Research Article

Statistics in Medicine

Received 7 December 2009, Accepted 28 April 2010 Published online 3 February 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.3988

Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures

Frank Bretz[a,b*†], Willi Maurer[a] and Gerhard Hommel[c]

A variety of powerful test procedures are available for the analysis of clinical trials addressing multiple objectives, such as comparing several treatments with a control, assessing the benefit of a new drug for more than one endpoint, etc. However, some of these procedures have reached a level of complexity that makes it difficult to communicate the underlying test strategies to clinical teams. Graphical approaches have been proposed instead that facilitate the derivation and communication of Bonferroni-based closed test procedures. In this paper we give a coherent description of the methodology and illustrate it with a real clinical trial example. We further discuss suitable power measures for clinical trials with multiple primary and/or secondary objectives and use a generic example to illustrate our considerations. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: Bonferroni; closure principle; gatekeeping procedure; multiple comparison procedure; multiple testing; shortcut procedure; step-down procedure

1. Introduction

Multiple test procedures are often used in the analysis of clinical trials addressing multiple objectives, such as comparing several treatments with a control, assessing the benefit of a new drug for more than one endpoint, etc. In the following we consider multiple test procedures control the familywise error rate (FWER) in the strong sense. That is, the probability to erroneously reject at least that one true null hypothesis is controlled at a pre-specified significance level $\alpha \in (0, 1)$ under any configuration of true and false null hypotheses. Over the past decades various powerful test procedures have been developed that allow one to map the relative importance of the different study objectives as well as their relation onto an appropriately tailored multiple test procedure. Examples of such procedures include the weighted or unweighted Bonferroni–Holm procedure [1], fixed sequence tests [2, 3], the fallback procedure [4, 5], and gatekeeping procedures [3, 6, 7].

In the meantime, some of these procedures have reached a level of complexity that makes it difficult to communicate the underlying test strategies to clinical teams. However, many of these procedures belong to a subclass of weighted Bonferroni-based closed test procedures [8] that fulfill a mild monotonicity condition on the weights [9]. Based on this result [10, 11] independently derived graphical approaches that facilitate the derivation and communication of Bonferroni-based closed test procedures. Graphical approaches are often easier to communicate to clinical teams than long and abstract decision tables. Using graphs, one can better explore different test strategies together with the clinical team and thus tailor the multiple test procedure to the given study objectives.

[a] *Statistical Methodology, Novartis Pharma AG, Lichstr. 35, CH-4002 Basel, Switzerland*
[b] *Institute for Biometry, Hannover Medical School, 30623 Hannover, Germany*
[c] *Institute of Medical Biostatistics, Epidemiology and Informatics, University of Mainz, Mainz, Germany*
[*] *Correspondence to: Frank Bretz, Statistical Methodology, Novartis Pharma AG, Lichstr. 35, 4002 Basel, Switzerland.*
[†] *E-mail: frank.bretz@novartis.com*

In this paper we review the current methodology of Bonferroni-based closed test procedures, including related shortcut procedures and construction of simultaneous confidence intervals. We illustrate the methods in detail with a real clinical trial. We describe how to determine a suitable multiple test procedure tailored to the study objectives and how to perform resulting shortcut procedures. A special focus lies on the graphical approaches mentioned above. Finally, we give some guidance on suitable power measures in clinical trials with multiple primary and/or secondary objectives and use a generic example to illustrate our considerations.

2. Bonferroni-based closed test procedures

Understanding the closure principle enables one to take full advantage of its flexibility and to tailor multiple test procedures to given study objectives. In the following we will (i) describe the class of Bonferroni-based closed test procedures; (ii) give a sufficient characterization to derive sequentially rejective multiple test procedures based on shortcuts of the closure tree and demonstrate that many common multiple test procedures are special cases thereof; (iii) construct simultaneous confidence intervals; and (iv) provide graphical tools that facilitate the derivation and communication of these procedures. In order to keep the description at a reasonable size, we omit the technical details and refer to the original publications instead, see also [12] for a recent review.

2.1. *Class of Bonferroni-based closed test procedures*

Consider the problem of testing m elementary hypotheses $H_1, \ldots, H_m$ and let $I = \{1, \ldots, m\}$ denote the associated index set. To keep the discussion simple, assume that the elementary hypotheses satisfy the free combination condition, i.e. for any subset $J \subseteq I$ the simultaneous truth of $H_i, i \in J$, and falsehood of the remaining hypotheses is possible [1]. For related results under restricted combinations, where the previous condition does not hold, we refer to [13].

Applying the closure principle [8] leads to the consideration of the intersection hypotheses $H_J = \bigcap_{j \in J} H_j$, $J \subseteq I$. For each intersection hypothesis H_J assume a collection of weights $w_j(J)$ such that $0 \leqslant w_j(J) \leqslant 1$ and $\sum_{j \in J} w_j(J) \leqslant 1$. These weights quantify the relative importance of the hypotheses H_j included in the intersection H_J. Finally, let p_j denote the unadjusted P-value for H_j, $j \in I$.

Assume that each intersection hypothesis H_J is tested with a weighted Bonferroni test. Consequently, one can obtain P-values

$$p_J = \min\{q_j(J) : j \in J\}$$

for the weighted Bonferroni test for H_J, where

$$q_j(J) = \begin{cases} \min\{1, p_j / w_j(J)\} & \text{if } w_j(J) > 0, \\ 1 & \text{if } w_j(J) = 0. \end{cases}$$

An intersection hypothesis H_J is rejected if $p_J \leqslant \alpha$, where $\alpha \in (0, 1)$ denotes the pre-specified overall significance level. This defines the class $\mathscr{B}$ of all closed test procedures that use weighted Bonferroni tests for each intersection hypothesis. Any collection of weights subject to the above constraints can be used. Thus, one can tailor the weights to the given study objectives and maximize the probability of a successful trial. Many standard multiple test procedures belong to the class $\mathscr{B}$, such as the weighted Bonferroni–Holm procedure [1], fixed sequence tests [2, 3], the fallback procedure [4, 5], and many standard Bonferroni-based gatekeeping procedures [9].

2.2. *Sequentially rejective Bonferroni-based closed test procedures*

With the class $\mathscr{B}$ defined in Section 2.1, one can further show that under a mild monotonicity condition on the weights $w_j(J)$ the closure principle leads to consonant multiple test procedures. Thus, shortcut versions can be derived that substantially simplify the implementation and interpretation of the related procedures. Hommel *et al.* [9] showed that all the specific approaches mentioned in Section 2.1 belong to a subclass $\mathscr{S} \subset B$ of shortcut procedures characterized by the monotonicity condition

$$w_j(J) \leqslant w_j(J') \quad \text{for all } J' \subseteq J \subseteq I \text{ and } j \in J'.$$

This condition ensures consonance, i.e. if an intersection hypothesis H_J is rejected, there is an index $j \in J$, such that the elementary hypothesis H_j can be rejected as well. Therefore, shortcut procedures of order m can be constructed such that the elementary hypotheses $H_1, \ldots, H_m$ are tested in m steps instead of testing all $2^m - 1$ intersection hypotheses as usually required by the closure principle. This simplification is a key characterization of the Bonferroni–Holm procedure and the results from [9] ensure that this remains true for *any* procedure in $\mathscr{S}$. As a consequence, shortcut procedures from $\mathscr{S}$ can be carried out with the following m-step procedure. Start testing the global intersection hypothesis $H_I, I = \{1, \ldots, m\}$. If it is rejected, there is an index $i \in I$ as described above such that H_i is rejected by the closed test procedure. At the next step, continue testing the global intersection $H_{I \setminus i}$ of the remaining, not yet rejected hypotheses, and so on, until the first non-rejection. Similar arguments were applied in [14, p. 55] to union–intersection tests and in [13] to restricted hypotheses. Moreover, in [15] it is shown that the parallel gatekeeping procedures proposed in [7] admit a shortcut.

2.3. Simultaneous confidence intervals

The previous characterization of the class $\mathscr{S}$ can also be used to construct compatible simultaneous confidence intervals [16, 17]. Consider the one-sided null hypotheses $H_i: \ \theta_i \leqslant \delta_i, \ i \in I = \{1, \ldots, m\}$, where $\theta_1, \ldots, \theta_m$ are the parameters of interest and $\delta_1, \ldots, \delta_m$ are pre-specified constants (e.g. noninferiority margins). Let $\alpha_j(J) = \alpha w_j(J)$ denote the local significance levels with $j \in J \subseteq I$. Further, let $L_i(\bar{\alpha})$ denote the marginal lower confidence limit for θ_i at level $1 - \bar{\alpha}$, $i = 1, \ldots, m$. Finally, let R denote the index set of hypotheses rejected by a multiple test procedure from $\mathscr{S}$. Then, lower one-sided confidence bounds for $\theta_1, \ldots, \theta_m$ with simultaneous coverage probability of at least $1 - \alpha$ are given by

$$\widetilde{L}_i = \begin{cases} \delta_i & \text{if } i \in R \text{ and } R \neq I, \\ L_i(\bar{\alpha}_i) & \text{if } i \notin R, \\ \max(\delta_i, L_i(\bar{\alpha}_i)) & \text{if } R = I, \end{cases}$$

where $\bar{\alpha}_i = \alpha_i(I \setminus R)$ if $i \notin R \neq I$. In the case $R = I$, where all hypotheses can be rejected, the choice of the local levels $\bar{\alpha}_i = \alpha_i(\emptyset)$ is free [16]. Thus, in order to compute the simultaneous confidence limits, one only needs to know the set R of rejected hypotheses and the corresponding local levels $\bar{\alpha}_i$ for all indices i of retained hypotheses. This construction method can be used to derive simultaneous confidence intervals for the Bonferroni, Bonferroni–Holm, fixed-sequence, fallback and other Bonferroni-based gatekeeping procedures.

Note that if not all hypotheses are rejected, the confidence bounds associated with the rejected hypotheses reflect the test decision $\theta_i > \delta_i$ and the confidence limits associated with the retained hypotheses are the marginal confidence limits at level $\alpha_i(I \setminus R)$. In other words, unless $R = I$, the simultaneous confidence intervals for the rejected hypotheses do not provide any further information beyond the test decision, which limits their practical use. However, it is possible to get more informative simultaneous confidence intervals if non-exhaustive tests are used. Such confidence intervals have been investigated in [12, 16, 18].

2.4. Graphical visualization

In Section 2.3 we mentioned that the class $\mathscr{S}$ includes various Bonferroni-based test procedures, including many fixed-sequence, fallback, and gatekeeping procedures. Using procedures in this class, one can map the difference in importance as well as the relationship between various study objectives onto an adequate multiple test procedure. However, since the procedures are based on the closure principle, one needs to specify the weights $w_j(J)$ for each of the $2^m - 1$ intersection hypotheses $H_J, J \subseteq I$. Unless these weights follow some simple and well-known specification rules (such as the Bonferroni–Holm procedure), the underlying test strategy may be difficult to communicate to clinical team members. Conversely, communication of weights reflecting relative importance of and relationship among several hypotheses is only possible in simple situations.

Graphical tools have been proposed instead that help to construct and visualize different sequentially rejective test strategies and thus to best tailor a multiple test procedure to given study objectives [10, 11]. Using a graphical approach, the hypotheses $H_1, \ldots, H_m$ are represented by vertices with associated weights denoting the local significance levels $\alpha_1, \ldots, \alpha_m$. The weight associated with a directed edge between any two vertices indicates the fraction of the (local) significance level that is shifted once the hypothesis at the tail of the edge has been rejected.

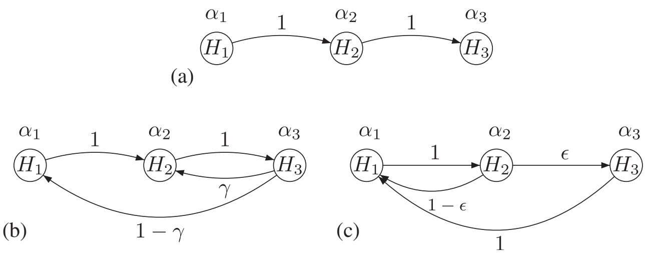

Figure 1. Graphical illustration of the fallback procedure (top) and two extensions (bottom).

For illustration, consider testing three null hypotheses H_1, H_2, and H_3. Figure 1(a) displays the fallback procedure of [4]. Each of the hypotheses is assigned its associated local significance level α_i, such that $\alpha_1+\alpha_2+\alpha_3=\alpha$. If H_1 is rejected, then the level α_1 is carried over to H_2, as indicated by the edge pointing from H_1 to H_2. If H_2 is rejected at its local significance level (either α_2 or $\alpha_1+\alpha_2$), then that level is carried over to H_3, as indicated by the edge pointing from H_2 to H_3.

It is important to note that graphical tools of this kind also help derive other, potentially more powerful test strategies. For example, one concludes from Figure 1(a) that if H_3 is rejected its local significance level α_3 is not passed on to any other hypothesis. Figure 1(b) shows a simple improvement by shifting α_3 along the two edges pointing back to H_1 and H_2, where $\gamma=\alpha_2/(\alpha_1+\alpha_2)$. The resulting test procedure is equivalent to the α-exhaustive extension of the fallback procedure introduced in [5]. Figure 1(c) displays yet another extension of the fallback procedure by shifting the significance level to the first hypothesis in the hierarchy that has not been rejected so far [19]. Here, ε denotes an infinitesimally small weight, indicating that the significance level is carried over from H_2 to H_3 only if both H_1 and H_2 are rejected. The motivation for this extension is that H_1 is deemed more important than H_3. Thus, once H_2 is rejected, its associated significance level should be carried over first to H_1 before continue testing H_3.

Graphical visualizations are available for many sequentially rejective multiple test procedures in $\mathscr{S}$. Any graph, like the ones shown in Figure 1, satisfies the monotonicity condition from Section 2.2 and can be used iteratively when applying the actual test procedure to a given set of observed P-values, as long as the update algorithm from [10] is used. This will be demonstrated in more detail with a case study in Section 3.

3. Application to a case study

In this section we use a real Phase III clinical study to illustrate some of the ideas of Section 2. In Section 3.1 we provide the relevant background of the study. In Section 3.2 we describe how to convert the requirements from the clinical team into a multiple test procedure tailored to the study objectives. In Section 3.3 we illustrate the iterative use of the graphical approach with a numerical example.

3.1. Background

This case study is a 24-week multicenter, randomized, double-masked, placebo controlled study, and is one of two pivotal trials from a Phase III development program. The purpose of this pivotal trial is to evaluate a new compound as an adjunctive therapy to reduce the relapse rate during the 24 weeks of study therapy as compared to standard-of-care alone.

Three dose regimens of the new compound are to be compared with placebo adjunctive to standard-of-care therapy. The primary efficacy variable is the difference in proportion of patients with recurrence at 24 weeks between the three active arms and placebo. The key secondary variable is the difference in mean change in total medication score from baseline to 24 weeks between the three active arms and placebo. This gives six statistical hypotheses H_{ij}, $i=1,2,3$ (three treatment–control comparisons), $j=1,2$ (two endpoints). Further secondary analyses will be conducted at study end, but are not accounted in the FWER control.

3.2. Multiple test procedure based on the graphical approach

The six null hypotheses H_{ij} are tested using a gatekeeping procedure based on the graphical approach described in Section 2.4. Significance levels α_{ij} are initially defined such that they sum up to α. As

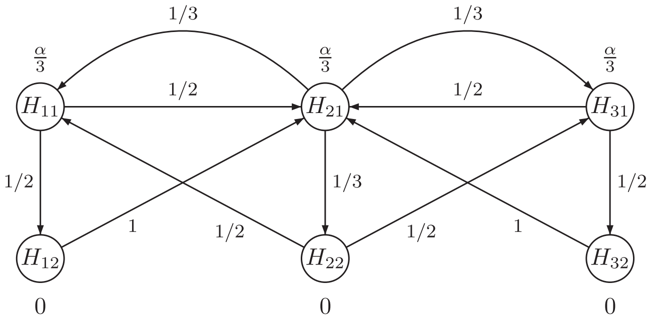

Figure 2. Graphical illustration of the multiple test procedure from the case study.

described in Section 2, the procedure is then as follows: Test the hypotheses H_{ij} each at its local significance level α_{ij}. If a hypothesis H_{ij} can be rejected, reallocate its level to one of the other hypotheses according to pre-specified rules represented by a weighted graph. Update the reallocation weights in the reduced graph and repeat the test step for the remaining, non-rejected hypotheses with the updated local significance levels. This possibly leads to further rejected null hypotheses with associated reallocation of the local significance levels. The procedure is repeated until no further hypothesis can be rejected. Details on the updating process are given further below.

The practically relevant question is how to derive a suitable multiple test procedure. In the following we describe how the initial requirements from the clinical team can be converted into a powerful multiple test procedure by determining a suitable graph tailored to the study objectives. The graph defining the finally selected multiple test procedure is given in Figure 2. The three primary hypotheses H_{11}, H_{21}, and H_{31} will be treated as the ‘gatekeepers’ and are allocated the levels $\alpha_{11}=\alpha_{21}=\alpha_{31}=\alpha/3$. The initial levels of the secondary hypotheses H_{12}, H_{22}, and H_{32} are $\alpha_{12}=\alpha_{22}=\alpha_{32}=0$, indicating that they will only be tested if at least one of the ‘gatekeeping’ primary hypotheses has been rejected before. This multiple test procedure is the result after regular discussions with the clinical team and motivated by the following considerations.

(I) The primary hypotheses H_{11}, H_{21}, and H_{31} are considered to be more important than the key secondary hypotheses H_{12}, H_{22}, and H_{32}. Thus, in the initial step only the primary hypotheses are tested. Once a primary hypothesis is rejected, the associated significance level is split: One part is passed on to the remaining primary hypotheses and the other part is passed on to the associated key secondary hypothesis. This reflects the preference to test additional primary hypotheses instead of testing only the key secondary hypotheses associated with an already rejected primary hypothesis at a higher significance level.

(II) The primary hypotheses H_{11}, H_{21}, and H_{31} are treated equally, except for the fact that it is preferred to have two adjacent doses becoming significant (i.e. reject H_{11}, H_{21} or H_{21}, H_{31}) rather than having two non-adjacent doses becoming significant (i.e. reject H_{11} and H_{31}). As a consequence, the initial local significance levels are equal (i.e. $\alpha_{11}=\alpha_{21}=\alpha_{31}=\alpha/3$), but there are no edges connecting H_{11} and H_{31}. Similarly, the secondary hypotheses are treated equal (i.e. $\alpha_{12}=\alpha_{22}=\alpha_{32}=0$) and the edges pointing back to the primary hypotheses again reflect the preference for testing adjacent doses.

(III) We do not want to reject key secondary hypotheses without having rejected the associated primary hypothesis. As a consequence, there are no edges from any primary hypothesis to any non-associated key secondary hypothesis (e.g. there is no edge connecting H_{11} and H_{22}). Also, there are no edges connecting the key secondary hypotheses among themselves in the initial graph.

The test procedure is fully determined by the initial weighted graph from Figure 2, where the elementary hypotheses are represented by vertices with associated weights representing the local significance levels. The weight associated with a directed edge between any two vertices indicates the fraction of the (local) significance level at the initial vertex (tail) that is added to the significance level at the terminal vertex (head) if the hypothesis at the tail is rejected. Note that weights of edges not displayed in a graph are 0. Together with Algorithm 1 from [10] for sequentially updating the graph after rejection of a hypothesis, this approach controls the FWER strongly at level α.

Assume, for example, that in Figure 2 the hypothesis H_{11} is rejected. Then, the associated significance level $\alpha_{11}=\alpha/3$ is split equally into two parts. One half is passed on to H_{21} (because of (II) from above), the other half is passed on to the associated key secondary hypothesis H_{12} (because of (I) and (III)). Similarly, if H_{31} is rejected, the associated significance level $\alpha_{31}=\alpha/3$ is split equally and passed on to H_{21} and H_{32}. Otherwise, if H_{21} is rejected, the associated level $\alpha_{21}=\alpha/3$ is split equally into three parts. Two parts are passed on equally to H_{11} and H_{31} (because of (I) and (II)), the third is passed on to the associated key secondary hypothesis H_{22} (because of (I) and (III)).

Note that the test procedure from Figure 2 satisfies by construction the monotonicity condition from Section 2.2 and therefore defines a shortcut procedure. As shown in [10], the test decisions are independent of the rejection sequence. That is, if at any point more than one hypothesis could be rejected, the choice of the hypothesis does not influence the total set of hypotheses that eventually can be rejected. The initial graph and the algorithm unequivocally define the multiple test strategy.

Finally, we note that alternative test strategies were discussed with the clinical team, but discarded for different reasons. For example, a fixed sequence test based on a-priori ordered hypotheses [2, 3] was rejected as the team wanted to mitigate the risk of stopping the analysis early in the hierarchy due to a non-significant result among the first hypotheses. In addition, it was considered using a Dunnett test for the primary hypotheses instead of the proposed Bonferroni test to account for the correlations among the test statistics for H_{11}, H_{21}, and H_{31}. However, the resulting procedure may not longer be consonant unless the critical values are adjusted properly.

3.3. Numerical example for the iterative use of the graphical approach

We now illustrate how to apply iteratively the graph from Figure 2 when analyzing a set of observed P-values. The results from Section 2 ensure that shortcuts are available which test the six hypotheses H_{ij} in at most six steps, despite the fact that the underlying closure tree contains $2^6-1=63$ intersection hypotheses.

Assume that the initial significance levels α_{ij} are $\alpha_{11}=\alpha_{21}=\alpha_{31}=\alpha/3=0.0083$, where $\alpha=0.025$ (one-sided), and $\alpha_{12}=\alpha_{22}=\alpha_{32}=0$. Assume further that at the final analysis unadjusted one-sided P-values p_{11}, p_{21}, and p_{31} for the three primary hypotheses and P-values p_{12}, p_{22}, and p_{32} for the three key secondary hypotheses are observed. At each step of the test procedure these P-values are compared to the local significance level. Hence, given the specific initial significance levels above, at least H_{11}, H_{21}, or H_{31} must be rejected at level $\alpha/3=0.0083$ in order to test the other hypotheses at updated levels. In the following we illustrate the sequential graphical approach assuming the one-sided P-values $p_{11}=0.1$, $p_{21}=0.008$, $p_{31}=0.005$, $p_{12}=0.15$, $p_{22}=0.04$, and $p_{32}=0.006$.

Starting from the initial graph in Figure 2, one can reject H_{31} because $p_{31}=0.005<0.0083=\alpha/3$. Note that $p_{21}<\alpha/3$ and one could start rejecting H_{21} instead of rejecting H_{31}, because at the first step one essentially applies a Bonferroni procedure to test the primary hypotheses H_{11}, H_{21}, or H_{31}. Either way, the final decisions remain the same, as noted in Section 3.2.

After rejecting H_{31}, the associated node can be removed in Figure 2 and the graph is updated to the one in Figure 3 with the remaining five hypotheses. To this end, the local significance level $\alpha_{31}=\alpha/3$ is split into half and shifted to H_{21} and H_{32}, resulting in the updated levels $\alpha_{21}=\alpha/3+\alpha/6=\alpha/2$ and $\alpha_{32}=\alpha/6$, respectively. The weights on the edges are updated accordingly; see Algorithm 1 from [10]. As an example, consider updating the weight for the edge $H_{21}\rightarrow H_{11}$ after having rejected H_{31}. For short, let $i=[21]$, $j=[11]$, and $r=[31]$. Then, the updated weight $g_{ij}=g_{[21],[11]}$ for the edge $H_{21}\rightarrow H_{11}$ is

$$\frac{g_{ij}+g_{ir}g_{rj}}{1-g_{ir}g_{ri}}=\frac{\frac{1}{3}+\frac{1}{3}0}{1-\frac{1}{3}\frac{1}{2}}=\frac{2}{5}.$$

Note that edges between hypotheses that have not been present in the original graph can now appear, such as the one from H_{22} to H_{32} in Figure 3, which results from connecting the previous edges $H_{22}\rightarrow H_{31}$ and $H_{31}\rightarrow H_{32}$.

Next, because $p_{21}=0.008<0.0125=\alpha/2$, one can also reject H_{21}. The graph is updated to the one displayed in Figure 4, where now the four remaining hypotheses H_{11}, H_{12}, H_{22}, and H_{32} can be tested simultaneously at the indicated levels. If in Figure 3 H_{21} would not have been significant, one could still test H_{11} at level $\alpha/3$ and H_{32} at level $\alpha/6$. If H_{11} could be rejected, one could test H_{21} at level $2\alpha/3$ according to the sequential graphical approach. If H_{32} could be rejected, one could test H_{21} at level $2\alpha/3$. Otherwise, the procedure stops at this stage.

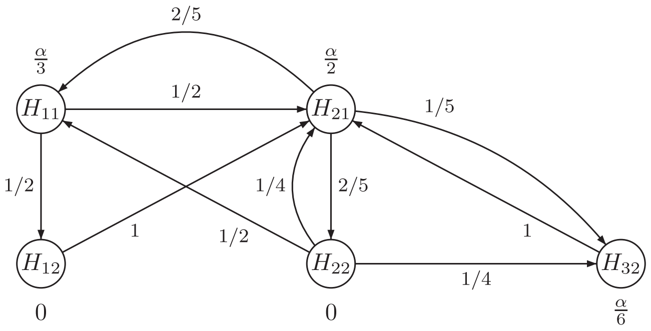

Figure 3. Updated graph after rejecting H_{31}.

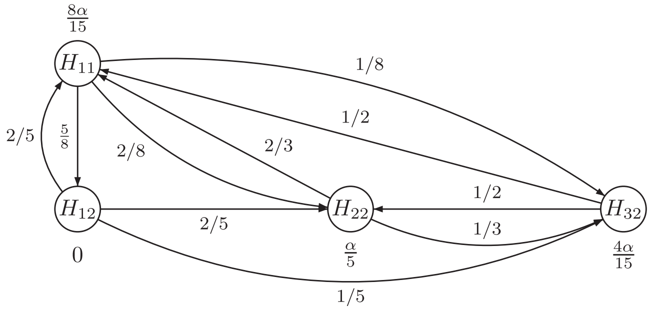

Figure 4. Updated graph after sequentially rejecting H_{31} and H_{21}.

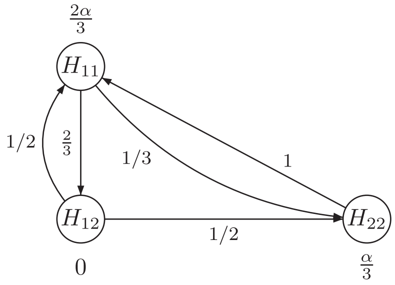

Figure 5. Updated graph after sequentially rejecting H_{31}, H_{21}, and H_{32}.

With the updated graph in Figure 4, one sees that $p_{32}=0.006<0.0067=4\alpha/15$ and can now reject H_{32}, which was not possible in the previous steps. After the next update (Figure 5), no further hypothesis can be rejected and the procedure stops.

It should be noted that in practice the iterations from Figures 3 through 5 will be conducted automatically by a simple program implementing Algorithm 1 from [10]. In the Appendix we provide SAS/IML code that can be used to reproduce the results from Figure 5. The value of the previous discussion is to get an understanding of the possible decision paths for a given multiple test procedure. By doing this, the characteristics of competing test strategies can be assessed and compared to each other. This will be further explored in Section 4.

We conclude this section by computing the simultaneous confidence intervals for the example above. For simplicity, let $H_{ij}\colon\theta_{ij}\leqslant 0$, $i=1,2,3$, and $j=1,2$. We assume the observed standardized treatment differences $X_{ij}=\Phi^{-1}(1-p_{ij})$ to be (asymptotically) normal distributed, i.e. $X_{ij}\sim N(\theta_{ij},1)$ and p_{ij} given above. We compute the simultaneous lower one-sided confidence bounds for θ_{ij} using the results from Section 2.3. According to Figure 5, H_{21}, H_{31}, and H_{32} are all rejected at level $\alpha=0.025$ and $\widetilde{L}_{21}=\widetilde{L}_{31}=\widetilde{L}_{32}=0$. Furthermore, the last local significance levels at which H_{11} and H_{22} were tested are $\bar{\alpha}_{11}=2\alpha/3=0.0167$ and $\bar{\alpha}_{22}=0.0083$, respectively. Thus, $\widetilde{L}_{11}=X_{11}-z_{1-2\alpha/3}=1.2816-2.128=-0.8466$, where $z_{\gamma}=\Phi^{-1}(\gamma)$. Similarly, $\widetilde{L}_{22}=X_{22}-z_{1-\alpha/3}=1.7507-2.394=-0.6433$. Because $\bar{\alpha}_{12}=0$, H_{12} is not tested and $\widetilde{L}_{12}=-\infty$.

4. Power considerations

A common requirement for any statistical test is to maximize the power for a given significance level α. Power considerations are thus an integral part of designing clinical studies [20–22]. However, the traditional power concept can be generalized in various ways when moving from single to multiple hypotheses test problems. Several authors have introduced a plethora of different power concepts [23–26] and it is not always clear which of these concepts is best suited in practice. In this section we provide some considerations for power calculation in clinical trials with multiple objectives that are divided into primary and secondary objectives.

Having multiple primary and secondary objectives in a single trial, it becomes important to distinguish between the probability for a successful trial, as driven by the primary objectives, and the power to reject the individual null hypotheses. To reflect these two objectives, we propose to (i) first select a general test strategy addressing the primary study objectives, and (ii) subsequently fine tune it based on the importance relationships among the primary and secondary hypotheses, as induced by the study objectives and the prior assumptions about the effect sizes for all primary and secondary variables.

To keep the discussion simple, consider a clinical trial comparing two doses of a new compound with placebo for a primary and a secondary endpoint, resulting in four null hypotheses of interest, $H_i:\ \theta_i \leqslant 0,\ i=1,\ldots,4$. It seems natural to declare a trial successful, if at least one of the two primary dose-control comparisons is statistically significant. Further significant results are nice to have, but not essential for claiming the success of a trial. Consequently, if a Bonferroni-based closed test procedure is used, it is sufficient to reject the global intersection hypothesis, which in turn depends only on the initial local significance levels α_i. If H_1, H_2 denote the two primary hypotheses and H_3, H_4 the two secondary hypotheses, one should therefore set $\alpha_3=\alpha_4=0$ to maximize the probability for a successful trial. Consequently, we argue that the relevant primary power measure should be the probability for a successful trial, i.e. the probability of rejecting either H_1 or H_2 at their local significance levels α_1 and $\alpha_2=\alpha-\alpha_1$, if they are in fact not true. Under standard ANOVA assumptions, this probability reduces to calculating a bivariate normal probability with correlation $1/2$. The values for α_1 and α_2 need to be carefully considered, as they impact both the probability for a successful trial as well as the power for the individual assessments, see further below.

When using a graphical approach, a multiple test strategy is essentially defined through the initial significance levels α_i and the weights for the directed edges connecting the nodes, as explained in Sections 2 and 3. Once the primary power measure has been defined, fine tuning the test strategy essentially reduces to finding suitable weights for the edges. A natural step is to account for the relative importance between the primary and secondary hypotheses, as induced by the study objectives. In the above example, it seems natural to avoid rejecting a secondary hypothesis without having rejected the associated primary hypothesis. If this motivation holds, the weights associated with the edges $H_1 \to H_4$ and $H_2 \to H_3$ are set to 0. These basic considerations are sufficient to derive a general test strategy, such as the one displayed in Figure 6, where $0 \leqslant \gamma_1, \gamma_2 \leqslant 1$ are yet to be determined.

Selecting γ_1 and γ_2 is critical for the power of the individual assessments. The main question is what proportion of the significance level should be shifted to the associated secondary hypothesis, once either H_1 or H_2 is rejected: Is it advisable to shift the entire level to the secondary hypothesis ($\gamma_1=\gamma_2=0$) or should one rather split the level and increase the probability of rejecting the other primary hypothesis ($\gamma_1, \gamma_2 > 0$)? If safety problems are not foreseen, one may prefer shifting the entire level to the secondary hypothesis instead of scarifying power by increasing the level for testing the other dose. In this case, it is sufficient to declare any of the two doses significant and $\gamma_1=\gamma_2=0$ seems to be a good choice. Note that if both primary and secondary hypotheses have been rejected for one dose, the strategy in Figure 6 still allows one to test the other dose at level α. However, if safety is of potential concern, it might be advisable to choose $\gamma_1, \gamma_2 > 0$ and therefore mitigate the risk of declaring a single dose significant, which potentially turns out to be not safe.

Once it has been decided to select $\gamma_1, \gamma_2 > 0$, it becomes necessary to finetune their values together with those for α_1 and α_2 based on the prior assumptions about the effect sizes for all primary and secondary variables. Westfall and Krishen [3] investigated this problem formally and derived optimal weights for various multiple test procedures. In practice, we recommend simulating the power under different realistic scenarios in order to understand the operating characteristics of a given multiple test procedure. Table I shows one possibility how such power simulations could be summarized. We simulated standard normally distributed variables $Z_i, i=1,\ldots,4$, such that $\text{corr}(Z_1, Z_2)=\text{corr}(Z_3, Z_4)=1/2$ (this is the structural correlation induced by comparing two doses with a common control in a balanced parallel group design), $\mathrm{corr}(Z_1, Z_3) = \mathrm{corr}(Z_2, Z_4) = \rho$ (the assumed correlation between the primary and secondary endpoint), and $\mathrm{corr}(Z_1, Z_4) = \mathrm{corr}(Z_2, Z_3) = \rho/2$. Table I summarizes the probability π for a successful trial and the power for the individual assessments $\pi_1, \ldots, \pi_4$ for different design options (i.e. choices for α_i and $\gamma_i, i = 1, 2$) and scenarios of ρ and $\theta_i, i = 1, \ldots, 4$.

We briefly summarize the key findings from Table I. By construction, the FWER is kept at level $\alpha = 0.025$ (cases #1 and #14). Cases #2 through #7 consider the power for different scenarios of $\theta_1, \ldots, \theta_4$ while keeping the design parameters and the correlation ρ fixed. The probability π for a successful trial depends only on θ_1 and θ_2 (#2, #3, #4). If both doses are effective for the primary endpoint, we have a power of 90 per cent to declare a successful trial (#7). Otherwise, if only one dose is effective, the dose drops to 78 per cent (#4) or even lower depending on the magnitude of the effect (#5 and #6). The correlation seems to have limited impact on the power for the primary hypotheses but does impact the power for the secondary hypotheses (#8, #9, #10). Cases #11 through #14 compare different alternative test strategies and the impact on power. In case #11, ε is an infinitesimally small weight (i.e. $\varepsilon \to 0$, see [10] for a formal description). With the choice of $\gamma_1 = \gamma_2 = 1 - \varepsilon$, both H_1 and H_2 have to be rejected before the secondary hypotheses H_3 and H_4 can be tested each at level $\alpha_1 = \alpha_2 = \alpha/2 = 0.0125$. Test strategies #12, #13, and #14 imply that if either H_1 or H_2 are rejected, the associated local significance level $\alpha/2$ is shuffled to H_3 and H_4. Clearly, if initially an unequal split of the overall significance level α is applied (e.g. $\alpha_1 = \alpha$ and $\alpha_2 = 0$), the power depends highly on the effect profile (cases #13 and #14).

A graph like Figure 6 offers a simple way of visualizing different test strategies that could be compared in simulation studies. Even from the limited results displayed in Table I we can conclude that a suitable choice of the test strategy is crucial to obtain a reasonable power and that no uniformly best test strategy exists. Moreover, the power depends highly on the assumed treatment effect assumption. The simulations for Table I are in no way sufficiently broad for clinical practice. The aim of Table I is to stress the importance of conducting clinical trial simulations at the design stage of a study and investigating the operating characteristics; see also [27–29].

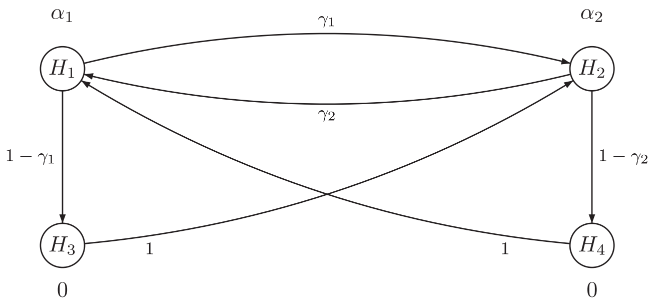

Figure 6. Multiple test strategy for two primary hypotheses H_1, H_2 and two secondary hypotheses H_3, H_4.

Table I. Probability π for a successful trial and individual power π_i for different design options and scenarios ($\alpha = 0.025$).

Case	Design parameters				Scenarios					Outcome measures				
#	α_1	α_2	γ_1	γ_2	ρ	θ_1	θ_2	θ_3	θ_4	π	π_1	π_2	π_3	π_4
1	0.0125	0.0125	0.5	0.5	0.5	0	0	0	0	0.025	0.015	0.014	0.002	0.001
2	0.0125	0.0125	0.5	0.5	0.5	3	0	0	0	0.773	0.773	0.018	0.006	0.003
3	0.0125	0.0125	0.5	0.5	0.5	3	0	3	0	0.774	0.774	0.022	0.596	0.003
4	0.0125	0.0125	0.5	0.5	0.5	3	0	3	3	0.78	0.78	0.026	0.606	0.025
5	0.0125	0.0125	0.5	0.5	0.5	2	0	3	3	0.404	0.403	0.023	0.351	0.022
6	0.0125	0.0125	0.5	0.5	0.5	1	0	3	3	0.111	0.108	0.018	0.102	0.017
7	0.0125	0.0125	0.5	0.5	0.5	3	3	0	0	0.897	0.806	0.806	0.014	0.015
8	0.0125	0.0125	0.5	0.5	0.5	3	3	2	2	0.896	0.808	0.809	0.409	0.402
9	0.0125	0.0125	0.5	0.5	0	3	3	2	2	0.899	0.812	0.81	0.359	0.353
10	0.0125	0.0125	0.5	0.5	0.99	3	3	2	2	0.897	0.812	0.812	0.448	0.44
11	0.0125	0.0125	$1-\varepsilon$	$1-\varepsilon$	0.5	3	0	3	0	0.774	0.774	0.024	0.131	0.004
12	0.0125	0.0125	0	0	0.5	3	0	3	0	0.779	0.779	0.026	0.663	0.005
13	0.025	0	0	0	0.5	3	0	3	0	0.85	0.85	0.023	0.759	0.004
14	0.025	0	0	0	0.5	0	3	3	0	0.025	0.025	0.024	0.024	0.002

5. Discussion

In this paper we summarized the current methodology on Bonferroni-based closed test procedures, including related shortcut procedures and construction of simultaneous confidence intervals. We focused on the description of recently developed graphical approaches that support the construction and visualization of different sequentially rejective test strategies. Using these methods one can tailor a multiple test procedure to given study objectives. We illustrated in detail the derivation of a reasonable test strategy using trial and project level considerations to guide the choice of the local significance levels α_i and weights g_{ij}. These approaches are easily implemented, see the Appendix for sample SAS/IML code.

It transpires that the α_i and g_{ij} fully define the graph and thereby the multiple test strategy. Although their choice is arbitrary (subject to some regularity conditions), in practice the study objectives specified in a protocol will guide setting-up a high-level strategy. Specifying the parameters is done based on prior assumptions about the effect profiles. Simulations are necessary to further fine tune these parameter specification and understand the operating characteristics of the resulting multiple test procedure, including its robustness properties against deviations of the initial assumptions. This leads to our recommendations in Section 4 on how to approach power calculations in trials with multiple primary and secondary objectives.

The examples in Sections 3 and 4 illustrate how the α_i and g_{ij} can be chosen to meet the study objectives, supported by proper clinical trial simulations. Many considerations presented in this paper try to address 'illogical problems' that often occur when using recently developed complex gatekeeping strategies [30, 31]. For example, in a clinical trial with multiple endpoints and more than one dose it would be difficult to claim efficacy for a secondary endpoint if the primary endpoint has not been declared significant for a same dose. Such procedures were denoted as *successive* in [32]. In addition, one should always be concerned about the need to obtain 'consistent results' [33]. For example, in a clinical trial with two primary endpoints it would be difficult to claim overall efficacy if one of the treatment effects is significantly negative, see also [34] for a clinical trial example and related discussion.

The methods proposed in this paper also apply to combined non-inferiority and superiority testing in active controlled trials with more than one endpoint or dose. A similar graph as in Figure 6 was introduced in [31] for this situation. The hierarchy of primary and secondary hypotheses arises naturally from non-inferiority being a prerequisite for showing superiority over the same control. Given there are hypotheses related to multiple (primary) endpoints or multiple treatment arms to be tested, the family of primary hypotheses consists of those related to non-inferiority and the secondary hypotheses are those related to superiority. Furthermore, [32] introduced partially hierarchical test procedures protecting the FWER, which equally apply to this test problem.

Note that the class of test procedures described by the graphical approach in [10] does not include all approaches from the class $\mathscr{S}$ of shortcut procedures defined in [9] and reviewed in Section 2.2. Examples include the truncated Holm procedures described in [16, 35] and the procedure from [36]. One could also address the case study from Section 3 by choosing a suitable procedure from the larger class $\mathscr{S}$, as illustrated with several examples in [9]. Such an approach would lead to even more possibilities of choosing suitable multiple test procedures, but determination of the weights for the resulting shortcut procedure may not be straightforward and more difficult to communicate. In our opinion, the graphical approach is a sufficient and very natural way to determine suitable weights.

Appendix

In the following we present relevant SAS code to perform the sequentially rejective multiple test procedure described in Section 2.4, see also Algorithm 1 in [10]. At the end, we include the call for the numerical example from Section 3.

```
proc iml;

/***************** Input parameters *******************************************
h: indicator whether a hypothesis is rejected (= 1) or not (= 0) (1 x n vector)
a: initial significance level allocation (1 x n vector)
w: weights for the edges (n x n matrix)
p: observed P-values (1 x n vector)
*****************************************************************************/
```

```
START mcp(h, a, w, p);
   n = NCOL(h);
   mata = a;

   crit = 0;
   DO UNTIL(crit = 1);
      test = (p < a);
      IF (ANY(test)) THEN DO;
         rej = MIN(LOC(test#(1:n)));
         h[rej] = 1;
         w1 = J(n, n, 0);
         DO i = 1 TO n;
            a[i] = a[i] + a[rej]*w[rej,i];
            IF (w[i,rej]*w[rej,i]<1) THEN DO j = 1 TO n;
               w1[i,j] = (w[i,j] + w[i,rej]*w[rej,j])/(1 - w[i,rej]*w[rej,i]);
            END;
            w1[i,i] = 0;
         END;
         w = w1; w[rej,] = 0; w[,rej] = 0;
         a[rej]  = 0;
         mata = mata // a;
      END;
      ELSE crit = 1;
   END;


   PRINT h; PRINT (ROUND(mata, 0.0001)); PRINT (ROUND(w,0.01));
FINISH;

/*** Numerical example from Section 3 ***/
h = {      0          0         0         0           0          0};
a = {0.00833    0.00833   0.00833         0           0          0};
w = {      0        0.5         0       0.5           0          0  ,
      0.3333          0    0.3333         0      0.3333          0  ,
           0        0.5         0         0           0        0.5,
           0          1         0         0           0          0  ,
         0.5          0       0.5         0           0          0  ,
           0          1         0         0           0          0};
p =     {0.1      0.008     0.005      0.15        0.04      0.006};

run mcp(h, a, w, p);
quit;

/*** Output ***

Final decisions (1 = rejected, 0 = not rejected):

    0      1      1      0      0      1

Matrix with local significance levels at each step:

 0.0083   0.0083   0.0083      0        0         0
 0.0111        0   0.0111      0   0.0028         0
 0.0133        0        0      0    0.005    0.0067
 0.0167        0        0      0   0.0083          0
Matrix with weigths for the edges after the last iteration (see also Figure 5):

    0      0       0      0.67     0.33      0
    0      0       0         0        0      0
    0      0       0         0        0      0
  0.5      0       0         0      0.5      0
    1      0       0         0        0      0
    0      0       0         0        0      0
*/
```

Acknowledgements

We are grateful to Nikos Sfikas for providing us the case study from Section 3 and for the continuous discussions. We thank H. M. James Hung for motivating the power discussion in Section 4. We also thank Alex Dmitrienko, Ekkehard Glimm, and Ajit Tamhane for their suggestions and comments. Finally, we thank the referee Steven Julious and two anonymous referees for their helpful comments, which improved the presentation of the paper.

References

1. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979; **6**:65–70.
2. Maurer W, Hothorn L, Lehmacher W. Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. *Biometrie in der chemisch-pharmazeutischen Industrie*, Vollmar J (ed.). Fischer Verlag: Stuttgart, 1995; 3–18.
3. Westfall PH, Krishen A. Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures. *Journal of Statistical Planning and Inference* 2001; **99**:25–40.
4. Wiens BL. A fixed sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics* 2003; **2**:211–215.
5. Wiens BL, Dmitrienko A. The fallback procedure for evaluating a single family of hypotheses. *Journal of Biopharmaceutical Statistics* 2005; **15**:929–942.
6. Bauer P, Röhmel J, Maurer W, Hothorn L. Testing strategies in multi-dose experiments including active control. *Statistics in Medicine* 1998; **17**:2133–2146.
7. 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.
8. Marcus R, Peritz E, Gabriel KR. On closed testing procedure with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
9. Hommel G, Bretz F, Maurer W. Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies. *Statistics in Medicine* 2007; **26**:4063–4073.
10. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* 2009; **28**:586–604.
11. Burman CF, Sonesson C, Guilbaud O. A recycling framework for the construction of Bonferroni-based multiple tests. *Statistics in Medicine* 2009; **28**:736–761.
12. Dmitrienko A, Bretz F, Westfall PH, Troendle J, Wiens BL, Tamhane AC, Hsu JC. Multiple testing methodology. In *Multiple Testing Problems in Pharmaceutical Statistics*, Dmitrienko A, Tamhane AC, Bretz F (eds). Chapman & Hall/CRC Biostatistics Series: Boca Raton, 2009.
13. Brannath W, Bretz F. Shortcuts for locally consonant closed test procedures. *Journal of the American Statistical Association* 2010; **28**:739–761. DOI: 10.1198/jasa.2010.tm08127.
14. Hochberg Y, Tamhane AC. *Multiple Comparison Procedures*. Wiley: New York, 1987.
15. Dmitrienko A, Tamhane A, Wang X, Chen X. Stepwise gatekeeping procedures in clinical trial applications. *Biometrical Journal* 2006; **48**:984–991.
16. Strassburger K, Bretz F. Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni based closed tests. *Statistics in Medicine* 2008; **27**:4914–4927.
17. Guilbaud O. Simultaneous confidence regions corresponding to Holm's stepdown procedure and other closed-testing procedures. *Biometrical Journal* 2008; **50**:678–692.
18. Guilbaud O. Alternative confidence regions for Bonferroni-based closed-testing procedures that are not alpha-exhaustive. *Biometrical Journal* 2009; **51**:721–735.
19. Hommel G, Bretz F. Aesthetics and power considerations in multiple testing—a contradiction? *Biometrical Journal* 2008 **50**:657–666.
20. Julious SA. Tutorial in Biostatistics: sample sizes for clinical trials with normal data. *Statistics in Medicine* 2004; **23**:1921–1986.
21. Julious SA. *Sample Sizes for Clinical Trials*. Taylor & Francis: Boca Raton, 2009.
22. International Conference on Harmonization. *ICH Topic E9: Statistical Principles for Clinical Trials*. www.ich.org/LOB/media/MEDIA485.pdf 1998.
23. Ramsey PH. Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association* 1978; **73**:479–487.
24. Maurer W, Mellein B. On new multiple tests based on independent p-values and the assessment of their power. In *Multiple Hypothesenprüfung*, Bauer P, Hommel G, Sonnemann E (eds). Springer: Berlin, 1988; 121–135.
25. Westfall PH, Tobias R, Rom D, Wolfinger R, Hochberg Y. *Multiple Comparisons and Multiple Tests using SAS*. SAS Institute Inc: Cary, 1999.
26. Senn S, Bretz F. Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics* 2007; **6**:161–170.
27. Westfall PH, Tsai K, Ogenstad S, Tomoiaga A, Moseley S, Lu Y. Clinical trials simulation: a statistical approach. *Journal of Biopharmaceutical Statistics* 2008; **18**:611–630.
28. Bretz F, Wang SJ. From adaptive design to modern protocol design for drug development: Part II. Success probabilities and effect estimates for Phase 3 development programs. *Drug Information Journal* 2010; **44**(3):333–342.
29. Benda N, Branson M, Maurer W, Friede T. Aspects of modernizing drug development using clinical scenario planning and evaluation. *Drug Information Journal* 2010; **44**:299–315.
30. Hung HMJ, Wang SJ. Some controversial multiple testing problems in regulatory applications. *Journal of Biopharmaceutical Statistics* 2009; **19**:1–11.
31. Hung HMJ, Wang SJ. Challenges to multiple testing in clinical trials. *Biometrical Journal* 2010; DOI: 10.1002/bimj.
32. Maurer W, Glimm E, Bretz F. Multiple and repeated testing of primary, co-primary and secondary hypotheses. *Statistics in Biopharmaceutical Reserach* 2010; DOI: 10.1198/sbr.
33. Alosh M, Huque MF. A consistency-adjusted alpha-adaptive strategy for sequential testing. *Statistics in Medicine* 2010; DOI: 10.1002/sim.3896.
34. Brannath W, Bretz F, Maurer W, Sarkar S. Trimmed weighted Simes' test for two one-sided hypotheses with arbitrarily correlated test statistics. *Biometrical Journal* 2009; **51**:885–898.

35. Dmitrienko A, Tamhane A, Wiens B. General multi-stage gatekeeping procedures. *Biometrical Journal* 2008; **50**: 667–677.

36. Hommel G, Kropf S. Tests for differentiation in gene expression using a data-driven order or weights for hypotheses. *Biometrical Journal* 2005; **47**:554–562.