STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:586–604
Published online 2 December 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3495

A graphical approach to sequentially rejective multiple test procedures

Frank Bretz[1,2], Willi Maurer[1], Werner Brannath[3] and Martin Posch[3,*,†]

[1]*Novartis Pharma AG, Lichtstrasse 35, 4002 Basel, Switzerland*
[2]*Department of Biometry, Medical University of Hannover, 30623 Hannover, Germany*
[3]*Section of Medical Statistics, Medical University of Vienna, Spitalgasse 23, 1090 Wien, Austria*

SUMMARY

For clinical trials with multiple treatment arms or endpoints a variety of sequentially rejective, weighted Bonferroni-type tests have been proposed, such as gatekeeping procedures, fixed sequence tests, and fallback procedures. They allow to map the difference in importance as well as the relationship between the various research questions onto an adequate multiple test procedure. Since these procedures rely on the closed test principle, they usually require the explicit specification of a large number of intersection hypotheses tests. The underlying test strategy may therefore be difficult to communicate. We propose a simple iterative graphical approach to construct and perform such Bonferroni-type tests. The resulting multiple test procedures are represented by directed, weighted graphs, where each node corresponds to an elementary hypothesis, together with a simple algorithm to generate such graphs while sequentially testing the individual hypotheses. The approach is illustrated with the visualization of several common gatekeeping strategies. A case study is used to illustrate how the methods from this article can be used to tailor a multiple test procedure to given study objectives. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: adjusted *p*-values; Bonferroni; closure principle; gatekeeping procedures; multiple comparison procedures; shortcut procedures; simultaneous confidence intervals

1. INTRODUCTION

Many clinical trials aim at multiple study objectives, such as comparing several treatments with a control, investigating multiple endpoints or subgroups, etc. Testing multiple hypotheses, however, may increase the familywise error rate (FWER), i.e. the probability to erroneously reject at least one true null hypothesis, beyond the pre-specified significance level $\alpha \in (0, 1)$. Adequate multiple

*Correspondence to: Martin Posch, Section of Medical Statistics, Core Unit for Medical Statistics and Informatics, Medical University of Vienna, Spitalgasse 23, A-1090 Wien, Austria.
†E-mail: martin.posch@meduniwien.ac.at, martin.posch@univie.ac.at

Received 26 June 2008
Accepted 15 October 2008

Copyright © 2008 John Wiley & Sons, Ltd.

10970258, 2009, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.3495 by Gilead Sciences, Inc., Wiley Online Library on [06/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

test procedures have to be used, which guarantee a strong control of the FWER at level α under any configuration of true and false null hypotheses. A variety of sequentially rejective, weighted Bonferroni-based test procedures have been proposed, which are powerful and at the same time flexible enough to address multiple study objectives while controlling the FWER. 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 based on Bonferroni adjustments [3, 6–8]. Such methods allow to map the relative importance of the different study objectives as well as their relation onto an appropriately tailored multiple test procedure. As shown by [9], all these procedures belong to a subclass of weighted Bonferroni-based closed test procedures [10], which fulfill a mild monotonicity condition on the weights.

Assume that we are interested in testing, for example, four elementary hypotheses $H_1, \ldots, H_4$, which are grouped into a family $\mathcal{F}_1 = \{H_1, H_2\}$ of primary objectives and a family $\mathcal{F}_2 = \{H_3, H_4\}$ of secondary objectives. Assume further that we are interested in testing $\mathcal{F}_2$ only if at least one of the hypotheses in the primary family $\mathcal{F}_1$ was rejected. The gatekeeping approach proposed by Dmitrienko *et al.* [7] is a reasonable test strategy in this situation. They proposed a specific weighting scheme for each of the $2^4 - 1 = 15$ intersection hypotheses in the full closure (Table I in [7]), leading to a decision matrix that facilitates the computation of p-values for each intersection hypothesis (Table II in [7]). Although the proposed gatekeeping procedure takes advantage of the flexibility of weighted Bonferroni-based closed test procedures, it is often difficult in practice (i) to communicate it to the clinical teams and (ii) to apply it to similar multiple test problems. In this article we propose graphical tools to overcome these problems.

The application of graphical tools to multiple test problems has already been investigated before [11, 12]. In this article we illustrate the use of graphical tools in the particular context of weighted Bonferroni-based closed test procedures. Using a graphical approach, the elementary hypotheses are represented by a set of vertices with associated weights representing 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. To illustrate the basic concepts, consider the weighted Bonferroni–Holm procedure for two hypotheses. Let α_1 and α_2 denote the initial significance levels allocated to H_1 and H_2, respectively, such that $\alpha_1 + \alpha_2 = \alpha$. If H_1 is rejected at level α_1, then H_2 is tested at level α. Vice versa, if H_2 is rejected at level α_2, then H_1 is tested at level α [1]. Figure 1 visualizes the weighted Bonferroni–Holm procedure. Note the initial allocation of the overall significance level α to the individual hypotheses. If, for example, H_1 is rejected, the initially allocated significance level (α_1 at the vertex H_1) is passed on fully to H_2 (as indicated by the directed edge with associated weight 1).

As a matter of fact, Figure 1 defines both (i) a test for the global intersection hypothesis in the full closure through the initial allocation of the significance level α to the individual

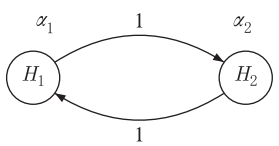

Figure 1. Graphical illustration of the weighted Bonferroni–Holm procedure with two hypotheses.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2009, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.3495 by Gilead Sciences, Inc., Wiley Online Library on [06/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

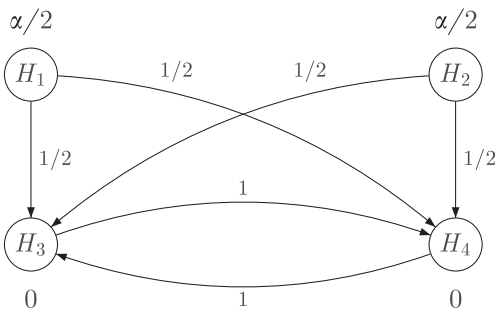

Figure 2. Graphical illustration of the gatekeeping procedure from [7] with four hypotheses.

hypotheses and (ii) a sequentially rejective multiple test procedure (since after rejecting, for example, H_1, only H_2 remains to be tested). In this sense Figure 1 defines an *iterative graph* for the weighted Bonferroni–Holm procedure. This will be illustrated later in more detail with further examples.

In a similar manner the gatekeeping procedure [7] is fully specified by the directed graph in Figure 2. The two hypotheses H_1 and H_2 from the primary family $\mathscr{F}_1$ are initially assigned the local level $\alpha/2$ each, whereas H_3 and H_4 from the secondary family $\mathscr{F}_2$ are assigned the local level 0. If H_1 and/or H_2 are rejected, the local level $\alpha/2$ is split into half and passed on to H_3 and H_4 as indicated by the directed edges with weights $\frac{1}{2}$. If H_3 (H_4) is rejected in the subsequence at its local significance level (either $\alpha/2$ or $\alpha/4$), this level is passed to H_4 (H_3) as indicated by the directed edges with weights 1. Thus, Figure 2 fully specifies the sequentially rejective procedure from [7].

A graphical approach as displayed in Figures 1 and 2 has several advantages. Graphs are easier to communicate with clinical teams than long and abstract decision tables, which typically are not intuitive. 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. Moreover, study protocols should be written in such a way that investigators and other personnel distant to the clinical team are still able to have a basic understanding of the underlying statistical design and analysis methods. Also note that the decision tables proposed in [7] will quickly become untractable if the number of hypotheses increases. If, for example, the study objective is to compare four dose levels with placebo for two endpoints (resulting in eight elementary hypotheses), the decision table of [7] requires 255 rows whereas the associated graph only requires eight vertices. Finally, using the results of [9], the iterated graphs proposed in this article always lead to shortcut procedures, thus making extensive computer programming unnecessary.

In the remainder of this article we formalize the ideas presented in this section and discuss several applications and extensions. In Section 2 we formalize the graphical approach and present the main methodological results, including a simple iterative algorithm to conduct a multiple test procedure derived from a directed graph. In Section 3 we provide various extensions, such as the computation of adjusted p-values, simultaneous confidence intervals, and further considerations on shuffling the significance level between families of hypotheses. Numerical examples are used to illustrate the key results. A case study is discussed in Section 4 to illustrate how the methods from this article can be used to best tailor a multiple test procedure to the given study objectives. Concluding remarks are given in Section 5.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2009, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.3495 by Gilead Sciences, Inc., Wiley Online Library on [06/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2. METHODOLOGY

In this section we describe the main results of this article. In Section 2.1 we give a heuristic justification to motivate the use of iterative graphs. This is formalized in Section 2.2, where we provide a simple algorithm, which essentially results in a sequentially rejective test procedure associated with an iterated graph. We prove that this algorithm (and thus the multiple test procedure derived from an iterated graph) strongly controls the FWER at a pre-specified level α. Illustrative examples are given in Section 2.3.

2.1. Heuristics

Assume that we are interested in testing m elementary null hypotheses $H_1, \ldots, H_m$. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ denote the initial allocation of the overall significance level to the m hypotheses, such that $\sum_{i=1}^{m} \alpha_i \leqslant \alpha$. Finally, assume that we observe the m unadjusted p-values $\mathbf{p} = (p_1, \ldots, p_m)$ for the elementary hypotheses H_i.

Consider the following heuristic approach. Test the m hypotheses each at its local significance level α_i. If a hypothesis H_i can be rejected, reallocate its level to one of the other hypothesis (according to a pre-specified rule). Repeat the testing step for the remaining, non-rejected hypotheses with the updated local significance levels, thus possibly leading to further rejected null hypotheses with associated reallocation of the local significance levels. This procedure is repeated until no further hypothesis can be rejected. In Section 2.2 we show (after a suitable formalization) that this heuristic approach indeed controls the FWER strongly at level α.

This heuristic approach is easily described by the directed graphs introduced in Section 1. In fact, most Bonferroni-based closed test procedures described in the literature are examples of this heuristic and can thus be displayed graphically. We have already mentioned the connection to the Bonferroni–Holm procedure and the gatekeeping procedure from [7].

To further elaborate on the (unweighted) Bonferroni–Holm procedure, recall that it rejects all null hypotheses $H_{(i)}$ with $i \leqslant r = \max\{i \in I : p_{(j)} \leqslant \alpha/(m-j+1) \text{ for all } j \leqslant i \in I\}$, where $p_{(1)} \leqslant \cdots \leqslant p_{(m)}$ denote the ordered p-values with associated ordered null hypotheses $H_{(i)}$ [1]. Figure 3 displays graphically the Bonferroni–Holm procedure for $m = 3$ hypotheses and equal initial allocation of the significance level (i.e. $\alpha_1 = \alpha_2 = \alpha_3 = \alpha/3$). To illustrate the connection between the graph from

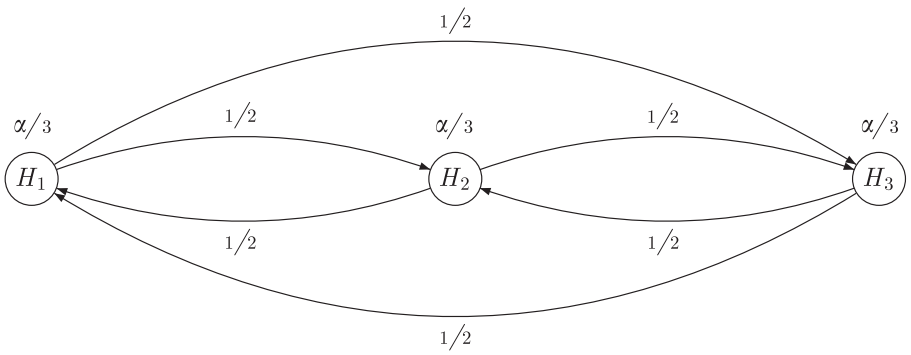

Figure 3. Graphical illustration of the Bonferroni–Holm procedure with $m = 3$ hypotheses and initial allocation $\boldsymbol{\alpha} = (\alpha/3, \alpha/3, \alpha/3)$.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2009, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.3495 by Gilead Sciences, Inc., Wiley Online Library on [06/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Figure 3 and the heuristic approach described above, assume that the p-values $p_1=0.02$, $p_2=0.055$, and $p_3=0.012$ were observed. Figure 4 displays the resulting sequentially rejective test procedure for $\alpha=0.05$. The hypothesis H_3 is rejected at the first step, since $p_3<0.01667=\alpha/3$. The associated local significance level $\alpha/3$ is split equally and passed on to the remaining (not yet rejected) hypotheses H_1 and H_2, as indicated by the directed edges in the left graph in Figure 4. This reallocation step results in the updated local significance levels $\alpha_1=\alpha_2=\alpha/3+\alpha/6=\alpha/2=0.025$, as depicted in the middle graph in Figure 4. Since $p_1=0.02<0.025$, H_1 is now rejected and the updated local significance level $\alpha_2=\alpha/2+\alpha/2=\alpha=0.05$. The right graph in Figure 4 (consisting of only one remaining vertex) displays the final step. Since $p_2>0.05$, H_2 is not rejected and the procedure stops.

Note that Figure 3 fully defines the sequentially rejective Bonferroni–Holm procedure (for $m=3$), while Figure 4 displays the individual steps of the resulting test procedure once the p-values were observed. The graph in Figure 3 therefore defines a stepwise test procedure, where the individual steps can also be visualized, thus leading to an iterative graphical approach to multiple test procedures, as already indicated in Section 1. Note further that the initial allocation of the overall significance level in Figure 3 is arbitrary (subject to some regularity conditions specified below). For example, choosing $\boldsymbol{\alpha}=(\alpha,0,0)$ in Figure 3 leads to a Bonferroni–Holm procedure applied to H_2 and H_3, once H_1 was rejected before at level α.

2.2. General result

We now formalize the heuristic approach from Section 2.1 and prove that it controls strongly the FWER at a pre-specified significance level α. As before, let $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_m)$ denote the local significance levels, such that $\sum_{i=1}^m \alpha_i \leqslant \alpha$. Let $\mathbf{G}=(g_{ij})$ denote an $m\times m$ transition matrix with freely chosen entries g_{ij} that are subject to the regularity conditions

$$0\leqslant g_{ij}\leqslant 1, \quad g_{ii}=0 \text{ and } \sum_{k=1}^{m} g_{ik}\leqslant 1 \quad \text{for all } i,j=1,\ldots,m \tag{1}$$

The weight g_{ij} determines the fraction of the local level α_i that is allocated to H_j in case H_i was rejected and the transition matrix $\mathbf{G}$ thus fully determines the directed edges. Based on the observed p-values $p_i, i\in M=\{1,\ldots,m\}$, we define a sequentially rejective test procedure through

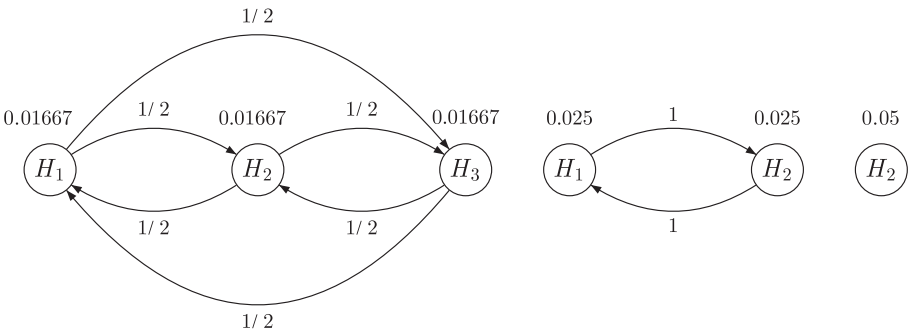

Figure 4. Bonferroni–Holm procedure with observed p-values $\mathbf{p}=(0.02,0.055,0.012)$ and overall significance level $\alpha=0.05$.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2009, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.3495 by Gilead Sciences, Inc., Wiley Online Library on [06/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the following algorithm:

Algorithm 1

0. Set $I = M$.
1. Let $j = \arg\min_{i \in I} p_i/\alpha_i$
2. If $p_j \leqslant \alpha_j$, reject H_j; otherwise stop.
3. Update the graph:

$$I \to I \setminus \{j\}$$

$$\alpha_\ell \to \begin{cases} \alpha_\ell + \alpha_j g_{j\ell}, & \ell \in I \\ 0 & \text{otherwise} \end{cases}$$

$$g_{\ell k} \to \begin{cases} \dfrac{g_{\ell k} + g_{\ell j} g_{jk}}{1 - g_{\ell j} g_{j\ell}}, & \ell, k \in I, \ \ell \neq k \\ 0 & \text{otherwise} \end{cases}$$

4. If $|I| \geqslant 1$, go to step 1; otherwise stop.

In the Appendix we show that a graph $\mathcal{G} = (\boldsymbol{\alpha}, \mathbf{G})$ together with the updating rules from Algorithm 1 defines a short cut for a consonant closed test procedure where each intersection hypothesis is tested with a weighted Bonferroni test. Together with Algorithm 1, a graph $\mathcal{G} = (\boldsymbol{\alpha}, \mathbf{G})$ thus defines a sequentially rejective multiple test procedure that strongly controls the FWER at level α, where $\boldsymbol{\alpha}$ and $\mathbf{G}$ are subject to the constraints above.

To illustrate the connection between Algorithm 1 and the proposed graphs, consider Figure 5 for an example involving $m = 3$ hypotheses. For the top left graph we have $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ and

$$\mathbf{G} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

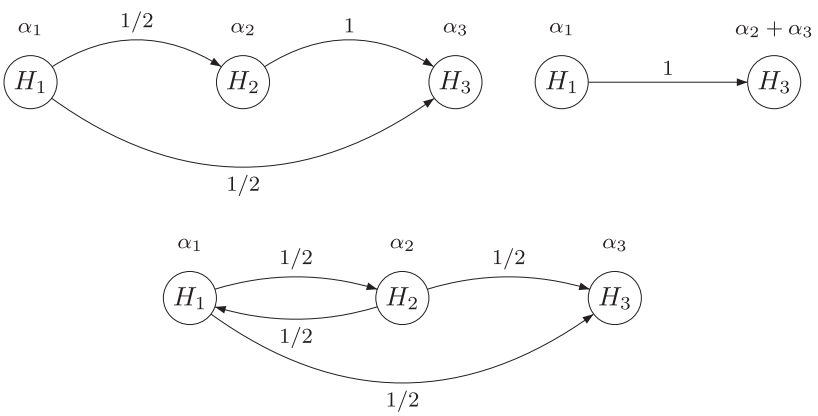

Figure 5. Example multiple test procedures to illustrate Algorithm 1.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2009, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.3495 by Gilead Sciences, Inc., Wiley Online Library on [06/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Assume that H_2 is rejected, i.e. $j=2$ and $p_2 \leqslant \alpha_2$. Applying the graph iteratively, the vertex H_2 is deleted and the associated significance level α_2 is passed along the outgoing edges, according to the associated weights $g_{2\ell}$. In our example, there is only one edge starting from H_2 and the updated vector of significance levels becomes $\alpha=(\alpha_1, 0, \alpha_3+\alpha_2)$. At the same time, 'loose' edges are connected and their weights are renormalized to fulfill the regularity conditions (1), ultimately leading to the top right graph in Figure 5. These updates in the graph are essentially reflected and formalized in Step 3 of Algorithm 1. In other cases, the graphical update needs more care. Consider the bottom graph in Figure 5. The only difference to the previous example is the additional edge from H_2 to H_1 such that

$$\mathbf{G}=\begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \end{pmatrix}$$

Note that H_1 and H_2 are now connected via a circular loop. However, applying Step 3 of Algorithm 1 ensures that after rejecting H_2 one essentially obtains the top right graph in Figure 5, but with the updated vector of significance levels $\boldsymbol{\alpha}=(\alpha_1+\alpha_2/2, 0, \alpha_3+\alpha_2/2)$.

We conclude this section with five remarks: (i) Algorithm 1 is mainly required to provide a rule for updating the initial graph $\mathcal{G}=(\boldsymbol{\alpha}, \mathbf{G})$ while testing sequentially the m elementary null hypotheses $H_1, \ldots, H_m$. In many cases sequentially rejective multiple test procedures defined through Algorithm 1 are characterized by a simple iterative graph $\mathcal{G}$, where the rejection of a null hypothesis leads to the deletion of a vertex with associated edges (and a related direct update of the remaining weights). Figure 2 displays such a simple graph, whereas Figure 3 displays an example, where the update is more complex due to the inherent loops induced by the edges. Such loops with two edges can occur during the iteration if there is a directed cycle in the graph even if it is not present initially. (ii) Algorithm 1 specifies weighted Bonferroni-based closed test procedures satisfying the monotonicity condition from [9] and applying it to the graph iteratively ensuring that one obtains a shortcut of length m, thus making extensive computer programming unnecessary. (iii) Sometimes, several hypotheses H_i with $p_i \leqslant \alpha_i$ might be rejected at the same iteration step. The resulting final set of rejected hypotheses is independent of how the single index j is chosen. Taking the argument of the minimum in Step 1 of Algorithm 1 is a convenient solution, but can be replaced by any other selection rule. (iv) In essence, the vector $\boldsymbol{\alpha}$ specifies a weighted Bonferroni test for the global intersection hypothesis $H_M=\bigcap_{i\in M} H_i$ and the directed graphs specify the weighted Bonferroni tests for the $(m-1)$-way intersection hypotheses $H_{M\setminus\{j\}}=\bigcap_{i\in M\setminus\{j\}} H_i$, $j=1, \ldots, m$. This leads to a specification of m^2 weights. Since the closed test procedure involves 2^m-1 intersection hypotheses, consonant Bonferroni-based closed test procedures can be constructed for $m \geqslant 4$, which are not covered by the graphs proposed so far. However, the graphs can be extended to include other test strategies, some of which are discussed in Section 3. (v) The graphs can also be used to tabulate the weights α_ℓ/α, $\ell \in I$, for all intersection hypotheses H_I, $I \subseteq M$, of the full closure by removing the vertices associated with the hypotheses H_i, $i \in M\setminus\{I\}$ and updating the graph accordingly. Such weights are the basis of the decision tables introduced in [7]. In the Appendix it is shown that with Algorithm 1 the resulting weights for the intersection hypotheses and related decision tables are independent of the sequence, when the vertices are removed.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2009, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.3495 by Gilead Sciences, Inc., Wiley Online Library on [06/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.3. Examples

We now illustrate the choice of $(\boldsymbol{\alpha}, \mathbf{G})$ for several commonly used multiple test procedures. Gatekeeping strategies are left out here and discussed separately in Sections 3.3 and 4.

Fixed sequence test: Consider first fixed sequence tests, where the test sequence of the hypotheses is fully specified in advance. Each hypothesis is tested at level α, where non-rejection at any step renders further testing unnecessary [2, 3]. Figure 6 illustrates the fixed sequence test with three hypotheses, where H_1 precedes H_2, which in turn precedes H_3. Note the initial allocation of the overall significance level α to the individual hypotheses. If, for example, H_1 is rejected, the initially allocated significance level (α at the vertex H_1) is passed on fully to H_2 (as indicated by the directed edge with associated weight 1). Accordingly, $\boldsymbol{\alpha} = (\alpha, 0, 0)$ and

$$\mathbf{G} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

where $g_{12} = 1$ ($g_{23} = 1$) denotes the transition of the local significance level α_1 (α_2) from H_1 to H_2 (H_2 to H_3).

Fallback procedure: Wiens [4] proposed a modification of the fixed sequence test, which overcomes the dependence on the order of the hypotheses (while sacrificing some power for the individual tests, since they are performed at local significance levels less than α). In the notation from Algorithm 1, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ and

$$\mathbf{G} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

An improvement was proposed by [5] using the closed test procedure. It can be shown that Figure 7 visualizes the improved procedure for $m = 3$.

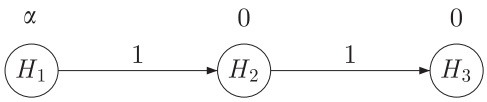

Figure 6. Graphical illustration of the fixed sequence test with three hypotheses.

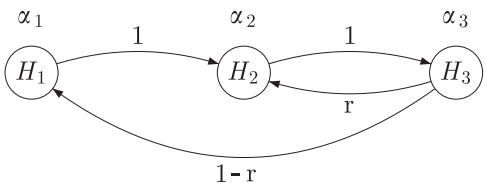

Figure 7. Improvement of the fallback procedure by [5] with $r = \alpha_2/(\alpha_1 + \alpha_2)$.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2009, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.3495 by Gilead Sciences, Inc., Wiley Online Library on [06/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

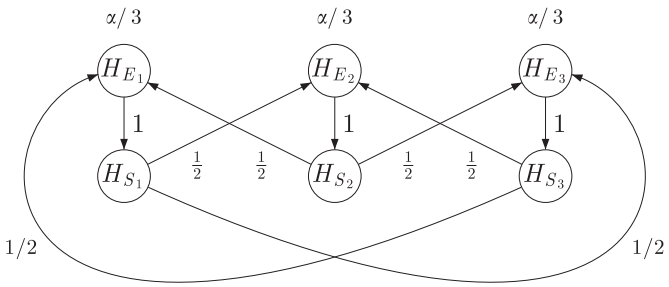

Figure 8. The step-down test without order constraints from [13]. H_{E_i} and H_{S_i} denote the efficacy and safety null hypotheses for treatment $i = 1, 2, 3$, respectively.

Bonferroni–Holm procedure: As seen from Figure 3, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ and

$$\mathbf{G} = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0 \end{pmatrix}$$

fully specify the weighted Bonferroni–Holm procedure. Note that weights other than 0.5 could be used as entries for $\mathbf{G}$, thus generalizing the Bonferroni–Holm procedure.

Step-down tests for identifying effective and safe treatments: Bauer *et al.* [13] considered the comparison of several treatments with a control for both an efficacy and a safety endpoint. They proposed several stepwise test procedures that can all be constructed with the iterative graphs proposed here. For example, the graph in Figure 8 corresponds to the step-down procedure without order constraints between treatments (Section 3 in [13]). Here, the significance level is equally split across treatments and within each treatment a fixed sequence test for the efficacy and safety is performed. If for one treatment both hypotheses can be rejected, the level allocated to that treatment is equally distributed among the other treatments.

3. EXTENSIONS

In this section we describe some extensions of the graphs considered so far. In Section 3.1 we describe how Algorithm 1 can be modified so that the adjusted p-values can be computed. In Section 3.2 we describe how to calculate simultaneous confidence intervals in the current framework. Finally, in Section 3.3 we re-visit the gatekeeping procedure described in Section 1 and propose some extensions of the graphs to include further Bonferroni-based closed test procedures.

3.1. Adjusted p-values

Adjusted p-values are often used to describe the outcome of a multiple test procedure, since after their calculation the test can be performed at any significance level α. Following [14], the adjusted p-value p_j^{adj} for the hypothesis H_j is the smallest significance level at which one can reject the hypothesis using the given multiple test procedure. Adjusted p-values thus incorporate the structure of the underlying decision rule that can be quite complex. In the following we show that a slight

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2009, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.3495 by Gilead Sciences, Inc., Wiley Online Library on [06/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

modification of Algorithm 1 allows the calculation of $p_1^{\text{adj}}, \ldots, p_m^{\text{adj}}$. To this end, define the weights $\mathbf{w} = (w_1, \ldots, w_m) = (\alpha_1, \ldots, \alpha_m)/\alpha$.

Algorithm 2

0. Set $I = M$ and $p_{\max} = 0$.
1. Let $j = \arg\min_{i \in I} p_i / w_i$.
2. Calculate $p_j^{\text{adj}} = \max\{p_j / w_j, p_{\max}\}$ and set $p_{\max} = p_j^{\text{adj}}$
3. Update the graph:

$$I \to I \setminus \{j\}$$

$$w_\ell \to \begin{cases} w_\ell + w_j g_{j\ell}, & \ell \in I \\ 0 & \text{otherwise} \end{cases}$$

$$g_{\ell k} \to \begin{cases} \dfrac{g_{\ell k} + g_{\ell j} g_{jk}}{1 - g_{\ell j} g_{j\ell}}, & \ell, k \in I, \ \ell \neq k \\ 0 & \text{otherwise} \end{cases}$$

4. If $|I| \geqslant 1$, go to step 1; otherwise stop.
5. Reject all hypotheses H_j with $p_j^{\text{adj}} \leqslant \alpha$.

To illustrate Algorithm 2, we revisit the numerical example from Figure 4. Here, $\mathbf{w} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. At the first iteration, $j = 3$ and $p_3^{\text{adj}} = \max\{0.012/0.333, 0\} = 0.036$. After updating the graph we obtain at the second iteration $j = 1$ with $p_1^{\text{adj}} = \max\{0.02/0.5, 0.036\} = 0.04$. Finally, $p_2^{\text{adj}} = \max\{0.055, 0.04\} = 0.055$. These are exactly the adjusted p-values that would be obtained when applying the algorithm described in [14]. Thus, we can reject H_3 for any significance level $\alpha \geqslant 0.036$, reject H_1 for any $\alpha \geqslant 0.04$, and reject H_2 for any $\alpha \geqslant 0.055$. It should be noted that the test decisions obtained from Algorithm 2 are exactly the same as those from Algorithm 1.

3.2. Simultaneous confidence intervals

In the Appendix we prove that the proposed graphs lead to consonant closed test procedures satisfying a natural monotonicity condition. Since [15, 16] derived compatible simultaneous confidence intervals for such procedures, we can apply them in a similar manner to the present framework. Consider the one-sided null hypotheses $H_i : \vartheta_i \leqslant \delta_i$, $i \in M = \{1, \ldots, m\}$, where ϑ_i are the parameters of interest (e.g. treatment means or contrasts thereof) and δ_i are the pre-specified constants (e.g. non-inferiority margin). Let further $L_i(\gamma)$ denote local (i.e. marginal) lower confidence bounds at level $1 - \gamma$. Finally, let $R \subseteq \{1, \ldots, m\}$ denote the index set of hypotheses H_i rejected by a multiple test procedure specified through a graph $\mathcal{G} = (\boldsymbol{\alpha}, \mathbf{G})$. Following [15], the one-sided lower confidence bounds $\bar{L}_i$ for $\vartheta_i, i \in M$ with coverage probability of at least $1 - \alpha$ are given by

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

where $\bar{\alpha}_i$ is the level for the hypothesis H_i in the final graph when applying Algorithm 1.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2009, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.3495 by Gilead Sciences, Inc., Wiley Online Library on [06/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

To illustrate the calculation of the simultaneous confidence intervals, we revisit the numerical example from Figure 4 assuming $\delta_i = 0$ for all i. Applying the Bonferroni–Holm procedure at level $\alpha = 0.05$ leads to the rejection of H_1 and H_3. Thus, $R = \{1, 3\} \subsetneq M$ and $\bar{L}_1 = \bar{L}_3 = 0$. As seen from Figure 4, $\bar{\alpha}_2 = \alpha$ and $\bar{L}_2$ reduces to the marginal confidence bound at level $1 - \alpha$.

3.3. Shifting significance levels between families of hypotheses

Consider a situation where families of hypotheses are given and where the rejection of hypotheses in one family is of interest only if all the hypotheses from another family were rejected. In such cases a multiple test procedure can be applied that allows for a reallocation of the significance level between families of hypotheses. Such a test strategy can be implemented with graphs that include edges with infinitesimally small weights. Along the vertices with an infinitesimally small weight ε no significance level is passed. However, if during the iterative procedure for a vertex only infinitesimal outgoing edges remain, they become non-infinitesimal edges after normalization (such that the sum of outgoing weights becomes one) and can pass the level to other hypotheses.

More formally, when updating the transition weights g_{ij} in the graph according to Algorithms 1 or 2, ε is treated in the calculations as a variable representing some fixed positive real number. In contrast, for the computation of the updated levels α_i (Algorithm 1) or weights w_i (Algorithm 2), we let $\varepsilon \to 0$. Thus, for all real numbers $x > 0$ we get the calculation rules $x + \varepsilon = x$, $x\varepsilon = 0$, $\varepsilon^0 = 1$, and for all non-negative integers k, l

$$\frac{\varepsilon^k}{\varepsilon^l} = \begin{cases} 0 & \text{if } k > l \\ 1 & \text{if } k = l \\ \infty & \text{if } k < l \end{cases}$$

As an example consider the test of three hypotheses H_1, H_2, and H_3, where H_1, H_2 are of primary interest and H_3 is of interest only if H_1 and H_2 can be both rejected. Consider the following test strategy. Perform the Bonferroni–Holm procedure at level α for H_1 and H_2. If both H_1 and H_2 can be rejected, then H_3 is tested at level α. Figure 9 shows the corresponding graph that differs from the graph of the Bonferroni–Holm procedure with two hypothesis (see Figure 1) only by the additional edge from H_2 to H_3 with weight ε. Additionally, to achieve weights that sum to one, the weight of the edge H_2–H_1 is set to $1 - \varepsilon$ (instead of 1 in the Bonferroni–Holm procedure). By the calculation rules set above, these modifications do not effect the Bonferroni–Holm procedure for H_1 and H_2. However, if both hypotheses H_1 and H_2 are rejected, the significance level α is shuffled to H_3. To illustrate the procedure numerically, assume that $\alpha = 0.05$ and the observed p-values are $p_1 = 0.04$, $p_2 = 0.01$, and $p_3 = 0.03$. As $p_2 < \alpha/2$, H_2 is rejected in the first step and its level $\alpha/2$ is shuffled to hypothesis H_1, since by the above calculation rules $(1 - \varepsilon)\alpha/2 = \alpha/2$

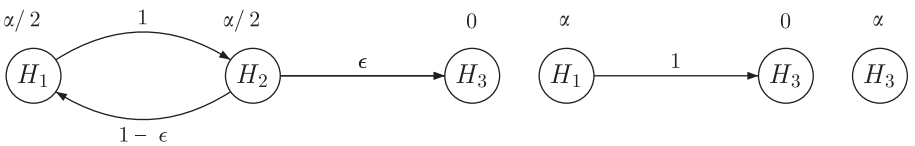

Figure 9. The Bonferroni–Holm procedure as gatekeeper and the iterated graphs with the observed p-values $p_1 = 0.04$, $p_2 = 0.01$, and $p_3 = 0.03$.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2009, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.3495 by Gilead Sciences, Inc., Wiley Online Library on [06/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and $\varepsilon\alpha/2=0$. Now, the vertex H_2 is dropped from the graph and the edges of H_1 are updated. In particular, the edge from H_1 to H_3 gets the weight $0+\varepsilon/(1-(1-\varepsilon))=1$ (see Figure 9). In the second step, H_1 is rejected ($p_1\leqslant\alpha$) and its level is passed on to H_3 that is rejected in the last step ($p_3\leqslant\alpha$). Note that instead of defining an ε-edge from H_2 to H_3 we could equivalently define such an edge from H_1 to H_3 or, as a third possibility, both ε edges. By the calculation rules all these graphs will lead to the same multiple test procedure.

The example can easily be extended to allow for different weights in the ε-graph. Assume that there is an additional fourth hypothesis H_4 and that H_3 and H_4 are of interest only if both H_1 and H_2 were rejected. We wish to perform the Bonferroni–Holm procedure at level α for the two hypotheses H_1 and H_2 of primary interest. If both H_1 and H_2 can be rejected, then a weighted Bonferroni–Holm test for H_3 and H_4 with weights r_1 and r_2 is foreseen, such that $r_1+r_2=1$. The left graph in Figure 10 visualizes the two Bonferroni–Holm procedures that are joined by edges with weights $r_1\varepsilon$ and $r_2\varepsilon$. As above, the weight of the edge between H_2 and H_1 is set to $1-\varepsilon$. If both hypotheses H_1 and H_2 are rejected, the significance level α is shuffled to H_3 and H_4 according to the weights r_1 and r_2: H_3 receives $r_1\alpha$ and H_4 receives $r_2\alpha$. To illustrate the resulting sequentially rejective procedure, set $\alpha=0.05$, $r_1=0.8, r_2=0.2$ and assume the p-values $p_1=0.04$, $p_2=0.01$, $p_3=0.03$, and $p_4=0.04$ have been observed. Since $p_2<\alpha/2$, H_2 is rejected in the first step and its level $\alpha/2$ is shuffled to hypothesis H_1. Now, the vertex H_2 is dropped from the graph and the edges of H_1 are updated (middle graph in Figure 10). For example, the edge from H_1 to H_3 gets the weight $0+r_1\varepsilon/(1-(1-\varepsilon))=r_1=0.8$. In the second step, H_1 is rejected since $p_1<\alpha$. Now, H_3 receives the level $r_1\alpha=0.8\cdot0.05=0.04$ and H_4 the level 0.01 (right graph in Figure 10). Accordingly, H_3 is rejected next ($p_3<0.04$) and passes its level to H_4, which is finally rejected with $p_4<0.05$.

More generally, assume that the hypotheses are grouped into families $\mathcal{F}_k, k=1,\ldots,K$, such that for all hypotheses $H_i, H_{i'}\in\mathcal{F}_k$, there exists a path from H_i to $H_{i'}$ along the edges that have positive (non infinitesimal) weights. Additionally, assume that for all hypotheses $H_i\in\mathcal{F}_k$, $H_{i'}\notin F_k$ all paths from H_i to $H_{i'}$ have at least one edge with an infinitesimal weight. The significance level for the hypotheses in F_k is then shifted to hypotheses outside of $\mathcal{F}_k$ only if all hypotheses in F_k have been rejected. Note that it makes no difference which hypotheses in $\mathcal{F}_k$ are chosen as the origin of the ε edges; all choices lead to the same multiple test procedure.

Adding ε edges may uniformly improve a multiple test as we show below for the gatekeeping procedure from Figure 2. As a general rule, a graph is complete (and thus cannot be improved by adding additional edges) if the weights of outgoing edges sum to one at each vertex and if the graph is irreducible, i.e. if every vertex is accessible from any of the other vertices. If initially a

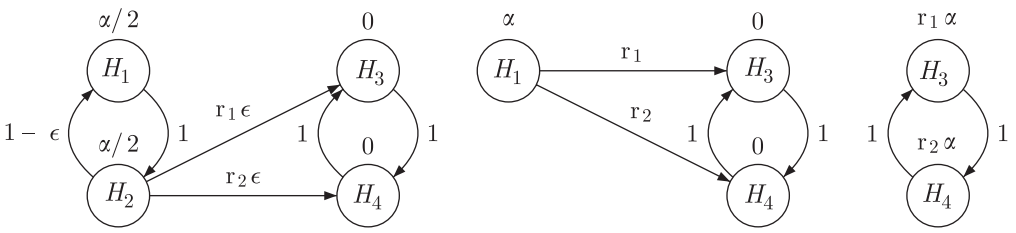

Figure 10. The Bonferroni–Holm procedure as gatekeeper and the iterated graphs with parameters $\alpha=0.05$, $r_1=0.8$, $r_2=0.2$ and the observed p-values $p_1=0.04$, $p_2=0.01$, $p_3=0.03$, $p_4=0.04$.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2009, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.3495 by Gilead Sciences, Inc., Wiley Online Library on [06/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

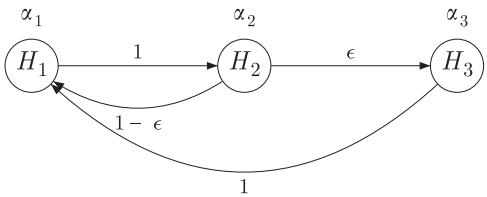

Figure 11. A modified fallback procedure.

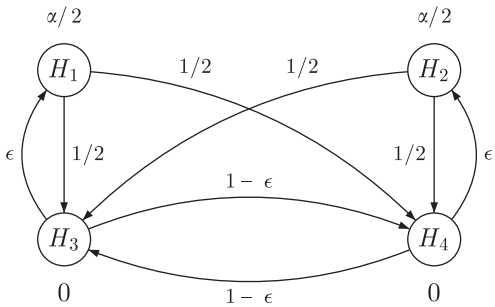

Figure 12. Improvement of the gatekeeping procedure from [7].

positive significance level is assigned to each hypothesis, i.e. $\alpha_i > 0, i = 1, \ldots, m$, irreducibility is also a necessary condition for completeness.

In the following we give two further examples of multiple test procedures with ε-edges.

Modified fallback procedure: After the rejection of a hypothesis in the original fallback procedure, the significance level is passed down to the next hypothesis in the hierarchy. This has been critically discussed [17] since it implies that the significance level of less important hypotheses may be increased while hypotheses that are higher in the hierarchy remain not rejected. Figure 11 (which is a simple extension of the test procedure defined in Figure 9) defines a fallback procedure that shifts the level to the first hypothesis in the hierarchy that has not been rejected so far.

Improved gatekeeping procedure. Infinitesimal weights can also be used to uniformly improve the gatekeeping procedure from [7], as seen in the following example. Let $\alpha = 0.05$ and assume that $p_1 = 0.02$, $p_2 = 0.04$, $p_3 = 0.01$, and $p_4 = 0.015$. According to Figure 2, first H_1 is rejected and the level $\alpha/2$ is distributed equally to H_3 and H_4, which are both now assigned the level $\alpha/4$. Next, H_3 is rejected and passes the level on to H_4, which then is rejected. Here the algorithm stops. Although H_3 and H_4 are both rejected, the significance level cannot be shuffled to H_2 (which has not been rejected yet) since a corresponding edge is missing. Thus, the gatekeeping procedure can be improved by adding ε edges from H_3 to H_1 and from H_4 to H_2 (see Figure 12). For the numerical example above this implies that in the improved gatekeeping procedure the only outgoing edge from H_4 after the rejection of H_1 and H_3 is the ε-edge to H_2 and is thus assigned the weight 1. After rejecting H_4 the level is passed to H_2, which then can also be rejected. In this numerical example we can therefore reject all four hypotheses with the improved procedure from Figure 12, but only H_1, H_2, and H_3 with the original procedure displayed in Figure 2. Note that this improvement has been described previously in [9, 18].

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2009, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.3495 by Gilead Sciences, Inc., Wiley Online Library on [06/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4. CASE STUDY

In this section we discuss a case study to illustrate how the methods from this article can be used to best tailor a multiple test strategy to given study objectives. This case study refers to the late phase development of a new drug for the indication of multiple sclerosis. The primary objective of the study is to compare two dose levels of the new drug with a control treatment for three hierarchically-ordered endpoints (annualized relapse rate, number of lesions in the brain, and disability progression). We have therefore six elementary hypotheses $H_{ij}:\theta_{ij}\leqslant 0$, where θ_{ij} denotes the mean difference of treatment $i=1$ (high dose), 2 (low dose), and control for endpoint $j=1,2,3$. In the following we describe several test strategies and use the graphical tools developed in this article to visualize them. It is not the purpose to recommend one strategy, since each has its advantages and disadvantages. The following discussion is rather meant to demonstrate the flexibility of Bonferroni-based closed test procedures and the need to understand the study objectives well in order to propose a reasonable test strategy with good operational characteristics (i.e. high probability of success for the study).

Strategy 1. A straightforward approach is to apply a fixed sequence test to the six hypotheses being and to test each hypothesis sequentially at level α. The sequence $H_{11}\rightarrow H_{21}\rightarrow H_{12}\rightarrow H_{22}\rightarrow H_{13}\rightarrow H_{23}$ is a reasonable possibility, see also Figure 13. In practice such a strategy is often not recommended because of the inherent risk to stop too early. If, for example, the observed p-value for H_{11} is larger than α, none of the subsequent hypotheses can formally be rejected, even if their p-values are very small.

Strategy 2. An alternative approach that avoids stopping too early if the hypotheses corresponding to the first dose cannot be rejected is to group the six elementary hypotheses according to the dose into the two families $\mathscr{F}_1=\{H_{11},H_{12},H_{13}\}$ and $\mathscr{F}_2=\{H_{21},H_{22},H_{33}\}$. Assuming that there is the wish to reject the secondary (tertiary) endpoint for dose $i=1,2$ only if the associated primary (primary and secondary) endpoints were rejected before, the left graph in Figure 14 visualizes a possible strategy. Within each family the endpoints are tested in a fixed sequence at bonferronized level $\alpha/2$. If for any dose level the three related null hypotheses can be rejected, the fixed sequence for the other dose level can be conducted at level α. The right graph of Figure 14 shows a modification of this test strategy that puts more weight on the hypotheses corresponding to the endpoints in the primary positions of the hierarchy. After each rejection the level is split between the two families and allocated to the first endpoint in each family that has not been rejected so far. If all the hypotheses are rejected in a family, the total level is allocated to the other family.

Strategy 3. In some situations it may be reasonable to order the dose levels, for example, because of safety concerns or because the higher dose level $i=1$ is expected to have a larger treatment effect than the lower dose level $i=2$. Such assumptions then may lead to different families of hypotheses than considered previously. For example, if one is indeed willing to assume $\theta_{1j}>\theta_{2j}$ for all j, it seems natural to start testing the high dose for the primary endpoint at level α. If H_{11} was rejected, the question is then whether one can argue that H_{12} is more important than H_{21}

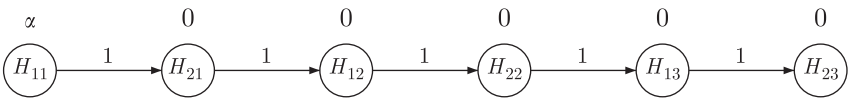

Figure 13. Visualization of Strategy 1.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2009, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.3495 by Gilead Sciences, Inc., Wiley Online Library on [06/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

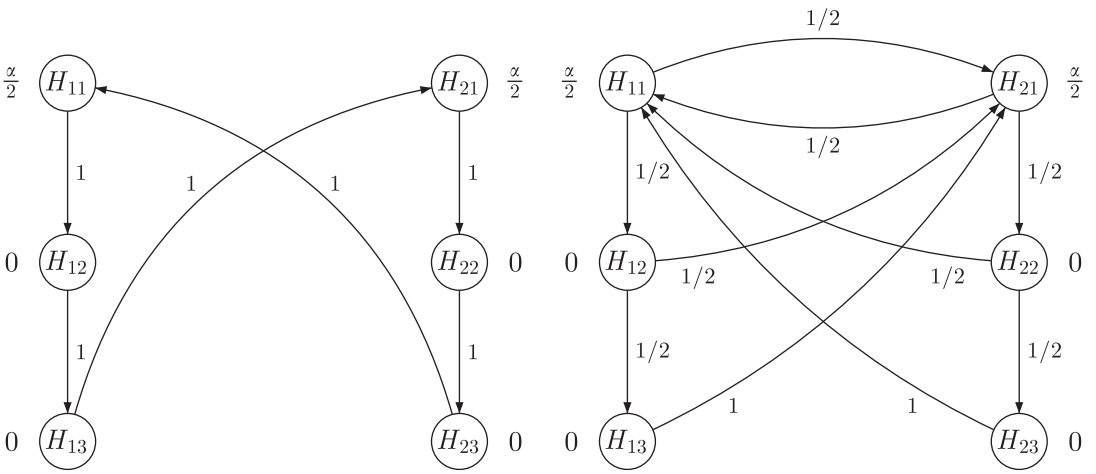

Figure 14. Visualization of different implementations for Strategy 2.

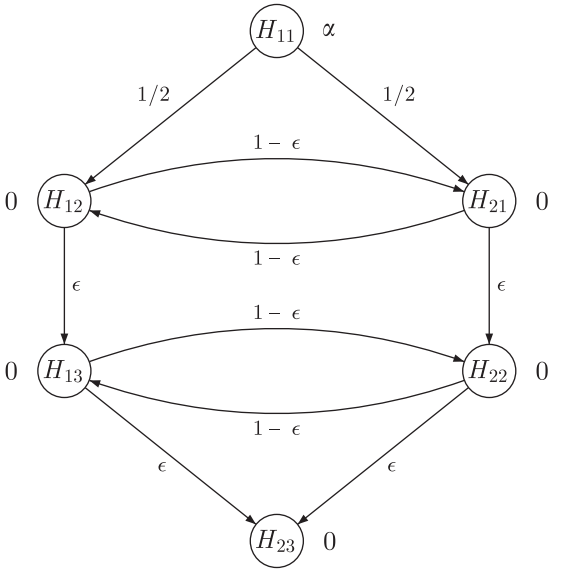

Figure 15. Visualization of a possible implementation for Strategy 3.

(or vice-versa), leading to a fixed sequence as discussed in Strategy 1, or whether both hypotheses H_{12} and H_{21} are equally important. In the latter case, this would lead naturally to the set of families $\mathscr{F}_1=\{H_{11}\}$, $\mathscr{F}_2=\{H_{12}, H_{21}\}$, $\mathscr{F}_3=\{H_{13}, H_{22}\}$, and $\mathscr{F}_4=\{H_{23}\}$, where $\mathscr{F}_i$ precedes $\mathscr{F}_j$ for $1\leqslant i<j\leqslant 4$. While there is no need for further discussion on how to test $\mathscr{F}_1$ and $\mathscr{F}_4$, different possibilities exist for $\mathscr{F}_2$ and $\mathscr{F}_3$. E.g. one could test $\mathscr{F}_3$ only if at least one hypothesis from $\mathscr{F}_2$ was rejected. An alternative approach is to test $\mathscr{F}_3$ only, if both hypotheses from $\mathscr{F}_2$ were rejected, as visualized in Figure 15.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2009, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.3495 by Gilead Sciences, Inc., Wiley Online Library on [06/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

5. DISCUSSION

In this article we propose a simple graphical tool to construct and compare weighted Bonferroni-based closed test procedures. We believe that the resulting graphs are easier to communicate to the clinical teams than extensive tables or matrices containing the weights for the intersection hypotheses. Because of this, different strategies can be constructed quite easily for a given problem and compared with each other in order to select the most suitable strategy, i.e. the one that best addresses the study objectives. As demonstrated with the case study in Section 4, Bonferroni-based closed test procedures are very flexible and lead to a variety of test strategies. The ultimate choice is always a compromise between the risk of stopping too early in the test sequence and a potential loss of power induced by splitting the significance level.

The proposed graphs are in fact iterative in the sense that together with a simple updating algorithm each of them fully describes a sequentially rejective test procedure. In addition to this, the results from [9] ensure that the resulting test procedures are always consonant. Thus, shortcut procedures of length m are obtained and any multiple test procedure constructed with the methods from this article can in essence be performed with the same amount of effort as the common Bonferroni–Holm procedure. In addition, adjusted p-values and simultaneous confidence intervals are readily available.

The graphs that are iteratively generated by Algorithm 1 can be interpreted as a sequence of finite Markov chains where the hypotheses (i.e. the vertices of the graph) correspond to states and the vector $\mathbf{w}=\boldsymbol{\alpha}/\alpha$ to the initial probability distribution on these states. For each Markov chain (graph) at a certain step of the algorithm all hypotheses not rejected so far are defined as absorbing states (that keep the level allocated to them) and the rejected hypotheses as transient states with transition probabilities defined by $\mathbf{G}$. The absorption probabilities multiplied by α are then the local significance levels associated with the not yet rejected hypotheses. It can be shown that Algorithm 1 is a variant of the state reduction algorithm for the computation of absorption probabilities of finite Markov chains [19].

The proposed graphs, together with an updating rule as Algorithm 1, define the weights for the 2^m-1 intersection hypotheses. Note that these intersection hypotheses do not need to be tested with weighted Bonferroni tests. In principle, any weighted α-level tests could be used, for example, resampling based max t, weighted Simes, Dunnett, or Sidak tests. However, for tests other than the Bonferroni test, the consonance property may be lost and instead of the shortcut procedure resulting from Algorithm 1 or Algorithm 2, the full closure may be required, testing through all 2^m-1 intersection hypotheses.

APPENDIX A

We verify the main result from Section 2.2 by showing that a graph $\mathcal{G}$ defines a closed test procedure with weighted Bonferroni tests for the intersection hypotheses and that Algorithm 1 is a shortcut for this closed test procedure. We thereby utilize results from [9], which give a necessary and sufficient monotonicity condition for the local significance levels of the weighted Bonferroni tests such that the resulting closed test is consonant and admits a shortcut. This shortcut is then shown to be equivalent to Algorithm 1. Accordingly, we prove the main result in three steps.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2009, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.3495 by Gilead Sciences, Inc., Wiley Online Library on [06/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(i) *Monotonicity conditions leading to consonant closed test and shortcut procedures*: Here, we state the necessary and sufficient monotonicity conditions for the local significance levels of the weighted Bonferroni tests applied to all intersection hypotheses such that the resulting closed test procedure is consonant. Following [9], we then describe the resulting shortcut procedure.

(ii) *A graph $\mathcal{G}$ with an appropriate updating rule generates a consonant closed test*: We show that any given initial graph $\mathcal{G}$ applied to subsets of vertices (hypotheses) together with the updating rules from Algorithm 1 generates a unique set of local significance levels for such subsets. These local significance levels define weighted Bonferroni tests for the corresponding intersection hypotheses. It is shown that the monotonicity condition still holds, thus resulting in a consonant closed test procedure.

(iii) *The shortcut is applied to the closed test generated by graph $\mathcal{G}$ and the updating rule is equivalent to Algorithm 1*: Here we show that the shortcut procedure from (i) is applied to the local significance levels generated by $\mathcal{G}$ and the updating rule leads to Algorithm 1 for any given set of local p-values and hence the resulting test procedure protects the FWER.

(i) *Monotonicity conditions leading to consonant closed test and shortcut procedures*: Following [9], let $\alpha_i(I), i \in I \subseteq M = \{1, \ldots, m\}$ denote the local significance levels for an intersection hypothesis such that

$$\sum_{i \in I} \alpha_i(I) \leqslant \alpha \tag{A1}$$

and the monotonicity condition

$$\alpha_i(I) \leqslant \alpha_i(J) \quad \text{for all } i, I, J \text{ with } i \in J \text{ and } J \subset I \subseteq M \tag{A2}$$

holds. Given the unadjusted p-values $p_i, i = 1, \ldots, m$, the corresponding closed test procedure based on local Bonferroni tests for the intersection hypotheses $H_I = \bigcap_{i \in I} H_i$ for all $I \subseteq M$ is consonant and equivalent to the following shortcut procedure [9, Theorem 1]:

0. Set $I = M$.
1. If there is a $j \in I$ such that $p_j \leqslant \alpha_j(I)$, then reject H_j; otherwise stop.
2. Set $I \to I \setminus \{j\}$.
3. If $|I| \geqslant 1$, go to step 1; otherwise stop.

(ii) *The graph $\mathcal{G}$ generates a consonant closed test*: Let $\mathcal{G}(M)$ be the initial weighted graph with transition matrix $\mathbf{G}(M)$ and initial significance levels $\boldsymbol{\alpha}(M)$ on all $m = |M|$ vertices representing the elementary hypotheses $H_1, \ldots, H_m$. We generate from $\mathbf{G}(M)$ and $\boldsymbol{\alpha}(M)$ a unique set of initial significance levels $\boldsymbol{\alpha}(J)$ for all nonempty subsets $J \subseteq M$ satisfying (A1) and (A2). This set is generated by the following inductive procedure: We start defining $\alpha_l(M) = \alpha_j$ and $g_{lk}(M) = g_{lk}$ for all $l, k \in M$. Given an index set $I \subseteq M$ and assuming the significance levels $\alpha_l(I)$ and the transition matrix $g_{lk}(I)$ for $l, k \in I$, we define significance levels and a transition matrix for all $J = I \setminus \{j\}$ with $j \in I$ by

$$\alpha_\ell(J) = \begin{cases} \alpha_\ell(I) + \alpha_j(I) g_{j\ell}(I), & \ell \in J \\ 0 & \text{otherwise} \end{cases} \tag{A3}$$

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2009, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.3495 by Gilead Sciences, Inc., Wiley Online Library on [06/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and for $|I|\geqslant 3$

$$g_{\ell k}(J)=\begin{cases}\dfrac{g_{\ell k}(I)+g_{\ell j}(I)g_{jk}(I)}{1-g_{\ell j}(I)g_{j\ell}(I)}, & \ell,k\in I,\ \ \ell\neq k\\ 0 & \text{otherwise}\end{cases} \tag{A4}$$

More generally, if $J=M\setminus\{j_1,\ldots,j_k\}$ for the k distinct indices $j_1,\ldots,j_k$ then we inductively apply (A3) and (A4) to the sequence $I_u=M\setminus\{j_1,\ldots,j_u\}$, $u=1,\ldots,k$, starting with $I_1=M\setminus\{i_1\}$. The final transition matrix $\mathbf{G}(I_k)$ and weights $\boldsymbol{\alpha}(I_k)$ are then assigned to J.

To end up with a unique set of transition matrices and weights, we need to show that they are independent from the ordering in which we have removed the indices $\{j_1,\ldots,j_u\}$. Certainly, this is only a question if more than one element is removed, i.e. for index sets of size $m-1$ uniqueness is out of question. Because every ordering can be obtained by successively interchanging neighboring indices in the sequence $j_1,\ldots,j_u$, independence from the ordering for $|J|<m-1$ would follow from $\boldsymbol{\alpha}[(I\setminus\{i\})\setminus\{j\}]=\boldsymbol{\alpha}[(I\setminus\{j\})\setminus\{i\}]$ for all $I\subset M$ with $|I|\geqslant 3$ and $i,j\in I$, and $\mathbf{G}[(I\setminus\{i\})\setminus\{j\}]=\mathbf{G}[(I\setminus\{j\})\setminus\{i\}]$ for all $|I|\geqslant 4$ with $i,j\in I$. Applying (A3) and (A4) we compute that for all $I\subset M$ with $|I|\geqslant 3$ and $i,j\in I, i\neq j$

$$\alpha_\ell[(I\setminus\{j\})\setminus\{i\}]=\begin{cases}\alpha_\ell+\dfrac{(\alpha_j g_{j\ell}+\alpha_i g_{i\ell})+(\alpha_i g_{ij}g_{j\ell}+\alpha_j g_{ji}g_{i\ell})}{1-g_{ij}g_{ji}}, & \ell\in I\setminus\{i,j\}\\ 0 & \text{otherwise}\end{cases} \tag{A5}$$

whereby we have put $g_{jl}=g_{jl}(I)$ for simplicity. Since this expression is symmetric in i and j we have $\boldsymbol{\alpha}(\{I\setminus\{j\}\}\setminus\{i\})=\boldsymbol{\alpha}(\{I\setminus\{i\}\}\setminus\{j\})$. Inverting (A3) shows that for all $I\subset M$ with $|I|\geqslant 4$ and $i,j,k,l\in I$, we have $g_{kl}[(I\setminus\{i\})\setminus\{j\}]=[\alpha_l(I\setminus\{i,j,k\})-\alpha_l(I\setminus\{i,j\})]/\alpha_k(I\setminus\{i,j\})$, which is also symmetric in i,j. This shows that $\boldsymbol{\alpha}(J)$ is uniquely defined for all $J\subseteq M$.

Monotonicity: By construction, property (A1) holds for the initial graph $\mathcal{G}(M)$ as well as the regularity conditions for the transition matrix $\mathbf{G}(M)$ stated in Section 2.2, i.e. $0\leqslant g_{ij}(M)\leqslant 1$, $g_{ii}(M)=0$, and $\sum_{j=1}^m g_{ij}(M)\leqslant 1$ for all $i=1,\ldots,m$. Assume that these conditions hold for any $I\subset M$ with $|I|\geqslant n, n<m$. Let $J=I\setminus\{j\}$, $j\in I$. Applying (A3) to $\alpha(I)$ and using the regularity condition for $\mathbf{G}(I)$ then guarantees conditions (A1) and (A2) to hold for all $J\subset M$ with $|J|=|I|-1$. Similarly, the regularity condition of $\mathbf{G}(J)$, for the same J, are easily verified by applying equation (A4) to $\mathbf{G}(I)$. By induction then the monotonicity and regularity conditions hold for all $I\subset M$.

(iii) *The shortcut applied to the closed test generated by graph $\mathcal{G}$ and the updating rule is equivalent to Algorithm 1*: Given an initial graph $\mathcal{G}(M)$ and univariate p-values $p_i, i\in M$, one can now apply the shortcut procedure described in (i). Given the initial weight vector $\boldsymbol{\alpha}(M)$, one performs step 1 of the algorithm, i.e. checks if a hypothesis H_j can be rejected at its local significance level $\alpha_j(M)$. If this is the case, one has to check whether any further rejection is possible with the significance levels $\boldsymbol{\alpha}(M\setminus\{j\})$. According to (ii), these levels can be computed via equation (A3). If again a rejection of one of the remaining hypotheses is possible, one computes the transition matrix $\mathbf{G}(M\setminus\{j\})$ using (A4) to obtain the weight vector $\boldsymbol{\alpha}$ for the remaining hypotheses, and so on. This, however, describes exactly Algorithm 1. Note that if at any step more than one hypothesis could be rejected, it does not matter which one to select. The rule given in Algorithm 1 is just one of the possible selection rules, see also Remark (iii) in Section 2.2.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2009, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.3495 by Gilead Sciences, Inc., Wiley Online Library on [06/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ACKNOWLEDGEMENTS

We thank Dejun Tang for providing us the case study and for continuing discussions during the progress of this work. We are grateful to Peter Bauer, Richardus Vonk and two anonymous referees for their helpful comments, which improved the presentation of the article.

REFERENCES

1. Holm S. A simple sequentally 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. In *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. Chen X, Luo X, Caprizzi T. The application of enhanced gatekeeping strategies. *Statistics in Medicine* 2005; **24**:1385–1397.
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. Marcus R, Peritz E, Gabriel KR. On closed testing procedure with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
11. Conforti M, Hochberg Y. Sequentally rejective pairwise testing procedures. *Journal of Statistical Planning and Inference* 1987; **17**:193–208.
12. Edwards D, Madsen J. Constructing multiple test procedures for partially ordered hypothesis sets. *Statistics in Medicine* 2007; **26**:5116–5124.
13. Bauer P, Brannath W, Posch M. Multiple testing for identifying effective and safe treatments. *Biometrical Journal* 2001; **43**:605–616.
14. Westfall PH, Young SS. *Resampling-based Multiple Testing*: *Examples and Methods for P-value Adjustment*. Wiley: New York, 1993.
15. 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.
16. Guilbaud O. Simultaneous confidence regions corresponding to Holm's stepdown procedure and other closed-testing procedures. *Biometrical Journal* 2008; **50**:678–692.
17. Hommel G, Bretz F. Aesthetics and power considerations in multiple testing—a contradiction? *Biometrical Journal* 2008; **50**:657–666.
18. Guilbaud O. Bonferroni parallel gatekeeping—transparent generalizations, adjusted *p*-values, and short direct proofs. *Biometrical Journal* 2007; **49**:917–927.
19. Sheskin TJ. Computing absorption probabilities for a Markov chain. *International Journal of Mathematical Education in Science and Technology* 1991; **22**:799–805.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

10970258, 2009, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/sim.3495 by Gilead Sciences, Inc., Wiley Online Library on [06/02/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License