \alpha'$, where $\alpha' = \Pr\{T_1^P > c_1 \text{ or } T_2^P > c_1\}$, whereas for the intersection hypotheses $H_{101111}, H_{101110}, H_{101011}$ and H_{101010}, c_2 is calculated so that $\Pr\{T_2^S > c_2\} = \alpha - \alpha'$, where $\alpha' = \Pr\{T_1^P > c_1 \text{ or } T_3^P > c_1\}$. Although (T_1^P, T_2^P) and (T_1^P, T_3^P) both follow a Dunnett-t distribution with the same df N^P (the error df corresponding to

DOI: 10.1002/pst

estimating σ^P), if n_2^P and n_3^P are different, $\Pr\{T_1^P > c_1$ or $T_2^P > c_1\}$ would be different from $\Pr\{T_1^P > c_1$ or $T_3^P > c_1\}$. The reason for this difference is that in calculating the probability of a Dunnett-t distribution using SAS [19], the PROBMC function involves not only the quantile c_1 but also the λ parameters that involve n_2^P and n_3^P. Hence, c_2 could be different depending on which intersection hypothesis H it corresponds to. The similar reasoning applies to c_3 and c_6 so that they could also be different depending on which H they correspond to.

For the purpose of simplicity, one can take the maximum of different values of c_2, c_3 and c_6 to reduce the number of critical values. In Section 6, we illustrate the implementation of the Dunnett–-Bonferroni-based procedure in an unbalanced case using a clinical study example. The Dunnett–Bonferroni-based procedure can easily be made to handle unbalanced cases.

6. APPLICATION OF THE DUNNETT–BONFERRONI-BASED PROCEDURE TO A CLINICAL STUDY

As an illustration for applying the Dunnett–Bonferroni-based procedure, we consider a placebo-controlled, parallel group clinical study that studies the efficacy of three active doses of a study drug. This example is based on an actual clinical trial, but for proprietary reasons, we cannot describe the purpose of the study and the actual names of the primary and secondary endpoints. The data that we present are a subset of our actual data. We call the primary endpoint the Behavior Rating Scale (BRS) total score and call the secondary endpoint the Overall Functioning (OF) score. The following sequential families of analyses are considered:

Family F_1: Each active dose is compared with the placebo in the primary endpoint (the change in the BRS total score). An analysis of variance (ANOVA) model with treatment groups as a factor is used. The Dunnett method is applied to adjust for multiplicity across multiple dose-to-control comparisons.

Family F_2: Each active dose is compared with the placebo in the secondary endpoint (the change in the OF score). An ANOVA model with treatment groups as a factor is used. A dose can be claimed superior to the placebo in the OF improvement only when this dose is claimed significantly superior to the placebo in the BRS improvement.

The summary statistics are given in Table I. The error df in the ANOVA model for the increase (improvement) in the BRS total score is equal to $df_1 = 153$, and the error df in the ANOVA model for the increase (improvement) in the OF score is equal to $df_2 = 151$. Assume that one-sided type I error level is $\alpha = 0.025$.

The critical values of our Dunnett–Bonferroni parallel gatekeeping procedure are given in Table II. Because of the unequal number of observations, some intersection hypotheses no longer share the same critical values as they do

Table I. Summary statistics for three active doses and placebo in two endpoints.

		Placebo	Low dose	Middle dose	High dose
BRS (primary)	Sample size	$n_0^P = 33$	$n_1^P = 39$	$n_2^P = 44$	$n_3^P = 41$
	Mean	16.3030	19.7436	20.3864	21.7073
	SE*	1.3894	1.2781	1.2033	1.2465
	T-statistic		$T_1^P = 1.8225$	$T_2^P = 2.2216$	$T_3^P = 2.8952$
OF (secondary)	Sample size	$n_0^S = 33$	$n_1^S = 38$	$n_2^S = 43$	$n_3^S = 41$
	Mean	16.4242	20.7368	25.0000	26.0976
	SE	1.7747	1.6539	1.5547	1.5922
	T-statistic		$T_1^S = 1.7777$	$T_2^S = 3.6347$	$T_3^S = 4.0571$

*Standard error.

DOI: 10.1002/pst

Table II. The critical values of the proposed Dunnett–Bonferroni-based parallel gatekeeping procedure ($n_0^P = 33$, $n_1^P = 39$, $n_2^P = 44$, $n_3^P = 41$; $n_0^S = 33$, $n_1^S = 38$, $n_2^S = 43$, $n_3^S = 41$).

Critical values (corresponding intersection hypotheses)	Original critical values	Unique critical values
c_1	2.3611	2.3611
c_2 (H_{110101}, H_{110011}, H_{110001}, H_{110111})	2.4805	2.4830
c_2 (H_{101111}, H_{101110}, H_{101011}, H_{101010})	2.4830	
c_2 (H_{011111}, H_{011110}, H_{011101}, H_{011100})	2.4785	
c_3 (H_{100111}, H_{100011})	2.4235	2.4253
c_3 (H_{010111}, H_{010101})	2.4253	
c_3 (H_{001111}, H_{001110})	2.4247	
c_4	2.1838	2.1838
c_5	2.3623	2.3623
c_6 (H_{000110})	2.2267	2.2275
c_6 (H_{000101})	2.2275	
c_6 (H_{000011})	2.2254	
c_7	1.9759	1.9759

Table III. p-Values for a dose–response study with multiple endpoints (three dose-to-control comparisons, two endpoints).

Individual hypothesis	Adjusted p-value based on the Dunnett–Bonferroni-based method	Nominal p-value
H_1^P	0.0829	0.0703
H_2^P	0.035	0.0278
H_3^P	0.0059	0.0043
H_1^S	0.0829	0.0775
H_2^S	0.035	0.0004
H_3^S	0.0059	<0.0001

under the balanced case. Hence, we can have different values of c_2, c_3 and c_6. To reduce the complexity of this problem and to be conservative, we take the maximum of these critical values to get a unique critical value. The third column in Table II gives the unique values for $c_1, c_2, \ldots, c_7$.

With the values of T_j^i and c from Tables I and II, the decision to reject an intersection hypothesis H is made by applying the decision rules in Table AI. An original hypothesis H_j^i can be rejected if all the intersection hypotheses H containing H_j^i are rejected. For this clinical study example, H_3^P and H_3^S are rejected. That means, only the high dose is found significantly superior to placebo in improving both endpoints: BRS and OF.

In practice, we also report adjusted p-values. Table III presents the adjusted p-values for the six original hypotheses in this clinical study example. For comparison purposes, Table III also presents the nominal p-values. Based on the adjusted p-values, we conclude that the superiority of the high dose over placebo is established in the improvement of both BRS and OF measurements at level 0.025 (adjusted p-values: $p_3^P = 0.0059$, $p_3^S = 0.0059$). Obviously, this result is the same as that obtained by comparing T_j^i's with the critical values. In addition, note that all p_j^S's are no larger than p_j^P, which is consistent with the logical restriction that no secondary hypothesis can be declared significant unless its corresponding primary hypothesis was rejected.

7. COMPARISONS OF METHODS

In this section, we compare in a dose–response clinical study setting, our proposed Dunnett–Bonferroni-based procedure, the Dunnett-based procedure [7] and a modification of Liu *et al.*'s procedure [17]. In their setting, Liu *et al.* [17] impose two layers of logical restriction: (1) requiring that the significance in the secondary endpoint for a dose level cannot be claimed unless its corresponding primary hypothesis was found

DOI: 10.1002/pst

significant and (2) requiring that the significance in a low dose level cannot be claimed unless the high dose level was claimed significant. The Dunnett–Bonferroni-based and the Dunnett-based procedures do not require the second logical restriction (but can be easily extended to incorporate it if one wanted to do so). To fairly compare Liu *et al.*'s [17] procedure with the other two, we modify it as follows: only the first logical restriction will be required or, more specifically, the same form of decision rules (Table AI) as those for our proposed procedure and the Dunnett-based procedure [7] will be used. This modification will be referred to as *the independent-T method* hereafter and denoted by us as the IT procedure.

For the purpose of simplicity, assume that the sample sizes across dose levels and endpoints are equal. The correlation between two endpoints is assumed to be ρ and estimated at $\hat{\rho}$.

Table IV. The critical values c_1, c_5, c_6 and c_7.

n	c_1	c_5	c_6	c_7
50	2.367	2.367	2.228	1.972
100	2.358	2.358	2.220	1.966
200	2.353	2.353	2.216	1.963

7.1. Comparison based on the critical values

The critical values c_1, c_5, c_6 and c_7 do not depend on the estimated correlation $\hat{\rho}$ between two endpoints since their calculation only involves test statistics from the same family. Their values with different sample sizes n (50, 100 and 200) are given in Table IV. It is not surprising to see that the larger the n, the smaller the c_1, c_5, c_6 and c_7.

The critical values c_2, c_3 and c_4 would differ for different values of $\hat{\rho}$ when the Dunnett-based parallel gatekeeping procedure [7] is applied. Their values under different combinations of $\hat{\rho}$ and sample size n are given in Table V.

From Table V, we first notice that, similar to c_1, c_5, c_6 and c_7, the larger the n, the smaller the c_2, c_3 and c_4. For the Dunnett-based (DM) parallel gatekeeping procedure, c_2, c_3 and c_4 increase as $\hat{\rho}$ decreases. This can be easily explained by Slepian's inequality (see [18, Appendix A.3]). Therefore, we can expect the testing results for the secondary hypotheses to be more significant

Table V. The critical values c_2, c_3 and c_4.

		Method					
		DM*				DB†	IT‡
		$\hat{\rho}$					
n	Critical values	−0.9	0	0.5	0.9	NA§	NA§
50	c_2	2.462	2.454	2.431	2.385	2.462	2.517
	c_3	2.416	2.412	2.400	2.377	2.417	2.436
	c_4	2.170	2.167	2.155	2.133	2.171	2.167
100	c_2	2.450	2.444	2.420	2.376	2.450	2.503
	c_3	2.404	2.403	2.390	2.366	2.406	2.424
	c_4	2.163	2.159	2.147	2.126	2.163	2.159
200	c_2	2.443	2.438	2.415	2.371	2.445	2.497
	c_3	2.401	2.397	2.384	2.362	2.401	2.419
	c_4	2.159	2.155	2.143	2.122	2.159	2.155

*DM, the Dunnett-based parallel gatekeeping procedure as introduced by Dmitrienko *et al.* [7]. The Genz and Bretz method [15] was used for calculating the critical values.

†DB, our proposed Dunnett–Bonferroni-based parallel gatekeeping procedure.

‡IT, the independent-T method, a modification of Liu *et al.*'s procedure [17].

§NA, not applicable.

DOI: 10.1002/pst

when the correlation between two endpoints is higher.

It is interesting to see that, for all three methods, c_1 is smaller than c_2 and c_3. As explained in [7], the larger values of c_2 and c_3 compared with c_1 are the price of sequential testing. For the DM procedure, the penalty becomes smaller with increasing correlation.

As explained in Section 5.2, for the Dunnett–Bonferroni-based (DB) parallel gatekeeping procedure, generally $c_2 \geq c_3$ and $c_3 \geq c_4$. From Table V, we find that this is also the case for the DM procedure and the IT procedure. This is easy to explain for the IT procedure. For the IT procedure, the critical values c_2 and c_3 are obtained by solving the equations that only depend on the one-dimensional marginal distributions of T_j^i, which under the equal sample size assumptions all follow a common t distribution. c_2 is obtained by solving

$$\Pr\{t > c_1\} * \Pr\{t > c_1\} * \Pr\{t > c_2\} = \alpha \quad (5)$$

and c_3 is obtained by solving

$$\Pr\{t > c_1\} * \Pr\{t > c_3\} * \Pr\{t > c_3\} = \alpha \quad (6)$$

where t denotes the common marginal variable. It can be easily shown that $c_2 \geq c_1$ and $c_3 \geq c_1$ so that from (5) and (6) it follows that $c_2 \geq c_3$. c_4 is obtained by solving

$$\Pr\{t > c_1\} * \Pr\{t > c_4\} = \alpha \quad (7)$$

By comparing equations (6) and (7), it is also easy to conclude that $c_3 \geq c_4$. A similar reasoning can be applied to explain that such ordering of c_2, c_3 and c_4 generally holds for the DM procedure.

As discussed in Section 5.1, when $\hat{\rho} = -1$, the DB procedure is equivalent to the DM procedure. When $\hat{\rho} \neq -1$, the DB procedure gives slightly larger critical values than the DM procedure. This is not surprising since the DB method partly replaces the multivariate-t-based calculation (with$\hat{\rho}$) of the DM procedure by the Bonferronization/α-splitting. However, for the cases of $\hat{\rho} = -0.9, 0$ and 0.5, the critical values of the DB procedure generally only differ from those of the DM procedure at the second decimal place. A second decimal place difference in terms of the critical values generally implies a third decimal place difference in terms of the p-values. Therefore, if one is not concerned about truly controlling type I error and uses $\hat{\rho}$ for the DM procedure, as long as $\hat{\rho}$ is not too large, the power loss of applying the DB procedure compared with the DM procedure is very minimal. Furthermore, to include the secondary endpoint in the drug label, regulatory agencies require that the secondary endpoint provides different information that does not duplicate the primary endpoint. If the correlation between the endpoints tends to be large, it implies that the two endpoints are measuring essentially the same construct. This further downgrades any possible disadvantage of using the DB procedure over the DM procedure in terms of the power loss. Besides the power comparison of the DB and DM procedures and associated type I error issues, we must keep in mind that the DB procedure avoids other hard-to-justify assumptions that are required in applying the DM procedure, which are discussed in Section 4.2.

An interesting finding is that, for the DB procedure, c_2 and c_3 are apparently smaller than those for the IT procedure, while c_4 is larger than that for the IT procedure. This is not hard to explain though. c_2 is the critical value for those decision rules involving two test statistics T_j^P from the primary family and one test statistic T_j^S from the secondary family. The DB procedure would first account for the joint distribution of two T_j^P's, which is a Dunnett-t distribution, and then apply the Bonferroni inequality for α-splitting and then find c_2 for T_j^S. However, when finding c_2, the IT procedure does not take into account the positive association between test statistics T_j^P's, which results from their common placebo factor, but just assumes the independence of them. This results in smaller values for c_2 when the DB procedure is applied compared with the IT procedure. A similar reasoning applies to c_3. Different from c_2 and c_3, c_4 is the critical value for those decision rules involving only one test statistic each from the primary and secondary families. It is obvious that the DB procedure gives larger values of c_4 than the IT procedure, since the DB procedure applies the Bonferroni inequality between two families without assuming any

DOI: 10.1002/pst

distributional assumptions, while the IT procedure assumes the independence of test statistics from two families. As explained before, for both the DB procedure and the IT procedure, generally $c_2 \geq c_3$ and $c_3 \geq c_4$. Therefore, the rejection of a secondary hypothesis H_j^S is more reliant on c_2 and c_3 than on c_4. Considering that for the DB procedure, c_2 and c_3 are smaller than those for the IT procedure, the DB procedure is generally more powerful than the IT procedure, even with the fact that c_4 for the DB procedure is greater than that for the IT procedure.

The IT method always gives larger critical values than the DM procedure when $\hat{\rho} \geq 0$. This is not unexpected since it has been proven by Liu *et al*. [17] that the IT method is more conservative than the DM procedure when $\hat{\rho} \geq 0$ in terms of controlling type I error. Obviously, when $\hat{\rho} = 0$, the IT method gives the same c_4 as the DM procedure.

7.2. Comparison of power through simulation

We performed a series of simulations to compare power of these three methods. The simulations were based on a hypothetical clinical trial comparing 3 doses of study drug to placebo with 50 subjects in each treatment group. The treatment groups were labeled P (placebo), L (low dose), M (median dose) and H (high dose). The efficacy of the study drug was studied using two normally distributed endpoints (primary and secondary) with correlation coefficient ρ. The standard deviation of the responses in the four treatment groups was set to 1. Following Dmitrienko *et al*. [7], we studied the three scenarios as described in Table VI.

The simulation considered three scenarios for the population means (see Table VI) and two possibilities for the estimated correlation coefficients $\hat{\rho} = 0.5$ and 0.9. The overall one-sided type I error was set to be 0.025 and 10 000 replications were done for each of the 6 scenario–correlation combinations.

Considering that the critical values for test statistics from the primary family are the same for the three methods, the power of declaring significant hypotheses in the primary family is the same for all three methods so that the power of passing the primary family to enter the secondary family must also be the same. The only difference in power lies in declaring significance in secondary hypotheses. Table VII summarizes the results of the simulation study. The DB procedure is only slightly less powerful than the DM procedure. The difference in power between these two procedures is negligible ($\leq 1\%$ in the presented scenarios). The DB procedure is a little more powerful than the IT procedure. For Scenario (3), the low dose is assumed to be effective for the secondary endpoint, but since it is not assumed to be effective for the primary endpoint, the chance of declaring significance in the secondary endpoint for this dose is less than 1% for all three methods.

8. DISCUSSION AND CONCLUSION

A parallel gatekeeping procedure deals with the multiple testing problems with hierarchically ordered hypotheses, which have become increasingly common in clinical studies. In this paper, we introduce a Dunnett–Bonferroni-based parallel gatekeeping procedure in the dose–response

Table VI. Means in each treatment group of primary and secondary endpoints for different scenarios (simulation study).

Scenario	P1	L1	M1	H1	P2	L2	M2	H2
(1)	0	0.65	0.65	0.65	0	0.65	0.65	0.65
(2)	0	0	0.65	0.65	0	0	0.65	0.65
(3)	0	0	0.65	0.65	0	0.65	0	0.65

Note: 1 and 2 denote the primary and secondary families, respectively.

DOI: 10.1002/pst

Table VII. Estimated power (%) of the Dunnett-based (DM), Dunnett–Bonferroni-based (DB) and independent-T (IT) parallel gatekeeping procedures in a clinical trial with two endpoints comparing three doses (labeled L, M and H) with placebo (P).

	Secondary endpoint							
	$\hat{\rho} = 0.5$				$\hat{\rho} = 0.9$			
Gatekeeping procedure	Scenario	L vs P	M vs P	H vs P	Scenario	L vs P	M vs P	H vs P
DM	(1)	73.06	73.32	73.56	(1)	78.07	78.52	78.43
DM	(2)		71.24	71.26	(2)		76.91	76.97
DM	(3)	0.98		69.4	(3)	0.97		75.55
DB	(1)	73.00	73.23	73.46	(1)	77.87	78.37	78.26
DB	(2)		71.05	71.02	(2)		76.57	76.58
DB	(3)	0.98		68.99	(3)	0.97		74.74
IT	(1)	72.92	73.15	73.36	(1)	77.77	78.25	78.14
IT	(2)		70.16	70.73	(2)		76.29	76.29
IT	(3)	0.98		68.48	(3)	0.97		74.28

The estimated correlation between the two endpoints $\hat{\rho}$ is 0.5 or 0.9, the overall one-sided type I error rate is 0.025 and the sample size per treatment group is 50. The cells corresponding to ineffective doses are left blank.

clinical study setting. This new procedure was constructed following the gatekeeping strategy of the Dunnett-based parallel gatekeeping procedure as introduced by Dmitrienko *et al*. [7].

We do note that there are other gatekeeping approaches that share in overlapping ways some of the ideas in our approach but are different in their entireties. For instance, the idea of applying the Dunnett adjustment within each family and splitting α using the Bonferroni inequality among families is conceptually similar to a gatekeeping procedure based on weighted hypotheses introduced by Hommel *et al*. [12, Section 3.1]. The procedure in [12, Section 3.1] groups hypotheses for high doses across all endpoints as the primary family and hypotheses for low doses across all endpoints as the secondary family. It recommends using a Dunnett test on hypotheses corresponding to the same endpoint within the same family. This procedure differs from our proposed procedure mainly in that: (1) this procedure does not apply a Dunnett test on the secondary hypotheses when any primary hypothesis is also part of an intersection hypothesis, while the Dunnett–Bonferroni-based procedure does not have such a restriction; (2) for this procedure, the p-value for each intersection hypothesis is a function of p-values for individual hypotheses and their pre-specified weights (representing relative importance), while the Dunnett–Bonferroni-based procedure does not adopt pre-specified weights in obtaining p-values.

In addition, the Dunnett–Bonferroni-based procedure shares some similarities with the multi-stage Dunnett-based procedure introduced in Dmitrienko *et al*. [20]. The multi-stage gatekeeping procedure splits α based on the observed number of accepted hypotheses in the primary family F_1, and therefore α' is unique. The Dunnett–Bonferroni-based procedure splits α under *each* intersection hypothesis, and therefore α' is not unique and depends on the number of test statistics from F_1 that are involved in a decision rule. As a result, the multi-stage gatekeeping procedure is simpler to implement, but may be less powerful than the Dunnett–Bonferroni-based procedure.

From a more general conceptual viewpoint, our approach of calculating type I error spent in the primary family is similar to calculating the conditional error function in adaptive designs (for example, the conditional error rate in an adaptive Dunnett test [21]), in the sense that they calculate the remaining type I error rate to be spent for the next family/stage.

DOI: 10.1002/pst

We also note that our proposed method can be easily extended to situations where the independence condition is dropped. Under such situations, the critical values for the primary hypotheses may not be kept the same for different intersection hypotheses. Nevertheless, the idea of applying Dunnett adjustment within each family and applying Bonferronization between families stays the same and can be easily applied to decision rules that are in line with the objectives of clinical studies.

Our proposed Dunnett–Bonferroni-based parallel gatekeeping procedure is similar to the Dunnett-based approach [7] in power performance, but relaxes some hard-to-justify assumptions that are required for the Dunnett-based procedure. Our proposed method strongly controls type I error rate, while the Dunnett-based procedure could be liberal in controlling type I error [17] due to using the Genz and Bretz method [15] in a way that is not intended to be used, such as in the case of unequal variance estimates. The proposed method is easier to implement since in most cases an SAS [19] function PROBMC is enough for calculating the critical values. However, for the Dunnett-based procedure [7], a relatively more complicated and computationally intensive algorithm for finding the quantiles of multivariate-t distributions is always required. Moreover, our proposed method, unlike the Dunnett-based procedure, does not require using the estimated correlation for calculating the critical values. From a regulatory point of view, the type I error is strictly controlled by the Dunnett–Bonferroni-based procedure, whereas the Dunnett-based approach does not because it is based on estimating the correlation between endpoints.

The independent-T approach, a parallel gatekeeping method that assumes the independence between the endpoints (a modification of Liu *et al.*'s method [17]), is generally less powerful than our proposed Dunnett–Bonferroni-based procedure. It is important to be aware that it does require non-negative correlation in order to be conservative in terms of type I error control, which may be an assumption that is inappropriate in a regulatory setting.

Overall, the newly proposed Dunnett–Bonferroni-based parallel gatekeeping method avoids assumptions that might be challenged by regulatory agencies and does so with virtually no cost.

ACKNOWLEDGEMENTS

The authors would like to thank two anonymous referees for their helpful comments.

REFERENCES

1. Dmitrienko A, Tamhane AC. Gatekeeping procedures with clinical trial applications. *Pharmaceutical Statistics* 2007; **6**:171–180.
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.), vol. 6. Fischer Verlag: Stuttgart, 1995; pp. 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. Westfall PH, Krishen A. Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *Journal of Statistical Planning and Inference* 2001; **99**:25–41.
5. 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.
6. Chen X, Luo X, Capizzi T. The application of enhanced parallel gatekeeping strategies. *Statistics in Medicine* 2005; **24**:1385–1397.
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. 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.
9. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* 1955; **50**:1096–1121.
10. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.

DOI: 10.1002/pst

11. Dmitrienko A, Molenberghs G, Chuang-Stein C, Offen W. *Analysis of clinical trials using SAS: a practical guide*. SAS Press: Cary, NC, 2005.
12. Hommel G, Bretz F, Maurer W. Powerful shortcuts for multiple testing procedures with special reference to gatekeeping strategies. *Statistics in Medicine* 2007; **26**:4063–4073.
13. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **63**:655–660.
14. Sarkar SK, Chang CK. Simes' method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* 1997; **92**:1601–1608.
15. Genz A, Bretz F. Methods for the computation of multivariate *t*-probabilities. *Journal of Computational and Graphical Statistics* 2002; **11**:950–971.
16. Pollard K, van der Laan M. Choice of a null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference* 2005; **125**:85–100.
17. Liu Y, Hsu J, Ruberg S. Partition testing in dose–response studies with multiple endpoints. *Pharmaceutical Statistics* 2007; **6**:181–192.
18. Hsu JC. *Multiple comparisons: theory and methods*. Chapman & Hall: London, 1996.
19. SAS Institute. *The SAS system for Windows, release 9.1*. SAS Institute: Cary, NC, 2005.
20. Dmitrienko A, Tamhane A, Wiens B. General multistage gatekeeping procedures. *Biometrical Journal* 2008; **5**:667–677.
21. Koenig F, Brannath W, Bretz F, Posch M. Adaptive Dunnett test for treatment selection. *Statistics in Medicine* 2008; **27**:1612–1625.

APPENDIX A

The decision rules for 63 intersection hypotheses are given in Table AI.

Table AI. Decision rules for 63 intersection hypotheses (two endpoints, three dose-to-control comparisons).

Intersection hypotheses	Decision rule
H_{111111}, H_{111110}, H_{111101}, H_{111100}, H_{111011}, H_{111010}, H_{111001}, H_{111000}	$T_1^P > c_1$ or $T_2^P > c_1$ or $T_3^P > c_1$
H_{110110}, H_{110100}, H_{110010}, H_{110000}	$T_1^P > c_1$ or $T_2^P > c_1$
H_{101101}, H_{101100}, H_{101001}, H_{101000}	$T_1^P > c_1$ or $T_3^P > c_1$
H_{011011}, H_{011010}, H_{011001}, H_{011000}	$T_2^P > c_1$ or $T_3^P > c_1$
H_{100100}, H_{100000}	$T_1^P > c_1$
H_{010010}, H_{010000}	$T_2^P > c_1$
H_{001001}, H_{001000}	$T_3^P > c_1$
H_{110101}, H_{110011}, H_{110001}, H_{110111}	$T_1^P > c_1$ or $T_2^P > c_1$ or $T_3^S > c_2$
H_{101111}, H_{101110}, H_{101011}, H_{101010}	$T_1^P > c_1$ or $T_3^P > c_1$ or $T_2^S > c_2$
H_{011111}, H_{011110}, H_{011101}, H_{011100}	$T_2^P > c_1$ or $T_3^P > c_1$ or $T_1^S > c_2$
H_{100111}, H_{100011}	$T_1^P > c_1$ or $T_2^S > c_3$ or $T_3^S > c_3$
H_{010111}, H_{010101}	$T_2^P > c_1$ or $T_1^S > c_3$ or $T_3^S > c_3$
H_{001111}, H_{001110}	$T_3^P > c_1$ or $T_1^S > c_3$ or $T_2^S > c_3$
H_{100110}, H_{100010}	$T_1^P > c_1$ or $T_2^S > c_4$
H_{100101}, H_{100001}	$T_1^P > c_1$ or $T_3^S > c_4$
H_{010110}, H_{010100}	$T_2^P > c_1$ or $T_1^S > c_4$
H_{010011}, H_{010001}	$T_2^P > c_1$ or $T_3^S > c_4$
H_{001101}, H_{001100}	$T_3^P > c_1$ or $T_1^S > c_4$
H_{001011}, H_{001010}	$T_3^P > c_1$ or $T_2^S > c_4$
H_{000111}	$T_1^S > c_5$ or $T_2^S > c_5$ or $T_3^S > c_5$
H_{000110}	$T_1^S > c_6$ or $T_2^S > c_6$
H_{000101}	$T_1^S > c_6$ or $T_3^S > c_6$
H_{000011}	$T_2^S > c_6$ or $T_3^S > c_6$
H_{000100}	$T_1^S > c_7$
H_{000010}	$T_2^S > c_7$
H_{000001}	$T_3^S > c_7$

APPENDIX B

This section gives a proof that, when the correlation ρ between test statistics from two families (say, T^P and T^S) is equal to -1, the Dunnett–-Bonferroni-based parallel gatekeeping procedure is equivalent to the Dunnett-based procedure [7].

When $\rho = -1$, T^S can be expressed as $a^* T^P + b$, where a and b are constants and $a<0$. Under the global null hypothesis, we know that the expected values of two test statistics are equal to 0, that is, $E(T^P) = E(T^S) = 0$. Therefore, obviously b must be equal to 0. Thus, we get $T^S = a^* T^P$, where $a<0$.

DOI: 10.1002/pst

Pharmaceutical
STATISTICS

Therefore, when $\rho = -1$,

$$\begin{aligned}
&\Pr\{T^P > c_1 \text{ or } T^S > c_2\} \\
&\quad = \Pr\{T^P > c_1 \text{ or } a * T^P > c_2\} \\
&\quad = \Pr\{T^P > c_1 \text{ or } T^P < c_2/a\} \qquad \text{(B1)} \\
&\quad = \Pr\{T^P > c_1\} + \Pr\{T^P < c_2/a\} \qquad \text{(B2)} \\
&\qquad\quad = \Pr\{T^P > c_1\} + \Pr\{T^S > c_2\}
\end{aligned}$$

Equation (B1) holds because $a < 0$. Equation (B2) holds because $c_1 \geq 0 \geq c_2/a$ since both c_1 and c_2 are non-negative.

We then have $\Pr_{\rho=-1}\{T^P > c_1 \text{ or } T^S > c_2\} = \Pr\{T^P > c_1\} + \Pr\{T^S > c_2\}$. The left side of this equation is used for finding the critical values for the Dunnett-based procedure. The right side of this equation is actually the upper bound of $\Pr_{-1 \leq \rho \leq 1}\{T^P > c_1 \text{ or } T^S > c_2\}$ when the Bonferroni inequality is applied, that is, it is the equation used for finding the critical values for the Dunnett–Bonferroni-based procedure. The statement that the Dunnett-based procedure is equivalent to the Dunnett–Bonferroni-based procedure when $\rho = -1$ is proved.

DOI: 10.1002/pst