Tutorial in Biostatistics

Statistics in Medicine

Received 2 December 2011, Accepted 26 August 2013 Published online 16 September 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5974

Advanced multiplicity adjustment methods in clinical trials

Mohamed Alosh,[a*†‡] Frank Bretz[b] and Mohammad Huque[c‡]

During the last decade, many novel approaches for addressing multiplicity problems arising in clinical trials have been introduced in the literature. These approaches provide great flexibility in addressing given clinical trial objectives and yet maintain strong control of the familywise error rate. In this tutorial article, we review multiple testing strategies that are related to the following: (a) recycling local significance levels to test hierarchically ordered hypotheses; (b) adapting the significance level for testing a hypothesis to the findings of testing previous hypotheses within a given test sequence, also in view of certain consistency requirements; (c) grouping hypotheses into hierarchical families of hypotheses along with recycling the significance level between those families; and (d) graphical methods that permit repeated recycling of the significance level. These four different methodologies are related to each other, and we point out some connections as we describe and illustrate them. By contrasting the main features of these approaches, our objective is to help practicing statisticians to select an appropriate method for their applications. In this regard, we discuss how to apply some of these strategies to clinical trial settings and provide algorithms to calculate critical values and adjusted p-values for their use in practice. The methods are illustrated with several numerical examples. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: multiple testing; α-propagation; adaptive alpha; gatekeeping; graphical methods

1. Introduction

Many clinical trials aim at multiple study objectives, which are formulated by suitable multiple hypotheses for establishing an efficacy claim. These objectives might include investigating treatment effects on multiple endpoints, time points, subgroups, doses, interim analyses, and sometimes a combination of these factors. It is well known that testing multiple hypotheses without any adjustment may increase the probability of erroneously rejecting at least one true null hypothesis. This error probability is commonly known as the familywise error rate (FWER). Various procedures have been proposed in the literature to guarantee strong control of the FWER at a pre-specified significance level α, that is, under any configuration of true and false null hypotheses. For a good exposure to these methods, we refer the interested reader to the books by Hochberg and Tamhane [1], Dmitrienko *et al.* [2], Bretz *et al.* [3] or Dmitrienko and D'Agostino [4] and the references therein.

This article discusses some recent advances in multiple testing strategies, which in many cases, provide more flexibility than traditional methods in addressing multiple study objectives. With these methods, one can, for example, test hierarchically ordered hypotheses even if a hypothesis in the beginning of the sequence is not rejected or test a hypothesis at a higher significance level than initially allocated by benefiting from decisions on hypotheses tested earlier in that sequence. Despite their flexibility, these methods strongly control the FWER. Because of space limitation, we focus on few selected topics based on our possibly biased preferences and experiences, with the risk that there are other important contributions not covered here.

[a] *Division of Biometrics III, Office of Biostatistics, OTS, CDER/FDA, Silver Spring, MD, U.S.A.*
[b] *Novartis Pharma AG, Basel, Switzerland and Hannover Medical School, Germany*
[c] *Office of Biostatistics, OTS, CDER, FDA, Silver Spring, MD, U.S.A.*
**Correspondence to: Mohamed Alosh, Division of Biometrics III, Office of Biostatistics, OTS, CDER/FDA, Silver Spring, MD, U.S.A.*
[†] *E-mail: Mohamed.Alosh@fda.hhs.gov*
[‡] *The views expressed in this paper are the authors' personal views and not necessarily those of the U.S. Food and Drug Administration*

The methodologies considered in this tutorial article are related to the following: (a) recycling local significance levels to test hierarchically ordered hypotheses; (b) adapting the significance level for testing a hypothesis to the findings of testing previous hypotheses within a given test sequence, with particular emphasis on certain consistency requirements; (c) grouping hypotheses into hierarchical families of hypotheses along with recycling the significance level between those families; and (d) graphical methods that permit repeated recycling of the significance level. These four methodologies are closely related to each other, and we point out some connections as we describe and illustrate them. For example, fallback procedures (a) can be regarded as a special case of gatekeeping (c), which in turn can often be visualized using the graphical methods (d).

It should be noted that with the FWER being strongly controlled at level α, the power differences between the various methods are frequently small and no method is expected to uniformly outperform other methods. However, understanding the subtle differences between these methods helps the practitioner to select an appropriate procedure for a specific situation at hand. Obviously, the more information is available about the magnitude of the treatment effects and the relationship among the multiple hypotheses, for example, information that might be obtained from previous trials, the better one would be equipped for selecting an appropriate testing method. However, in the light of the frequently limited information available when conducting confirmatory trials, the flexibility in designing a trial to address efficiently its objectives and the ability to derive suitable multiple testing procedures becomes critical to the success of the clinical development program at hand.

2. Clinical trial examples

In this section, we describe several clinical trial examples that will be re-visited later to illustrate some of the methods described in the sequel. The following examples motivate the need for advanced multiple testing procedures in the context of real problems. We also describe some basic concepts, before introducing them more formally in Section 3.

2.1. Two doses and two hierarchical endpoints

Consider a three-armed trial comparing two doses of a new compound (e.g., low and high) with placebo for two endpoints (e.g., relapse rate and total medication score), resulting in two levels of multiplicity and four null hypotheses H_1, H_2, H_3, H_4. In addition to the given set of null hypotheses, clinical considerations often lead to a structured hypotheses testing problem subject to certain logical constraints. Assume for our example that relapse rate is more important than total medication score, thus grouping the hypotheses into two primary hypotheses H_1, H_2 (both dose–placebo comparisons for relapse rate) and two secondary hypotheses H_3, H_4 (both dose–placebo comparisons for total medication score). Both doses are considered equally important, which rules out a full hierarchy of testing first the high dose and, conditional on its significance, then the low dose. In addition, it is required that a secondary hypothesis is not tested without having rejected the associated primary hypothesis (successiveness property, see [5, 6]). The objective is to test all four hypotheses under strong FWER control while reflecting the clinical considerations above and without leading to illogical decisions [7, 8]. Standard multiple comparison procedures, such as those by Bonferroni, Holm, or Dunnett, are not suitable here, because they treat all four hypotheses equally and do not address the underlying structure of the test problem. Instead, one might consider intuitive approaches that combine standard multiple testing procedures in a seemingly sensible way to reflect the structured trial objectives. For example, one could consider testing the two primary hypotheses H_1, H_2 using the Holm procedure [9] at level α. If at least one hypothesis is rejected, one would continue testing the descendant secondary hypothesis at level $\alpha/2$. However, this and many similar procedures do not control the FWER at level α. Instead, tailored test procedures are needed to fulfill the trial objectives, such as those described in Section 4.

2.2. Composite endpoint with possible claim for 'hard' endpoints

Consider a cardiovascular clinical trial with four outcomes of interest: all-cause mortality, non-fatal myocardial infarction, stroke, and hospitalization. The mortality outcome is of primary interest. However, as mortality event rates are usually very low, it is common for such trials to investigate the treatment effect on a composite endpoint constructed from the individual components in order to keep the trial size and duration manageable; see, for example, [10] for a related discussion.

The use of composite endpoints may lead to a reduction of trial size and duration and avoids at first glance the multiplicity issue arising from testing the individual outcomes. However, interpretation of study findings based on a composite endpoint can be problematic. This is in particular true if a statistically significant treatment effect for the composite endpoint was driven mainly by a 'soft' endpoint, such as hospitalization, and treatment effects for the 'hard' endpoints are small or even in the opposite direction than that for the composite endpoint. Consequently, for an appropriate interpretation of study results, it is important to analyze the treatment effects for the individual components.

The analysis of individual components is merely descriptive if it is based on reporting mean responses, nominal confidence intervals, and p-values or forest plots, together with the findings for the composite endpoint. In this case, no formal claim is intended for the component endpoints.

In contrast, it is sometimes of interest to establish an efficacy claim for a key individual component, such as all-cause mortality, by testing its corresponding hypothesis, say H_2, after the composite endpoint hypothesis, say H_1, is either rejected or missed slightly the significance level α. In such cases, multiplicity needs to be formally taken into account. Standard multiple testing procedures, such as the Bonferroni test, may not be appropriate as they treat both hypotheses H_1 and H_2 as equally important and test them individually regardless of the findings of the other hypothesis. A hierarchical test procedure might not be appropriate either, as failure to reject the hypothesis of the composite endpoint H_1 prohibits testing the individual endpoint hypothesis H_2. In Section 5.2, we revisit this problem and consider multiple testing strategies that allow testing H_2 as long as the result of testing H_1 establishes a pre-specified minimum level of efficacy, a so-called consistency criterion. Furthermore, the significance level for testing the mortality hypothesis H_2 can be adapted to the findings of testing the composite endpoint hypothesis H_1.

2.3. Testing for non-inferiority and superiority for multiple doses and endpoints

Consider a four-armed clinical trial for treatment of psoriasis. The trial is designed to compare two doses of a new compound (e.g., low and high) with placebo and an active control. Establishing an efficacy claim for this indication is usually based on two endpoints: a score of 0 or 1 on a 5-point Investigator Global Assessment (IGA) scale and a success criterion defined as change from baseline of at least 75% on the Psoriasis Area and Severity Index scale (PASI 75). These two endpoints are co-primary in the sense that efficacy has to be shown on both for a given dose. In addition, PASI 90 is often considered as a key secondary endpoint for this indication, that is, a change of at least 90% from baseline PASI score. For this example, we assume that the objectives of the clinical trial are to establish (1) treatment efficacy for either dose in comparison with placebo; (2) non-inferiority of the winning dose(s) relative to the active control; and (3) superiority to active after establishing non-inferiority for the winning dose(s).

These trial objectives result in a total of 10 null hypotheses: for each of the two doses, we have three superiority hypotheses against placebo (PASI 75, IGA 0 or 1, PASI 90) and two hypotheses against active control (non-inferiority and superiority for PASI 75). These hypotheses have to be tested in a clinically meaningful way while controlling the FWER in the strong sense at level α. For this example, we assume that the two doses are considered equally important. Once efficacy has been established for a certain dose on the basis of the two co-primary endpoints, testing the key secondary endpoint, PASI 90, for superiority against placebo and PASI 75 for non-inferiority against the active comparator is equally important. Furthermore, if the non-inferiority hypothesis is rejected for PASI 75, the next objective is to test the same endpoint for superiority. Section 5.3 proposes a gatekeeping strategy that appropriately handles the hierarchical ordering of the hypotheses and their structural relationship. In addition, Section 5.3 presents a graphical display of the resulting multiple testing strategy to visualize the sequential testing scheme that takes into account the aforementioned clinical considerations through the initial split of the significance level α across the 10 hypotheses and a suitable set of α-propagation rules.

3. Notations, definitions, and some standard multiple test procedures

We consider the general problem of testing k null hypotheses $H_1, H_2, \ldots, H_k$, which could be related to multiple hierarchical endpoints, multiple doses or treatments, non-inferiority or superiority, multiple populations, or any combination thereof. Let $\bigcap_{i=1}^{k} H_i$ denote the global null hypothesis. If the hypotheses are given in a pre-specified hierarchical order, denote them by $H^{(1)}, H^{(2)}, \ldots, H^{(k)}$. Let $\alpha_1, \alpha_2, \ldots, \alpha_k$ denote the initial allocation of the significance level α to the individual null hypotheses, such that $\alpha_i = w_i\alpha$ for given non-negative weights w_i subject to $\sum w_i \leqslant 1$. Let further $T_1, T_2, \ldots, T_k$ denote the test statistics and $p_1, p_2, \ldots, p_k$ the corresponding unadjusted p-values. Some methods rely

on the ordered p-values $p_{(1)} \leqslant p_{(2)} \leqslant \ldots \leqslant p_{(k)}$, where $H_{(1)}, H_{(2)}, \ldots, H_{(k)}$ denote the accordingly ordered hypotheses. Following the standard conventions for confirmatory clinical trials, we aim at controlling the FWER in the strong sense at level α. That is, if V denotes the number of rejected true null hypotheses, we aim at controlling the probability $P(V > 0)$ under any configuration of true and false null hypotheses, that is,

$$\max_{J \subseteq \{1,\ldots,k\}} P\left(V > 0 \left| \bigcap_{i \in J} H_i \right.\right) \leqslant \alpha.$$

Most multiple testing procedures described in this paper can be formulated as closed testing procedures, which operationally can be defined as follows. Consider the $2^k - 1$ intersection hypotheses $H_J = \cap_{i \in J} H_i$ for all non-empty subsets $J \subseteq 1, \ldots, k$. Each H_J is tested at level α or less. An individual null hypothesis H_i is rejected if every intersection hypothesis H_J with $i \in J$ is rejected by its local level α test. This controls the FWER in the strong sense at level α [11]. Following [12], we denote a closed testing procedure α-exhaustive, if the size of each intersection hypothesis test equals α, that is, $P(\text{reject } H_J) = \alpha$ under the null for all $J \subseteq 1, \ldots, k$.

To illustrate the closure principle, suppose we want to compare two doses of a new compound with placebo for a single primary endpoint. Thus, we want to test the $k = 2$ null hypotheses $H_1\colon \mu_1 - \mu_0 \leqslant 0$ and $H_2\colon \mu_2 - \mu_0 \leqslant 0$ that dose $i = 1, 2$ is not better than placebo $i = 0$, where μ_i denotes the mean effect of treatment $i = 0, 1, 2$. The two null hypotheses H_1 and H_2 are not disjoint: The intersection of both is given by $H_{12}\colon \mu_1 - \mu_0 \leqslant 0$ and $\mu_2 - \mu_0 \leqslant 0$, that is, both doses are not better than placebo. Testing H_{12} requires a multiplicity adjustment, such as the Bonferroni procedure introduced in Section 3.1. Figure 1 visualizes the hypotheses H_1, H_2, and H_{12} using a Venn diagram. The closure principle then follows intuitively from the Venn diagram: Because H_{12} is a subset of H_1, it has to be tested and rejected first. The same relationship holds also for H_2. Consequently, applying the closure principle, one would first test the single intersection hypothesis H_{12} (using, for example, the Bonferroni test at level α). If H_{12} is rejected, one would then continue testing the individual null hypotheses H_1 and H_2 (using, for example, standard t-tests at level α). If H_{12} is not rejected, one can stop proceeding further because neither H_1 nor H_2 can be rejected. For $k > 2$ hypotheses, the principle remains the same, but we have to consider many more intersection hypotheses. Further illustration and discussions are given, for example, in Section 2.2.3 in Bretz *et al.* [3].

Bauer [13] pointed out that closed testing procedures are optimal whenever the FWER has to be controlled. The reason is that any closed testing procedure is by construction coherent. That is, if a null hypothesis H_i is rejected, all subsets $H_J \subseteq H_i$ are rejected as well [14]. Coherence is an important requirement for any multiple testing procedure, as otherwise problems with the interpretation of the test results may occur. If, for example, H_1 is rejected in Figure 1 but H_{12} is not, we would have problems in interpreting these results. The closure principle avoids such problems as it implies coherence by construction: It first tests H_{12}, and only if this is rejected, it continues testing the individual null hypotheses. Sonnemann and Finner [15] showed in addition that any non-coherent multiple testing procedure can be replaced by a coherent procedure that never rejects less but may reject more hypotheses. Furthermore, any coherent multiple test controlling the FWER is a closed test. Thus, for standard clinical trial applications, there is no need to consider a multiple test that is not a closed testing procedure.

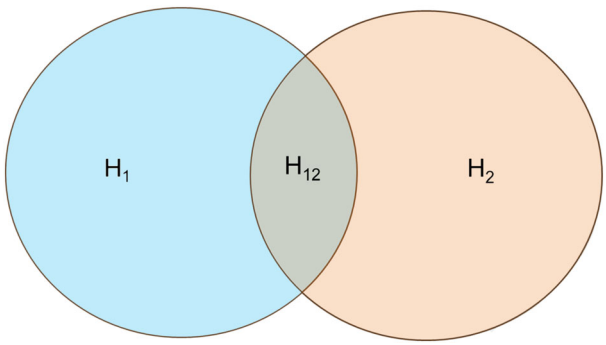

Figure 1. Venn diagram for two hypotheses H_1 and H_2 and their intersection H_{12}.

3.1. Bonferroni procedure, with extensions

The Bonferroni procedure is the simplest method among all p-value-based procedure. In its original form, the significance level α is split equally among the k hypotheses, and each is tested at level α/k. That is, the individual null hypothesis H_i is rejected if $p_i < \alpha/k$, and the global null hypothesis H is rejected if $\min p_i < \alpha/k$. A weighted version of the Bonferroni procedure allows unequal allocations of the significance level, which is useful for testing unequally important hypotheses. In this more general case, the hypothesis H_i is tested at level $\alpha_i = w_i\alpha$, with $w_i \geqslant 0$ and $\sum w_i \leqslant 1$.

The Bonferroni procedure always controls the FWER in the strong sense, and every hypothesis is tested regardless of the findings for the other hypotheses. However, the latter property can cause difficulties when interpreting the overall study findings if the individual test results trend into different directions, for example, a new treatment being superior to placebo for one endpoint but inferior for another endpoint. In addition, the Bonferroni procedure is conservative and consequently leads to a power loss if the number of hypotheses k is large and/or if the corresponding test statistics are highly correlated.

Simes [16] proposed a sharpened Bonferroni inequality. Accordingly, the Simes procedure rejects the global null hypothesis H if there exists an $i \in 1, \ldots, k$ such that $p_{(i)} \leqslant i\,\alpha/k$. The Simes procedure has higher power than the Bonferroni procedure to reject the global null H, but its control of the type 1 error rate for the global test is guaranteed only if the test statistics are independent or positive regression dependence holds [17]. A weighted version of the Simes procedure is given by Benjamini and Hochberg [18].

To illustrate the Bonferroni and Simes procedures as well as other procedures described later in this tutorial, we use a numerical example with three p-values $p_1 = 0.02$, $p_2 = 0.005$, and $p_3 = 0.01$. Using $\alpha = 0.025$ (one sided), we reject H_2 with the Bonferroni test because $p_2 = 0.005 < 0.0083 = \alpha/3$, but no other hypothesis. Using the Simes test, we reject the global null hypothesis H because $p_{(1)} = p_2 = 0.005 < 0.0083 = \alpha/3$, but we cannot test any individual null hypothesis.

3.2. Holm and Hochberg testing procedures

Both the Bonferroni and Simes procedures described in the previous section can be improved, resulting in more powerful testing procedures. We first consider the Holm procedure [9], which extends the Bonferroni procedure. It starts by testing $H_{(1)}$ at level α/k and continues testing the subsequent ordered hypotheses only if $H_{(1)}$ is rejected. Specifically, $H_{(i)}$ is tested by comparing $p_{(i)}$ with $\alpha/(k-i+1)$ as long as all previous ordered hypotheses are rejected. Once an ordered hypothesis is not rejected, no further testing is possible. After each rejection, the Holm procedure essentially 'updates' the remaining significance levels by accounting for the number of remaining hypotheses yet to be tested. As a result, the Holm procedure has higher power than the Bonferroni procedure, that is, it rejects all hypotheses rejected by the Bonferroni procedure and possibly others. In fact, the Holm procedure is a closed testing procedure using Bonferroni procedure for each intersection hypothesis; see Figure 1 and the related discussion. The Holm procedure remains conservative as it does not exploit the correlation between the test statistics. Among the advantages of the Holm procedure is that it is simple to apply and it controls the FWER in the strong sense.

Next, we consider the Hochberg procedure [19], which is a conservative shortcut to a closed procedure (i.e., the Hommel procedure) that uses the Simes test for interaction hypotheses and allows assessing the individual null hypotheses H_i. The Hochberg procedure uses the same local significance levels as the Holm procedure but starts by comparing the largest p-value $p_{(k)}$ with α. If $p_{(k)} \leqslant \alpha$, it rejects all hypotheses and testing stops; otherwise, it retains $H_{(k)}$ and continues testing $H_{(k-1)}$ by comparing $p_{(k-1)}$ with $\alpha/2$ and so on. By construction, the Hochberg procedure rejects all hypotheses rejected by the Holm procedures and possibly others. The Hochberg procedure is further improved by the Hommel procedure [20], which uses the full closed testing procedure by employing the Simes test for each intersection hypothesis. In fact, the Hochberg procedure is a conservative shortcut of the Hommel closed testing procedure and thus less powerful. That is, the Hommel procedure rejects all hypotheses rejected by the Hochberg procedure and possibly others. As both the Hochberg and Hommel procedures rely on the Simes test, they share its limitations on the type I error rate control [17]. Weighted versions of the Holm procedure were discussed by Holm [9] and Zhang *et al.* [21], among others. Tamhane and Liu [22] investigated weighted versions of the Hochberg procedure. We refer to Hommel *et al.* [23] for a recent review of weighted and unweighted multiple testing procedures based on ordered p-values.

We continue with the numerical example from Section 3.1 to illustrate the Holm and Hochberg procedures. Consider first the Holm procedure. We start with the most significant p-value and reject H_2

because of $p_{(1)} = p_2 = 0.005 < 0.0083 = \alpha/3$. Next, $p_{(2)} = p_3 = 0.01 < 0.0125 = \alpha/2$, and we reject H_3 as well. Finally, $p_{(3)} = p_1 = 0.02 < 0.025 = \alpha$, and we reject H_1 as well. Using the Hochberg procedure, we start with the least significant p-value and immediately reject all three null hypotheses because $p_{(3)} = p_1 = 0.02 < 0.025 = \alpha$.

3.3. *Fixed sequence procedure*

If the sequence of hypotheses is fixed *a priori* on the basis of the relative importance of the individual hypotheses or expected treatment effects, one can apply a fixed sequence procedure [24, 25]. The procedure is carried out sequentially where $H^{(i)}$ is tested at level α as long as all previous hypotheses $H^{(1)}, \ldots, H^{(i-1)}$ are rejected; testing stops with the first non-significant result. Effectively, if the fixed sequence procedure rejects a hypothesis, its significance level α is propagated to the next hypothesis in the sequence, which in turn can be tested at level α.

The fixed sequence procedure has the advantage of being simple and controls the FWER strongly. It is optimal when the hypotheses early in the sequence are associated with large effects and perform poorly otherwise [25]. A major drawback of the fixed sequence procedure is that once a hypothesis is not rejected, no further testing is permitted. Thus, great care is advised when specifying the testing sequence prior to the study. In general, long chains of hypotheses tested in sequence lead to low power and often do not reflect the underlying study objectives. In the next section, we discuss several procedures that alleviate these concerns. For example, the fallback procedure reserves a fraction of α for the later hypotheses in the sequence and thus allows one to test those even if the initial hypotheses in the sequence are not rejected.

We now illustrate the fixed sequence procedure by revisiting the numerical example from Section 3.1. Assume that $H^{(1)} = H_1$, $H^{(2)} = H_2$, $H^{(3)} = H_3$ is the pre-specified sequence of hypotheses. We then start testing $H^{(1)}$ at level α and reject H_1 because of $p_1 = 0.02 < 0.025 = \alpha$. Next, we can test $H^{(2)}$ at level α to reject H_2 and similarly reject H_3 in the end as well. Note that if instead $p_1 = 0.03$ and thus only slightly larger than the observed 0.02, we would not have been able to reject any of the three hypotheses, regardless of the very small p_2 value.

4. Recent developments in multiple testing

This section discusses recent advances in multiple testing strategies, as outlined in Section 1. Section 4.1 introduces fallback procedures that recycle unspent significance levels to test subsequent hypotheses within a pre-specified hypothesis sequence. Section 4.2 discusses adaptations of the significance level for testing subsequent hypotheses based on the level of evidence from testing previous hypotheses. It also considers the requirement of establishing a certain minimum level of evidence, called consistency criterion, as a requisite for testing other hypotheses. Section 4.3 considers grouping multiple hypotheses into families of hypotheses, based on the hierarchical objectives of a clinical trial, along with recycling the significance level between those families. Section 4.4 discusses graphical methods that permit repeated recycling of the significance level.

4.1. *Recycling unspent significance levels when testing hierarchical hypotheses*

The fallback procedure introduced by Wiens [26] addresses the major drawback of fixed sequence procedure by allowing one to test all hypotheses in the pre-specified sequence even if the initial hypotheses are not rejected. The significance level α is split among the k hypotheses with $\alpha_i = w_i\alpha$, allocated for testing $H_i (i = 1, 2, \ldots, k)$ such that $\sum w_i = 1$ (as in the Bonferroni test). However, the hypotheses are ordered prospectively, and testing is carried out sequentially (as in the fixed sequence procedure). Then, $H^{(i)}$ is tested at level $\acute{\alpha}_i$, where $\acute{\alpha}_i = \alpha_i$ if $H^{(i-1)}$ is not rejected and $\acute{\alpha}_i = \alpha_i + \acute{\alpha}_{i-1}$ if $H^{(i-1)}$ is rejected. Testing proceeds in a sequential fashion similar to the fixed sequence procedure, with the distinction that later hypotheses in the sequence can be tested at a larger local significance level than initially assigned, even if one or more hypotheses in the beginning of the sequence are not rejected. Because of this α propagation, the fallback procedure has more power than the Bonferroni procedure. Note that the fixed sequence tests is obtained as a special case from the fallback procedure by setting $w_1 = 1$ and $w_i = 0$ for $i > 1$.

Wiens and Dmitrienko [27] showed that the fallback procedure from Wiens [26] is a closed testing procedure and thus controls the FWER strongly. Furthermore, they noted that the original fallback procedure can be made more powerful by exhausting the significance level using α-exhaustive tests (see

Section 3 for a formal definition). However, this extension of the fallback procedure is at the expense of its simplicity. Hommel and Bretz [28] introduced an alternative α-exhaustive extension, where significance levels are recycled first to hypotheses higher in the hierarchy, thus enhancing the power for those hypotheses that are deemed to be more important.

We illustrate the fallback procedure [26] by revisiting the numerical example from Section 3.1. Assume that $H^{(1)} = H_1$, $H^{(2)} = H_2$, $H^{(3)} = H_3$ is the pre-specified sequence of hypotheses and equal weights $w_1 = w_2 = w_3 = 1/3$. Consequently, the initial significance level for testing each of the three hypotheses is $\alpha/3$. As $p_1 = 0.02 > 0.0083 = \alpha/3$, we cannot reject H_1 and no propagation of its significance level to H_2 is possible. Next, we reject H_2 because $p_2 = 0.005 < 0.0083 = \alpha/3$ and update the significance level for testing H_3 to $\acute{\alpha}_3 = \alpha_3 + \acute{\alpha}_2 = 2\alpha/3$. Now, we can reject H_3 as well because $p_3 = 0.01 < 0.0167 = 2\alpha/3$. Note that without the propagation of α_2, we cannot reject H_3. Figure 4 provides a graphical visualization of the rejection sequence and iterated graphs for this example. Note also that we can reject H_1 using the α-exhaustive fallback procedure from [27].

Huque and Alosh [29] proposed a parametric extension of the fallback procedure, denoted as flexible fixed sequence (FFS) or parametric fallback procedure. The FFS procedure accounts for the correlation among the test statistics when calculating the local significance levels. The authors presented an algorithm for calculating these significance levels when the test statistics have a multivariate normal distribution. For example, for $k = 2$ hypotheses, $H^{(1)}$ is tested at level α_1. Next, $H^{(2)}$ is tested at level α if $H^{(1)}$ is rejected and at level γ_2 otherwise, where γ_2 is obtained by solving

$$\alpha - \alpha_1 = \Pr\left(T_1 < z_{\alpha 1}, T_2 \geqslant z_{\gamma 2} | H^{(1)} \cap H^{(2)}\right) \tag{1}$$

and z_x denotes the $(1-x)$ percentile of the standard normal distribution. The authors noted that the gain in power by accounting for the correlation can range from 0% to 8% and depends on the prospective ordering of the hypotheses, the magnitude of treatment effects, the initial allocation of the significance level, and the correlation. The Supporting information S.1[§] lists a SAS/IML program to compute γ_2 for $k = 2$ and given design parameters α_1 and $\alpha - \alpha_1$ (for one-sided tests) when the distribution of the test statistics T_1 and T_2 is bivariate normal with correlation ρ. The Supporting information S.2 lists a similar SAS/IML program to compute γ_3 for $k = 3$ for given design parameters α_1, α_2, and $\alpha - \alpha_1 - \alpha_2$, when the distribution of the test statistics T_1, T_2, and T_3 is trivariate normal with correlation matrix R.

We illustrate the FFS procedure through its application to the numerical example in Section 3.1. As this procedure accounts for the correlation between the endpoints, we assume for simplicity that the endpoints are equally correlated with $\rho = 0.6$ and assume, as that mentioned earlier, $H^{(1)} = H_1$, $H^{(2)} = H_2$, $H^{(3)} = H_3$. Because $p_1 = 0.02 > 0.0083 = \alpha/3$, we cannot reject H_1 and consequently cannot propagate its significance level to H_2. Next, we run the SAS code from the Supporting information S.1 to calculate the significance level γ_2 for testing H_2, with $alpha1 = alpha2 = 0.025/3$ (i.e., for 0.05/3 for 2-sided tests) and $r = 0.6$. We used *italics* to distinguish between the input and output parameters for the SAS code and rounded outputs to three decimal places. Running the code results in $C2 = 2.326$, which consequently can be imputed in the last section of the SAS code to give $\gamma_2 = 0.010$. Now, the significance level for testing H_3, γ_3, can be calculated by running the SAS code in the Supporting information S.2 with $C2 = 2.326$, $alpha1 = alpha2 = alpha3 = 0.025/3$ and $r = 0.6, 0.6, 0.6$, which gives $C3 = 2.273$. This value can be transferred back to the last section of SAS code to give $\gamma_3 = 0.012$. With these calculations, we can reject H_2 because $p_2 = 0.005 < 0.010$. Then, with the rejection of H_2, we test H_3 at the significance level $\gamma_3^* = \max(\alpha_2 + \alpha_3,\ \gamma_3) = 2\alpha/3$ and reject it because $p_3 = 0.01 < 0.0167 = 2\alpha/3$.

Wiens and Dmitrienko [30] compared the fallback, alpha-exhaustive fallback, weighted Holm, and fixed sequence procedures with two or three hypotheses by expressing the individual decision rules in terms of closed testing. The findings of their comparisons indicate that there is no universally superior procedure and selection among these procedures is driven by knowledge of treatment effects. The fallback procedure outperforms the fixed sequence procedure when early hypotheses in the test sequence have low power. If the treatment effects are unknown, the authors concluded that an optimal procedure might not be available and compromise procedures that retain some level of power for a broad range of plausible scenarios are preferable.

[§] *Supporting information may be found in the online version of this article.*

4.2. Adapting alpha to the findings of previous testing and consistency requirement

While the significance level for testing a hypothesis with the fallback depends on whether the previous hypothesis in the sequence was rejected, a more general approach is to allow the significance level for testing a hypothesis later in the sequence to depend on the level of evidence of testing earlier hypotheses. The first procedure in this direction was introduced by Li and Mehrotra [31] to deal with potentially underpowered endpoints, denoted as adaptive alpha allocation approach (4A).

The 4A method was developed for testing multiple endpoints. Consider the case of $k = 2$ endpoints. Like the fallback procedure, it allocates $\alpha_i = w_i\alpha$ for testing $H^{(i)}, i = 1, 2$, where $w_1 + w_2 = 1$. If the endpoints are independent, the procedure starts by testing $H^{(1)}$ at level α_1; if rejected, $H^{(2)}$ is tested at level α by accumulating the unspent significance levels. Otherwise, $H^{(2)}$ is tested at level $\acute{\alpha}_2 = \min\left(\alpha_t / p_1^2, \alpha_1\right)$, where

$$\alpha_t = \begin{cases} \alpha_1 \frac{\alpha - \alpha_1}{1 - \alpha_1}, & \text{if } \alpha_1 + \alpha_1^2 - \alpha_1^3 > \alpha, \\ \alpha_1 \left(1 - \sqrt{\frac{2\alpha_1 - \alpha - \alpha_1^2}{\alpha_1}}\right)^2, & \text{otherwise.} \end{cases}$$

For two correlated endpoints, the authors proposed adjusting α_t to control the FWER and provided values for $\acute{\alpha}_2$ under various configurations of $0 \leqslant p_1 \leqslant 0.90$ and $0 \leqslant \rho \leqslant 0.9$. Li and Mehrotra noted that the 4A method can be extended to the case of three or more independent endpoints classified into two sets. In such cases, the p-values for the second set of the endpoints, which might be underpowered, are assessed at an adaptive overall level $\acute{\alpha}_2$, which is a function of the maximum p-value of the first set of endpoints along with α_1, α and the number of endpoints in the first set. The authors proved that the 4A method strongly controls the FWER for two or more independent endpoints and noted that further research is needed to extend this procedure to correlated endpoints.

According to the 4A method, the significance level for testing $H^{(2)}$ is a non-increasing function of the level of evidence of testing $H^{(1)}$ through the value of p_1 (Figure 2). Specifically, the weaker the evidence against the first null hypothesis is, the stronger the evidence against the second null hypothesis would have to be for a positive trial outcome. Although this seems intuitive, for a clinically meaningful interpretation of the overall study results, it is necessary that the finding of the individual endpoints trend in the same direction, at a minimum. The 4A method does not impose such a consistency requirement, as it allows one to test $H^{(2)}$ even for relatively large values of p_1, although at smaller significance levels. In the remainder of this section, we discuss other approaches that require that the results of $H^{(1)}$ meet a certain level of evidence, called consistency criterion, as a pre-requisite for testing $H^{(2)}$. In addition, these methods allow any non-increasing function in the p_1-values to be used for adapting the significance level for testing $H^{(2)}$ and thus provide more flexibility than the 4A method.

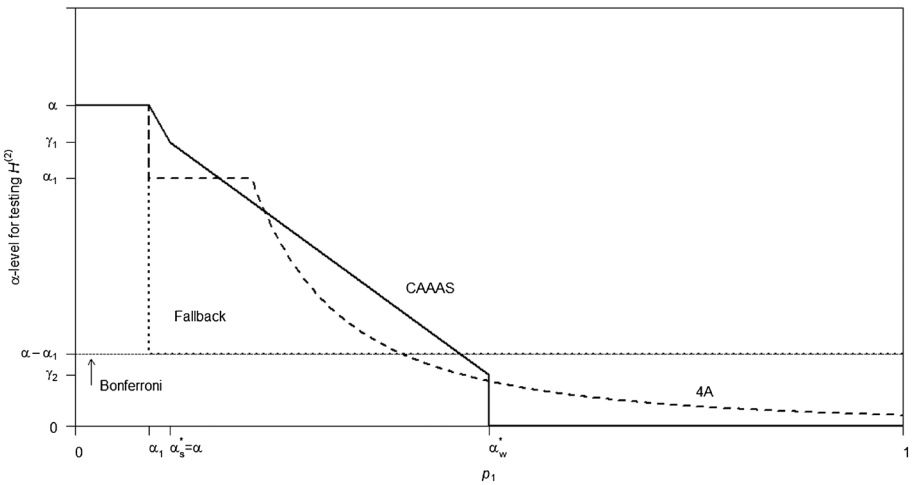

Figure 2. Comparison of the Bonferroni, fallback, adaptive alpha allocation approach (4A), and consistency-adjusted alpha-adaptive strategy (CAAAS) procedures (adapted from [39]).

The consistency concept was first introduced in the context of clinical trials aiming at testing both an overall population and a pre-specified subgroup. For such trials, Song and Chi [32] and Alosh and Huque [33,34] required that efficacy in the overall population should meet a pre-specified degree of consistency, denoted as α^*, before testing efficacy in the subgroup. For one-sided tests, one should select $\alpha^* \leqslant 0.5$ to ensure that results of $H^{(1)}$ are not favoring the control treatment.

Consistency can be also applied to test a designated primary endpoint and an alternative endpoint for establishing an efficacy claim. In this case, the findings of the two endpoints should be at least trending in the same direction. Otherwise, it would be difficult to interpret the overall study findings if the results of the alternative endpoint were statistically significant but the results of the designated primary endpoint do not show a trend in efficacy, or even worse, if they were in the opposite direction (i.e., a p-value larger than 0.5 for a one-sided test). Note that introducing a consistency requirement does not only ensure a clinically meaningful interpretation of study findings but also improves the significance level for testing $H^{(2)}$ when the consistency requirement for $H^{(1)}$ is met.

The aforementioned consistency concept is unidirectional, as it assigns more weight to the overall population or the designated primary endpoint as compared with the subgroup or the alternative endpoint, respectively. In other applications, it might be meaningful to consider a symmetric consistency requirement. Examples include the investigation of two equally important endpoints or subgroups. In this case, one might require that findings of testing each of the endpoints or subgroups meet certain consistency thresholds. This can be accomplished by requiring $p_{(2)} \leqslant \alpha^*$ [35]. The concept can be generalized to more than two endpoints, but it is beyond the scope of this tutorial.

The consistency criterion can be coupled with other properties of recently proposed procedures to provide a more general approach for multiple testing. The flexible strategy for testing a subgroup and the overall population (FSTS) by Alosh and Huque [34] can be viewed as a generalization of FFS by including the consistency requirement. With the initial allocation of the significance levels α_1 and $\alpha - \alpha_1$ for testing the hypotheses H_1 and H_2 (corresponding to the overall and a pre-specified subgroup, respectively), and for a consistency criterion α^* such that $\alpha_1 < \alpha < \alpha^*$, the testing strategy can be outlined as follows. Test H_1 at level α_1. If H_1 is rejected, then test H_2 at level α. If H_1 is not rejected and its p-value meets the consistency criterion, that is, $p_1 \leqslant \alpha^*$, then test H_2 at the updated significance level α_2, where α_2 is obtained by the following relations after solving first for α_s:

$$\alpha - \alpha_1 = \Pr([\alpha_1 \leqslant P_1 < \alpha^*)] \cap [P_2 \leqslant \alpha_s] | H_1 \cap H_2) \quad \text{and} \quad \alpha_2 = \min(\alpha_s, \alpha). \tag{2}$$

If p_1 does not meet the consistency criterion, no further testing is allowed.

Huque and Alosh [36] extended the aforementioned procedure, denoted as consistency-adjusted strategy (CAS), by including an extra step that allows rejecting both H_1 and H_2 if $\alpha_1 \leqslant p_1 < \alpha$ and $p_2 \leqslant \alpha_2 \leqslant \alpha$. A SAS/IML program for calculating the significance level for testing H_2 in CAS, when the distribution of the test statistics T_1 and T_2 for testing H_1 and H_2 has a bivariate normal distribution with correlation ρ, is given in Huque *et al.* [37]. In addition, the authors introduced a modification of the 4A method to address the consistency requirement and provided the significance level for testing H_2 in this case.

While the 4A method uses a specific functional form to adapt the significance level of H_2 as a function of p_1, in practice, it might be more convenient to use a simpler non-increasing function. In this regard, two approaches have been introduced in the literature. Zhao *et al.* [38] introduced the feedback procedure in the context of subgroup analysis, which allows the significance level for testing a targeted subgroup to be determined through a general non-increasing α–spending function in the p-value for the total population. The authors noted that several existing procedures, including the parametric fallback, 4A, the procedure introduced by Song and Chi [32] and the FFS, arise as special cases by selecting the form of their α-spending function. The authors also discussed restricting the function to meet a consistency criterion. The second approach is an extension of the CAS and is denoted as consistency-adjusted alpha-adaptive strategy (CAAAS) introduced by Alosh and Huque [39].

The CAAAS is a general approach for addressing multiplicity, which allows two levels of consistency as well as general non-increasing functions in different regions of the p_1-value for adapting the significance level for testing the next hypothesis. Basically, the CAAAS distinguishes between three levels of evidence in the findings of the designated primary endpoint within the consistency region and considers different forms of adaptation of the significance level for testing the second hypothesis H_2 depending on the p-value from testing H_1. Specifically, if H_1 is rejected, then its allocated level α_1 is recycled and H_2 is tested at level α. Otherwise, if H_1 is not rejected at level α_1, but its p-value, p_1, is still relatively

small and meets a strong consistency requirement, α_s^*, with $\alpha_1 < \alpha_s^* \leqslant \alpha$, then H_2 is tested at level $\alpha_{22} = f_2(p_1)$; if $p_2 \leqslant \alpha_{22}$, then both H_2 and H_1 would be rejected. Finally, if H_1 is not rejected, but its p-value meets a weak consistency requirement α_w^*, with $\alpha \leqslant \alpha_w^* \leqslant 0.5$, then H_2 is tested at level $\alpha_{23} = f_3(p_1)$. No testing for H_2 is permitted if $p_1 \geqslant \alpha_w^*$. Clearly, the strength of evidence in interpreting the overall study findings depends on the consistency level observed from testing H_1 and the results of testing H_2.

The values α_{22} and α_{23} are determined through

$$\begin{aligned}\alpha - \alpha_1 = &\Pr\left(\left[\alpha_1 \leqslant P_1 < \alpha_s^*\right] \cap \left[P_2 \leqslant f_2(p_1)\right] | H_1 \cap H_2\right) \\ &+ \Pr\left(\left[\alpha_s^* \leqslant P_1 < \alpha_w^*\right] \cap \left[P_2 < f_3(p_1)\right] | H_1 \cap H_2\right),\end{aligned} \tag{3}$$

where the joint distribution of P_1 and P_2 is derived under the global null hypothesis $H_1 \cap H_2$ and $\alpha_{22} = \min(f_2(p_1), \alpha), \alpha_{23} = \min(f_3(p_1), \alpha)$ and $\alpha_{23} \leqslant \alpha_{22} \leqslant \alpha$.

Alosh and Huque [39] have shown that several multiple testing procedures, such as fallback, FFS, 4A, CAS, FSTS, Bonferroni, and the fixed sequence, are special cases of the CAAAS method. In addition, the feedback procedure arises as a special case of CAAAS as well by setting $\alpha_w^* = \alpha$ and selecting $\alpha_{21} = \alpha_{22} = \alpha$. Figure 2 presents a graphical display of the CAAAS procedure in comparison with other procedures.

For application of CAAAS, we recommend selecting the simplest adaptation functions for determining the significance level for testing H_2 (e.g., linear forms), unless there is strong justification for using complicated functions. In practice, we do not expect much difference between various forms of non-increasing functions (Figure 2). The Supporting information S.4 lists SAS/IML code for CAAAS for one-sided tests taken as piecewise non-increasing linear functions, $f_2(p_1)$ and $f_3(p_1)$, in the p_1-value within their respective consistency region. This implies that the strength of evidence required for establishing an efficacy claim for the second endpoint is linearly related to the decrease of the level of evidence from the first endpoint. These linear forms should be reasonable for most applications unless there is a strong justification for more complicated functions.

We now illustrate the 4A method and CAAAS procedures by revisiting the numerical example from Section 3.1 and restricting it to the first two hypotheses to make it applicable for the adaptation and consistency concepts. With this setting, let us assume that the pre-specified weights for $H^{(1)} = H_1$, $H^{(2)} = H_2$, are $w_1 = 2/3$ and $w_2 = 1/3$ and assume that the observed p-values for testing H_1 and H_2 are 0.02 and 0.005, respectively, and that the correlation between the test statistics is 0.6, as before. As for testing H_1, each of the two methods would not reject H_1 as its p-value $= 0.02 > 0.0167 (= 2\alpha/3)$ and consequently no recycling of α_1 for testing H_2 can be made. Next, although $p_2 = 0.005 < 0.0083 (= \alpha/3)$, we consider for illustration purposes calculation of the significance level for testing H_2 according to the specified adaptation function for each method. For the 4A method, we run the SAS code in the Supporting information S.3 to obtain values for *alphat* and w_1, which will be used to find alpha2 for any value of p_1 in the range (0.02, 0.50). With the inputs $r = 0.6, alphas = 0.50$ (as no consistency requirement), $alpha = 0.025$ and *alphad* (i.e., α_1) $= 0.0167$ and initial value $w = 0.11533$; the results are $alphat = 0.002$ and the corresponding convergence parameter value $y = -0.00013$. Now, we consider increasing the value to $w = 0.12138$, to obtain smaller value for y, which leads to $y = -3.656\text{E}{-8}$. Now, alpha2 $= \min(w_1.\ alphat/p^{**}2, \text{alpha1}) = \min(0.470, 0.0167) = 0.0167$ when $p_1 = 0.02$. As $p_2 = 0.005 < 0.0167$, we reject H_2.

For the application of CAAAS, first we note that because the values of p_1 and p_2 satisfy the relations $\alpha_1 \leqslant p_1 = 0.02 < \alpha$ and $p_2 = 0.005 < \alpha$. Then, according to this procedure, we reject both H_1 and H_2, and testing is completed. However, to show the computation of α_2, let us assume that $p_1 = 0.03 > \alpha$ and calculate α_2 for testing H_2. For this case, we run the SAS code in the Supporting information S.4 with $r = 0.6, \alpha_s^* = (alphas) = 0.02$ and $\alpha_w^* = (alphass) = 0.10$, and $gamma2 = 0.001$ (a pre-specified value for α_2 when $p_1 = \alpha_w^*$). This leads to $gamma1 = 0.039$ as the solution to Equation (3). This can be transferred to the last section of the program to solve for $u = \text{alpha} - ((\text{alpha} - g1)/(\text{alphas} - \text{alpha1})) * (p_1 - \text{alpha1})$, which gives $u = 0.081$; and $v = g1 - ((g1 - g2)/(\text{alphass} - \text{alphas})) * (p_1 - \text{alpha1})$, which gives $v = 0.033$. Hence, the three significance levels for testing H_2, depending on the location of p_1 (Figure 2), are respectively $alpha21 = 0.025, alpha22 = \min(0.081, alpha) = 0.025$ and $alpha23 = \min(0.033, alpha) = 0.0.025$. Now, because $\alpha_s^* \leqslant p_1 = 0.03 < \alpha_w^*$, we use the value *alpha23* for testing H_2, and therefore, the procedure rejects H_2. It is worth mentioning that for application of CAS, there is a SAS code given in [37].

4.3. Gatekeeping strategies

Gatekeeping procedures gained popularity in recent years as they provide a novel approach for handling multiplicity problems in clinical trials with hierarchical objectives. The null hypotheses are grouped into ordered families, $F_i (i = 1, 2, \ldots, m)$, where a family consists of all hypotheses corresponding to a hierarchical objective of the trial. For example, in oncology trials, we could have the sequence F_1 (overall survival) $\rightarrow$ F_2 (progression free survival) $\rightarrow$ F_3 (time to treatment failure). In general, the families $F_1, F_2, F_3, \ldots, F_m$ might represent primary, secondary, tertiary, and so on objectives. Testing is carried out sequentially for the m families of hypotheses, starting with the first family F_1, where the result of testing the hypotheses of a given family serves as a gatekeeper for testing the hypotheses in the next family. The FWER should be controlled strongly at level α across all families $F_1 \cup \ldots \cup F_m$. To simplify the ideas and notations, we restrict the discussion in the remainder of this section to $m = 2$ families, where F_1 denotes the family of k_1 primary null hypotheses and F_2 denotes the family of the remaining secondary k_2 null hypotheses such that $k_1 + k_2 = k$.

Maurer *et al.* [24], Bauer *et al.*[40], and Westfall and Krishen [25] considered the case of serial gatekeeping in which hypotheses in F_2 are tested if and only if all hypotheses in the previous family F_1 are rejected. Here, each family of hypotheses is tested at level α, and testing stops once at least one hypothesis in F_1 is not rejected. In contrast, Dmitrienko *et al.* [41] proposed a parallel gatekeeping procedure, where the hypotheses in F_2 are tested if at least one hypothesis in F_1 is rejected. The authors used the closure principle of Marcus *et al.* to test intersection hypotheses within each family and used weighted Bonferroni tests for testing the intersection hypotheses. Dmitrienko *et al.* [42] introduced tree-structured gatekeeping that unifies and generalizes serial and parallel gatekeeping procedures to address multiple hypotheses testing problems. This procedure uses weighted Bonferroni procedure for testing each intersection hypothesis in a closure method. Dmitrienko *et al.* [43] gave a set of sufficient conditions on the weights assigned to the hypotheses in each intersection hypothesis to ensure validity of the gatekeeping property.

Dmitrienko *et al.* [44] proposed a general multistage gatekeeping procedure (GMGP) to address arbitrary parallel gatekeeping problems. The procedure is a generalization of Guilbaud [45] who showed that a stepwise parallel gatekeeping procedure can be directly constructed without appeal to the closure principle. Guilbaud [45] proposed the application of a two-stage procedure, where the Bonferroni procedure is used for the first family and any multiple testing procedure that controls the FWER can be used for the second family. Dmitrienko *et al.* [44] noted that the key property of the Bonferroni procedure used for F_1 is that it does not exhaust the significance level α, unless all hypotheses are retained, and thus, one can carry over the remaining significance level to test F_2. The authors denoted an MTP with such a property as separable; see [44] for a formal definition of the separability condition. These authors advocated using any multiple testing procedure that does not exhaust α unless all hypotheses are retained in place of the Bonferroni procedure.

The GMGP uses a stepwise testing procedure where one first specifies the multiple testing procedure, termed component procedure, for each family. Then, the gatekeeping procedure is constructed by performing the component procedure sequentially starting with F_1. For the two-family setting, the method uses a two-stage procedure where in stage 1, the hypotheses in family F_1 are tested using a procedure $\mathcal{P}_1$ at level $\alpha_1 = \alpha$. If at least one hypothesis is rejected, then the hypotheses in F_2 are tested using a procedure $\mathcal{P}_2$ at the level α_2, which is determined by the number of hypotheses rejected by $\mathcal{P}_1$. The GMGP controls the FWER under the assumptions that (i) each procedure controls the FWER locally within the corresponding family and (ii) $\mathcal{P}_1$ is separable.

Single-step procedures such as Bonferroni and Dunnett are separable, but others such as Holm and Hochberg are not. Dmitrienko *et al.* [44] proposed a modification of these procedures to satisfy the separability condition by taking a convex combination of their critical constants with the Bonferroni critical constant. For example, for the truncated Holm procedure, the critical constant for comparing $p_{(i)}$ is $w_i\alpha = [\gamma/(k - i + 1) + (1 - \gamma)/k]\alpha$ for a pre-specified truncation parameter $\gamma (0 \leqslant \gamma < 1)$. For $\gamma = 0$ and $\gamma = 1$, the truncated Holm test reduces to the Bonferroni and Holm procedures, respectively.

Concerning the performance of the procedure, Dmitrienko *et al.* noted that if a truncated multiple test procedure is used for F_1, the power of the test is an increasing function in the truncation factor γ. However, the relationship between the power of the tests in F_2 and the truncation fraction used in F_1 is more complicated as it would depend on the number of hypotheses in each family, their weights, and effect sizes for false hypotheses. For other properties of the truncated testing procedures including the Holm approach, the reader is referred to the article by Dmitrienko *et al.* [44].

Other developments in gatekeeping procedures include the mixture parallel gatekeeping procedures introduced by Dmitrienko *et al.* [46], which can be more powerful than the GMGP. These procedures utilize the mixture method developed in [47] and are a direct application of the closure principle [11]. It should be noted that GMGP was derived using a method that was not explicitly based on the closure principle. In addition, Dmitrienko *et al.* [46] showed that the power of the GMGP can be improved also by using α-exhaustive tests [12] for the component procedures; however, this would be applicable only for cases in which one can relax the requirement that inferences on the primary endpoints are not influenced by those on the secondary endpoints, known as the independence condition.

Li and Mehrotra [48] considered a parallel gatekeeping procedure that serves as an extension of the 4A procedure. The method is based on a two-stage algorithm, where in stage 1, the k_1 hypotheses in F_1 (family of primary hypotheses) are tested using the Hochberg procedure at level $\alpha_1 < \alpha$. Then, the significance level for testing the $k_2(= k - k_1)$ hypotheses in F_2 (secondary hypotheses) is determined adaptively, on the basis of the maximum p-value, $p_{(k_1)}$, from the primary hypotheses using a similar formula to that of the 4A method. Following determination of the value of α_2, the hypotheses in F_2 can be tested using Hochberg's procedure at level α_2 if at least one primary hypothesis in F_1 is rejected. It should be noted that α_2 is determined by $p_{(k_1)}$ from testing k_1 hypotheses in family 1; this is in contrast to the GMGP method in which the significance level for testing the hypotheses in family 2 is determined by the number of hypotheses rejected in the family 1. In addition, the 4A-based method assumes a full parametric model and consequently requires knowledge of the joint distribution of the hypothesis test statistics in family 1 and family 2 under the global null hypothesis, in order to calculate α_2 along with other parameters. This can be a potential limitation for its application in practice; however, the authors provided tables for these parameters for selected values of α_1 and number of hypotheses (k_1) in the first family for one-sided and two-sided tests. Other gatekeeping procedures include the approaches by Neuhaeuser [49], Quan *et al.* [50], Xu *et al.* [51], and Kim *et al.* [52], among many others.

Dmitrienko *et al.* [53] compared the main features of three procedures proposed in the literature, which can be used for building parallel gatekeeping procedures in clinical trials with multiple objective grouped into two or more families, namely, GMGP, Bonferroni-based parallel gatekeeping procedure based on the methodology developed in [54, 55], and the 4A-based methods. The findings of their comparison show that the GMGP has the highest power for the primary tests and also leads to power advantage in the secondary family compared with the 4A-based method in problems with three or more secondary tests. The Bonferroni-based method is less powerful than the 4A-based method in the primary family, but this is offset by higher power in the secondary family for three or more secondary hypotheses.

4.4. Graphical approaches and recycling significance levels

Bretz *et al.* [54] and Burman *et al.* [55] independently introduced graphical approaches to construct and visualize complex multiple testing strategies. In addition, their methods serve as a tool for iteratively recycling unspent significant levels. Initially, the approach was introduced to construct and perform Bonferroni-based multiple testing procedures by splitting α among the different null hypotheses with $\alpha_i = w_i\alpha$ allocated for testing $H_i (i = 1, 2, \ldots, k)$. Whenever a null hypothesis H_i is rejected, its local level alpha α_i may be recycled to test the remaining, not yet rejected hypotheses. The Bonferroni-based graphical approaches have the advantage of leading to sequentially rejective test procedures and thus reduce the number $2^k - 1$ of hypotheses considered by the full closed test procedures to k [56].

In the approach by Bretz *et al.* [54], the elementary hypotheses are represented by a set of vertices i with associated local significance levels α_i. The weight $g_{ij} \geqslant 0$ associated with the directed edge between any two vertices i and j indicates the fraction of the local significance level α_i at the initial vertex i that is added to the significance level α_j at the terminal index j if the hypothesis H_i is rejected. The weights are chosen freely subject to $\sum_j g_{ij} \leqslant 1$. In addition, $g_{ii} = 0$, and by convention, edges with $g_{ij} = 0$ are not displayed graphically. Together with a suitable update algorithm (given later), this provides an iterative graph for a valid multiple testing procedure, that is, a graphical approach to sequentially rejective Bonferroni-based multiple test procedures. Because these graphical methods are closed testing procedures, they control the FWER strongly.

To complete the description, we give here the update algorithm (based on [5]) that ensures a sequentially rejective Bonferroni-based multiple testing procedure, where $I = 1, \ldots, k$ and $|\mathrm{A}|$ denotes the cardinality of a set A.

(1) Select an i such that $p_i \leqslant \alpha_i$. If no such i exists, stop; otherwise, reject H_i. If $|I| = 1$, stop.
(2) Update the graph:

(a) $I \to I \backslash i$

(b) $\alpha_j \to \begin{cases} \alpha_j + \alpha_i g_{ij}, & \text{if } j \in I, \\ 0, & \text{otherwise.} \end{cases}$

If $|I| = 1$, go to step 1.

(c) $g_{jl} \to \begin{cases} \frac{g_{jl} + g_{ji} g_{il}}{1 - g_{ji} g_{il}}, & \text{if } j, l \in I,\ j \neq l,\ g_{ji} g_{ij} < 1 \\ 0, & \text{otherwise.} \end{cases}$

(3) If $|I| > 1$, go to step 1; otherwise, stop.

This resulting class of sequentially rejective Bonferroni-based multiple test procedure is the same as the class of Bonferroni-based chain procedure introduced in [57]. The Supporting information S.5 lists SAS/IML code to calculate adjusted p-values, based on Algorithm 2 in [54].

Extensions of the original graphical approach using weighted and trimmed Simes tests [5] or weighted parametric tests [57, 58] are possible but generally do not lead to sequentially rejective multiple test procedures. Kordzakhia and Dmitrienko [59] developed superchain procedures that have graphical representations, where each node in a graph corresponds to a family of hypotheses and directed edges represent connections among the families, which can be regarded as conservative shortcuts of the mixture procedures developed by Dmitrienko *et al.* [46]. Another extension is the use of graphical approaches for multiple testing problems in group-sequential designs [60]. We finally note that the class of multiple test procedures described in [55] and [54] are equivalent, if convex combinations of the latter are considered [61].

In the following, we give graphical representations of various methods discussed previously in this paper. In Section 5, we illustrate how the graphical approaches can be tailored to address complex clinical trial objectives. The top graph of Figure 3 displays the fixed sequence test from Section 3.3 for $k = 3$ hypotheses with $w_1 = 1$ and $w_2 = w_3 = 0$. Thus, the first hypothesis H_1 is tested at its local significance level $\alpha_1 = w_1 \alpha = \alpha$. If rejected, its level is reallocated to the second hypothesis H_2, and so on. The bottom graph of Figure 3 displays the fallback procedure from Section 4.1 for $k = 3$ hypotheses and weights $w_1, w_2, w_3 \geqslant 0$ subject to $w_1 + w_2 + w_3 = 1$. For illustration, assume three unadjusted p-values p_1, p_2, and p_3 for the three elementary hypotheses H_1, H_2, and H_3. Each of the hypotheses H_i is tested at its local significance level $\alpha_i = w_i \alpha$. If the hypothesis H_i can be rejected, that is, $p_i \leqslant \alpha_i$, its level α_i is reallocated to the next hypothesis H_{i+1} according to the pre-specified sequence. The testing is repeated for the remaining, non-rejected hypotheses with the updated local significance levels, thus possibly leading to further rejected null hypotheses. This procedure is repeated until no further hypothesis can be rejected.

Figure 4 displays the rejection sequence and iterated graphs resulting from the fallback procedure as applied to the numerical example from Section 3.1. Alternatively, using the SAS/IML code from the Supporting information S.5, one obtains the adjusted p-values 0.06, 0.015, 0.015 for the three hypotheses H_1, H_2, H_3, respectively, and we obtain the same test decision as in Section 4.1 at the significance level $\alpha = 0.025$. Graphical visualization of the exhaustive fallback procedures described in Section 4.1 are given in [54].

Figure 5 displays a generalized version of the Holm procedure [9]. For equal significance levels $\alpha_1 = \alpha_2 = \alpha_3 = 1/3$ and equal transition weights $g_{12} = g_{21} = g_{31} = 1/2$, one obtains the regular unweighted Holm procedure described in Section 3.2. Assuming the p-values $p_1 = 0.02$, $p_2 = 0.005$, $p_3 = 0.01$ from that mentioned earlier, one checks that the procedure in Figure 5 rejects sequentially

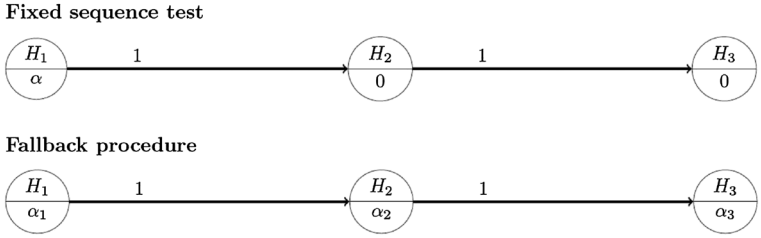

Figure 3. Graphical visualization of the fixed sequence test (top) and the fallback procedure (bottom) for three hypotheses.

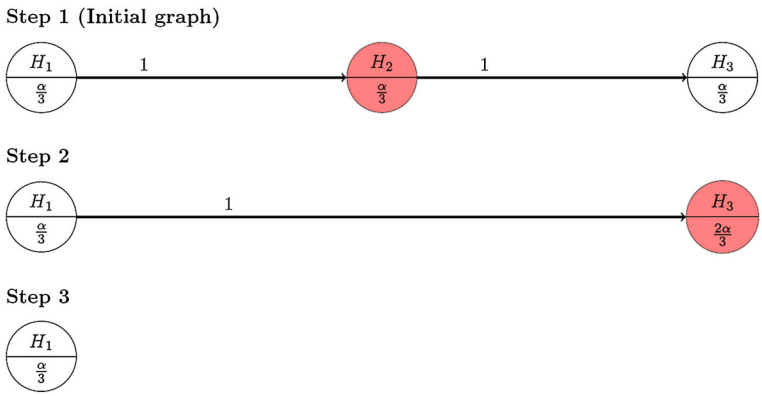

Figure 4. Graphical visualization of the rejection sequence and iterated graphs for the fallback procedure with weights $w_1 = w_2 = w_3 = 1/3$, p-values $p_1 = 0.02$, $p_2 = 0.005$, $p_3 = 0.01$, and significance level $\alpha = 0.025$. Rejected hypotheses are shaded.

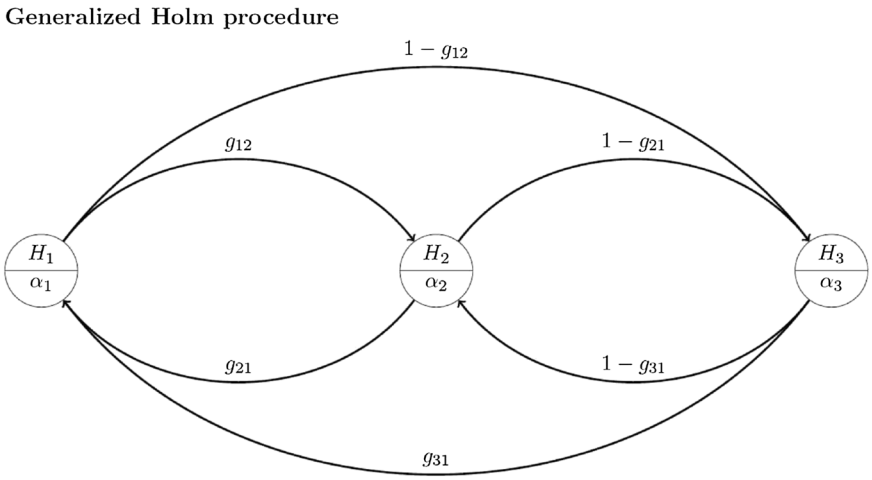

Figure 5. Graphical visualization of the rejection sequence and iterated graphs for the generalized Holm procedure.

H_2, H_3, and finally H_1 at level $\alpha = 0.025$. The same result is obtained when applying the Holm procedure described in Section 3.2. A weighted version described already by Holm [9] is obtained for general $\alpha_i = w_i\alpha$ and setting $g_{12} = w_2/(w_2 + w_3)$, $g_{21} = w_1/(w_1 + w_3)$, and $g_{31} = w_1/(w_1 + w_2)$, which ensures that the local significance levels remain proportional after each iteration. The graph in Figure 5, however, is more general as it allows the weights g_{12}, g_{21}, g_{31} to take any values subject to the regularity conditions mentioned earlier.

Similarly, truncated versions of the Holm procedure, such as described in Section 4.3, can be displayed graphically as well. Given multiple families of hypotheses in a pre-specified hierarchical order, truncated tests avoid re-shuffling the complete significance level within a family until all its hypotheses are rejected in order to proceed testing the next family in the hierarchy. Instead, once at least one hypothesis is rejected in a given family, a fraction of the significance level is reserved to test subsequent families of hypotheses. Consider as an example two primary hypotheses H_1, H_2 and two secondary hypotheses H_3, H_4. The secondary hypotheses are only tested if at least one of the primary hypotheses is rejected. Figure 6 displays graphically the resulting multiple testing procedure, if H_1, H_2 are tested using the truncated Holm procedure from Section 4.3 with truncation parameter $g_{12} = g_{21} = \gamma \in [0, 1]$, and H_3, H_4 are tested with the regular Holm procedure; see also [58] for a numerical example. For more than two hypotheses, the original graphical approach leads to a different version of truncated Holm procedure than proposed by Dmitrienko *et al.* [44]. Instead, convex combination of several graphs as described in [61] has to be used to display the truncated Holm procedure.

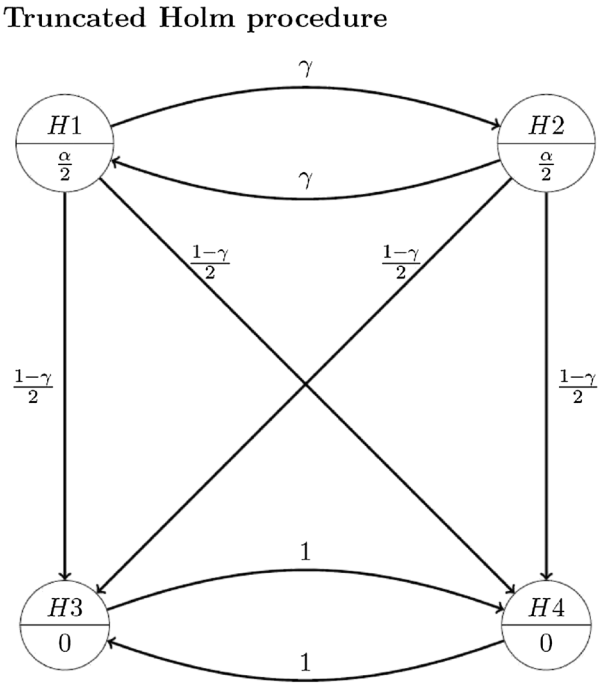

Figure 6. Graphical visualization of the truncated Holm procedure.

5. Case studies from Section 2 re-visited

In this section, we re-visit the clinical trial examples from Section 2 to illustrate some of the methods described in the previous sections.

5.1. Two doses and two hierarchical endpoints

One possibility to address the clinical trial objectives in Section 2.1 is to use the graphical approach from Section 4.4. As both doses are considered equally important, we split the significance level α equally across both doses but initially assign weight 0 to the secondary hypotheses to reflect the hierarchy between the two endpoints within a given dose. Therefore, $\alpha_1 = \alpha_2 = \alpha/2$ and $\alpha_3 = \alpha_4 = 0$.

There are in total 12 possible edges to connect any two nodes in order to specify how the significance levels are propagated after a hypothesis is declared significant. However, this number of edges can be reduced substantially by accounting for the given clinical considerations. In our example, the edges $H_3 \to H_1$ and $H_4 \to H_2$ receive weight 0 because of the hierarchy among the two endpoints within a given dose. In addition, if successiveness is required, there are no edges $H_1 \to H_4$, $H_2 \to H_3$, $H_3 \to H_4$, and $H_4 \to H_3$, as otherwise one can always construct examples where, for a given dose, the secondary hypothesis is rejected but the associated primary is not. This leaves us with the six edges displayed in the left graph of Figure 7. As the sum of the weights over all outgoing edges for a given node has to be less than or equal to 1, this gives $g_{41} = g_{32} = 1$, and we are left with two remaining parameters $g_{12} = \gamma_1$, $g_{21} = \gamma_2$ to be determined. Their choice can be based on different considerations. If, for example, safety is a major concern, one might prefer testing the primary endpoint for the low dose, in case high dose is significant but not safe (leading to a large value of γ_2). Otherwise, if safety is not of major concern, one might prefer giving the secondary endpoint more weight instead of passing additional significance level to the other primary hypothesis, thus leading to smaller values of γ_1 and γ_2. The right graph in Figure 7 displays the resulting graph for the extreme case $\gamma_1 = \gamma_2 = 0$ and can be interpreted as a Holm-type procedure applied to the families of hypotheses per dose, H_1, H_3 and H_2, H_4 : if both the primary and secondary endpoints for a given dose are rejected at $\alpha/2$, then the other dose can be tested a full α. Finally, if no preference for the choice of γ_1 and γ_2 is at hand, on the basis of the available clinical considerations, numerical optimization can be used to determine their values in order to maximize the power of the multiple testing procedure (e.g., [62]).

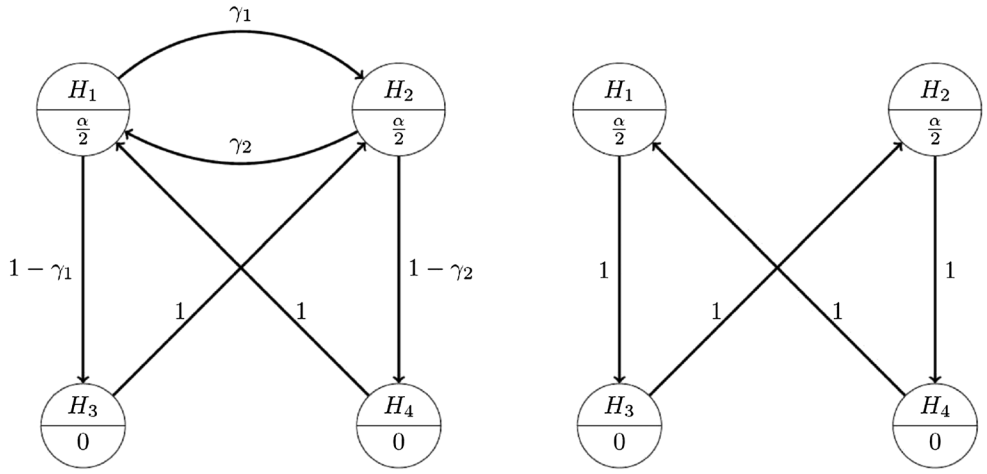

Figure 7. Left: Graphical visualization of a possible multiple testing procedure for the case study from Section 2.1. Right: resulting graph for the case $\gamma_1 = \gamma_2 = 0$.

5.2. Composite endpoint with possible claim for 'hard' endpoints

As the composite endpoint in this example is of higher priority than the individual component endpoints, multiple testing strategies that do not consider this structural priority, such as the Bonferroni and Holm procedures discussed in Sections 3.1 and 3.2, would not be appropriate as they treat all hypotheses at the same level of importance. On the other hand, the fixed sequence approach (Section 3.3) and the serial gatekeeping for families of hypotheses (Section 4.3) assume a full hierarchy of the composite endpoint and its individual hard component, death. These methodologies, however, do not allow testing the hypothesis H_2 unless H_1, the hypothesis for the composite endpoint, is rejected. As such, these approaches do not allow testing the hypothesis H_2 if the results of testing H_1 were reasonably strong but missed the threshold of the significance level ($\alpha = 0.05$, say), as it is the goal in this example.

In contrast, the fallback method discussed in Section 4.1, by allocating a fraction of α to H_2, would allow testing H_2 in such situations, possibly at higher significance level than initially allocated if H_1 is rejected. Similarly, the 4A method from Section 4.2 would allow testing H_2 at a significance level adapted to the findings of testing H_1. However, each of these approaches allows one to test H_2 even if the result of testing H_1 does not show efficacy for the composite endpoint. This may lead to difficulties in interpreting the overall study findings if the result for the hard endpoint, death, is significant but the result for the composite endpoint does not show efficacy or even trends in the opposite direction.

Taking into account that the composite endpoint is of higher priority than the individual component endpoints and that testing the individual endpoints requires the results of testing H_1 either significant or at least meet a certain threshold, at a minimum, suggest that consistency-based procedures, such as CAS, restricted feedback, or CAAAS (as discussed in Section 4.2), are reasonable procedures for addressing the multiplicity issue in this example. In order to apply any of these procedures, one would (i) pre-specify the minimum level of evidence in testing the composite endpoint H_1 as a prerequisite for testing H_2; (ii) allocate $\alpha_1 < \alpha$ for testing H_1 and leave the remaining fraction of α, that is, $\alpha - \alpha_1$, for testing H_2; and (iii) pre-specify the form of a non-increasing function for adapting the significance level for testing H_2 as function of the p-value obtained from testing H_1 along with an estimate of the correlation between the test statistics for the composite and individual component endpoints.

As for selecting the value for the consistency criterion, this should be driven by the relative importance of the two endpoints along with clinical considerations including toxicities and adverse events. In particular, for treatment with serious toxicities and adverse events, one should be inclined to require a stronger consistency to allow testing for a hard endpoint if the treatment did not make on the composite endpoint. Concerning splitting of α between the composite endpoint and its hard component endpoint, this can be driven by the treatment effects and relative population sizes. Also, estimates for the correlations between the two endpoints might be driven by the relative contribution of the component endpoint in the composite endpoint. In general, the results of early clinical trials on composite endpoint along with discussion with the clinical team should enable setting reasonable estimates for these parameters. As for the selection of the adaptation functions, we consider, following the discussion of CAAAS, linear forms in the p_1-value.

Let us assume that, on the basis of the aforementioned discussion concerning selection of the parameters for application of CAAAS, the weights for the $H^{(1)} = H_1$ and $H^{(2)} = H_2$ are $w_1 = 4/5$ and $w_2 = 1/5$ and that the correlation between the test statistics is 0.4, $\alpha_s^* = \alpha$, $\alpha_w^* = 0.10$ and that a linear non-increasing function for adaptation would be adequate for adapting the significance level for testing H_2, α_2, as a function of the observed p-value from testing H_1. Also, it is desired that the significance level for testing H_2 should be at least 0.001 in the consistency region (i.e., $\gamma_2 = 0.001$ at the point $\alpha_w^* = 0.10$; Figure 2). Further, we will assume that the observed p-values for testing H_1 and H_2 are 0.09 and 0.007, respectively. With this information, although we cannot reject H_1, however, as $p_1 < 0.10(= \alpha_w^*)$, we can proceed to test H_2. The significance level for testing H_2 can be calculated by running the SAS code in the Supporting information S.4, with the following inputs $\alpha = 0.025$, $\alpha_1 = (4/5)\ \alpha = 0.02$, $\alpha_s^* = \alpha$, $\alpha_w^* = 0.10$, and $\gamma_2 = 0.001$. This leads to $\gamma_1 = 0.036078$, which following the discussion of this method in Section 4.2 gives $u = 0.180092$ and $v = 0.0033385$; consequently $alpha21 = 0.025, alpha22 = \min(0.180092, alpha) = 0.025$ and $alpha23 = \min(0.0033385, alpha) = 0.0033385$. Now, because $\alpha_s^* \leqslant p_1 = 0.09 < \alpha_w^*$, we use the value *alpha23* for testing H_2, and therefore, the procedure cannot reject H_2.

5.3. *Non-inferiority and superiority for multiple doses and multiple endpoints*

The complexity in this example stems from the fact that we have to test 10 null hypotheses in a clinically meaningful way while controlling the FWER in the strong sense at level α. The test strategy described later is an extension of the approach proposed by Hung and Wang [8] for combined non-inferiority and superiority testing in three-armed trials with two doses and an active control.

By convention, let H_i with odd integers $i = 1, 3, 5, 7$, and 9 denote the five null hypotheses for the low dose. Let H_1 and H_3 denote the two co-primary hypotheses (superiority over placebo for PASI 75 and IGA 0 or 1, respectively). Let further H_5 denote the superiority hypothesis for PASI 90 over placebo and, finally, H_7 and H_9 the non-inferiority and superiority hypothesis for PASI 75 against the active control, respectively. The hypotheses H_i with even integers $i = 2, 4, 6, 8$, and 10 denote the corresponding five null hypotheses for the high dose.

On the basis of the clinical considerations in this example, we can conclude the following points in addressing the multiplicity issues. (i) As the two doses are considered equally important, there is no hierarchical ordering between them, similar to the example in Section 2.1. Consequently, splitting the significance level α evenly across the two doses, so that each is tested at the level $\alpha/2$, seems appropriate. (ii) As the win criterion for each dose requires rejecting the null hypothesis for each of the two co-primary endpoints (superiority over placebo for PASI 75 and IGA 0 or 1, respectively), each of these hypotheses is tested at level $\alpha/2$. (iii) Because testing superiority for the key secondary endpoint (PASI 90) and non-inferiority against the active comparator for PASI 75 are considered equally important, we test each hypothesis at the same significance level, once the two co-primary hypotheses have been rejected for a given dose. (iv) Finally, if the non-inferiority hypothesis for PASI 75 is rejected for one

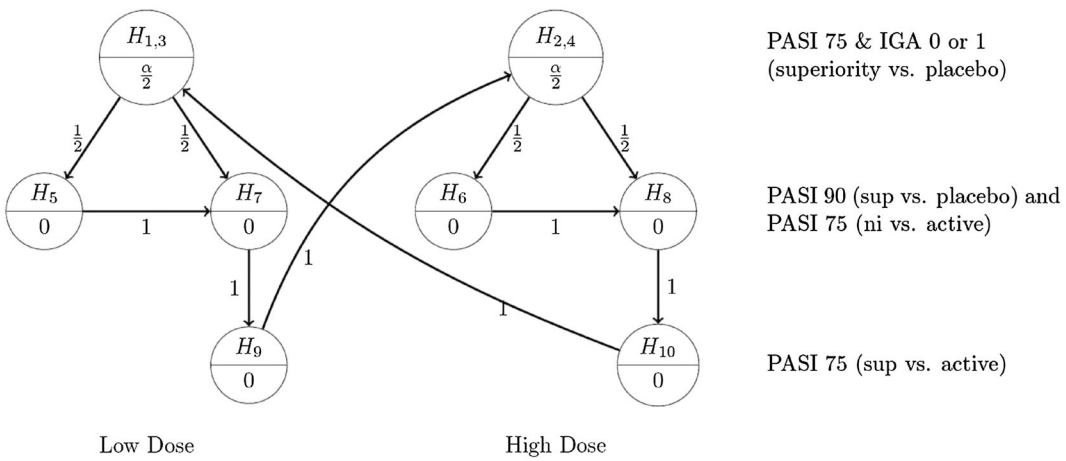

Figure 8. Graphical visualization of a possible multiple testing procedure for the case study from Section 2.3. sup, superiority; ni, non-inferiority.

of the two doses, we continue testing the related superiority hypothesis against the active comparator. This leads to a gatekeeping approach, based on the methods discussed in Section 4.3, which seems to provide a reasonable strategy for considering the structural hierarchy of the hypotheses. Furthermore, the graphical approach from Section 4.4 provides a visual presentation of the testing strategy taking into account the hierarchical structure among the hypothesis test and allowing repeated cycling of the significance levels once a hypothesis is rejected (Figure 8), where for brevity, the co-primary hypotheses are denoted as $H_{i,j} = H_i \cup H_j$. It should be noted that alternative test strategies are possible and a graphical visualization might support the discussion within the clinical team. For example, one might consider adding edges from H_7 to H_5 and from H_8 to H_6, respectively, in order to increase the power of rejecting the PASI 90 hypothesis before proceeding down the sequence. Note that no significance level should be propagated directly from H_5 to H_9 and from H_6 to H_{10}, respectively, as otherwise one would test the PASI 75 superiority hypotheses (against the active control) without necessarily having rejected the corresponding non-inferiority hypotheses.

6. Concluding remarks

This tutorial discusses recent methodologies for complex multiplicity issues arising in clinical trials. As no single procedure is expected to outperform other procedures, the focus of this tutorial is on contrasting the merits and limitations of these methodologies so that the reader, equipped with the information available about the logical and stochastic dependencies between the endpoints as well as the magnitude of treatment effects from previous clinical trials, can select an appropriate procedure for addressing the multiplicity issues at hand.

The underlying theme of most of these procedures is to spend the overall significance level α more efficiently, compared with traditional procedures, such as Bonferroni, Holm, and fixed sequence. Generally speaking, a more efficient spending of the significance level is expected to increase the study power, thus leading to a larger probability of the trial to be positive. The new methodologies extend the traditional procedures by building upon their merits and avoiding their disadvantages. This can be seen clearly from the fallback method, which—in comparison with the Bonferroni and fixed sequence procedures—recycles the local significance levels more efficiently and at the same time allows for the possibility that all hypotheses are tested at some local significance level. Other procedures, such as the 4A method and its extensions, attempt to take the concept of recycling further by adapting the significance level for testing the next hypothesis in the sequence to the evidence level from the previous test, which might not be significant. The CAAAS procedure considers recycling and adapting the significance level along with setting pre-specified consistency criterion so that the overall results of the study remain interpretable.

The popularity of the gatekeeping strategies emerges from the fact that they account for the relationship among the hypotheses and group them into hierarchical families of hypotheses depending on their relative importance. Gatekeeping strategies recycle the significance level between families of hypotheses, in contrast to previous approaches that recycle it between individual hypotheses. For multiple test procedures that exhaust α and are not separable, modified versions are available by mixing them with the Bonferroni method (say), which makes them amenable to recycling un-used significance levels.

The graphical approach for multiple testing is a general framework that encompasses several of the multiple testing procedures considered in this tutorial. The approach visualizes the underlying multiple testing strategy in a transparent way, which can thus be easily communicated to clinical teams. In addition, graphical methods allow repeated recycling to accumulate further un-used significance levels for subsequent testing.

The examples given in this tutorial illustrate the application of some of these methodologies reviewed earlier. However, even if one has decided to apply a certain methodology on the basis of the available information, the selection of the remaining design parameters remains critical. This includes allocating the significance level among the hypotheses and defining the transition weights to decide on how the local significance levels are propagated once a hypothesis is rejected as well as selecting a consistency criterion, if any. Selection of these parameters and their impact on study power remain a topic under research for several of the methodologies described in this tutorial. The computer programs provided with this paper are intended to support the reader in the process of fine tuning a given multiple test procedure.

Some of the methods described in this paper allow the construction of compatible simultaneous confidence intervals, which have a joint coverage probability of at least $1 - \alpha$ for the parameters of interest.

Here, compatibility between a multiple testing procedure and a set of simultaneous confidence intervals means that if a null hypothesis is rejected with the test procedure, then the associated multiplicity-corrected confidence interval excludes all parameter values for which the null hypothesis is true [63]. Such simultaneous confidence intervals are easily obtained for the Bonferroni method. For stepwise testing procedures, however, the currently available compatible simultaneous confidence intervals have limited practical use because the confidence bounds for those parameters whose hypotheses are rejected reflect only the stepwise test decision (unless all hypotheses are rejected); see the results from Hsu and Berger [64] for fixed sequence procedure, Strassburger and Bretz [65] and Guilbaud [45] for the Holm procedure and other Bonferroni-based closed test procedures introduced by Hommel *et al.* [56], and Guilbaud [66] and Guilbaud and Karlsson [67] for further research in this area.

Acknowledgements

We are very thankful to two anonymous referees, Dr. Alex Dmitrienko, Associate Editor, and Dr. Ralph D'Agostino, Editor, for their constructive comments that led to significant improvements in the paper. Also, we would like to thank our colleague, Dr. Kathleen Fritsch, from the FDA, for the helpful comments on the manuscript.

References

1. Hochberg Y, Tamhane AC. *Multiple Comparison Procedures*. Wiley: New York, 1987.
2. Dmitrienko A, Tamhane AC, Bretz F. *Multiple Testing Problems in Pharmaceutical Statistics*. Chapman & Hall/CRC Press: New York, 2009.
3. Bretz F, Hothorn T, Wesfall P. *Multiple Comparison using R*. Chapman & Hall/CRC: Boca Raton, FL, 2010.
4. Dmitrienko A, D'Agostino R. Tutorial in biostatistics: methods for addressing multiplicity problems in clinical trials. *Statistics in Medicine* 2013. In press.
5. Maurer W, Glimm E, Bretz F. Multiple and repeated testing of primary, co primary and secondary hypotheses. *Statistics in Biopharmaceutical Research* 2011; **3**:336–352.
6. O'Neill RT. Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance. *Controlled Clinical Trials* 1997; **18**:550–556.
7. Hung J, Wang SJ. Some controversial multiple testing problems in regulatory applications. *Journal of Biopharmaceutical Statistics* 2009; **19**:1–11.
8. Hung J, Wang SJ. Challenges to multiple testing in clinical trials. *Biometrical Journal* 2010; **52**:747–756.
9. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979; **6**:65–70.
10. Moyé LA. *Multiple Analysis in Clinical Trials: Fundamental for Investigators*. Springer-Verlag: New York, Inc., 2003.
11. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
12. Grechanovsky E, Hochberg Y. Closed procedures are better and often admit a shortcut. *Journal of Statistical Planning and Inference* 1999; **76**:79–91.
13. Bauer P. Multiple testing in clinical trials. *Statistics in Medicine* 1991; **10**:871–890.
14. Gabriel KR. Simultaneous test procedures some theory of multiple comparisons. *The Annals of Mathematical Statistics* 1969; **40**:224–250.
15. Sonnemann E, Finner H. Vollständigkeitssätze für multiple test probleme. In *Multiple Hypothesenprüfung*, Bauer P, Hommel G, Sonnemann E (eds). Springer: Berlin, 1988; 121–135.
16. Simes JR. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **73**:75–754.
17. Sarkar SK. Some probability inequalities for censored MTP2 random variables: a proof of the Simes conjecture. *The Annals of Statistics* 1998; **26**:494–504.
18. Benjamini Y, Hochberg Y. Multiple hypothesis testing with weights. *Scandinavian Journal of Statistics* 1997; **24**:407–418.
19. Hochberg Y. A sharper Bonferroni procedure for multiple significance testing. *Biometrika* 1988; **75**:800–802.
20. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 1988; **75**:383–386.
21. Zhang J, Quan H, Ng J, Stepanavage ME. Some statistical methods for multiple endpoints in clinical trials. *Controlled Clinical Trials* 1997; **18**:204–221.
22. Tamhane AC, Liu L. On weighted Hochberg procedures. *Biometrika* 2008; **95**:279–294.
23. Hommel G, Bretz F, Maurer W. Multiple hypotheses testing based on ordered p-values – a historical survey with applications to medical research. *Journal of Biopharmaceutical Statistics* 2011; **21**:595–609.
24. 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.). Stuttgart: Fischer Verlag: New York, 1995; 3–18.
25. Westfall PH, Krishen A. Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures. *Journal of Statistical Planning and Inference* 2001; **99**:25–40.
26. Wiens BL. A fixed sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics* 2003; **2**:211–215.

27. Wiens BL, Dmitrienko A. The fallback procedure for evaluating a single family of hypotheses. *Journal of Biopharmaceutical Statistics* 2005; **15**:929–42.
28. Hommel G, Bretz F. Aesthetics and power considerations in multiple testing – a contradiction? *Biometrical Journal* 2008; **50**:657–666.
29. Huque MF, Alosh M. A flexible fixed-sequence testing method for hierarchically ordered correlated multiple endpoints in clinical trials. *Journal of Statistical Planning and Inference* 2008; **138**:321–335.
30. Wiens BL, Dmitrienko A. On selecting a multiple comparison procedure for analysis of a clinical trial: fallback, fixed sequence, and related procedures. *Statistics in Biopharmaceutical Research* 2010; **2**:22–32.
31. Li J, Mehrotra D. An efficient method for accommodating potentially underpowered primary endpoints. *Statistics in Medicine* 2008; **27**:5377–5391.
32. Song Y, Chi GY. A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine* 2007; **26**:3535–3549.
33. Alosh M, Huque MF. A flexible strategy for testing subgroups and overall population. *Proceedings of the Biopharmaceutical Section of Joint Statistical Meetings*, Salt Lake City, UT, 2007; 198–203.
34. Alosh M, Huque MF. A flexible strategy for testing subgroups and overall population. *Statistics in Medicine* 2009; **28**:3–23.
35. Alosh M, Huque MF. Multiplicity considerations for subgroup analysis subject to consistency constraints. *Biometrical Journal* 2013; **55**:444–462.
36. Huque MF, Alosh M. A consistency-adjusted strategy for accommodating an underpowered primary endpoint. *Journal of Biopharmaceutical Statistics* 2012; **22**:160–179.
37. Huque MF, Alosh M, Bhore R. Addressing multiplicity issues of a composite endpoint and its components in clinical trials. *Journal of Biopharmaceutical Statistics* 2011; **21**:610–634.
38. Zhao YD, Dmitrienko A, Tamura R. Design and analysis considerations in clinical trials with a sensitive subpopulation. *Statistics in Biopharmaceutical Research* 2010; **2**:72–83.
39. Alosh M, Huque MF. A consistency-adjusted alpha-adaptive strategy for sequential testing. *Statistics in Medicine* 2010; **29**:1559–1571.
40. Bauer P, Roehmel J, Maurer M, Hothorn L. Testing strategies in multi-dose experiments including active control. *Statistics in Medicine* 1998; **17**:2133–2146.
41. 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.
42. Dmitrienko A, Wiens B, Tamhane AC, Wang X. Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Statistics in Medicine* 2007; **26**:2465–2478.
43. Dmitrienko A, Tamhane AC, Liu L, Wiens B. A note on tree gatekeeping procedures in clinical trials. *Statistics in Medicine* 2007; **06**:1–6.
44. Dmitrienko A, Tamhane AC, Wiens B. General multistage gatekeeping procedures. *Biometrical Journal* 2008; **50**:667–677.
45. Guilbaud O. Bonferroni parallel gatekeeping-transparent generalization, adjusted p-values and short direct proofs. *Biometrical Journal* 2007; **49**:917–927.
46. Dmitrienko A, Kordzakhia G, Tamhane AC. Multistage and mixture parallel gatekeeping procedures in clinical trials. *Journal of Biopharmaceutical Statistics* 2011; **21**:726–747.
47. Dmitrienko A, Tamhane AC. Mixture of multiple testing procedures for gatekeeping applications in clinical trials. *Statistics in Medicine* 2011; **30**:1473–1488.
48. Li J, Mehrotra D. Gatekeeping testing via adaptive alpha allocation. *Biometrical Journal* 2008; **5**:704–715.
49. Neuhaeuser M. How to deal with multiple endpoints in clinical trials. *Fundamental and Clinical Pharmacology* 2006; **20**:515–523.
50. Quan H, Capizzi T, Zhang J. Multiplicity adjustment for clinical trials with two doses of an active treatment and multiple primary and secondary endpoints. *Statistics in Biopharmaceutical Research* 2009; **1**:258–267.
51. Xu H, Nuamah I, Liu J. A Dunnett-Bonferroni-based parallel gatekeeping procedure for dose-response clinical trials with multiple endpoints. *Pharmaceutical Statistics* 2009; **8**:301–316.
52. Kim HJ, Entsuah R, Shults J. The union closure method for testing a fixed sequence of families of hypotheses. *Biometrika* 2011; **98**:391–401.
53. Dmitrienko A, Soulakova J, Millen B. Three methods for constructing parallel gatekeeping procedures in clinical trials. *Journal Pharmaceutical Statistic* 2011; **21**:769–786.
54. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* 2009; **28**:586–604.
55. Burman CF, Sonesson C, Guilbaud O. A recycling framework for the construction of Bonferroni-based multiple tests. *Statistics in Medicine* 2009; **28**:739–761.
56. 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.
57. Millen B, Dmitrienko A. Chain procedure: a class of flexible closed testing procedures with clinical trial application. *Statistics in Biopharmaceutical Research* 2011; **3**:14–30.
58. Bretz F, Posch M, Glimm E, Klinglmueller F, Maurer W, Rohmeyer K. Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. *Biometrical Journal* 2011; **5353**:894–913.
59. Kordzakhia G, Dmitrienko A. Superchain procedures in clinical trials with multiple objectives. *Statistics in Medicine* 2013; **32**:486–508.
60. Maurer W, Bretz F. Multiple testing in group-sequential designs using graphical approaches. *Statistics in Biopharmaceutical Research*. DOI: 10.1080/19466315.2013.807748.

61. Maurer W, Bretz F. Sequentially rejective graphical multiple test procedures with memory 2012. *Statistics in Medicine* 2013; **32**:1739–1753.
62. Bretz F, Maurer W, Hommel G. Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures. *Statistics in Medicine* 2011; **30**:1489–1501.
63. Hayter AJ, Hsu JC. On the relationship between stepwise decision procedures and confidence sets. *Journal of the American Statistical Association* 1994; **89**:128–136.
64. Hsu JC, Berger RL. Stepwise confidence intervals without multiplicity adjustment for dose-response and toxicity studies. *Journal of the American Statistical Association* 1999; **94**:468–482.
65. 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.
66. Guilbaud O. Alternative confidence regions for Bonferroni-based closed-testing procedures that are not alpha-exhaustive. *Biometrical Journal* 2009; **51**:721–735.
67. Guilbaud O, Karlsson P. Confidence regions for Bonferroni-based closed tests extended to more general closed tests. *Journal of Biopharmaceutical Statistics* 2011; **21**:682–707.