


A unified framework for weighted parametric multiple test procedures

Dong Xi^{*,1} , Ekkehard Glimm^{2,3}, Willi Maurer², and Frank Bretz^{2,4}

¹ Novartis Pharmaceuticals Corporation, East Hanover, NJ 07936, USA

² Novartis Pharma AG, 4002 Basel, Switzerland

³ Otto-von-Guericke-University Magdeburg, Medical Faculty, Institute for Biometrics and Medical Informatics, Germany

⁴ School of Statistics and Management, Shanghai University of Finance and Economics, People's Republic of China

Received 18 November 2016; revised 23 January 2017; accepted 27 January 2017

We describe a general framework for weighted parametric multiple test procedures based on the closure principle. We utilize general weighting strategies that can reflect complex study objectives and include many procedures in the literature as special cases. The proposed weighted parametric tests bridge the gap between rejection rules using either adjusted significance levels or adjusted p -values. This connection is made by allowing intersection hypotheses of the underlying closed test procedure to be tested at level smaller than α . This may be also necessary to take certain study situations into account. For such cases we introduce a subclass of exact α -level parametric tests that satisfy the consonance property. When the correlation is known only for certain subsets of the test statistics, a new procedure is proposed to fully utilize this knowledge within each subset. We illustrate the proposed weighted parametric tests using a clinical trial example and conduct a simulation study to investigate its operating characteristics.

Keywords: Adjusted p -value; Closure principle; Consonance; Multiple test procedure; Non-exhaustiveness.



Additional supporting information including source code to reproduce the results may be found in the online version of this article at the publisher's web-site

1 Introduction

Scientific experiments are often faced with simultaneous inference problems when addressing multiple objectives, such as assessing the differences between several experimental conditions. However, testing multiple hypotheses simultaneously may lead to an increased Type I error rate if no multiplicity adjustment is foreseen. Starting from the work of Holm (1979), weighted multiple test procedures (MTPs) are commonly used to control the overall Type I error rate by assigning weights to different hypotheses in order to reflect the relative importance of objectives in the test strategy. Early references on weighted min- p tests include the resampling-based tests from Westfall and Young (1993, Chapter 6), and Westfall *et al.* (1998). Weighted MTPs based on specific parametric models have been investigated using hierarchical tests (Huque and Alosh, 2008) and graphical approaches (Bretz *et al.*, 2011). These procedures discuss weighted parametric MTPs using the closure principle (Marcus *et al.*, 1976) where each intersection hypothesis is tested at exact level α . For the step-down parametric procedure by

*Corresponding author: e-mail: dong.xi@novartis.com, Phone: +1-862-778-7498, Fax: +1-973-781-8265

Dunnett and Tamhane (1991), a weighted step-down procedure has been introduced by Xie (2012) based on adjusted p -values.

MTPs are usually carried out by comparing either adjusted significance levels with unadjusted p -values or adjusted p -values with the unadjusted level α . Although various weighted parametric tests have been proposed in the literature, the link between rejection rules using either adjusted significance levels or adjusted p -values has not been systematically explored. In addition, the majority of the procedures in the literature focus on the case where each intersection hypothesis is tested at exact level α . It remains unclear how to deal with the nontrivial case where the significance level is strictly less than α for some of the intersection hypotheses. This is a relevant question for certain study considerations. For example, in the phase III clinical trial of buparlisib in patients with advanced and metastatic breast cancer, the analysis of progression-free survival (PFS) endpoints happens much earlier in time than the analysis of the overall survival (OS) endpoints (Goteti *et al.*, 2014). Thus, testing of PFS hypotheses does not benefit from rejecting the OS hypotheses at a later time point. Besides, various parallel and k -out-of- n gatekeeping procedures test certain intersection hypotheses involving primary hypotheses at levels smaller than α to allow testing secondary hypotheses, even if not all primary hypotheses are rejected (Dmitrienko *et al.*, 2008; Xi and Tamhane, 2014).

We propose a unified framework for weighted parametric MTPs by applying the closure principle to general weighting strategies that includes many procedures in the literature as special cases. When some intersection hypotheses are tested at levels smaller than α , we describe a special property of a class of parametric tests that proportionally increases the hypothesis weights to ensure exact α -level tests. When the parametric assumptions only apply to subsets of hypotheses, we propose a new procedure that utilizes the parametric assumptions within each subset. We derive analytic expressions for the adjusted p -values to avoid numerical root finding under multidimensional integration.

The paper is structured as follows. In Section 2, we introduce relevant notations as well as the general concepts of closed testing and weighting schemes. In Section 3, we provide a motivating example to illustrate the need for weighted parametric tests. In Section 4, we propose a general framework for weighted parametric tests based on adjusted significance levels and adjusted p -values when the parametric assumption can be applied to either all hypotheses under consideration or only a subset thereof. In Section 5, we discuss the consonance property for the proposed general weighted parametric tests. In Section 6, we revisit the motivating example and illustrate the proposed procedure with numerical results and a simulation study. We conclude the paper with a discussion in Section 7.

2 Notation

Consider testing m elementary null hypotheses $H_i, i \in I = \{1, \dots, m\}$ against one-sided or two-sided alternatives simultaneously. Under the closure principle (Marcus *et al.*, 1976), we test each nonempty intersection hypothesis $H_J = \cap_{j \in J} H_j, J \subseteq I$, at level α . We reject an elementary hypothesis $H_i, i \in I$, if every intersection hypothesis H_J with $i \in J \subseteq I$ is rejected by its associated α -level test. The closed test procedure controls the familywise error rate (FWER) at level α in the strong sense (Hochberg and Tamhane, 1987).

Because some hypotheses among H_1, \dots, H_m may be more important than others, we assign weights for different hypotheses to reflect the relative importance. Using the notation from Maurer and Bretz (2013), let $\mathbf{w}_J = (w_j(J), j \in J)$ denote a vector of weights for an index set $J \subseteq I$. A weighting scheme $\mathbf{W} = \{\mathbf{w}_J, J \subseteq I\}$ is called valid if for every $J \subseteq I$ and $j \in J$ we have $w_j(J) \geq 0$ and $0 < \sum_{j \in J} w_j(J) \leq 1$. Throughout this paper, we assume that weighting schemes are valid. In addition, \mathbf{W} is called exhaustive if for every $J \subseteq I$ we have $\sum_{j \in J} w_j(J) = 1$. Exhaustiveness is a desirable property but not required in this paper, although weighting schemes underlying many common MTPs such as the step-down procedures by Holm (1979) and Dunnett and Tamhane (1991) are both valid and exhaustive. Both procedures have the same weighting scheme $w_j(J) = 1/|J|$ for $j \in J \subseteq I$, where $|J|$ denotes the number

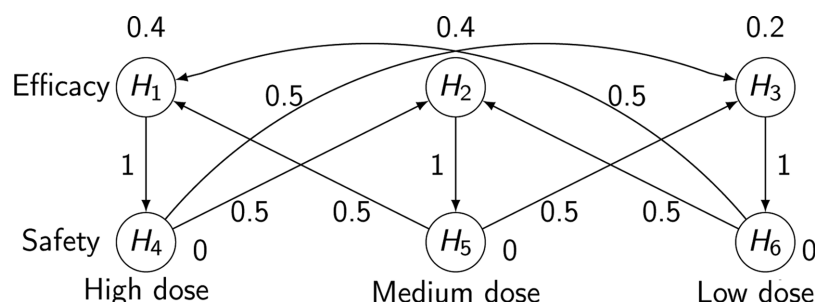


Figure 1 Graphical multiple test procedure for the motivating example.

of indices in J . While the former procedure is based on the Bonferroni test, the latter procedure takes the correlations between the test statistics into account.

Let p_i denote the unadjusted p -value for H_i , $i \in I$. Consider the weighted Bonferroni test that rejects H_j at level α if $p_j \leq w_j(J)\alpha$ for any $j \in J$. An equivalent way of testing H_j is to use its p -value $\tilde{p}_J = \min[1, \min_{j \in J} \{p_j/w_j(J)\}]$. Accordingly, we can reject H_j if $\tilde{p}_J \leq \alpha$. Applying the closure principle, we can then reject the elementary hypothesis H_i if its adjusted p -value $p_i^{\text{adj}} = \max_{\{J: i \in J \subseteq I\}} \tilde{p}_J \leq \alpha$. In the following, $w_j(J)$ and $w_j(J)\alpha$ are called the local weight and the local significance level, respectively.

Throughout this paper, the p -value is a function of the test statistic through an appropriate cumulative distribution function, which could be, for example, an (asymptotically) normal or a t distribution. We assume that under the null hypothesis H_i the unadjusted p -value p_i is uniformly distributed over $[0, 1]$, $i = 1, \dots, m$. The joint distribution of the p_i 's can be derived if the multivariate probability distribution of the corresponding test statistics is known or estimated, such as an (asymptotically) multivariate normal distribution with a known or estimated correlation matrix.

3 Motivating example

Consider the clinical trial example from Bauer *et al.* (2001) to test for the superiority of three doses of an investigational treatment against a control regarding an efficacy and a safety endpoint. There are three efficacy hypotheses H_1, H_2, H_3 and three safety hypotheses H_4, H_5, H_6 for the comparison of high, medium, and low dose against control, respectively. The overall significance level is one-sided $\alpha = 0.025$. We modify the step-down procedure without order constraints between the doses from Section 3 in Bauer *et al.* (2001) as follows. Initially, α is split across the three efficacy hypotheses, which means that the safety hypotheses do not receive any portion of α . Assume that high and medium dose receive a larger weight 0.4 so that the associated null hypotheses have a higher chance to be rejected than the case when α had been split equally across all three efficacy hypotheses, which leaves the weight 0.2 for the remaining efficacy hypothesis. Thus the weights for all hypotheses are $w_I = (0.4, 0.4, 0.2, 0, 0, 0)$. Within each dose-control comparison, the hypothesis on the efficacy endpoint is tested first. If rejected, the test on the safety endpoint is performed at the same local significance level as the corresponding test on the efficacy endpoint. If both hypotheses can be rejected for the same dose, the associated local level is equally distributed among the other two doses.

To facilitate our discussion, we use the graphical approach by Bretz *et al.* (2009, 2011). In this framework, hypotheses are denoted by nodes associated with their local weights. A directed edge from H_i to H_j means that when H_i is rejected, its local weight can be propagated to H_j . The weight associated with that edge quantifies the proportion of the local weight of H_i that is propagated to H_j . Accordingly, the graphical representation of the proposed MTP for the motivating example is shown in Fig. 1. Algorithm 1 in Bretz *et al.* (2011) describes how to derive the weighting scheme based on any graph.

Table 1 The weighting scheme of the motivating example.

J	Local weights $w_j(J)$						J	Local weights $w_j(J)$					
	H_1	H_2	H_3	H_4	H_5	H_6		H_1	H_2	H_3	H_4	H_5	H_6
{1, 2, 3, 4, 5, 6}	0.4	0.4	0.2	0	0	0	{2, 3, 4}	—	0.4	0.2	0.4	—	—
{1, 2, 3, 4, 5}	0.4	0.4	0.2	0	0	—	{2, 3, 5}	—	0.6	0.4	—	0	—
{1, 2, 3, 4, 6}	0.4	0.4	0.2	0	—	0	{2, 3, 6}	—	0.6	0.4	—	—	0
{1, 2, 3, 5, 6}	0.4	0.4	0.2	—	0	0	{2, 4, 5}	—	0.5	—	0.5	0	—
{1, 2, 4, 5, 6}	0.4	0.4	—	0	0	0.2	{2, 4, 6}	—	0.4	—	0.4	—	0.2
{1, 3, 4, 5, 6}	0.4	—	0.2	0	0.4	0	{2, 5, 6}	—	0.6	—	—	0	0.4
{2, 3, 4, 5, 6}	—	0.4	0.2	0.4	0	0	{3, 4, 5}	—	—	0.2	0.4	0.4	—
{1, 2, 3, 4}	0.4	0.4	0.2	0	—	—	{3, 4, 6}	—	—	0.4	0.6	—	0
{1, 2, 3, 5}	0.4	0.4	0.2	—	0	—	{3, 5, 6}	—	—	0.4	—	0.6	0
{1, 2, 3, 6}	0.4	0.4	0.2	—	—	0	{4, 5, 6}	—	—	—	0.4	0.4	0.2
{1, 2, 4, 5}	0.5	0.5	—	0	0	—	{1, 2}	0.5	0.5	—	—	—	—
{1, 2, 4, 6}	0.4	0.4	—	0	—	0.2	{1, 3}	0.6	—	0.4	—	—	—
{1, 2, 5, 6}	0.4	0.4	—	—	0	0.2	{1, 4}	1	—	—	0	—	—
{1, 3, 4, 5}	0.4	—	0.2	0	0.4	—	{1, 5}	0.5	—	—	—	0.5	—
{1, 3, 4, 6}	0.6	—	0.4	0	—	0	{1, 6}	0.6	—	—	—	—	0.4
{1, 3, 5, 6}	0.4	—	0.2	—	0.4	0	{2, 3}	—	0.6	0.4	—	—	—
{1, 4, 5, 6}	0.4	—	—	0	0.4	0.2	{2, 4}	—	0.5	—	0.5	—	—
{2, 3, 4, 5}	—	0.4	0.2	0.4	0	—	{2, 5}	—	1	—	—	0	—
{2, 3, 4, 6}	—	0.4	0.2	0.4	—	0	{2, 6}	—	0.6	—	—	—	0.4
{2, 3, 5, 6}	—	0.6	0.4	—	0	0	{3, 4}	—	—	0.4	0.6	—	—
{2, 4, 5, 6}	—	0.4	—	0.4	0	0.2	{3, 5}	—	—	0.4	—	0.6	—
{3, 4, 5, 6}	—	—	0.2	0.4	0.4	0	{3, 6}	—	—	1	—	—	0
{1, 2, 3}	0.4	0.4	0.2	—	—	—	{4, 5}	—	—	—	0.5	0.5	—
{1, 2, 4}	0.5	0.5	—	0	—	—	{4, 6}	—	—	—	0.6	—	0.4
{1, 2, 5}	0.5	0.5	—	—	0	—	{5, 6}	—	—	—	—	0.6	0.4
{1, 2, 6}	0.4	0.4	—	—	—	0.2	{1}	1	—	—	—	—	—
{1, 3, 4}	0.6	—	0.4	0	—	—	{2}	—	1	—	—	—	—
{1, 3, 5}	0.4	—	0.2	—	0.4	—	{3}	—	—	1	—	—	—
{1, 3, 6}	0.6	—	0.4	—	—	0	{4}	—	—	—	1	—	—
{1, 4, 5}	0.5	—	—	0	0.5	—	{5}	—	—	—	—	1	—
{1, 4, 6}	0.6	—	—	0	—	0.4	{6}	—	—	—	—	—	1
{1, 5, 6}	0.4	—	—	—	0.4	0.2							

Table 1 shows the local weight vector w_j for each intersection hypothesis H_J , $J \subseteq I$ using the gMCP R package (Rohmeyer and Klinglmueller, 2015). For example, the local weights for $H_{123} = H_1 \cap H_2 \cap H_3$ are $w_1(\{1, 2, 3\}) = 0.4$, $w_2(\{1, 2, 3\}) = 0.4$ and $w_3(\{1, 2, 3\}) = 0.2$. For $H_{234} = H_2 \cap H_3 \cap H_4$ they are $w_2(\{2, 3, 4\}) = 0.4$, $w_3(\{2, 3, 4\}) = 0.2$ and $w_4(\{2, 3, 4\}) = 0.4$. Because of the common control in the three dose-control comparisons, we assume the correlations are known between test statistics for the three efficacy hypotheses H_1, H_2, H_3 and propose a weighted parametric test. For the sake of illustration, we assume that the joint distribution of test statistics for the safety hypotheses H_4, H_5, H_6 is unknown and thus the Bonferroni method is applied.

4 Weighted parametric tests for intersection hypotheses

4.1 Joint distribution fully known

Let P_j denote the random variable whose realization is the observed unadjusted p -value p_j for H_j , $j \in J$, for some $J \subseteq I$. If the joint distribution of P_j , $j \in J$, is fully known, the weighted min- p test rejects H_J if $p_j \leq c_J w_j(J) \alpha$ for any $j \in J$, where c_J is dependent on α and is calculated such that

$$\Pr_{H_J} \left[\bigcup_{j \in J} \{P_j \leq c_J w_j(J) \alpha\} \right] = \alpha \sum_{j \in J} w_j(J). \quad (1)$$

Setting $c_J = 1$ results in the weighted Bonferroni test with an inequality in (1). Otherwise, $c_J > 1$ and the resulting weighted parametric test is more powerful than the weighted Bonferroni test. Let $q_J = \min_{j \in J} \{p_j / w_j(J)\}$ denote the smallest observed weighted p -value for H_j , $j \in J$. The p -value \tilde{p}_J for the intersection hypothesis H_J subject to $\sum_{j \in J} w_j(J) \leq 1$ is then given by

$$\tilde{p}_J = \min \left[1, \frac{1}{\sum_{j \in J} w_j(J)} \Pr_{H_J} \left[\bigcup_{j \in J} \left\{ \frac{P_j}{w_j(J)} \leq q_J \right\} \right] \right]. \quad (2)$$

Therefore, we reject H_J if $p_j \leq c_J w_j(J) \alpha$ for any $i \in J$ with c_J determined in (1) or, equivalently, if $\tilde{p}_J \leq \alpha$. By the closure principle, we reject an elementary hypothesis H_i , $i \in I$, if every H_J with $i \in J \subseteq I$ is rejected. Equivalently, the adjusted p -value p_i^{adj} of H_i is the maximum of \tilde{p}_J , $i \in J \subseteq I$, and we reject H_i if it is less than or equal to α . Together with a general weighting scheme W , the proposed weighted parametric test (1) and (2) includes many procedures in the literature as special cases, such as the step-down parametric procedure by Dunnett and Tamhane (1991), the parametric fallback procedure (Huque and Alos, 2008), and the graphical approaches with parametric assumptions (Bretz *et al.*, 2011).

To see how \tilde{p}_J is derived in (2), rewrite the left hand side of (1) as

$$\Pr_{H_J} \left[\bigcup_{j \in J} \left\{ \frac{P_j}{w_j(J)} \leq c_J \alpha \right\} \right] = \Pr_{H_J} \left[\min_{j \in J} \left\{ \frac{P_j}{w_j(J)} \right\} \leq c_J \alpha \right] = \alpha \sum_{j \in J} w_j(J).$$

Then $c_J \alpha$ is the $\{\alpha \sum_{j \in J} w_j(J)\}$ -th quantile of the distribution of the minimum weighted p -value $Q_J = \min_{j \in J} \{P_j / w_j(J)\}$. Under the null hypothesis H_J , the probability of observing an equally or more extreme outcome is $\Pr_{H_J} \{Q_J \leq q_J\}$. The p -value for H_J subject to $\sum_{j \in J} w_j(J) \leq 1$ is then given by (2), after truncation at 1. Note that it is computationally more efficient to derive rejection rules using \tilde{p}_J because it avoids solving numerically for c_J from an equation involving multidimensional integration.

4.2 Parametric tests that enforce exhaustiveness

In Section 4.1, we investigated weighted parametric tests that preserve the significance level for H_J , $J \subseteq I$, at level $\alpha \sum_{j \in J} w_j(J)$. However, it may be tempting to always increase the sum of the local weights

to 1. Xie (2012) considered the case when the initial weights $w_i = w_i(I) > 0$ for all $i \in I$. If the joint distribution among the p -values is fully known, they proposed a closed procedure using

$$\tilde{p}_J = \Pr_{H_J} \left[\bigcup_{j \in J} \left\{ \frac{P_j}{w_j} \leq q_J \right\} \right], \quad (3)$$

where $q_J = \min_{j \in J} \{p_j/w_j\}$. Here, (3) is stated more generally because we do not assume the ordering of weighted p -values as in Section 2.4 of Xie (2012). Compared to (2), the factor $1/\sum_{j \in J} w_j(J)$ is missing, which implies that $\sum_{j \in J} w_j(J)$ is always increased to 1.

Note that Xie (2012) did not provide rejection rules based on adjusted significance levels but it follows from (1) and (2) that we can reject H_J if $p_j \leq c_J w_j(J) \alpha$ for any $j \in J$, where c_J satisfies

$$\Pr_{H_J} \left[\bigcup_{j \in J} \left\{ P_j \leq c_J w_j(J) \alpha \right\} \right] = \alpha = \Pr_{H_J} \left[\bigcup_{j \in J} \left\{ P_j \leq c_J \frac{w_j}{\sum_{j \in J} w_j} \alpha \right\} \right]. \quad (4)$$

If $w_j(J) = w_j / \sum_{j \in J} w_j$, the leftmost and the rightmost expressions in (4) are the same. We then reject H_J if $p_j \leq c_J \alpha w_j / \sum_{j \in J} w_j$ for any $j \in J$. Thus, the procedure by Xie (2012) actually tests H_J in the following two steps. First, set $w_j(J)$ to $w_j / \sum_{j \in J} w_j$, that is to increase w_j proportionally such that $\sum_{j \in J} w_j(J) = 1$. Second, reject H_J if $p_j \leq c_J \alpha w_j / \sum_{j \in J} w_j$ for any $j \in J$ as in (4) or equivalently if $\tilde{p}_J \leq \alpha$ as in (3).

The resulting weighting scheme is always exhaustive when $w_i > 0$ for all $i \in I$. However, it requires that all local weights $w_j(J)$ are completely determined by the initial local weights, that is $w_j(J) = w_j / \sum_{j \in J} w_j$, $j \in J \subseteq I$. It does not apply to general weighting schemes, especially when some initial local weights are 0. Nevertheless, the idea by Xie (2012) can be generalized to any valid weighting scheme by dropping $\sum_{i \in J} w_j(J)$ and $1/\sum_{i \in J} w_j(J)$ from the right hand side of (1) and (2), respectively. The resulting closed procedure then always increases the local weight $w_j(J)$ proportionally to $w_j(J)/\sum_{i \in J} w_j(J)$.

It is not trivial to determine whether a weighting scheme generated from an MTP is exhaustive or not, even if the initial local weights sum to 1. In addition, it may be desirable to use a nonexhaustive weighting scheme for practical considerations. For these reasons, we recommend working on the weighting scheme separately to incorporate trial design considerations, and then using a weighted parametric test that preserves the significance level for each intersection hypothesis as in (1) and (2). For instance, when $\sum_{i \in I} w_i < 1$, the procedure by Xie (2012) can be implemented by first proportionally increasing local weights so that the weighting scheme is $\mathbf{W} = \{w_J = (w_j / \sum_{j \in J} w_j, j \in J), J \subseteq I\}$. Then we can apply the weighted parametric test in Section 4.1 within the closed procedure.

4.3 Joint distribution not fully known

If the joint distribution is only known for subsets of p -values, we can extend the parametric test in (1) and (2) using similar ideas as Bretz *et al.* (2011). Assume that I can be partitioned into ℓ mutually exclusive subsets I_h such that $I = \bigcup_{h=1}^{\ell} I_h$. For each subset I_h , $h = 1, \dots, \ell$, we assume that the joint distribution of the p -values p_i , $i \in I_h$, is fully known, but the joint distribution of p -values from different subsets is not necessarily known. For any $J \subseteq I$, let $J_h = J \cap I_h$, $h = 1, \dots, \ell$. Then we reject H_J if $p_j \leq c_J w_j(J) \alpha$ for any $j \in J$, where c_J satisfies

$$\sum_{h=1}^{\ell} \Pr_{H_J} \left[\bigcup_{j \in J_h} \left\{ P_j \leq c_J w_j(J) \alpha \right\} \right] = \alpha \sum_{i \in J} w_j(J). \quad (5)$$

The approach from Bretz *et al.* (2011) is a special case of (5) when $\sum_{j \in J} w_j(J) = 1$.

Note that (5) uses a common c_J for all subsets J_h , $h = 1, \dots, \ell$. Hence, the test decisions in J_h are affected by the distribution in other subsets although the joint distribution between subsets is not necessarily known. For example, if $J_h = \{j\}$ contains only one index, we reject H_j if $p_j \leq c_J w_j(J) \alpha$, which is no longer the rejection rule if the Bonferroni test were applied. Instead, we propose to use different c_{J_h} 's for different subsets J_h , $h = 1, \dots, \ell$, to fully utilize the parametric assumptions for J_h within J_h . The Bonferroni split is utilized among the subsets. Specifically, for any $J \subseteq I$, we reject H_J if $p_j \leq c_{J_h} w_j(J) \alpha$ for any $j \in J$, where c_{J_h} satisfies

$$\Pr_{H_J} \left[\bigcup_{j \in J_h} \left\{ P_j \leq c_{J_h} w_j(J) \alpha \right\} \right] = \alpha \sum_{j \in J_h} w_j(J) \quad (6)$$

for $h = 1, \dots, \ell$. If we take the sum of the left hand side in (6) over $h = 1, \dots, \ell$, we have $\alpha \sum_{j \in J} w_j(J)$ on the right hand side, which is the significance level for H_J .

Another advantage of using different c_{J_h} 's for J_h is that we can derive the p -values analytically. First, the p -value for each subset J_h is derived using (2) and then the p -value for H_J is the minimum over $h = 1, \dots, \ell$. Specifically, let $q_{J_h} = \min_{j \in J_h} \{p_j / w_j(J)\}$ such that the p -value for H_J becomes $\tilde{p}_J = \min_{h=1}^{\ell} \{\tilde{p}_{J_h}\}$, where

$$\tilde{p}_{J_h} = \min \left[1, \frac{1}{\sum_{j \in J_h} w_j(J)} \Pr_{H_J} \left[\bigcup_{j \in J_h} \left\{ \frac{P_j}{w_j(J)} \leq q_{J_h} \right\} \right] \right]. \quad (7)$$

5 Consonance

A closed procedure is called consonant (Gabriel, 1969) if the rejection of H_J , $J \subseteq I$, further implies that at least one H_j , $j \in J$, is rejected. Consonance is a desirable property leading to short-cut procedures that give the same rejection decisions as the original closed procedure but with fewer operations in the order of m or m^2 (Grechanovsky and Hochberg, 1999). Hommel *et al.* (2007) proved that the monotonicity condition $w_j(J) \leq w_j(J')$ for all $j \in J' \subseteq J \subseteq I$, guarantees consonance if weighted Bonferroni tests are applied to all intersection hypotheses.

If a weighted parametric test is applied as in (1), Bretz *et al.* (2011) showed that

$$c_J w_j(J) \leq c_{J'} w_j(J') \text{ for all } j \in J' \subseteq J \quad (8)$$

ensures consonance. If (8) is satisfied, Algorithm 3 in Bretz *et al.* (2011) carries out the short-cut procedure. But (8) is not always satisfied even if $w_j(J) \leq w_j(J')$ for all $j \in J' \subseteq J \subseteq I$. In such cases, Bretz *et al.* (2011) proposed to modify the weighting scheme such that (8) is satisfied for a particular significance level α . However, to calculate the p -values (2) for H_J , this modification has to be satisfied for c_J under all $\alpha \in [0, 1]$, which is difficult to achieve.

The procedure by Xie (2012) considers a special weighting scheme that ensures consonance. As in Section 4.2, it assumes $w_i = w_i(I) > 0$ for all $i \in I$ and defines the weighting scheme as $w_j(J) = w_j / \sum_{j \in J} w_j$, $j \in J \subseteq I$. If the joint distribution of all test statistics is fully known, we calculate c_J and $c_{J'}$ such that

$$\Pr_{H_J} \left[\bigcup_{j \in J} \left\{ p_j \leq c_J \frac{w_j}{\sum_{j \in J} w_j} \alpha \right\} \right] = \alpha = \Pr_{H_{J'}} \left[\bigcup_{j \in J'} \left\{ p_j \leq c_{J'} \frac{w_j}{\sum_{j \in J'} w_j} \alpha \right\} \right].$$

Because $J' \subseteq J$, the above equalities can only hold if $c_J / \sum_{j \in J} w_j \leq c_{J'} / \sum_{j \in J'} w_j$, which leads to (8). Xie (2012) also made a similar assessment using p -values but did not refer to consonance explicitly. In fact, (8) continues to hold if $w_j(J) = w_j / \sum_{j \in J} w_j$, $j \in J \subseteq I$, even if the joint distribution of all test statistics is not fully known, as in (5).

Xie (2012) provided a short-cut procedure to calculate the adjusted p -value for each elementary hypothesis. Here, we simplify the algorithm and do not assume the ordering in the weighted unadjusted p -values. For the overall intersection hypothesis H_{J_1} , $J_1 = I$, we calculate its p -value \tilde{p}_{J_1} according to (3). If $\tilde{p}_{J_1} \leq \alpha$, reject H_{J_1} with the adjusted p -value $p_{J_1}^{\text{adj}} = \tilde{p}_{J_1}$ and proceed to the next step, where $j_1 = \operatorname{argmin}_{j \in J_1} \{p_j/w_j\}$; otherwise stop. In general, for $i = 2, \dots, m$, let $J_i = J_{i-1} \setminus \{j_{i-1}\}$ and calculate the p -value \tilde{p}_{J_i} for H_{J_i} . If $\tilde{p}_{J_i} \leq \alpha$, reject H_{J_i} with the adjusted p -value $p_{J_i}^{\text{adj}} = \max\{p_{J_{i-1}}^{\text{adj}}, \tilde{p}_{J_i}\}$, and proceed to the next step (as long as $i < m$), where $j_i = \operatorname{argmin}_{j \in J_i} \{p_j/w_j\}$; otherwise stop. This short-cut procedure is performed in at most m operations and can be viewed as a weighted version of the step-down parametric procedure by Dunnett and Tamhane (1991).

If we generalize the procedure by Xie (2012) to any valid weighting scheme, a sufficient condition for (8) is that $w_j(J)$ can be written as $c_J w_j$, where c_J is a constant for all $j \in J$. As a simple example, we derive a weighted version of the single-step parametric test by Dunnett (1955). The weighting scheme for the associated closed procedure is $W = \{w_J = (c_I w_j, j \in J), J \subseteq I\}$ such that H_J is rejected if $p_j \leq c_I w_j \alpha$ for any $j \in J \subseteq I$. Here, c_I is a constant for every $j \in J \subseteq I$ such that $\Pr_{H_I} \{\cup_{j \in I} (P_j \leq c_I w_j \alpha)\} = \alpha$. Assuming $\sum_{i \in I} w_i = 1$, the weighted single-step parametric test rejects H_i if $p_i \leq c_I w_i \alpha$ for any $i \in I$. The adjusted p -value for H_i is $p_i^{\text{adj}} = \Pr_{H_I} \{\cup_{j \in I} (P_j/w_j \leq p_i/w_i)\}$. The single-step parametric test by Dunnett (1955) is seen to be the special case with $w_i = 1/m$, $i \in I$.

6 Numerical results

6.1 Motivating example revisited

In this section, we illustrate the proposed approaches using the motivating example from Section 3 and compare them with the existing approaches in the literature. The R program for the key functions and numerical results is provided in the Supporting Information. We assume that the joint distribution of the test statistics for the efficacy hypotheses H_1, H_2, H_3 is trivariate normal with a mean vector of 0's. Assuming equal group sizes, the pairwise correlations between the associated test statistics are 0.5 among the efficacy hypotheses. As mentioned in Section 3, we assume that the correlations between the test statistics for the safety hypotheses H_4, H_5, H_6 are unknown. The correlations of the test statistics between the efficacy and safety hypotheses are also assumed to be unknown. Given these assumptions, the index set $I = \{1, \dots, 6\}$ is therefore partitioned into four subsets $\{1, 2, 3\}$, $\{4\}$, $\{5\}$, $\{6\}$.

We calculate the local significance levels for all intersection hypotheses using (A) the weighted Bonferroni test, (B) the weighted parametric test (5), and (C) the weighted parametric test (6). The full results are provided in the Appendix. Here, we consider H_{234} as an example. Using (A), the local significance levels for H_{234} are $(0.4, 0.2, 0.4) \times 0.025 = (0.01, 0.005, 0.01)$. Using (B), we calculate $c_{234} = 1.033$ from (5) via the `mvtnorm` package in R (Genz *et al.*, 2016). The resulting local significance levels are $(0.4, 0.2, 0.4) \times 1.033 \times 0.025 = (0.0103, 0.0052, 0.0103)$. Using (C), we calculate $c_{23} = 1.057$ and $c_4 = 1$ from (6). The resulting local significance levels are $(0.4 \times 1.057, 0.2 \times 1.057, 0.4 \times 1) \times 0.025 = (0.0106, 0.0053, 0.01)$.

From the above calculations, we can see that both parametric tests (B) and (C) produce higher local significance levels than the Bonferroni test (A). Thus, they can reject at least as many hypotheses as the Bonferroni test. Differences between (B) and (C) arise for intersection hypotheses involving two efficacy hypotheses and at least one safety hypothesis. For example, (C) preserves the level for the safety hypotheses at the level of the Bonferroni test (A) but produces higher level for the efficacy hypotheses

Table 2 Unadjusted p -values (p) and adjusted p -values (p^{adj}) (in %) from (A) Bonferroni test, (B) parametric test using (5), and (C) parametric test using (7).

Method	H_1		H_2		H_3		H_4		H_5		H_6	
	p_1	p_1^{adj}	p_2	p_2^{adj}	p_3	p_3^{adj}	p_4	p_4^{adj}	p_5	p_5^{adj}	p_6	p_6^{adj}
(A)	0.9	2.25	1.1	2.75	0.9	3.25	1.3	3.25	1.6	3.25	0.4	3.25
(B)	0.9	2.19	1.1	2.66	0.9	3.25	1.3	3.25	1.6	3.25	0.4	3.25
(C)	0.9	2.14	1.1	2.60	0.9	3.25	1.3	3.25	1.6	3.25	0.4	3.25

than (B) when testing H_{234} . On the other hand, (B) has a higher level for the safety hypotheses but a lower level for the efficacy hypotheses. These conclusions apply to all intersection hypotheses.

Given the unadjusted p -values, we can calculate the adjusted p -values and make the same rejection decisions. For (A), a short-cut procedure exists and Algorithm 2 in Bretz *et al.* (2009) can be used to implement it. For (C), we can calculate the p -value for each intersection hypothesis using (7) and take the maximum over all intersections involving a given elementary hypothesis. For (B), solutions based on numerical search are provided since no analytic formulae are available. Table 2 provides a hypothetical set of unadjusted p -values and their adjusted p -values for each of the three methods. Both parametric tests (B) and (C) give less conservative adjusted p -values than the Bonferroni test (A).

6.2 Simulation study

We conduct a simulation study to compare Type I error rates and power for the proposed weighted parametric test and the Bonferroni test. We consider three hypotheses $H_i : \theta_i \leq 0, i = 1, 2, 3$. As in the motivating example, the initial weights for H_1, H_2, H_3 are 0.4, 0.4, 0.2, respectively. The weighted Bonferroni test rejects H_i if $p_i \leq w_i \alpha, i = 1, 2, 3$, where $\alpha = 0.025$. To compare, we use the weighted single-step parametric test proposed in Section 5, which only differs from the weighted Bonferroni test by incorporating a common correlation ρ between the test statistics. In the simulations we assume $\rho = 0, 0.5, 0.9$ to represent no, moderate and high correlations, respectively. For the mean parameter θ of the test statistics, we assume multiple scenarios to include parameters under the null and the alternative hypothesis. We report the probability of rejecting each individual hypothesis as well as the probability of rejecting any of the three hypotheses. Under the null hypothesis, this is the Type I error rate while in the alternative parameter space, this is the power to reject a hypothesis. The simulation results based on 1,000,000 replicates are shown in Table 3.

We consider three scenarios of the mean parameter θ . Cases 1, 2, and 3 represent the situation where all three null hypotheses are false. Cases 4, 5, and 6 represent the situation where the first two null hypotheses are false but the third is true. Cases 7, 8, and 9 represent the situation where all three null hypotheses are true. For all mean parameters and correlations, the weighted single-step parametric test has a higher probability to reject each hypothesis compared to the weighted Bonferroni test. Under the alternative hypothesis, the power advantage to reject each hypothesis increases up to 5% as the correlation changes from 0 to 0.9. For the power to reject any of the three hypotheses, the power advantage also increases up to 3% as the correlation changes from 0 to 0.9. Under the null hypothesis (cases 7, 8, 9), the probability to reject any hypothesis π_{any} is the FWER. We can see that the weighted single-step parametric test always exhausts the nominal level $\alpha = 0.025$ but the weighted Bonferroni test is quite conservative for $\rho = 0.9$ as the actual FWER is only 0.0154. Overall, the weighted parametric test takes into account the correlation and thus is more powerful than the

Table 3 Probability of being rejected (π) for different hypotheses and scenarios ($\alpha = 0.025$) for (A) weighted Bonferroni test and (B) weighted single-step parametric test from Section 5.

Case	Method	Scenario				H_1	H_2	H_3	Any
		θ_1	θ_2	θ_3	ρ	$\pi_1(\%)$	$\pi_2(\%)$	$\pi_3(\%)$	$\pi_{\text{any}}(\%)$
1	(A)	3.4	3.4	3.4	0	85.92	85.79	79.54	99.58
	(B)	3.4	3.4	3.4	0	85.99	85.86	79.62	99.59
2	(A)	3.4	3.4	3.4	0.5	85.83	85.85	79.44	96.58
	(B)	3.4	3.4	3.4	0.5	86.81	86.82	80.57	96.94
3	(A)	3.4	3.4	3.4	0.9	85.80	85.76	79.44	90.41
	(B)	3.4	3.4	3.4	0.9	89.78	89.72	84.19	93.39
4	(A)	3.4	3.4	0	0	85.83	85.90	0.49	98.00
	(B)	3.4	3.4	0	0	85.90	85.97	0.50	98.02
5	(A)	3.4	3.4	0	0.5	85.92	85.79	0.52	94.68
	(B)	3.4	3.4	0	0.5	86.88	86.76	0.58	95.19
6	(A)	3.4	3.4	0	0.9	85.87	85.86	0.50	89.87
	(B)	3.4	3.4	0	0.9	89.83	89.86	0.83	93.00
7	(A)	0	0	0	0	0.98	1.01	0.50	2.47
	(B)	0	0	0	0	0.99	1.02	0.51	2.49
8	(A)	0	0	0	0.5	1.00	1.00	0.49	2.23
	(B)	0	0	0	0.5	1.12	1.12	0.55	2.50
9	(A)	0	0	0	0.9	1.00	0.99	0.49	1.55
	(B)	0	0	0	0.9	1.68	1.65	0.83	2.52

weighted Bonferroni test. This conclusion can be generalized to all multiple test procedures based on the Bonferroni test such as the Holm (1979) procedure and the graphical approach (Bretz *et al.*, 2009).

7 Discussion

This paper provides a unified framework under the closure principle for weighted min- p tests with generally correlated parameters. It is based on general weighting schemes and weighted min- p tests for exhaustive and nonexhaustive cases. Many procedures in the literature are special cases, including the single-step parametric test by Dunnett (1955), the step-down parametric procedure by Dunnett and Tamhane (1991), the parametric fallback procedure (Huque and Alosch, 2008), and the graphical approach with parametric assumptions (Bretz *et al.*, 2011). Within the closed procedure, it expresses the rejection decision in terms of adjusted significance levels and adjusted p -values. If a nonexhaustive weighting scheme is enforced to be exhaustive, a class of procedures is identified that increases local weights proportionally to ensure exhaustiveness.

If the parametric assumption is applied only to some subsets of hypotheses, a new method is proposed to fully utilize the parametric assumptions within each subset. An analytic formula to calculate adjusted p -values is provided that avoids solving for the critical value from an equation with multidimensional integration. With the Bonferroni split among subsets of hypotheses, the idea of using the weighted parametric test within each subset can be generalized to non-min- p tests (e.g. F or χ^2 tests) when the local significance level is preserved for that subset.

In general, consonance is not necessarily fulfilled when using weighted parametric tests such that short-cut procedures are not available and the full closure has to be tested. In the single-step and step-down procedures using weighted parametric tests, however, consonance can be enforced and

short-cut procedures are derived. For more complex MTPs, we illustrated via an example how to derive a weighting scheme using the graphical approach, for which there exists an efficient algorithm to generate the weighting scheme. In such cases, the closed parametric procedure is computationally easy to perform using the provided analytic formulae for adjusted p -values. It would be of interest to characterize flexible weighting schemes that maintain consonance when using weighted parametric tests. We leave this topic for future research.

Acknowledgments The authors are grateful to the editor, the associate editor, and the anonymous referee for their comments and suggestions that led to improvements in the manuscript.

Conflict of interest

The authors have declared no conflict of interest.

Appendix

A.1 Local significance levels for the example in Section 6.1

We provide the complete results of the local significance levels for all intersection hypotheses in the motivating example from Section 6.1. We compare three approaches: (A) the weighted Bonferroni test, (B) the weighted parametric test (5), and (C) the weighted parametric test (6). We assume that the joint distribution of the test statistics for the efficacy hypotheses H_1, H_2, H_3 is trivariate normal with a mean vector of 0's. Assuming equal group sizes, the pairwise correlations between the test statistics are 0.5 among the efficacy hypotheses. As mentioned in Section 6.1, all other correlations are assumed to be unknown. The full correlation matrix in this setting is provided in the following matrix, with the unknown correlations denoted as NA.

$$\begin{bmatrix} 1 & 0.5 & 0.5 & \text{NA} & \text{NA} & \text{NA} \\ 0.5 & 1 & 0.5 & \text{NA} & \text{NA} & \text{NA} \\ 0.5 & 0.5 & 1 & \text{NA} & \text{NA} & \text{NA} \\ \text{NA} & \text{NA} & \text{NA} & 1 & \text{NA} & \text{NA} \\ \text{NA} & \text{NA} & \text{NA} & \text{NA} & 1 & \text{NA} \\ \text{NA} & \text{NA} & \text{NA} & \text{NA} & \text{NA} & 1 \end{bmatrix}$$

Given this joint distribution, we calculate the local significance levels for every intersection hypothesis $H_J, J \subseteq I$, in Tables A1 and A2.

Table A1 Local significance levels (in %) from A: Bonferroni test (B: parametric test using (5), C: parametric test using (6)) where $\alpha = 0.25$.

J	Local significance levels $w_J(J)\alpha$					
	H_1	H_2	H_3	H_4	H_5	H_6
{1, 2, 3, 4, 5, 6}	1(1.12,1.12)	1(1.12,1.12)	0.5(0.56,0.56)	0*	0	0
{1, 2, 3, 4, 5}	1(1.12,1.12)	1(1.12,1.12)	0.5(0.56,0.56)	0	0	—
{1, 2, 3, 4, 6}	1(1.12,1.12)	1(1.12,1.12)	0.5(0.56,0.56)	0	—	0
{1, 2, 3, 5, 6}	1(1.12,1.12)	1(1.12,1.12)	0.5(0.56,0.56)	—	0	0
{1, 2, 4, 5, 6}	1(1.06,1.07)	1(1.06,1.07)	—	0	0	0.5(0.53,0.5)
{1, 3, 4, 5, 6}	1(1.03,1.06)	—	0.5(0.52,0.53)	0	1(1.03,1)	0

Table A1 Continued

J	Local significance levels $w_j(J)\alpha$					
	H_1	H_2	H_3	H_4	H_5	H_6
{2, 3, 4, 5, 6}	–	1(1.03,1.06)	0.5(0.52,0.53)	1(1.03,1)	0	0
{1, 2, 3, 4}	1(1.12,1.12)	1(1.12,1.12)	0.5(0.56,0.56)	0	–	–
{1, 2, 3, 5}	1(1.12,1.12)	1(1.12,1.12)	0.5(0.56,0.56)	–	0	–
{1, 2, 3, 6}	1(1.12,1.12)	1(1.12,1.12)	0.5(0.56,0.56)	–	–	0
{1, 2, 4, 5}	1.25(1.35,1.35)	1.25(1.35,1.35)	–	0	0	–
{1, 2, 4, 6}	1(1.06,1.07)	1(1.06,1.07)	–	0	–	0.5(0.53,0.5)
{1, 2, 5, 6}	1(1.06,1.07)	1(1.06,1.07)	–	–	0	0.5(0.53,0.5)
{1, 3, 4, 5}	1(1.03,1.06)	–	0.5(0.52,0.53)	0	1(1.03,1)	–
{1, 3, 4, 6}	1.5(1.61,1.61)	–	1(1.08,1.08)	0	–	0
{1, 3, 5, 6}	1(1.03,1.06)	–	0.5(0.52,0.53)	–	1(1.03,1)	0
{1, 4, 5, 6}	1	–	–	0	1	0.5
{2, 3, 4, 5}	–	1(1.03,1.06)	0.5(0.52,0.53)	1(1.03,1)	0	–
{2, 3, 4, 6}	–	1(1.03,1.06)	0.5(0.52,0.53)	1(1.03,1)	–	0
{2, 3, 5, 6}	–	1.5(1.61,1.61)	1(1.08,1.08)	–	0	0
{2, 4, 5, 6}	–	1	–	1	0	0.5
{3, 4, 5, 6}	–	–	0.5	1	1	0
{1, 2, 3}	1(1.12,1.12)	1(1.12,1.12)	0.5(0.56,0.56)	–	–	–
{1, 2, 4}	1.25(1.35,1.35)	1.25(1.35,1.35)	–	0	–	–
{1, 2, 5}	1.25(1.35,1.35)	1.25(1.35,1.35)	–	–	0	–
{1, 2, 6}	1(1.06,1.07)	1(1.06,1.07)	–	–	–	0.5(0.53,0.5)
{1, 3, 4}	1.5(1.61,1.61)	–	1(1.08,1.08)	0	–	–
{1, 3, 5}	1(1.03,1.06)	–	0.5(0.52,0.53)	–	1(1.03,1)	–
{1, 3, 6}	1.5(1.61,1.61)	–	1(1.08,1.08)	–	–	0
{1, 4, 5}	1.25	–	–	0	1.25	–
{1, 4, 6}	1.5	–	–	0	–	1
{1, 5, 6}	1	–	–	–	1	0.5

Notes: No parenthesis means that the local significance levels produced by the three procedures are the same.

Table A2 Local significance levels (in %) from A: Bonferroni test (B: parametric test using (5), C: parametric test using (6)) where $\alpha = 0.25$.

J	Local significance levels $w_j(J)$					
	H_1	H_2	H_3	H_4	H_5	H_6
{2, 3, 4}	–	1(1.03,1.06)			–	–
			0.5(0.52,0.53)	1(1.03,1)		
{2, 3, 5}	–	1.5(1.61,1.61)	1(1.08,1.08)	–	0	–
{2, 3, 6}	–	1.5(1.61,1.61)	1(1.08,1.08)	–	–	0
{2, 4, 5}	–	1.25	–	1.25	0	–

Table A2 Continued

J	Local significance levels $w_j(J)$					
	H_1	H_2	H_3	H_4	H_5	H_6
{2, 4, 6}	–	1	–	1	–	0.5
{2, 5, 6}	–	1.5	–	–	0	1
{3, 4, 5}	–	–	0.5	1	1	–
{3, 4, 6}	–	–	1	1.5	–	0
{3, 5, 6}	–	–	1	–	1.5	0
{4, 5, 6}	–	–	–	1	1	0.5
{1, 2}	1.25(1.35,1.35)	1.25(1.35,1.35)	–	–	–	–
{1, 3}	1.5(1.61,1.61)	–	1(1.08,1.08)	–	–	–
{1, 4}	2.5	–	–	0	–	–
{1, 5}	1.25	–	–	–	1.25	–
{1, 6}	1.5	–	–	–	–	1
{2, 3}	–	1.5(1.61,1.61)	1(1.08,1.08)	–	–	–
{2, 4}	–	1.25	–	1.25	–	–
{2, 5}	–	2.5	–	–	0	–
{2, 6}	–	1.5	–	–	–	1
{3, 4}	–	–	1	1.5	–	–
{3, 5}	–	–	1	–	1.5	–
{3, 6}	–	–	2.5	–	–	0
{4, 5}	–	–	–	1.25	1.25	–
{4, 6}	–	–	–	1.5	–	1
{5, 6}	–	–	–	–	1.5	1
{1}	2.5	–	–	–	–	–
{2}	–	2.5	–	–	–	–
{3}	–	–	2.5	–	–	–
{4}	–	–	–	2.5	–	–
{5}	–	–	–	–	2.5	–
{6}	–	–	–	–	–	2.5

Notes: No parenthesis means that the local significance levels produced by the three procedures are the same.

References

- Bauer, P., Brannath, W. and Posch, M. (2001). Multiple testing for identifying effective and safe treatments. *Biometrical Journal* **43**, 605–616.
- Bretz, F., Maurer, W., Brannath, W. and Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* **28**, 586–604.
- Bretz, F., Posch, M., Glimm, E., Klinglmueller, F., Maurer, W. and Rohmeyer, K. (2011). Graphical approaches for multiple comparison procedures using weighted Bonferroni, simes, or parametric tests. *Biometrical Journal* **53**, 894–913.
- Dmitrienko, A., Tamhane, A. C. and Wiens, B. L. (2008). General multistage gatekeeping procedures. *Biometrical Journal* **50**, 667–677.
- Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* **50**, 1096–1121.
- Dunnett, C. W. and Tamhane, A. C. (1991). Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Statistics in Medicine* **10**, 939–947.

- Gabriel, K. R. (1969). Simultaneous test procedures—some theory of multiple comparisons. *The Annals of Mathematical Statistics* **40**, 224–250.
- Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F. and Hothorn, T. (2016). mvtnorm: multivariate normal and t distributions. R package version 1.0-5.
- Goteti, S., Hirawat, S., Massacesi, C., Fretault, N., Bretz, F. and Dharan, B. (2014). Some practical considerations for phase iii studies with biomarker evaluations. *Journal of Clinical Oncology* **32**, 854–855.
- Grechanovsky, E. and Hochberg, Y. (1999). Closed procedures are better and often admit a shortcut. *Journal of Statistical Planning and Inference* **76**, 79–91.
- Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. John Wiley & Sons, Inc., New York, NY, USA.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.
- Hommel, G., Bretz, F. and Maurer, W. (2007). Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies. *Statistics in Medicine* **26**, 4063–4073.
- Huque, M. F. and Alosh, M. (2008). A flexible fixed-sequence testing method for hierarchically ordered correlated multiple endpoints in clinical trials. *Journal of Statistical Planning and Inference* **138**, 321–335.
- Marcus, R., Peritz, E. and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.
- Maurer, W. and Bretz, F. (2013). Memory and other properties of multiple test procedures generated by entangled graphs. *Statistics in Medicine* **32**, 1739–1753.
- Rohmeyer, K. and Klinglmueller, F. (2015). *gMCP: Graph Based Multiple Test Procedures*. R package version 0.8–10.
- Westfall, P. H., Krishen, A. and Young, S. S. (1998). Using prior information to allocate significance levels for multiple endpoints. *Statistics in Medicine* **17**, 2107–2119.
- Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, Volume 279. John Wiley & Sons, New York, NY, USA.
- Xi, D. and Tamhane, A. C. (2014). A general multistage procedure for k-out-of-n gatekeeping. *Statistics in Medicine* **33**, 1321–1335.
- Xie, C. (2012). Weighted multiple testing correction for correlated tests. *Statistics in Medicine* **31**, 341–352.