Featured Article

Statistics in Medicine

Received 4 September 2012, Accepted 5 September 2012 Published online 9 October 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5642

Key multiplicity issues in clinical drug development

Alex Dmitrienko,[a*†] Ralph B. D'Agostino, Sr.[b] and Mohammad F. Huque[c]

Much progress has been made over the past decade with the development of novel methods for addressing increasingly more complex multiplicity problems arising in confirmatory Phase III clinical trials. This includes traditional problems with a single *source* of multiplicity, for example, analysis of multiple endpoints or dose–placebo contrasts. In addition, more advanced problems with several *sources* of multiplicity have attracted attention in clinical drug development. These problems include two or more families of objectives such as multiple endpoints evaluated at multiple dose levels or in multiple patient populations. This paper provides a review of concepts that play a central role in defining and solving multiplicity problems (error rate definitions) and introduces main classes of multiple testing procedures widely used in clinical trials (nonparametric, semiparametric, and parametric procedures). The paper also presents recent advances in multiplicity research, including gatekeeping procedures for clinical trials with multiple sets of objectives. The concepts and methods introduced in the paper are illustrated using several case studies on the basis of real clinical trials. Software implementation of commonly used multiple testing and gatekeeping procedures is discussed. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: clinical trials; multiple comparisons; type I error rate control; mltiple testing procedures; gatekeeping procedures

1. Introduction

The last decade has witnessed a surge of statistical methods for addressing multiplicity issues in clinical trials, including traditional multiplicity problems related to analysis of multiple endpoints (primary and/or secondary), multiple treatment arms, multiple subgroups, and composite endpoints and their components, as well as more advanced problems. The increasing interest in this research area reflects important trends in clinical drug development. Much emphasis is placed on increasing efficiency of clinical drug development through sophisticated designs and comprehensive characterization of the efficacy and safety profiles of new treatments. Modern clinical trials use complex sets of objectives, including multiple sets of endpoints and combinations of multiple endpoints evaluated at several dose levels or in several patient populations, which gives rise to increasingly more complex multiplicity problems.

Appropriate treatment of multiplicity issues is a key component of the regulatory definition of an adequate well-controlled clinical trial. Importance of addressing multiplicity problems in confirmatory Phase III clinical trials has been recognized in regulatory guidance documents. This includes the International Conference on Harmonization guidelines as well as the guidance document released by the European Medicines Agency [1]. The U.S. Food and Drug Administration is about to release a comprehensive guidance on this issue.

The main regulatory requirement in confirmatory clinical trials with multiple objectives is formulated in terms of controlling the rate of false-positive inferences (known formally as the familywise error rate

[a]*Quintiles, Inc., Durham, NC, U.S.A.*
[b]*Boston University, Boston, MA, U.S.A.*
[c]*Office of Biostatistics, OTS, CDER, U.S. Food and Drug Administration, Silver Spring, MD, U.S.A.*
**Correspondence to: Alex Dmitrienko, Quintiles, Inc., 4820 Emperor Blvd, Durham, NC 27703, U.S.A.*
[†]*E-mail: alex.dmitrienko@quintiles.com*

(FWER), see Section 3), so that conclusions can be made on individual hypotheses of interest without inflating the error rate. The error rate is controlled at a desired level by utilizing a multiplicity adjustment. A number of multiplicity adjustment methods are typically available given the multiplicity problem. To define the most appropriate multiplicity adjustment method for a particular setting, it is important to gather and utilize all available clinical and statistical information. This includes information on clinically meaningful relationships among the individual objectives, for example, whether the objectives are naturally ordered, nested, or can be examined independently of each other. In addition, it is important to evaluate available information on the joint distribution of the test statistics associated with the objectives. Given this information, available multiplicity adjustment options can be evaluated to arrive at the most powerful solution tailored to a specific setting.

This paper describes general methodology and practical considerations to facilitate the process of developing most relevant adjustments for multiplicity problems arising in confirmatory clinical trials. The intent of this paper is to provide an overview of principles and methods for biostatisticians involved in the design and analysis of clinical trials with multiple objectives and also discuss relevant regulatory considerations. The reader will find a more detailed overview of specific methodological topics mentioned in the paper in multiple comparison textbooks and book chapters. For example, multiple testing procedures used in the traditional settings are described in Hochberg and Tamhane [2] and Dmitrienko *et al*. [3]. Advanced multiplicity problems arising in clinical trials with inherent hierarchical structures are discussed in Dmitrienko and Tamhane [4]. Review papers on multiplicity problems arising in clinical trials include D'Agostino *et al*. [5], Sankoh, Huque and Dubey [6], and Sankoh, D'Agostino and Huque [7].

The outline of the paper is as follows. We will begin with a general discussion of multiplicity problems arising in confirmatory clinical trials. Sections 2 and 3 cover introductory topics, including clinical decision rules (win criteria) used in clinical trials with multiple objectives, error rate definitions, and classification of multiplicity adjustment methods (multiple testing procedures). The advanced reader can begin with Sections 4–6. These sections introduce main classes of procedures used in traditional multiplicity problems. Further, a more complex setting with multiplicity problems defined by a combination of several factors, for example, analysis of the efficacy profile of several doses in multiple patient populations, is considered in Section 7. Section 8 defines common methods for presenting multiplicity adjustments, including adjusted p-values and adjusted confidence intervals (simultaneous confidence intervals). Section 9 provides information on software implementation of popular multiple testing procedures. Finally, the Glossary defines important terms appearing in the text (these terms are shown in boldface type), and Appendix A includes an example that illustrates key properties of error rate definitions.

2. Multiplicity problems in confirmatory clinical trials

As was pointed out in the introduction, multiplicity commonly arises in confirmatory clinical trials. This includes, for example, the following clinical trial objectives:

- Investigate treatment effects of several endpoints.
- Evaluate treatments at several dose levels.
- Compare treatment to control for noninferiority and superiority on multiple endpoints.
- Perform subgroup analysis.
- Carry out analysis by baseline and demographic factors.
- Assess regional differences.

In general, multiplicity can be defined as simultaneous evaluation of different aspects of the efficacy profile (and in some cases the safety profile) of a treatment. In this paper, we provide a review of common multiplicity problems, including more straightforward problems with a single source of multiplicity and more advanced problems with several sources of multiplicity.

Various solutions to multiplicity problems have been developed and will be discussed in this paper (see Sections 4–7). However, it is important to realize that multiplicity does not always require a formal multiplicity adjustment, and methods for addressing multiplicity vary from one class of problems to another. It is critical to fully understand the objectives of a particular analysis before a multiplicity adjustment is developed. Clinical trial researchers need to be aware of settings in which multiplicity does not lead to error rate inflation, can be addressed without a statistical adjustment, or is introduced for purposes other than making an efficacy or safety claim. The first and second settings will be discussed

Statistics in Medicine

later in this section. The last setting refers to multiplicity problems arising in the context of exploratory analyses. Examples include sensitivity assessments or alternative evaluations of the treatment effect routinely performed as part of the primary analysis in confirmatory clinical trials. The primary analysis may be performed in several patient populations, including the intention-to-treat population as well as other populations defined by excluding patients who failed the inclusion criteria, did not take the medication as prescribed, and so on. Further, the treatment effect may be evaluated using a prespecified statistical method as well as several other methods as part of a general robustness assessment. Multiple tests are not subject to multiplicity adjustments in this setting because of their exploratory nature.

2.1. Notation

The following notation will be used throughout this paper. Consider a general multiplicity problem in a confirmatory clinical trial with several objectives, for example, several endpoints or dose–control comparisons. Let m denote the number of objectives, and let θ_i denote the true value of the treatment difference for the ith objective, $i = 1, \ldots, m$. For example, in a trial with multiple continuous endpoints, θ_i represents the true mean difference for the ith endpoint. Further, the θ_is may be the differences in proportions in the binary case or log-hazard ratios in the context of time-to-event analysis. Assume that positive values of θ_i define a beneficial treatment effect.

The null hypotheses related to the objectives are denoted by

$$H_1, \ldots, H_m.$$

Each null hypothesis is defined as the set of treatment differences corresponding to lack of expected treatment effect. For example, if the ith objective is a superiority objective, the null hypothesis H_i is defined as the set of negative treatment differences, that is,

$$H_i : \theta_i < 0, \ i = 1, \ldots, m.$$

Similarly, in a noninferiority setting, the null hypotheses are defined as

$$H_i : \theta_i < -\lambda, \ i = 1, \ldots, m,$$

where λ is a positive noninferiority margin.

The univariate p-values for testing the null hypotheses $H_1, \ldots, H_m$ are denoted by $p_1, \ldots, p_m$. Further, $p_{(1)} < \cdots < p_{(m)}$ denote the ordered p-values, and the associated ordered null hypotheses are denoted by $H_{(1)}, \ldots, H_{(m)}$. Finally, a two-sided setting is assumed throughout the paper, and the overall two-sided error rate is set to $\alpha = 0.05$.

2.2. Win criteria

It was emphasized earlier in this section that multiplicity does not always translate into a multiplicity adjustment, and different approaches to multiplicity adjustment are used in different types of multiplicity problems. Thus, it is helpful to begin with a general classification of multiplicity problems in clinical trial applications. The classification scheme considered here is built around the concept of clinical decision rules in confirmatory Phase III clinical trials. These rules are commonly referred to as **win criteria**. In a clinical trial with multiple objectives, it is critical to understand how conclusions about the efficacy profile (and potentially safety profile) of the treatment will be made or, in other words, how to specify the win criterion. The win criterion defines the outcomes that need to be observed to conclude that the treatment provides a clinically meaningful treatment benefit compared with a control. Choice of statistical methods for addressing multiplicity is driven by the win criterion.

Three types of win criteria are commonly utilized in clinical trials:

- At-least-one win criteria.
- All-or-none win criteria.
- Global win criteria.

We will use following three case studies to help illustrate the different types of win criteria:

- Case study 1. The efficacy profile of a single dose of a treatment is evaluated in a clinical trial by using two endpoints.

- Case study 2. Three doses of a new treatment are tested in a clinical trial versus a common control, for example, placebo, on a single primary endpoint.
- Case study 3. The primary analysis in a clinical trial with two arms (single dose of a new treatment versus placebo) and a single endpoint is performed in the overall patient population as well as in a prospectively defined subpopulation.

At-least-one win criteria are used in clinical trials where multiple objectives are treated as independent entities and each individual objective leads to a successful outcome for the trial. Consider, for example, Case study 1, and suppose that the trial's objective is to study the effects of a novel treatment on the risk of new vertebral fractures and invasive breast cancer in postmenopausal women with osteoporosis. The efficacy of this treatment is evaluated using incidence of vertebral fractures and incidence of breast cancer. Each endpoint provides independent evidence of a positive effect, and the trial's objective is to detect a beneficial effect on at least one endpoint, which serves as an example of an at-least-one win criterion. Similarly, at-least-one win criteria are used in Case studies 2 and 3. In particular, it is sufficient to establish a beneficial effect of the treatment at one or more dose levels in Case study 2. In Case study 3, the trial's sponsor is interested in developing a tailored therapy, and the win criterion is defined in terms of demonstrating superior efficacy in the general population or subpopulation (as a side note, decision rules used in the development of tailored therapies are discussed in Wang, O'Neill and Hung [8] and Millen *et al*. [9]).

Multiplicity problems arising in clinical trials with at-least-one win criteria are addressed by using multiple testing procedures on the basis of using the **union–intersection method** [3]. As a quick example, if the null hypotheses of interest are not ordered, the **Bonferroni procedure** can be applied. This procedure relies on the following simple rules:

$$\text{Reject the null hypothesis } H_i \text{ if } p_i \leqslant \alpha/m,\ i = 1, \ldots, m.$$

For instance, in a clinical trial with three dose–placebo comparisons (Case study 2), a null hypothesis is rejected if its p-value is no greater than $\alpha/3 = 0.0167$. Another commonly used procedure is the **fixed-sequence procedure**. If the hypothesis ordering is prospectively defined, for example, $H_1, \ldots, H_m$, the fixed-sequence procedure is applied by sequentially testing the null hypotheses in the prespecified sequence:

$$\begin{aligned}&\text{Reject } H_1 \text{ if } p_1 \leqslant \alpha.\\ &\text{Reject } H_2 \text{ if } p_2 \leqslant \alpha \text{ and } H_1 \text{ is rejected.}\\ &\text{Reject } H_3 \text{ if } p_3 \leqslant \alpha \text{ and } H_1 \text{ and } H_2 \text{ are both rejected, and so on.}\end{aligned}$$

To illustrate, if a positive dose–response relationship is expected in Case study 2, the three tests can be carried out sequentially beginning with the highest dose. If the p-value for the comparison between the dose and placebo does not exceed $\alpha = 0.05$, the associated null hypothesis of no effect is rejected, and the effect of the next dose in the sequence is tested. This test is also carried out at the full α and so on. It is important to note that the Bonferroni and fixed-sequence procedures are very basic multiple testing procedures, and more powerful/flexible methods will be discussed in Section 3.2.

An *all-or-none win criterion* can also be applied to the two-endpoint problem in Case study 1. This criterion states that all objectives must be simultaneously met for the trial to achieve its primary objective. For example, if the trial is conducted to evaluate the effects of a treatment on cognition and global changes in patients with mild to moderate Alzheimer's disease, the two efficacy endpoints can be defined on the basis of the Alzheimer's Disease Assessment Scale-Cognitive subscale and Clinician's Interview-based Impression of Change Plus. To obtain a regulatory claim, the trial's sponsor must demonstrate a beneficial effect on both endpoints rather than only one, which means that the win criterion used in this setting is an all-or-none criterion.

It is important to note that multiplicity problems with all-or-none win criteria are different from the usual multiplicity problems associated with type I error rate inflation. There is no error rate inflation in this setting, and thus, it is not appropriate to apply the Bonferroni or other multiple testing procedures. The solution to problems with all-or-none win criteria can be found by the **intersection-union method** [3] and is based on carrying out each individual test at the full α level, that is, the trial's objective is met if $p_i \leqslant \alpha$ simultaneously for all $i = 1, \ldots, m$. For example, in the Alzheimer's disease trial, an efficacy claim can be made only if the p-values for both endpoints are significant at $\alpha = 0.05$. For a discussion

of clinical and regulatory issues arising from this type of multiplicity problems, see Offen *et al*. [10] and Chuang-Stein, Dmitrienko and Offen [11]. Multiplicity problems based on all-or-none win criteria will not be discussed further in this paper.

It is worth mentioning that hybrid win criteria based on a combination of at-least-one and all-or-none criteria may be utilized in clinical trials with more complex sets of objectives. For example, in an epilepsy trial with the primary endpoints based on seizure rate, drop attack rate and seizure severity, the win criterion may be defined in terms of a beneficial effect on seizure rate only or a beneficial effect on both drop attack rate and seizure severity. Hybrid win criteria may also be developed in trials with composite endpoints, see Huque, Alosh and Bhore [12].

Global win criteria may be used in clinical trials with biologically related endpoints. For example, improvement in patients with osteoporosis may be studied by evaluating the treatment effect on multiple endpoints such as timed up-and-go test, 6-min walking distance test, and pain score. Because the endpoints are closely related to each other, they may be viewed as components of a single clinical variable and may be analyzed using an appropriate global testing procedure, for example, **O'Brien global procedures**. For more information on statistical methods used in this setting, see Sankoh *et al*. [13] and Tamhane and Dmitrienko [14]. Global win criteria are used in early stage development but are fairly uncommon in confirmatory trials and will not be discussed further in this paper.

3. Multiple inferences in problems with at-least-one win criteria

Starting with this section, we will focus on multiplicity problems in clinical trials with at-least-one win criteria. These criteria define the traditional settings in which multiplicity induces error rate inflation. This section presents key considerations such as the definition of the error rate used in confirmatory clinical trials and a classification of methods for addressing error rate inflation.

3.1. Type I error rate control

One of the most important properties of a traditional multiplicity setting is potential for an inflated probability of incorrect conclusions. When inferences in problems with multiple outcomes are performed without any adjustments and the most favorable outcome is selected, the false-positive rate will no longer be controlled at the prespecified level. A multiplicity adjustment is required to preserve the error rate. However, to define the correct adjustment, it is critical to first select the correct definition of the error rate.

A variety of error rate definitions, including the **FWER**, **generalized FWER**, or **false discovery rate** [3], has been proposed in the literature and found applications in clinical drug development. For example, multiple testing procedures that control the false discovery rate have been used in the analysis of adverse event data [15] and experiments in drug discovery.

In confirmatory clinical trials, the error rate is defined in terms of the probability of incorrectly finding a beneficial treatment effect for at least one objective among those that have no beneficial effects. In other words, in a multiplicity problem with m null hypotheses, the error rate is the probability of incorrectly rejecting *at least one* true null hypothesis. This probability depends on the true values of the treatment differences $\theta_1, \ldots, \theta_m$, and it is important to distinguish between the following definitions of error rate control.

First of all, the error rate may be controlled under a single set of the θs, for example, under the assumption that all null hypotheses are *simultaneously* true. This is known as **weak FWER control**. Second, the error rate may be controlled under all configurations of true and false null hypotheses, or equivalently, the maximum error rate may need to be protected. This is known as **strong FWER control**. The maximum error rate used in the definition of strong FWER control can differ markedly from the error rate computed under the weak assumption. A counter example given in Appendix A shows that weak FWER control does not provide adequate protection of the probability of an incorrect conclusion. For this reason, strong FWER control is mandated in all confirmatory clinical trials. Error rate control for all configurations of true and false null hypotheses enables the trial's sponsor to make specific claims for each objective or set of objectives. This is due to the fact that the probability of a false claim for any set of null hypotheses is guaranteed not to exceed the nominal level, for example, two-sided $\alpha = 0.05$, regardless of the magnitude of the treatment effect for the other null hypotheses. All multiple testing procedures introduced in this paper provide strong control of the FWER.

3.2. *Multiple testing procedures*

As was pointed earlier in this section, in traditional multiplicity settings based on at-least-one decision rules, unadjusted inferences cause FWER inflation. Multiplicity adjustments in the form of multiple testing procedures are used to address this problem. Commonly used multiple testing procedures for multiplicity problems will be defined in Sections 4–7. The first three sections consider more straightforward multiplicity problems. There is no underlying hierarchical structure in these problems, and thus, the multiple objectives considered in a clinical trial represent the same source of multiplicity, or in other words, the null hypotheses of interest can be placed in a single family. Further, procedures introduced in Section 7 are designed to handle more complex multiplicity problems with several sources of multiplicity. In this setting, multiple families of null hypotheses are defined to account for clinically relevant relationships among the individual objectives.

Given the number of multiple testing procedures used in clinical trials, it should not be surprising that one can consider several classification schemes for these procedures. In this section, we will focus on a classification scheme based on logical and distributional relationships among the null hypotheses of interest. This scheme is generally the most relevant one because the most critical consideration in the process of selecting an appropriate procedure for a multiplicity problem is full utilization of all information on clinically relevant connections among the null hypotheses (logical relationships) as well as the joint distribution of the hypothesis test statistics (distributional relationships).

We will begin with logical relationships. In a general problem of testing m null hypotheses, two types of logical relationships can exist among the null hypotheses:

- Prespecified hypothesis ordering. The null hypotheses are ordered and tested according to their clinical importance or other criteria.
- Data-driven hypothesis ordering. No predetermined hypothesis ordering exists, and the null hypotheses are tested in the order determined by the significance of the hypothesis test statistics.

To illustrate these concepts, consider the multiplicity problem with three null hypotheses based on three dose–control comparisons in Case study 2. In the case of prespecified hypothesis ordering, the testing sequence must be fixed before the data are unblinded. For example, the null hypothesis corresponding to the highest dose may be placed at the beginning of the sequence followed by the middle dose and then the lowest dose. After that, an appropriate procedure, for example, the fixed-sequence procedure can be applied, see Figure 1. In general, to prespecify the hypothesis ordering, the trial's sponsor needs to have strong historical data. The null hypotheses may be ordered on the basis of strong evidence of a positive dose–response relationship in previously conducted clinical trials. If little or no historical information is available, the sponsor can select a testing strategy on the basis of a data-driven hypothesis ordering. If a positive dose–response relationship is unlikely to be observed in the trial, a sensible strategy is to test the null hypotheses beginning with the null hypothesis corresponding to the most significant test statistic. Multiple testing procedures used in a data-driven setting are discussed in Sections 4 – 6.

The next important consideration is related to distributional relationships among the null hypotheses in a multiplicity problem. By utilizing all available information on the joint distribution of the hypothesis test statistics, a clinical trial's sponsor can improve power of a multiple testing procedure. Thus, it is in the sponsor's best interest to gather all information on distributional relationships among the null hypotheses and select procedures that fully utilize this information. Consider, for example, the multiplicity problems arising in Case studies 1 and 3. In the first case study, multiplicity is induced by the analysis of two endpoints. If the responses in this clinical trial are normally distributed, the test statistics for the two endpoints follow a bivariate normal distribution. Because the correlation between the endpoints is not generally known (in fact, it may not be clear whether or not the endpoints are positively correlated), the bivariate normal distribution cannot be fully specified. Thus, the trial's sponsor

Figure 1. Visual representation of the decision rules used in the fixed-sequence procedure in Case study 2 (H_1, high dose versus placebo; H_2, medium dose versus placebo; H_3, low dose versus placebo). Decisions: —, null hypothesis is rejected.

Statistics in Medicine

cannot utilize the joint distribution of the two test statistics to address the multiplicity problem in this trial. On the other hand, more information on the correlation between the test statistics is available in Case study 3. Under normality assumptions, the correlation between the test statistics in the overall population and subpopulation is $\sqrt{f}$, where f is the prevalence of classifier-positive patients. For example, if half of the patients in the population are classifier positive ($f = 0.5$), the correlation is $\sqrt{0.5} = 0.707$. If the prevalence is known in a trial, the joint distribution of the two test statistics is fully specified, and this distributional information can be used to set up a powerful multiple testing procedure in this multiplicity problem.

With the amount of distributional information utilized by a multiple testing procedure, three classes of procedures can be defined:

- **Nonparametric procedures** (Section 4) do not make any assumptions about the joint distribution of the hypothesis test statistics.
- **Semiparametric procedures** (Section 5) are based on decision rules that do not explicitly depend on the joint distribution of the hypothesis test statistics, but additional distributional assumptions are needed to establish FWER control.
- **Parametric procedures** (Section 6) make explicit assumptions about the joint distribution of the hypothesis test statistics.

Both nonparametric and semiparametric procedures can be thought of as model-free procedures that rely only on univariate p-values. For this reason, these procedures are also known as p-value-based procedures. By contrast, a distributional model must be specified to define the decision rules of a parametric procedure. Other classes of multiple testing procedures, including resampling-based procedures [16], are beyond the scope of this paper. Resampling-based procedures do not make distributional assumptions and instead approximate the joint distribution of the test statistics via bootstrap or permutation methods. Resampling-based procedures have not found many applications in confirmatory clinical trials mostly because they only provide asymptotic control of the error rate when the sample size approaches infinity.

Nonparametric, semiparametric, and fully parametric procedures can also be used in more advanced multiplicity problems with several families of null hypotheses. These procedures, commonly known as gatekeeping procedures, are discussed in Section 7.

4. Nonparametric procedures in single-family problems

This section and the next two sections introduce commonly used multiple testing procedures in traditional multiplicity problems with a single family of null hypotheses. More information on multiple testing procedures in single-family problems can be found in [3].

As was stated in Section 3.2, nonparametric multiple testing procedures assume no data models and thus utilize no distributional assumptions. To control the FWER, nonparametric procedures rely on the Bonferroni inequality. Nonparametric procedures tend to be conservative in the sense the actual FWER is generally less than the nominal value. This effect is pronounced in problems with a large number of null hypotheses and strongly positively correlated hypothesis test statistics.

In this section, we focus on a class of nonparametric procedures known as Bonferroni-based **chain procedures** [17] that were derived as an extension of flexible **fallback procedures** [18, 19]. The class of chain procedures includes the Bonferroni procedure introduced in Section 2 and all commonly used Bonferroni-based procedures, that is, the **Holm procedure** [20]. Multiple testing procedures in this class are based on the **closure principle** [21] and preserve the FWER in the strong sense for any distribution of the test statistics.

To motivate chain procedures, consider first the basic fixed-sequence procedure mentioned in Section 2. Focusing on Case study 2, let H_i, $i = 1, 2, 3$, denote the null hypotheses of no effect for the dose–control comparisons in this trial. The fixed-sequence procedure assumes that the null hypotheses are prospectively ordered, for example, H_1, H_2, and H_3. Each null hypothesis is tested at the full α as shown in Figure 1. Testing begins with the first null hypothesis in the sequence (H_1, high dose versus placebo). The next null hypothesis (H_2, medium dose versus placebo) is examined only if there is evidence of a significant effect at the high dose, that is, H_2 is tested only if H_1 is rejected. Similarly, H_3 (low dose versus placebo) is tested only if H_2 is rejected.

The fixed-sequence procedure utilizes a simple set of rigid rules and performs best when the monotonicity condition is met, that is, when the tests are ordered from the highest marginal power (largest effect size) to lowest marginal power (smallest effect size). Power loss is generally expected if the

monotonicity condition is not met. For example, the fixed-sequence procedure will likely fail to detect a significant effect at the medium and low doses in Case study 2 if the magnitude of the treatment effect at the high dose is lower than expected, which may be due to toxicity problems. This will create a situation that may seem paradoxical to nonstatisticians. Specifically, if there is no evidence of a beneficial effect at the high dose, no efficacy claims can be made for the other two doses even though the test statistics are highly significant.

When there is no guarantee that the monotonicity condition is satisfied, it is generally risky to apply the fixed-sequence procedure, and more flexible procedures are recommended, for example, chain procedures. To illustrate key features of chain procedures, consider again Case study 2. If a nonmonotonic dose–response function is expected in this trial, the treatment effect at the high dose may be smaller than that at the medium dose. To accommodate this uncertainty, the following two-step testing strategy can be utilized:

- Step 1. Instead of putting all the eggs in one basket and testing H_1 first, test H_1 and H_2 simultaneously.
- Step 2. If at least one null hypothesis is rejected in Step 1, proceed to testing the remaining null hypothesis H_3.

This testing strategy can be implemented via a flexible chain procedure. This chain procedure serves as an extension of the basic Bonferroni procedure. A key feature of the Bonferroni procedure is that it distributes the overall error rate α among the individual null hypotheses. For example, in a problem with three hypotheses, each hypothesis is tested at $\alpha/3 = 0.0167$. Likewise, the chain procedure distributes the overall error rate among the hypotheses, but an important difference is that this distribution is governed by a predefined **α allocation rule**. It was stated earlier that only two null hypotheses (H_1 and H_2) are tested in Step 1, which means that the initial α allocation is restricted to those hypotheses, that is, H_1 and H_2 will each be tested at $\alpha/2 = 0.025$. The significance level for H_3 will be set to 0 to indicate that the low dose cannot be tested unless at least one other dose is shown to provide a significant improvement over placebo.

Second, the significance levels can be transferred from a hypothesis to another. This is known as **α propagation**. In the two-step chain procedure, the significance level initially assigned to H_1, that is, $\alpha/2$, will be carried over to H_3 if H_1 is rejected in Step 1. Similarly, the significance level assigned to H_2 will be carried over to H_3 provided H_2 is rejected.

Finally, a very important feature of chain procedures is an option to retest nonrejected null hypotheses at a higher significance level, which leads to a power gain compared with more basic multiple testing procedures such as the Bonferroni procedure that do not use retesting. It is worth noting that the retesting option is a direct consequence of the closure principle and does not compromise strong error rate control. To enable retesting, another step is added to the testing algorithm, and if H_3 is rejected, the algorithm can return to H_1 and H_2 in Step 3. If both null hypotheses are rejected in Step 1, there is no need to retest them. However, if only one of these null hypotheses is rejected, the available error rate can be transferred from H_3 to the nonrejected null hypothesis. This null hypothesis is retested at a higher significance level in Step 3, which improves the probability of achieving a significant result.

The general principles described earlier can be implemented using a serial, cyclical, or simultaneous testing approach. Briefly, a serial algorithm is based on a prespecified testing sequence [22], a cyclical algorithm relies on a data-driven hypothesis ordering, and the last algorithm relies on simultaneous testing of all hypotheses [23]. In general, the serial testing approach facilitates the communication of decision rules to drug development teams because it uses a clinically meaningful ordering of the null hypotheses. Figure 2 illustrates the serial testing approach for the Bonferroni-based chain procedure. The circles in this figure denote the null hypotheses, and the value displayed above a circle defines the level used in the corresponding significance test. For example, the significance level used for testing the null hypothesis H_1 in Step 1 is 0.025, and thus, H_1 is rejected if $p_1 \leqslant 0.025$. Finally, the solid and dashed arrows are used to indicate rejection and failure to reject a particular null hypothesis, respectively.

As shown in Figure 2, the Bonferroni-based chain procedure relies on a three-step testing algorithm:

- Step 1. Bonferroni-based α-splitting strategy is applied to the null hypotheses H_1 and H_2, that is, each test is carried out at $0.05/2 = 0.025$.
- Step 2. If at least one test is significant in Step 1, the error rates from H_1 and H_2 are carried over to H_3 according to the α propagation rule. If both null hypotheses are rejected in Step 1, H_3 is tested at $0.025 + 0.025 = 0.05$, and if only one null hypothesis is rejected in Step 1, H_3 is tested at 0.025.

Statistics in Medicine

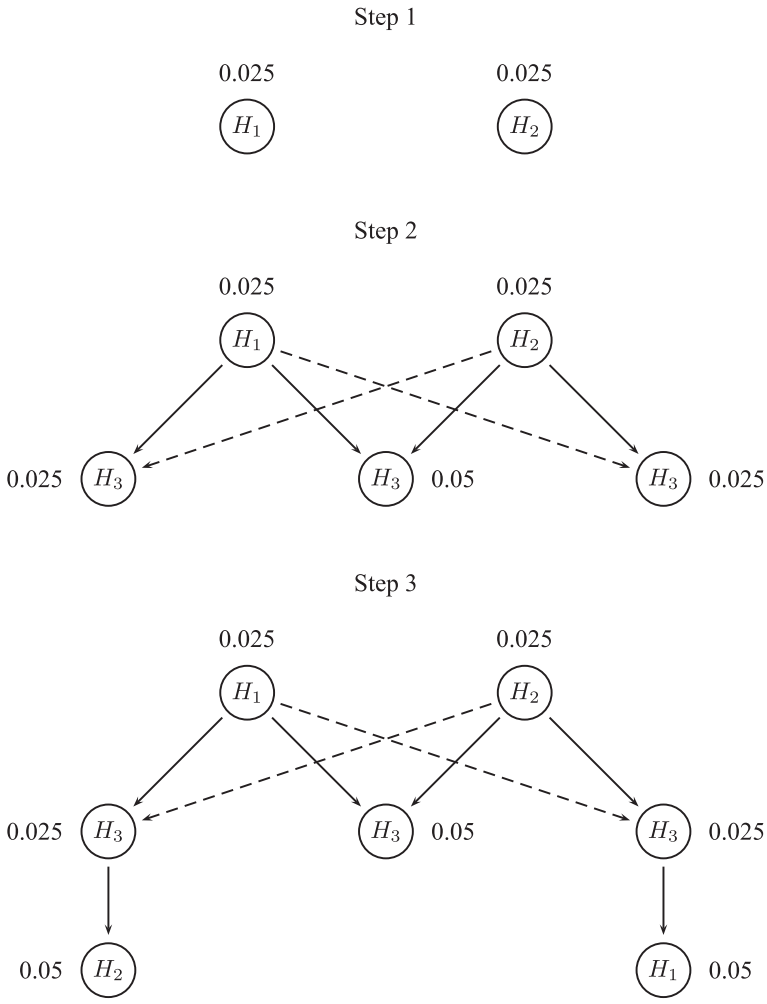

Figure 2. Implementation of the nonparametric Bonferroni-based chain procedure in Case study 2 with two-sided $\alpha = 0.05$ (H_1, high dose versus placebo; H_2, medium dose versus placebo; H_3, low dose versus placebo). Decisions: —, null hypothesis is rejected; - - -, null hypothesis is not rejected.

- Step 3. If the test for H_3 produces a significant result and either H_1 or H_2 is not rejected in Step 1, the error rate released after the rejection of H_3 is applied to the remaining nonrejected null hypothesis. For example, if H_1 is rejected in Step 1 but H_2 is not, the algorithm returns to H_2, and this null hypothesis is retested at the full 0.05 level. Because a higher significance level is used, H_2 can be rejected in this step even though it was not rejected earlier. A similar strategy is applied if H_2 is rejected but H_1 is not rejected in Step 1.

To define Bonferroni-based chain procedures in general problems with m null hypotheses, α allocation and propagation rules are defined using matrix notation. In general, an α allocation rule defines initial weights of the null hypotheses denoted by $w_1, \ldots, w_m$. In the problem discussed earlier, positive weights are assigned to H_1 and H_2 ($w_1 = 1/2$ and $w_2 = 1/2$), and a zero weight is assigned to H_3 ($w_3 = 0$). An α propagation rule defines a set of weighted connections among the null hypotheses. The weight of the (directed) connection from H_i to H_j is denoted by g_{ij}, and the connection weights are chosen such that

$$g_{ii} = 0,\ g_{i1} + g_{i2} + g_{i3} = 1,\ i = 1, 2, 3.$$

Thus, with the null hypothesis H_1, the weights of the connections from H_1 to H_2 (g_{12}) and from H_1 to H_3 (g_{13}) specify the fractions of the error rate available after rejecting H_1 that are transferred to H_2 and H_3, respectively. Further, the rejection of H_1 only opens a possibility to test H_3, and therefore, the weights of the connections from H_1 to H_2 and from H_1 to H_3 are given by $g_{12} = 0$ and $g_{13} = 1$.

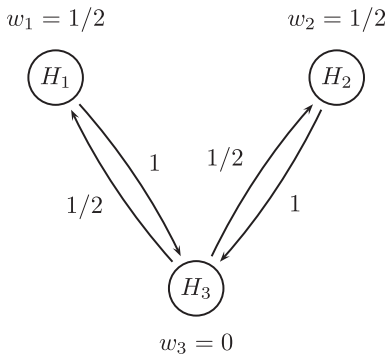

Figure 3. Visual representation of the α allocation and propagation rules used in the nonparametric Bonferroni-based chain procedure in Case study 2 (H_1, high dose versus placebo; H_2, medium dose versus placebo; H_3, low dose versus placebo).

Similarly, the weights of the out-going connections for H_2 are given by $g_{21} = 0$ and $g_{23} = 1$. A visual representation of the α allocation and propagation rules used in the Bonferroni-based chain procedure in Case study 2 is given in Figure 3.

The chain procedures defined in this section are nonparametric procedures that cannot utilize available distributional information. The flexible chain approach can be extended to multiplicity problems with a fully specified joint distribution of the hypothesis test statistics, see Section 6. The resulting parametric chain procedures are uniformly more powerful than nonparametric Bonferroni-based chain procedures.

5. Semiparametric procedures in single-family problems

The key feature of nonparametric Bonferroni-based procedures defined in Section 4 is that they make no assumptions about the joint distribution of the hypothesis test statistics in a multiplicity problem. These multiple testing procedures can be improved to achieve higher power if additional distributional information is available. This includes the semiparametric and parametric settings. In this section, we focus on the semiparametric setting and consider p-value-based procedures under the assumption that the joint distribution of the hypothesis test statistics is only partially specified. For example, it is common to assume that the test statistics follow a multivariate normal distribution, but the correlation matrix is unknown.

Semiparametric multiple testing procedures include procedures derived from the **Simes procedure** [24] and **Šidák procedure** [25]. The Simes procedure is used for testing the global null hypothesis on the basis of the intersection of the individual null hypotheses. This procedure can determine that at least one null hypothesis is false, but this false null hypothesis cannot be identified. The Simes procedure rejects the global null hypothesis if at least one of the following m conditions is met:

$$p_{(1)} \leqslant \alpha/m,\ p_{(2)} \leqslant 2\alpha/m,\ \ldots,\ p_{(m)} \leqslant \alpha.$$

Further, the Šidák procedure is a regular multiple testing procedure, and thus, it can be used for testing the individual null hypotheses. This procedure rejects the null hypothesis H_i if $p_i \leqslant 1-(1-\alpha)^{1/m}$, $i = 1,\ldots,m$. It is easy to verify that the critical values of the Simes and Šidák procedures are greater than those of the Bonferroni procedure and thus the two semiparametric procedures are uniformly more powerful than the nonparametric Bonferroni procedure. In particular, the Bonferroni-adjusted significance level for each of the three dose–placebo comparisons in Case study 2 is $\alpha/3 = 0.0167$, and the corresponding Šidák-adjusted level is $1-(1-\alpha)^{1/3} = 0.0170$.

The Simes and Šidák procedures are not by themselves of much interest in clinical trial applications because their power can be easily improved. It is accomplished by constructing stepwise Simes-based and Šidák-based procedures that utilize a data-driven hypothesis ordering. Stepwise procedures are applied to the ordered hypotheses $H_{(1)},\ldots,H_{(m)}$ corresponding to the ordered p-values $p_{(1)} < \cdots < p_{(m)}$ and rely on step-down or step-up testing algorithms:

- **Step-down procedures** test $H_{(1)},\ldots,H_{(m)}$ sequentially beginning with the null hypothesis corresponding to the most significant p-value.

Statistics in Medicine

- **Step-up procedures** test $H_{(1)}, \ldots, H_{(m)}$ sequentially beginning with the null hypothesis corresponding to the least significant p-value.

The step-down Šidák procedure [26] illustrates the step-down approach. This procedure is based on the following testing algorithm:

- Steps $i = 1, \ldots, m-1$. The null hypothesis $H_{(i)}$ is rejected if $p_{(i)} \leqslant 1-(1-\alpha)^{1/(m-i+1)}$. If $H_{(i)}$ is rejected, proceed to the next step. Otherwise, accept all remaining null hypotheses $H_{(j)}$, $j = i, \ldots, m$.
- Step m. The null hypothesis $H_{(m)}$ is rejected if $p_{(m)} \leqslant \alpha$. Otherwise, it is accepted.

It is worth noting that the step-down Šidák procedure uses the same significance level as the regular Šidák procedure in Step 1, that is, $1-(1-\alpha)^{1/m}$, and successively higher significance levels later in the sequence. This important feature of the step-down procedure is a direct consequence of α propagation discussed in Section 4. The significance levels increase in a monotone fashion as the step-down procedure progresses through the sequence of null hypotheses because a fraction of the error rate is transferred from each rejected null hypothesis to the remaining nontested null hypotheses. As a result, the step-down Šidák procedure is uniformly more powerful than the regular Šidák procedure.

As an illustration, we will apply the step-down Šidák procedure to the problem of testing three null hypotheses in Case study 2. First, the three p-values are ordered from the most significant to the least significant, and the ordered null hypotheses $H_{(1)}$, $H_{(2)}$, and $H_{(3)}$ are defined. Testing begins with the smallest p-value. The significance level in Step 1 is given by $1-(1-\alpha)^{1/3} = 0.0170$, and thus, the corresponding ordered null hypothesis is rejected if $p_{(1)} \leqslant 0.0170$. If the step-down procedure rejects this null hypothesis, a higher significance level is used for the next null hypothesis in the sequence, that is, $1-(1-\alpha)^{1/2} = 0.0253$. If this null hypothesis is rejected, the last ordered null hypothesis is tested at the full $\alpha = 0.05$. The resulting decision rules are shown in Figure 4. It is clear that the step-down Šidák procedure rejects all null hypotheses rejected by the regular Šidák procedure and potentially more.

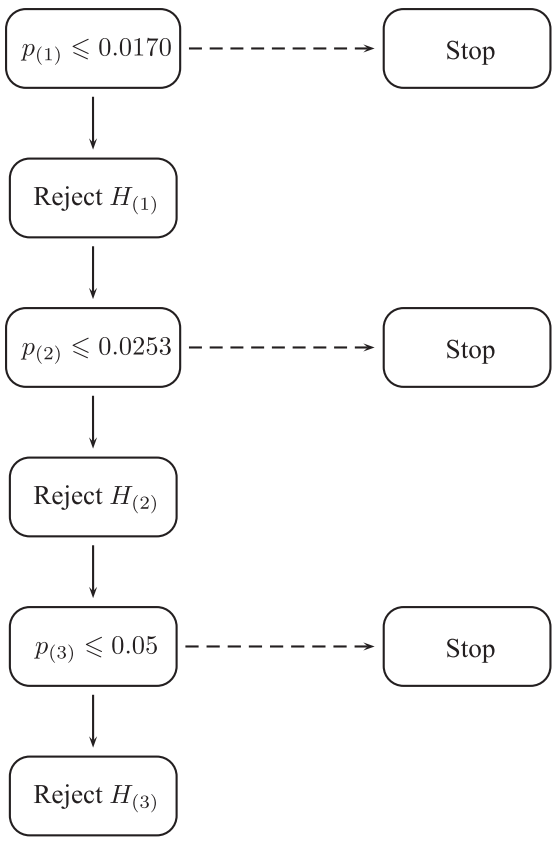

Figure 4. Visual representation of the decision rules used in the step-down Šidák procedure in Case study 2 with two-sided $\alpha = 0.05$. Decisions: —, condition is met; - - -, condition is not met.

It is instructive to compare α propagation rules utilized by Bonferroni-based chain procedures introduced in Section 4 and the α propagation rule used in the step-down Šidák procedure. Because the Bonferroni-based chain procedures rely on a simple α-splitting approach, the associated α propagation rules are additive, see, for example, the decision rules presented in Figure 2. By contrast, the step-down Šidák procedure relies on a more complex nonlinear α propagation rule. Similarly, α propagation rules used by other powerful semiparametric introduced in this section and parametric procedures introduced in Section 6 exhibit nonadditive behavior.

Powerful extensions of the Simes procedure include the **Hochberg procedure** [27] and **Hommel procedure** [28]. The Hochberg procedure utilizes a simple step-up testing algorithm that begins with the largest p-value:

- Steps $i = 1, \ldots, m-1$. The null hypothesis $H_{(m-i+1)}$ is accepted if $p_{(m-i+1)} > \alpha/i$. If $H_{(m-i+1)}$ is accepted, proceed to the next step. Otherwise, reject the null hypotheses $H_{(j)}$, $j = 1, \ldots, m-i+1$.
- Step m. The null hypothesis $H_{(1)}$ is accepted if $p_{(1)} > \alpha/m$. Otherwise, it is rejected.

The Hommel procedure is also based on a step-up testing algorithm [29]:

- Step $i = 1$. The null hypothesis $H_{(m)}$ is accepted if $p_{(m)} > \alpha$. If $H_{(m)}$ is accepted, go to Step 2. Otherwise, reject all null hypotheses.
- Steps $i = 2, \ldots, m-1$. The null hypothesis $H_{(m-i+1)}$ is accepted if $p_{(m-i+j)} > j\alpha/i$ for all $j = 1, \ldots, i$. If $H_{(m-i+1)}$ is accepted, proceed to the next step. Otherwise, reject the null hypotheses $H_{(j)}$, $j = 1, \ldots, m-i+1$, provided $p_{(j)} \leqslant \alpha/(i-1)$.
- Step m. The null hypothesis $H_{(1)}$ is accepted if $p_{(j)} > j\alpha/m$ for all $j = 1, \ldots, m$ or $p_{(1)} > \alpha/(m-1)$. Otherwise, $H_{(1)}$ is rejected.

The main difference between the two step-up testing algorithms is that the decision rule for a null hypothesis in the Hochberg procedure depends only on the p-value associated with that null hypothesis, whereas the Hommel procedure uses the p-value for the current null hypothesis as well as all preceding null hypotheses. This means that the Hommel procedure *borrows* power from the other null hypotheses, which leads to a power gain. In fact, the Hommel procedure is known to be uniformly more powerful than the Hochberg procedure.

To illustrate this property of the Hommel procedure, we return to the multiplicity problem in Case study 2. The Hochberg and Hommel procedures both rely on step-up algorithms with a data-driven hypothesis ordering, that is, the ordered null hypotheses $H_{(1)}$, $H_{(2)}$, and $H_{(3)}$ are tested beginning with $H_{(3)}$. Suppose that the ordered p-values are given by $p_{(1)} = 0.0190$, $p_{(2)} = 0.0306$, and $p_{(3)} = 0.0582$. The decision rules used in the two procedures in this multiplicity problem are shown in Figures 5 and 6. The Hochberg and Hommel procedures both fail to reject the null hypotheses $H_{(2)}$ and $H_{(3)}$ because $p_{(3)} > \alpha = 0.05$ and $p_{(2)} > \alpha/2 = 0.025$. Further, consider the decision rules for $H_{(1)}$. It follows from Figure 5 that the Hochberg procedure does not reject this null hypothesis because $p_{(1)} > \alpha/3 = 0.0167$.

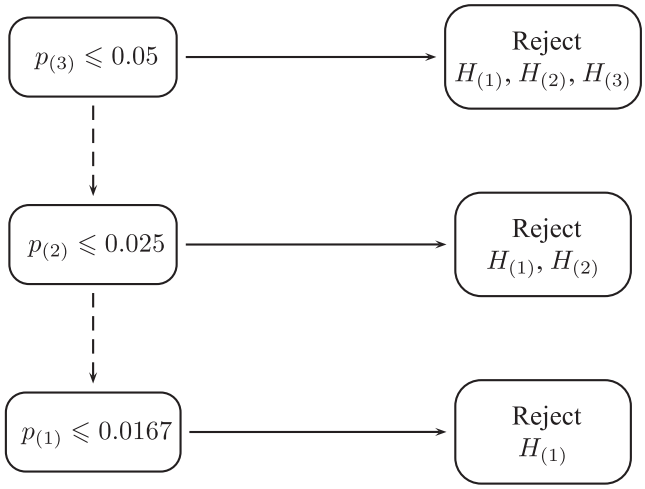

Figure 5. Visual representation of the decision rules used in the Hochberg procedure in Case study 2 with two-sided $\alpha = 0.05$. Decisions: —, condition is met; - - -, condition is not met.

Statistics in Medicine

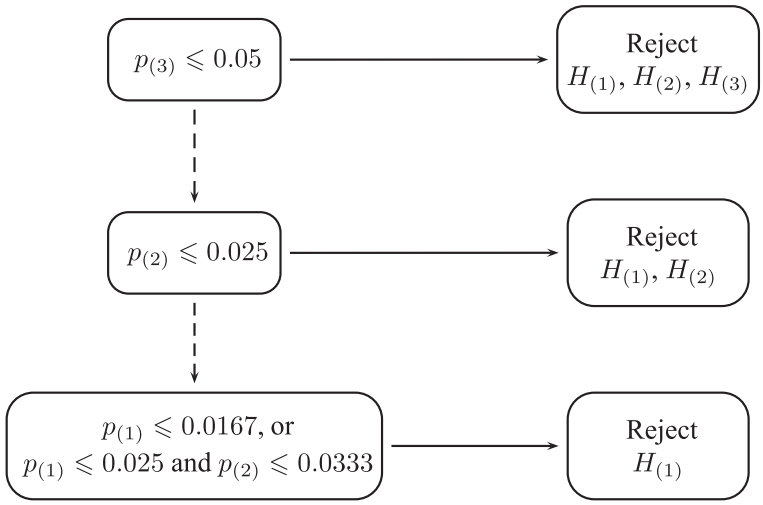

Figure 6. Visual representation of the decision rules used in the Hommel procedure in Case study 2 with two-sided $\alpha = 0.05$. Decisions: —, condition is met; - - -, condition is not met.

The Hommel procedure, on the other hand, *borrows* power from the other null hypotheses, and because of this, this procedure can reject $H_{(1)}$ even if $p_{(1)} > \alpha/3$. In this particular case, Figure 6 shows that the null hypothesis $H_{(1)}$ is rejected because $p_{(1)} \leqslant \alpha/2 = 0.0250$ and $p_{(2)} \leqslant 2\alpha/3 = 0.0333$. The Hommel procedure clearly provides a power advantage over the Hochberg procedure, and for this reason, the Hommel procedure is recommended for wider adoption in clinical trial applications.

An important feature of the Hochberg and Hommel procedures is that, unlike the Bonferroni-based or Šidák-based procedures, they *reward* consistency among the test outcomes. Consistency considerations play an important role in several classes of problems where secondary objectives are evaluated to provide supportive evidence for the primary analysis. To illustrate this concept, consider the following case study:

- Case study 4. The primary analysis in a cardiovascular trial is based on a composite endpoint based on three components (cardiovascular mortality, stroke, and myocardial infarction). The primary analysis is accompanied by the analysis of the three individual components that are analyzed after the primary objective is met.

As discussed in Huque, Alosh and Bhore [12], scientific credibility of the new treatment in Case study 4 is greatly improved if this treatment has a consistent effect on all component endpoints, for example, if all component p-values are simultaneously small. If the Bonferroni or Šidák procedures are applied to analyze the component endpoints, the same significance level is used for each p-value regardless of the magnitude of the other p-values. With the Bonferroni adjustment, for example, the first component is evaluated at $\alpha/3$ even though the treatment effect on the other two components may be highly nonsignificant. By contrast, the Hochberg and Hommel procedures favor the configurations of p-values that exhibit a certain degree of concordance. If the p-values for all component endpoints are simultaneously small, each one of them is tested at the full α level. However, if one or more p-values are far from being significant, much stricter significance levels are used for the other p-values, which represents a penalty for lack of concordance.

The Simes-based and Šidák-based procedures defined earlier are semiparametric procedures, which means that they control the FWER only under additional distributional assumptions. For example, if the hypothesis test statistics follow a multivariate normal distribution, the Simes-based procedures (Hochberg and Hommel procedures) preserve the error rate if the common correlation coefficient is nonnegative or all partial correlations among the test statistics are nonnegative [30, 31]. Thus, if the two test statistics in Case studies 1 or 3 follow a bivariate normal distribution, the error rate is controlled as long as the correlation coefficient is nonnegative. This will be the case in Case study 1 if the two endpoints are independent or positively correlated, and in Case study 3, the correlation is always nonnegative because the two patient populations overlap. If a trivariate normal distribution can be assumed in Case study 2,

the Hochberg and Hommel procedures control the error rate in any balanced design (if a common sample size is used across the treatment arms) because a balanced design implies a common positive correlation among the three test statistics.

The Šidák-based procedures impose less restrictive conditions on the joint distribution of the hypothesis test statistics. These procedures protect the error rate in any problem as long as the test statistics follow a multivariate normal distribution [26]. Thus, the regular and step-down Šidák procedures can be safely used in Case studies 1–3 under multivariate normality.

Finally, it was shown in Section 4 that nonparametric procedures are easily set up for multiplicity problems with unequally weighted null hypotheses. Similarly, it is easy to extend the regular and step-down Šidák procedures as well as the Hommel procedure to this general setting. However, the problem of defining a weighted version of the Hochberg procedure is more challenging, see Tamhane and Liu [32].

6. Parametric procedures in single-family problems

Unlike the nonparametric and semiparametric multiple testing procedures discussed in Sections 4 and 5, parametric procedures are based on an explicitly defined model that generates a fully specified joint distribution for the hypothesis test statistics. Multiplicity problems with multiple dose–control comparisons (Case study 2) or multiple patient populations (Case study 3) provide examples of settings with fully specified joint distributions.

Because parametric procedures make stronger distributional assumptions, they are more powerful than nonparametric and semiparametric procedures. However, improved power comes with a price. Parametric procedures guarantee error rate control only for specific models or distributions of the hypothesis test statistics. For this reason, parametric procedures can only be applied in the settings where the underlying joint distribution under the global null hypothesis is known and specified at the design stage of a clinical trial. Examples include multivariate normal or t distributions with known parameters (for example, the correlation coefficients must be specified).

A basic parametric procedure can be constructed as follows. Assume that the joint distribution for the test statistics $t_1, \ldots, t_m$ is known and a null hypothesis is rejected if the associated test statistic is large. For example, H_i is rejected if $t_i \geqslant c$, where c is an appropriately defined critical value, $i = 1, \ldots, m$. The FWER for this procedure is given by

$$P(t_1 \geqslant c \text{ or } \ldots \text{ or } t_m \geqslant c),$$

and the critical value c is found by equating this probability to α. This parametric procedure is exact in the sense that the critical value is computed directly from the joint distribution for the test statistics rather than approximations or probabilistic inequalities.

This general idea is used in the derivation of the Dunnett procedure [33] for clinical trials with several treatment groups and a single control arm. The most basic version of this procedure relies on a one-way analysis of variance model

$$y_{ij} = \mu_i + \varepsilon_{ij},$$

where y_{ij} denotes the response of the jth patient in the ith treatment group, $i = 0, \ldots, m$ ($i = 0$ is the index of the control group) and $j = 1, \ldots, n$. The errors in this model are normally distributed with mean 0 and a common standard deviation. The null hypotheses are defined as $H_i : \mu_i \leqslant \mu_0, i = 1, \ldots, m$, and the hypothesis test statistics follow a fully specified multivariate t distribution with known correlation coefficients. The critical value of the Dunnett procedure is computed from this multivariate distribution. Because the Dunnett procedure fully utilizes available distributional information, it is uniformly more powerful than the similar nonparametric and semiparametric procedures. As a quick comparison, by using Case study 2 with a balanced design and a common sample size of $n = 180$, the Dunnett-adjusted significance level for each of the three dose–placebo comparisons is 0.0196. This level is appreciably higher than the Bonferroni-adjusted level of 0.0167 or Šidák-adjusted level of 0.0170. Note that the sum of the Dunnett-adjusted significance levels across the three hypotheses is 0.0588, and it exceeds the full α level of 0.05. As was emphasized in Section 5, efficient multiple testing procedures do not rely on simplistic additive or α-splitting rules. Simple rules of this kind arise only when approximate approaches are used, for example, when the Bonferroni approach is applied [5].

Numerous extensions of the regular Dunnett ocedure have been proposed in the literature to form a family of Dunnett-type parametric procedures. This includes an extension based on a step-down algorithm [34], which is conceptually similar to the step-down Šidák procedure and a step-up extension [35] similar to the Hochberg procedure. Further, Dunnett-type parametric procedures have been developed for multiplicity problems with multiple dose–control contrasts, all pairwise contrasts, and arbitrary contrasts in general linear models with fixed and random effects and other models [36]. These procedures enable exact multiplicity adjustments for analysis models with covariates, for example, baseline values of the outcome variable and stratification factors, as well as longitudinal models. Note, however, that challenges arise in unbalanced designs, for example, in the presence of missing data. Voss and Hsu [37] showed that it is impossible to build an exact parametric procedure for mixed-effects models in an unbalanced design. Given this, clinical trial sponsors have to resort to approximate parametric procedures in this and similar settings.

Examples of other commonly used parametric multiple testing procedures include the class of parametric chain procedures (Millen and Dmitrienko [22]), which includes some parametric procedures from the Dunnett family as well as the parametric fallback procedure (Huque and Alosh [38]). The parametric chain approach is illustrated in the following. Another class of parametric procedures, termed feedback procedures, was developed in Zhao, Dmitrienko and Tamura [39]. This class includes as special cases the 4A procedure (Li and Mehrotra [40]) and CAS (consistency-adjusted strategy) procedures (Alosh and Huque [41, 42]). Feedback procedures have found applications in settings similar to that considered in Case study 3 where the efficacy profile of a new treatment is evaluated in the overall population of patients with the condition of interest and one or more prespecified subpopulations. The parametric approach used in these procedures is based on an intuitively appealing feedback principle, which adjusts the critical value for the subpopulation tests depending on the magnitude of the overall treatment effect (and thus the overall population test provides feedback to the subpopulation tests).

In what follows, we define a parametric chain procedure for the flexible testing strategy for Case study 2 introduced in Section 4. This procedure serves as an extension of the nonparametric Bonferroni-based chain procedure defined in that section and illustrates advantages of the parametric testing approach. The parametric chain procedure is on the basis of the same decision rules as the nonparametric procedure (see Figure 2). Specifically, testing begins with the null hypotheses H_1 and H_2 corresponding to the high and medium doses. If a significant treatment effect is detected at either dose or both doses, the treatment effect is evaluated at the low dose, that is, the null hypothesis H_3 is tested. Finally, a retesting option, used by all powerful multiple testing procedures, is applied, and the null hypotheses H_1 and H_2 are revisited if H_3 is rejected. The parametric procedure for a balanced design with 180 patients per group is implemented using a simple three-step testing algorithm (see Figure 7):

- Step 1. H_1 and H_2 are tested at the significance levels computed from the joint distribution of the two test statistics, that is, each test is carried out at 0.028.
- Step 2. If at least one test is significant in Step 1, H_3 is tested at 0.05 if both null hypotheses are rejected in Step 1 and at 0.028 if only one null hypothesis is rejected in Step 1.
- Step 3. If H_3 is rejected in Step 2 and one null hypothesis is not rejected in Step 1, the remaining nonrejected null hypothesis is retested at 0.05.

The key difference between the nonparametric and parametric chain procedures is that the latter utilizes the fact that the three test statistics follow a multivariate t distribution with positive pairwise correlations. Because of the positive correlations, the significance levels used by the parametric procedure are greater than those used by the nonparametric procedure in Figure 2, which results in a uniform power gain. The parametric procedure is guaranteed to reject as many and potentially more null hypotheses compared with the nonparametric procedure.

7. Multiplicity problems with multiple families of null hypotheses

The multiplicity problems discussed in Sections 4–6 were defined as problems with a single source of multiplicity or a single family of null hypotheses. In the context of confirmatory drug development, multiplicity problems of this kind are typically encountered in trials with multiple primary objectives. A more complex setting will be considered in this section. We will focus on multiplicity problems with several sources of multiplicity or several families of null hypotheses defined in terms of hierarchically ordered objectives, for example, primary, secondary, and other objectives in confirmatory clinical trials. Classification of clinical trial objectives, including hierarchies of multiple endpoints, is discussed

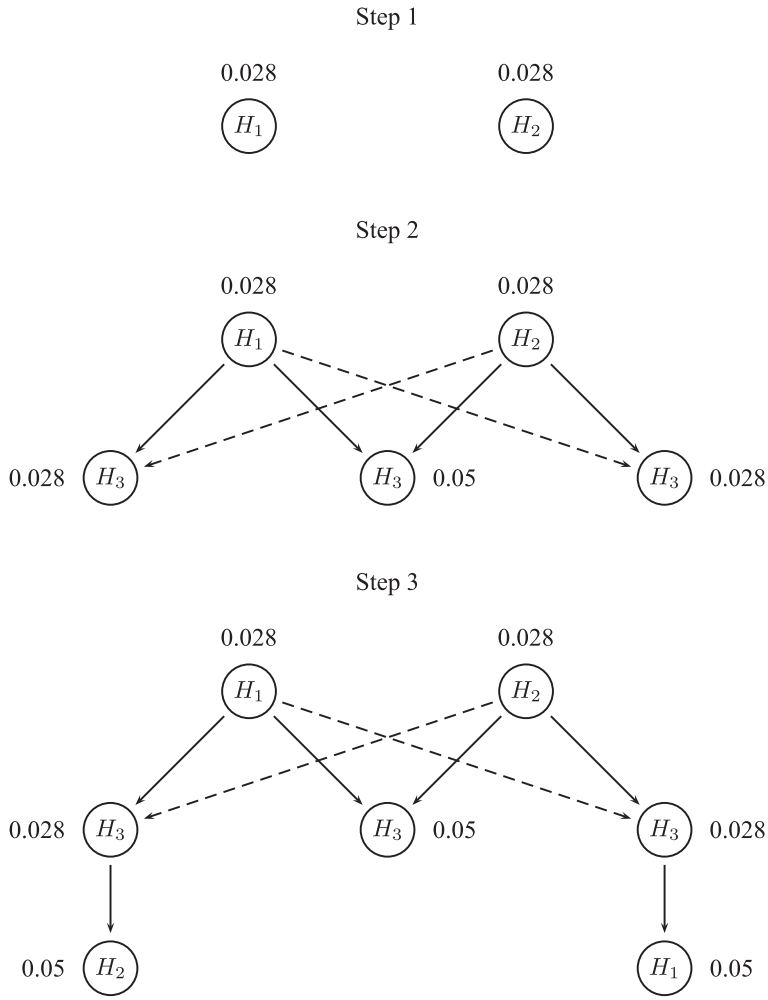

Figure 7. Implementation of the parametric chain procedure in Case study 2 with two-sided $\alpha = 0.05$ (H_1, high dose versus placebo; H_2, medium dose versus placebo; H_3, low dose versus placebo). Decisions: —, null hypothesis is rejected; - - -, null hypothesis is not rejected.

in D'Agostino and Heeren [43] , O'Neill [44], D'Agostino [45], Huque and Röhmel [46], and Hung and Wang [47].

The following two case studies will be used to illustrate key issues arising in problems with several sources of multiplicity:

- Case study 5. The efficacy profile of a new treatment versus placebo is evaluated at two different dose levels in the overall patient population and a predefined subpopulation.
- Case study 6. The efficacy profile of a new treatment versus placebo is evaluated at two different dose levels with respect to two endpoints (one primary and one secondary).

Multiplicity problems arising in Case studies 5 and 6 serve as examples of advanced problems with multiple families of null hypotheses. In both clinical trials, the objectives are naturally grouped into two families. The first family includes the most important analyses that define the overall outcome of the trial, and the second family includes the analyses that provide supportive evidence of treatment benefit. In Case study 5, the main regulatory claim will follow if the new treatment is shown to be effective in the overall population of patients, and thus the dose–placebo comparisons in this population are placed in the first family. The second family contains the subpopulation tests. In the other case study, the primary objective is formulated in terms of establishing a beneficial effect of the new treatment on the primary endpoint. Thus, the first family includes the dose–placebo tests based on the primary endpoint, and the secondary endpoint tests are put in the second family.

Statistics in Medicine

7.1. *Gatekeeping procedures*

Advanced multiplicity problems in clinical trials with several families of null hypotheses have received much attention in the literature. A general class of multiple testing procedures designed specifically for problems with multiple families, termed **gatekeeping procedures**, has been developed [4,5,43,48]. The main regulatory requirement for gatekeeping procedures in confirmatory clinical trials with multiple families of objectives is strong control of the global FWER at a two-sided $\alpha = 0.05$. The word *global* indicates that the error rate must be protected across all families. It is not sufficient to provide local FWER control within each family.

It is worth noting that in certain cases, multiple testing procedures such as the chain procedures (see Sections 4 and 6) may be applied to advanced multiplicity problems considered in this section. However, to accommodate multiple families, various artificial constructs typically need to be introduced. For example, as pointed out in Bretz *et al*. [49], *placeholder* connection weights need to be defined in multifamily problems within the chain framework. Consider a two-family problem based on Case study 2 with three null hypotheses. Assume a setting similar to that considered in Section 4, and suppose that the trial's sponsor expects a nonmonotonic dose–response function in the trial. The null hypotheses H_1 and H_2 are placed in Family 1 and tested first using the Holm procedure. The remaining null hypothesis is included in Family 2 and is tested if both null hypotheses are rejected in Family 1. A gatekeeping procedure for this problem is constructed in a straightforward manner as shown in Figure 8. The left panel in this figure defines a gatekeeping procedure that utilizes the Holm procedure in Family 1 and transfers all of the error rate to Family 2 if both null hypotheses are rejected in Family 1. By contrast, if families are not explicitly specified and a basic nonparametric chain procedure is applied, an infinitesimal weight ε for the connections between the two null hypotheses in Family 1 and single null hypothesis in Family 2 must be introduced to build a valid procedure (see the right panel of Figure 8). Infinitesimal weights are defined as infinitely small quantities that are still positive. They serve as placeholders and can turn into *real* weights if additional conditions are satisfied. Placeholder weights may be difficult to interpret even in simple problems. More complex and less transparent extensions of standard chain procedures need to be introduced in problems with three or more families of null hypotheses.

In what follows, we will describe three approaches to building gatekeeping procedures that satisfy the regulatory requirement of global error rate control and provide clinical trial sponsors with powerful methods for addressing advanced multiplicity problems. Assuming a simple setting with two families of null hypotheses, as in Case studies 5 and 6, the three approaches include the following:

- Sequential testing. Family 1 is tested first, and if a prespecified condition is satisfied, Family 2 is tested.
- Sequential testing with retesting. Family 1 is tested first. If a prespecified condition is satisfied in Family 1, Family 2 is tested. Further, if another prespecified condition is satisfied in Family 2, the null hypotheses in Family 1 are retested.
- Simultaneous testing. Families 1 and 2 are tested simultaneously. If a prespecified condition is satisfied, the two families are retested. Retesting is performed multiple times until all null hypotheses are rejected or the testing algorithm reaches a stable point.

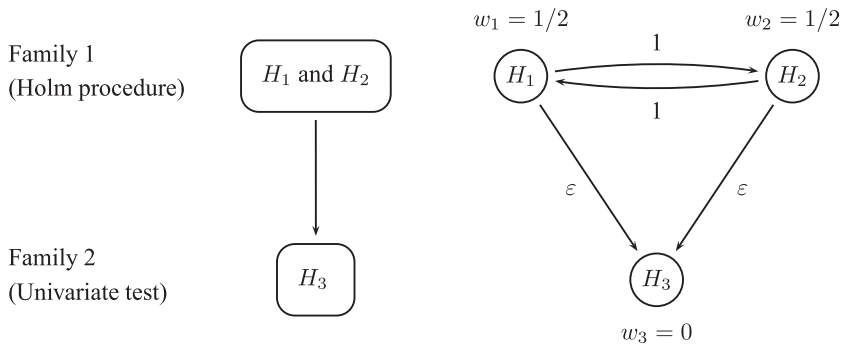

Figure 8. Gatekeeping strategies in Case study 2. Family-based gatekeeping procedure (left panel) and chain-based gatekeeping procedure (right panel). An infinitesimally small weight ε needs to be used in the right panel to define a valid chain procedure.

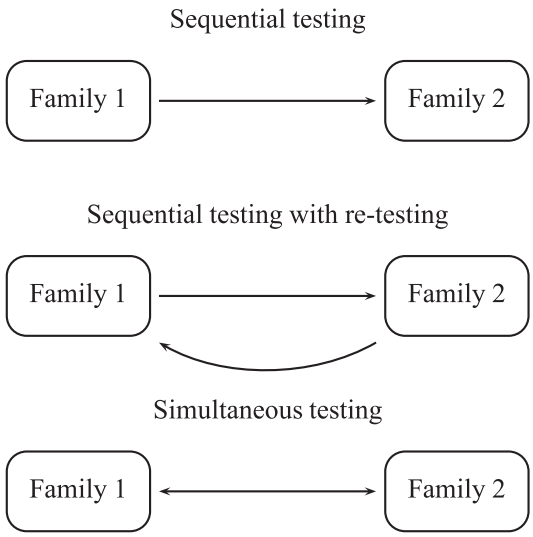

Figure 9. Three approaches to building gatekeeping procedures.

The three testing strategies are represented graphically in Figure 9. An important concept that unites the three strategies is the concept of α propagation, which also plays a key role in multiple testing procedures discussed in Sections 4–6. In the current setting, α propagation is applied across families of null hypotheses rather than individual null hypotheses. The concept of α propagation enables clinical trial sponsors to set up flexible gatekeeping procedures that support a wide variety of clinical trial objectives. For example, in the context of sequential testing with retesting, if a prespecified condition is met in Family 1, all available error rate is carried forward to Family 2, and under additional conditions, the remaining error rate may be transferred back to Family 1. The prespecified conditions take form of serial, parallel, or general gatekeeping conditions [4, 50] and will be defined and illustrated in the following.

7.2. Sequential testing

To motivate the sequential testing approach, consider Case study 5 with two families of null hypotheses:

- Family 1 contains the null hypotheses in the overall patient population (H_1, high dose versus placebo; H_2, low dose versus placebo).
- Family 2 contains the null hypotheses in the subpopulation (H_3, high dose versus placebo; H_4, low dose versus placebo).

On the basis of the objectives of this clinical trial, Family 1 is treated as the primary family, and the two families are tested sequentially as shown in the top panel of Figure 9. As pointed out in Hung and Wang [51], to set up a clinically meaningful gatekeeping procedure for this multiplicity problem, it is critical to first understand the logical relationships among the null hypotheses. In this clinical trial, Family 1 serves as a **gatekeeper** for Family 2, that is, certain conditions must be met in Family 1 before the sponsor can carry out the tests in for Family 2. Examples of gatekeepers include **serial gatekeepers** and **parallel gatekeepers**. With serial gatekeeping, one passes the gatekeeper and tests the null hypotheses in Family 2 only if there is evidence of a beneficial treatment effect at both dose levels in the overall population, that is, H_1 and H_2 are both rejected. This restriction is not consistent with the trial's objective because, from the sponsor's perspective, the primary objective is met if a positive effect is established at only one dose level in the overall population. In this case, it is more appropriate to define Family 1 as a parallel gatekeeper, which means that the gatekeeper is passed and both subpopulation tests are carried out if either or both overall population tests are significant. The relationships among the null hypotheses in the parallel gatekeeping case are displayed in Figure 10.

A general method for constructing gatekeeping procedures with parallel restrictions was introduced in Dmitrienko, Tamhane and Wiens [52] and Dmitrienko, Kordzakhia and Tamhane [53]. These gatekeeping procedures utilize straightforward stepwise testing algorithms that facilitate their

Statistics in Medicine

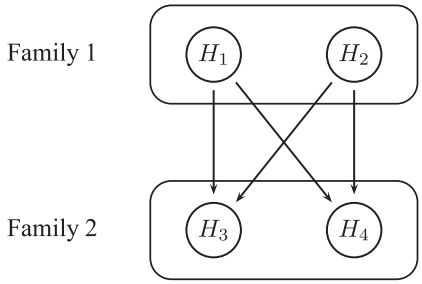

Figure 10. Logical relationships in the parallel gatekeeping procedure used in Case study 5 (H_1, high dose versus placebo in the overall population; H_2, low dose versus placebo in the overall population; H_3, high dose versus placebo in the subpopulation; H_4, low dose versus placebo in the subpopulation).

development and implementation. A gatekeeping procedure is built by specifying the procedures that will be applied in each family, known as *component procedures*, and an α propagation rule that determines how much error rate is transferred from Family 1 to Family 2 when the parallel gatekeeping condition is satisfied in Family 1.

The selection of component procedures is driven by the available information on the joint distribution of the hypothesis test statistics within each family. When the joint distribution is unknown, nonparametric component procedures are used. In this particular case, the test statistics within Family 1 and Family 2 follow a bivariate normal distribution with a positive correlation. For example, the correlation coefficient is 0.5 in a balanced design. As was explained in Section 5, semiparametric procedures control the error rate in this setting and provide a uniform power improvement over basic nonparametric procedures. The sponsor can set up a gatekeeping procedure with semiparametric Hochberg-type component procedures (note that the Hochberg procedure is equivalent to the Hommel procedure because there are two null hypotheses in each family). This Hochberg-based gatekeeping procedure is more powerful than a gatekeeping procedure constructed from Bonferroni components.

It is important to note that, if the regular Hochberg procedure is applied in Family 1, a positive fraction of the α is transferred from Family 1 to Family 2 only if both H_1 and H_2 are rejected. No error rate can be propagated to Family 2 if only one test is significant in Family 1. This is due to the fact that the regular Hochberg procedure is a *nonseparable* (greedy) procedure, which *spends* all of the available error rate. This procedure can be used only if Family 1 is a serial gatekeeper. To enable parallel gatekeeping, a modified version of the Hochberg procedure, termed the **truncated Hochberg procedure**, needs to be utilized in Family 1. This procedure is *separable* (generous) and is defined as a convex combination of the Bonferroni and Hochberg procedures. Specifically, let $p_{(1)} < p_{(2)}$ denote the ordered p-values in Family 1. As was explained in Section 5, the regular Hochberg procedure uses the following decision rules:

- Both null hypotheses in Family 1 are rejected if $p_{(2)} \leqslant \alpha$.
- Only the null hypothesis corresponding to $p_{(1)}$ is rejected if $p_{(1)} \leqslant \alpha/2$ and $p_{(2)} > \alpha$.

The truncated procedure relies on a similar set of rules, that is,

- Both null hypotheses in Family 1 are rejected if $p_{(2)} \leqslant (1+\gamma)\alpha/2$.
- Only the null hypothesis corresponding to $p_{(1)}$ is rejected if $p_{(1)} \leqslant \alpha/2$ and $p_{(2)} > (1+\gamma)\alpha/2$.

The parameter $0 \leqslant \gamma < 1$ is known as a truncation parameter and denotes the degree of *greediness*. It is easy to verify that the truncated procedure simplifies to the Bonferroni procedure (generous procedure) if $\gamma = 0$ and to the regular Hochberg procedure (greedy procedure) if $\gamma = 1$. Considering Family 2, the regular Hochberg procedure can be utilized in this family because the second family does not serve as a gatekeeper for any other family.

The resulting Hochberg-based parallel gatekeeping procedure is based on a straightforward two-step algorithm with the following α propagation rule:

- Step 1. The truncated Hochberg procedure is applied in Family 1 at $\alpha_1 = \alpha$. If no null hypotheses are rejected, testing stops. Otherwise, go to Step 2.
- Step 2. The regular Hochberg procedure is applied in Family 2 at α_2. This level equals α if both null hypotheses are rejected in Step 1 and $(1-\gamma)\alpha/2$ if only one null hypothesis is rejected in Step 1.

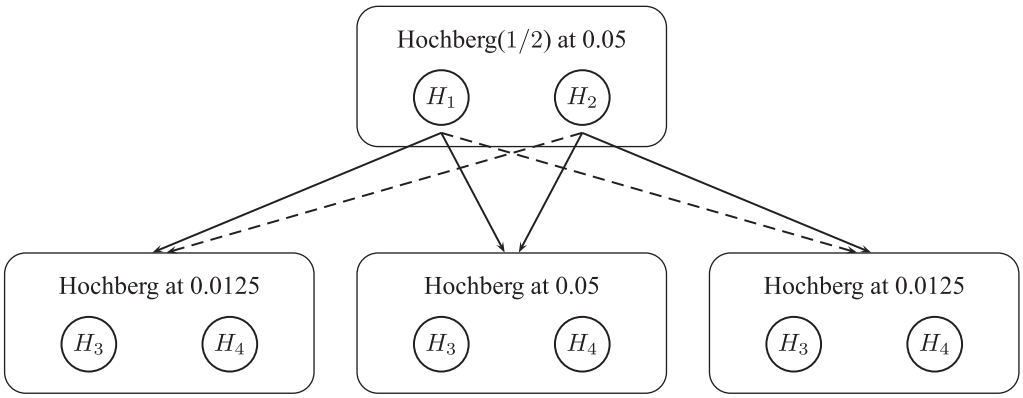

Figure 11. Parallel gatekeeping procedure in Case study 5 (H_1, high dose versus placebo in the overall population; H_2, low dose versus placebo in the overall population; H_3, high dose versus placebo in the subpopulation; H_4, low dose versus placebo in the subpopulation). Decisions: —, null hypothesis is rejected; - - -, null hypothesis is not rejected.

To illustrate the process of performing multiplicity adjustments on the basis of the parallel gatekeeping procedure, suppose that the p-values in Case study 5 are given by

$$\text{Family 1: } p_1 = 0.017,\ p_2 = 0.041;\ \ \text{Family 2: } p_3 = 0.011,\ p_4 = 0.008.$$

The decision rules used in the Hochberg-based gatekeeping procedure with the truncation parameter $\gamma = 1/2$ in Family 1 are displayed in Figure 11. Note that the truncated Hochberg procedure with $\gamma = 1/2$ is denoted by Hochberg(1/2). First, when the truncated procedure is applied in Family 1, the null hypothesis H_2 is not rejected because it corresponds to the larger p-value within this family and $p_2 > (1+\gamma)\alpha/2 = 0.0375$. The null hypothesis H_1 is rejected because $p_1 \leqslant \alpha/2 = 0.025$. This indicates that only one dose is effective in the overall population. As a result, as shown in the lower left panel of the figure, only a fraction of the overall error rate is transferred from Family 1 to Family 2. Specifically, the subpopulation tests in Family 2 are carried out on the basis of the regular Hochberg procedure at $(1-\gamma)\alpha/2 = 0.0125$. The reduced level represents a penalty for not rejecting both null hypotheses in Family 1. When the regular Hochberg procedure is applied in Family 2, both H_3 and H_4 are rejected because the larger of the two p-values in the family is less than $(1-\gamma)\alpha/2 = 0.0125$. This means that both doses provide evidence of efficacy in the subpopulation.

In general, in multiplicity problems with several families of null hypotheses, the error rate transferred from one family to the next family in the sequence is computed from the error rate function of the component procedure used in the current family [52]. For example, if the truncated Hochberg procedure with the truncation parameter γ is applied in Family 1 with m null hypotheses, the overall significance level for Family 2 is given by

$$\alpha_2 = \frac{(1-\gamma)r\alpha}{m} \text{ if } r < m \text{ and } \alpha_2 = \alpha \text{ if } r = m,$$

where r denotes the number of null hypotheses rejected in Family 1. This means that the null hypotheses in Family 2 are tested at the full α if all null hypotheses are rejected in Family 1 ($r = m$). On the other hand, no error rate is carried over to Family 2 if there are no significant tests in Family 1 ($r = 0$). Also, it is clear from this formula and the example based on Case study 5 that prespecified truncation parameters play an important role in multistep gatekeeping procedures. In Case study 5, the truncation parameter γ determines the significance level used in Family 2, and thus, it affects the balance of power between the two families. Criteria for choosing truncation parameters in multistep gatekeeping procedures are discussed in Dmitrienko *et al.* [54].

The sequential testing approach can also be applied to multiplicity problems with more general relationships among the families or individual null hypotheses. Case study 6 defines a problem where neither the serial nor parallel gatekeeping condition is satisfied. Rather, the gatekeeping condition is more

Statistics in Medicine

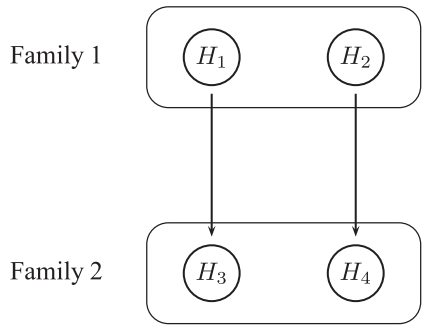

Figure 12. Logical relationships in the general gatekeeping procedure used in Case study 6 (H_1, high dose versus placebo based on the primary endpoint; H_2, low dose versus placebo based on the primary endpoint; H_3, high dose versus placebo based on the secondary endpoint; H_4, low dose versus placebo based on the secondary endpoint).

general than those two types. The four null hypotheses of no effect defined in this trial are grouped into two families:

- Family 1 includes the null hypotheses related to the primary endpoint (H_1, high dose versus placebo; H_2, low dose versus placebo).
- Family 2 includes the null hypotheses related to the secondary endpoint (H_3, high dose versus placebo; H_4, low dose versus placebo).

Family 1 serves as a gatekeeper, and the two families are tested sequentially as shown in Figure 12.

It is easy to see that Family 1 does not serve as a serial gatekeeper because the effect of the secondary endpoint at a dose level can still be analyzed in Family 2 even if the primary test for the other dose is not significant. Further, Family 1 is not a parallel gatekeeper for Family 2. Indeed, assume that only one dose is significantly different from placebo on the primary endpoint, for example, H_1 is rejected, but H_2 is not rejected. If Family 1 was a parallel gatekeeper, both null hypotheses would then be tested in Family 2. However, H_4 corresponds to the secondary endpoint test at the other dose, and it is not clinically meaningful to carry out this test given that the primary objective is not met at this dose level. The gatekeeper in Case study 6 is an example of a general gatekeeper. Testing decisions are made individually for each null hypothesis in Family 2 rather than for the entire family as in the case of serial or parallel gatekeeping. Because of the relationship between the primary and secondary endpoints, H_3 becomes testable only if H_1 is rejected, and similarly, H_4 is testable only if H_2 is rejected.

Gatekeeping procedures for multiplicity problems with general gatekeepers can be set up using the mixture method [50]. This method supports arbitrary logical relationships among the individual null hypotheses and efficiently incorporates distributional information to set up powerful gatekeeping procedures. For example, Brechenmacher *et al*. [29] applied this method to derive powerful Hommel-based gatekeeping procedures that were successfully utilized in the lurasidone Phase III trials [55].

In the context of Case study 6, the mixture method can be used to define a Hochberg-based gatekeeping procedure that accounts for the logical relationships among the four null hypotheses in this clinical trial (note again that the Hochberg procedure is equivalent to the Hommel procedure in this problem). As in Case study 5, a truncated version of the Hochberg procedure with the truncation parameter $0 \leqslant \gamma < 1$ must be used in Family 1 because this family is not a serial gatekeeper. The regular Hochberg procedure is applied in Family 2 because it is the last family in the sequence. The gatekeeping procedure is based on an easy-to-implement two-step algorithm:

- Step 1. The truncated Hochberg procedure is applied in Family 1 at $\alpha_1 = \alpha$. If no null hypotheses are rejected, testing stops. Otherwise, go to Step 2.
- Step 2. The regular Hochberg procedure is applied in Family 2 at $\alpha_2 = \alpha$ if both null hypotheses are rejected in Step 1. If exactly one null hypothesis is rejected in Step 1, the logically related null hypothesis in Family 2 is tested at $\alpha_2 = (1-\gamma)\alpha/2$.

The underlying testing algorithm used in this Hochberg-based gatekeeping procedure with the truncation parameter $\gamma = 1/2$ in Family 1 is defined in Figure 13. As in Figure 11, the truncated Hochberg procedure with $\gamma = 1/2$ is denoted by Hochberg(1/2). It is easy to see that the gatekeeping procedure is

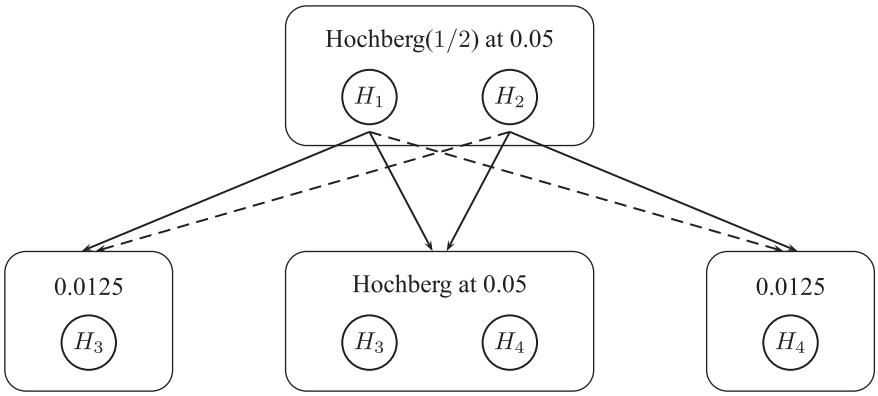

Figure 13. General gatekeeping procedure in Case study 6 (H_1, high dose versus placebo based on the primary endpoint; H_2, low dose versus placebo based on the primary endpoint; H_3, high dose versus placebo based on the secondary endpoint; H_4, low dose versus placebo based on the secondary endpoint). Decisions: —, null hypothesis is rejected; - - -, null hypothesis is not rejected.

consistent with the logical restrictions shown in Figure 12. For example, as shown in the lower left and central panels, the null hypothesis H_3 will be tested only if H_1 is rejected. In addition, this gatekeeping procedure utilizes an efficient α propagation rule, that is, no multiplicity penalty is paid if H_1 and H_2 are rejected, and thus both doses are significantly different from placebo on the primary endpoint. In this case, the secondary tests are carried out on the basis of the regular Hochberg procedure at the full $\alpha = 0.05$. Further, as shown in Figure 13, if only one dose is significant in Family 1, the corresponding null hypothesis in Family 2 is carried out using a univariate test at a reduced level. For example, if H_1 is rejected and H_2 is not rejected, the null hypothesis H_3 is tested at $(1-\gamma)\alpha/2 = 0.0125$, whereas H_4 is not tested at all.

The resulting Hochberg-based gatekeeping procedure is more powerful than a basic Bonferroni-based gatekeeping procedure that simply splits the FWER α between the two sets of null hypotheses. In other words, H_1 and H_2 are each tested at $\alpha/2 = 0.025$. Then, H_3 is tested at $\alpha/2 = 0.025$ provided H_1 is rejected, and similarly, H_4 is tested at $\alpha/2 = 0.025$ if H_2 is rejected.

7.3. Sequential testing with retesting

The testing approach defined in Section 7.2 emphasizes unidirectional α propagation rules. The error rate in the gatekeeping procedures based on the decision rules displayed in Figures 10 and 12 can only be transferred from Family 1 to Family 2 but not in the other direction. This approach is relevant if the primary inferences in Family 1 are required to be independent of the outcomes of the secondary tests in Family 2 [44]. If the independence condition can be relaxed, the strictly sequential testing strategy can be augmented by including an option to revisit Family 1 after certain conditions are met in Family 2. This strategy is shown in the middle panel of Figure 9 and is similar to the retesting strategies used in powerful multiple testing procedures introduced in Sections 4–6. It is important to note that the retesting approach is based on the closure principle and gatekeeping procedures with retesting still protect the global FWER in the strong sense.

The sequential testing approach used in Case study 5 is easily extended to include a retesting option. Another step is added in the two-step parallel gatekeeping procedure defined in Section 7.2 as follows:

- Step 3. The regular Hochberg procedure is applied in Family 1 at $\alpha_3 = \alpha$ if both null hypotheses are rejected in Step 2.

In other words, the family of subpopulation tests (Family 2) is treated as a serial gatekeeper for the overall population tests (Family 1). If the serial gatekeeping condition is met in Step 2 and there is a null hypothesis in Family 1 that is not rejected in Step 1, the procedure returns to the overall population tests and retests the nonrejected null hypothesis. This null hypothesis is retested at a higher significance level, and thus, the three-step gatekeeping procedure rejects as many and potentially more null hypotheses than the two-step gatekeeping procedure.

Statistics in Medicine

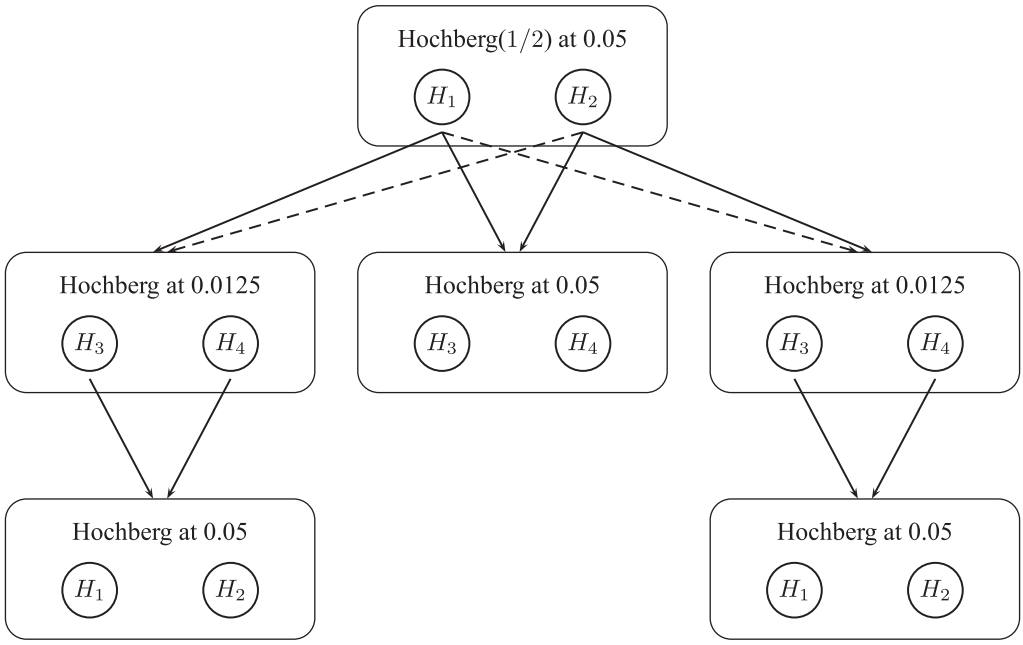

Figure 14. Parallel gatekeeping procedure with retesting in Case study 5 (H_1, high dose versus placebo in the overall population; H_2, low dose versus placebo in the overall population; H_3, high dose versus placebo in the subpopulation; H_4, low dose versus placebo in the subpopulation). Decisions: —, null hypothesis is rejected; - - -, null hypothesis is not rejected.

The parallel gatekeeping procedure with a retesting option is defined in Figure 14. Using the set of p-values introduced in Section 7.2, that is,

$$\text{Family 1: } p_1 = 0.017,\ p_2 = 0.041;\ \ \text{Family 2: } p_3 = 0.011,\ p_4 = 0.008,$$

recall that, with the two-step parallel gatekeeping procedure, only one test is found to be significant in Family 1 (H_1 is rejected) and both tests are significant in Family 2 (H_3 and H_4 are rejected). Because H_3 and H_4 are both rejected, Figure 14 shows that the three-step procedure returns to Family 1 in Step 3 and retests the two null hypotheses by using the regular Hochberg procedure at the full $\alpha = 0.05$. The regular Hochberg procedure is uniformly more powerful than the truncated Hochberg procedure applied in Step 1, and therefore, H_1 is guaranteed to be rejected in Step 3. In addition, H_2 is also rejected because $p_2 \leqslant \alpha = 0.05$. Thus, an application of the three-step procedure results in an additional significant outcome in Family 1.

Other methods for constructing powerful gatekeeping procedures with retesting options are discussed in Dmitrienko, Kordzakhia and Tamhane [53] and Dmitrienko *et al*. [54]. This includes, for example, two-step gatekeeping procedures with separable (generous) component procedures in Family 2. With a separable component procedure, Family 2 turns into a parallel gatekeeper, and thus, the gatekeeping procedure can return to Family 1 when only one test is significant in Family 2. This leads to more flexible gatekeeping methods that are conceptually similar to the methods defined in Section 7.4.

Operating characteristics of gatekeeping procedures with retesting options were studied in Dmitrienko, Soulakova and Millen [56]. In general, the retesting step leads to a material power gain only if less powerful component procedures, for example, Bonferroni procedures, are used in earlier steps of the testing algorithm. In the context of Case study 5, if a powerful Hochberg-type procedure (e.g., truncated Hochberg procedure with a large truncation parameter) is applied in Family 1, there is little room for improving power of the tests in that family, and the operating characteristics of gatekeeping procedures with retesting are very similar to the operating characteristics of gatekeeping procedures that do not utilize retesting.

7.4. Simultaneous testing

An important feature of the sequential testing approach with a retesting option introduced in Section 7.3 is that it provides an asymmetric treatment of the individual families of null hypotheses. Within this

framework, the families are ordered on the basis of their clinical relevance, and testing always begins with the first family in the sequence. An α propagation rule defines the process of distributing the available error rate across the families according to their position in the sequence. When the families are interchangeable, a testing approach based on symmetric α propagation rules is more appropriate. Consider, for example, the multiplicity problem arising in Case study 5. When this problem was discussed in Section 7.2, it was assumed that the primary objective of the trial was formulated in terms of establishing a beneficial treatment effect in the overall population of patients. The family of subpopulation tests was viewed as a secondary family. A sequential testing strategy for this setting is presented in the top panel of Figure 9. However, depending on the objectives of this clinical trial, the two families may be treated as coprimary families, and it may be sufficient to demonstrate a positive effect in either population to obtain a regulatory claim. If the two families are interchangeable, they can be tested simultaneously rather than sequentially as shown in the bottom panel of Figure 9.

A general method for constructing gatekeeping procedures on the basis of the simultaneous testing approach was developed in Kordzakhia and Dmitrienko [57]. Gatekeeping procedures in this class are termed **superchain procedures** to emphasize that they provide a direct extension of nonparametric and parametric chain procedures with flexible α propagation rules. Superchain procedures are based on straightforward multistep testing algorithms and are easy to implement in advanced multiplicity problems with several families of null hypotheses.

To build a superchain procedure for the multiplicity problem in Case study 5 under the assumption of interchangeable families, the sponsor needs to quantify the relative importance of Families 1 and 2 by defining family weights denoted by w_1 and w_2 (these weights are nonnegative values that add up to 1). If it is more desirable to establish treatment benefit in the overall population, a greater weight is assigned to Family 1.

Further, to efficiently utilize available information on positive correlations between the hypothesis test statistics within each family, the sponsor can perform multiplicity adjustments on the basis of powerful Hochberg-type component procedures in Families 1 and 2. The families are tested simultaneously, and each one serves as a parallel gatekeeper for the other family. If at least one null hypothesis is rejected in a family, the error rate from that family is transferred to the other family, and the two families are retested. Therefore, as was explained in Section 7.2, the truncated Hochberg procedures rather than regular Hochberg procedures need to be applied in both families to satisfy the parallel gatekeeping condition. Let $0 \leqslant \gamma_1 < 1$ and $0 \leqslant \gamma_2 < 1$ denote the parameters of the truncated Hochberg procedures used in Families 1 and 2, respectively.

The resulting superchain procedure is based on the following testing algorithm:

- Step 1. The truncated Hochberg procedure is applied in Family 1 at $\alpha_1 = w_1\alpha$ and in Family 2 at $\alpha_2 = w_2\alpha$. If no null hypotheses are rejected or all null hypotheses are rejected, testing stops. If both null hypotheses are rejected in one family but there are nonrejected null hypotheses in the other family, go to Step 3. Otherwise, go to Step 2.
- Step 2. Because there are nonrejected null hypotheses in both families, the families are retested using the truncated Hochberg procedures. Let r_1 and r_2 denote the number of null hypotheses rejected in Families 1 and 2, respectively, in the preceding step. Family 1 is retested using the procedure with the truncation parameter $w_1\gamma_1/\kappa_1 + (1 - w_1/\kappa_1)$ at $\kappa_1\alpha$. Similarly, Family 2 is retested using the procedure with the truncation parameter $w_2\gamma_2/\kappa_2 + (1 - w_2/\kappa_2)$ at $\kappa_2\alpha$. Here,

$$\kappa_1 = w_1 + w_2(1-\gamma_2)\left(1 - \frac{r_2}{2}\right),\ \kappa_2 = w_2 + w_1(1-\gamma_1)\left(1 - \frac{r_1}{2}\right).$$

  If no new null hypotheses are rejected or all null hypotheses are rejected, testing stops. If both null hypotheses are rejected in one family but there are nonrejected null hypotheses in the other family, go to Step 3. Otherwise, repeat Step 2.
- Step 3. The family with nonrejected null hypotheses is retested using the regular Hochberg procedure at the full α level.

It follows from this algorithm that the superchain procedure supports efficient α propagation with an option to transfer the error rate from the family of the overall population tests (Family 1) to the family of subpopulation tests (Family 2) and vice versa. For example, assume that significant findings in the overall population are valued higher than those in the subpopulation and thus the family weights are set to $w_1 = 2/3$ and $w_2 = 1/3$. The initial values of the truncation parameters in the truncated Hochberg

Statistics in Medicine

procedures are $\gamma_1 = 1/2$ and $\gamma_2 = 1/2$. Further, the p-values for the four null hypotheses are given by

$$\text{Family 1: } p_1 = 0.028,\ p_2 = 0.011;\ \text{ Family 2: } p_3 = 0.018,\ p_4 = 0.006.$$

A graphical summary of the decision rules used in the superchain procedure in this multiplicity problem is given in Figure 15. As before, Hochberg(γ) denotes the truncated Hochberg procedure with the truncation parameter γ. First of all, in Step 1, the superchain procedure splits the overall error rate between the families. The two families are tested simultaneously using the truncated Hochberg procedures with the truncation parameters of $\gamma_1 = 1/2$ and $\gamma_2 = 1/2$ at the significance levels given by $w_1\alpha = 0.0333$ and $w_2\alpha = 0.0167$, respectively. The significance levels for the larger p-values in the two families are given by

$$\frac{(1+\gamma_1)w_1\alpha}{2} = \frac{\alpha}{2} = 0.025 \text{ and } \frac{(1+\gamma_2)w_2\alpha}{2} = \frac{\alpha}{4} = 0.0125.$$

The larger p-value in Family 1 exceeds 0.025, and similarly, the larger p-value in Family 2 is greater than 0.0125. Thus, the corresponding null hypotheses of Families 1 and 2 (H_1 and H_3) cannot be rejected in Step 1. Further, the significance levels for the smaller p-values in the two families are

$$\frac{w_1\alpha}{2} = \frac{\alpha}{3} = 0.0167 \text{ and } \frac{w_2\alpha}{2} = \frac{\alpha}{6} = 0.0083.$$

Because the smaller p-values in the two families are both less than the corresponding levels, the superchain procedure rejects the null hypotheses H_2 and H_4 in Step 1 (these hypotheses are represented by closed circles in Figure 15). There are still nonrejected null hypotheses in both families, and the testing algorithm proceeds to Step 2.

In Step 2, the error rate released after the rejection of H_2 in Family 1 is transferred to Family 2, and similarly, the error rate released after the rejection of H_4 in Family 2 is carried over to Family 1. Because exactly one null hypothesis is rejected in each family in Step 1, $r_1 = r_2 = 1$. The overall significance level used in the truncated Hochberg procedure in Family 1 is then defined as a sum of two components. The first one is the level initially allocated to this family, and the second component is the error rate transferred from Family 2 after the null hypothesis H_4 is rejected in Step 1:

$$\kappa_1\alpha = w_1\alpha + w_2\alpha(1-\gamma_2)\left(1-\frac{1}{2}\right) = \frac{3\alpha}{4} = 0.0375.$$

Recall that the overall significance level for Family 1 in Step 1 is $2\alpha/3 = 0.0333$ and thus the null hypotheses in Family 1 are tested at a higher significance level in Step 2 compared with the preceding step because of a positive contribution from Family 2. Similarly, the overall significance level for the truncated Hochberg procedure in Family 2 in Step 2 is

$$\kappa_2\alpha = w_2\alpha + w_1\alpha(1-\gamma_1)\left(1-\frac{1}{2}\right) = \frac{\alpha}{2} = 0.025.$$

This means that the overall significance level for Family 2 increases from $\alpha/3 = 0.0167$ in Step 1 to 0.025 in Step 2.

Furthermore, the component procedures used in Families 1 and 2 are *upgraded* in the sense that their truncation parameters are adjusted upward to account for the rejections made in Step 1. The truncation parameter for the procedure in Family 1 increases from $1/2$ in Step 1 to

$$\frac{w_1\gamma_1}{\kappa_1} + \left(1-\frac{w_1}{\kappa_1}\right) = \frac{5}{9}$$

in Step 2, which results in improved power within Family 1. A power gain also occurs within Family 2 because the truncation parameter increases from $1/2$ in Step 1 to $2/3$ in Step 2. This feature is known as *power pumping*, and when it is combined with efficient α propagation rules, a synergistic effect is created, which improves power across the families of null hypotheses. In this particular case, when the truncated Hochberg procedures are applied in Families 1 and 2 in Step 2, the significance levels for the larger p-values in Families 1 and 2 are

$$\left(1+\frac{5}{9}\right)\frac{3\alpha/4}{2} = 0.0292 \text{ and } \left(1+\frac{2}{3}\right)\frac{\alpha/2}{2} = 0.0208,$$

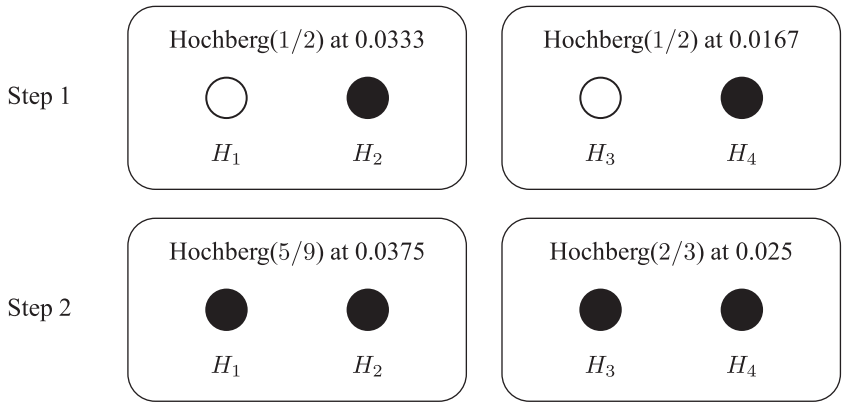

Figure 15. Superchain procedure in Case study 5 (H_1, high dose versus placebo in the overall population; H_2, low dose versus placebo in the overall population; H_3, high dose versus placebo in the subpopulation; H_4, low dose versus placebo in the subpopulation). Closed circle, rejected null hypothesis; Open circle, nonrejected null hypothesis.

respectively. It is easy to see that both p-values are significant at those levels within Family 1 as well as Family 2. This implies that the superchain procedure rejects all four null hypotheses in this multiplicity problem in Step 2 (the rejected null hypotheses are represented by closed circles in Figure 15).

Other examples of powerful and flexible gatekeeping procedures based on simultaneous testing are given in [57].

8. Adjusted p-values and confidence intervals

Multiplicity adjustments in problems with a single family or multiple families of null hypotheses are commonly summarized using adjusted p-values and adjusted confidence intervals (also known as simultaneous confidence intervals). Although clinical trial reports and publications normally focus on adjusted p-values, simultaneous confidence intervals play a potentially more important role because they provide valuable information on the magnitude of the treatment effect and are useful in risk–benefit assessments in clinical trials.

Adjusted p-values are easily computed for closed testing procedures, including all multiple testing procedures defined in this paper. The computational algorithm is based on the definition given by Westfall and Young [16]. An adjusted p-value for a given multiple testing procedure and a given null hypothesis is defined as the lowest overall error rate at which the procedure rejects the hypothesis. Note that an adjusted p-value does not depend on the prespecified error rate. Once adjusted p-values are found for all null hypotheses of interest, each null hypothesis is rejected if its adjusted p-value is no greater than the prespecified α level, for example, two-sided $\alpha = 0.05$.

Computation of simultaneous confidence intervals associated with multiple testing procedures is generally a more difficult task. Although it is easy to set up simultaneous confidence intervals for basic procedures, for exaqmple, the Bonferroni and Dunnett procedures, challenges arise when stepwise procedures are considered, including the step-down and step-up procedures introduced in Sections 4 – 6. A general method for computing simultaneous confidence intervals for a class of nonparametric Bonferroni-based chain procedures was proposed in Guilbaud [58] and Strassburger and Bretz [59]. The case of semiparametric procedures, including the Hochberg and Hommel procedures, was discussed in Guilbaud [60]. Simultaneous confidence intervals for parametric step-down procedures were derived in Stefansson, Kim and Hsu [61], and a new method applicable to nonparametric and parametric procedures in a single-family setting as well as selected gatekeeping procedures in a multifamily setting was introduced in Guilbaud and Karlsson [62].

An important property of stepwise multiple testing procedures is that the more powerful a procedure is, the less meaningful associated simultaneous confidence intervals generally are. To give examples of potential challenges in the interpretation of simultaneous confidence intervals for powerful stepwise procedures, consider a general problem of testing a set of m null hypotheses. The hypotheses are defined in terms of parameters of interest denoted by $\theta_1, \ldots, \theta_m$. The sample estimates, denoted by $\widehat{\theta}_1, \ldots, \widehat{\theta}_m$, are

Statistics in Medicine

assumed to be asymptotical normal with mean θ_i and variance σ_i^2, $i = 1, \ldots, m$. The null hypotheses, defined as

$$H_i : \theta_i < 0, \ i = 1, \ldots, m,$$

are tested versus the alternatives,

$$K_i : \theta_i \geqslant 0, \ i = 1, \ldots, m.$$

A univariate one-sided confidence interval for θ_i with the confidence level $1 - \alpha$ is given by (L_i, ∞) with the lower limit L_i defined as follows:

$$L_i = \widehat{\theta}_i - z_\alpha s_i, \ i = 1, \ldots, m.$$

Here, z_α is the upper $100\alpha\%$ critical point of the standard normal distribution, and s_i is the sample standard error.

A standard requirement for simultaneous confidence intervals associated with a multiple testing procedure is twofold. First, the joint coverage probability should be no less than $1 - \alpha$. Second, the confidence limits must be consistent with the procedure's decision rules. If the procedure rejects a null hypothesis, the lower limit for the corresponding parameter must lie within the alternative space (in this particular setting, the lower limit must be nonnegative). The Bonferroni-adjusted one-sided confidence limits, given by

$$\widetilde{L}_i = \widehat{\theta}_i - z_{\alpha/m} s_i, \ i = 1, \ldots, m,$$

satisfy both conditions. For example, the Bonferroni procedure rejects a null hypothesis, say, H_i, if and only if $p_i \leqslant \alpha/m$, or in other words,

$$1 - \Phi\left(\widehat{\theta}_i / s_i\right) \leqslant \alpha/m.$$

This immediately implies that the one-sided confidence limit $\widetilde{L}_i$ is greater than 0.

The Bonferroni-adjusted confidence limits are easy to compute because the basic Bonferroni procedure tests each null hypothesis independently of the others. With stepwise procedures, the algorithms for computing simultaneous confidence limits are more complex. For example, a two-step algorithm needs to be utilized to derive simultaneous confidence limits associated with the step-down Holm procedure, which is uniformly more powerful than the Bonferroni procedure. In the first step, the Holm procedure is applied to the null hypotheses, and the set of rejected null hypotheses is found. In the second step, the confidence limits are defined depending on the number of hypotheses rejected in the first step. To define the lower confidence limit for θ_i, denoted by $\widetilde{L}_i$, $i = 1, \ldots, m$, consider the following three cases:

- Case 1. If H_i is rejected and some other null hypotheses are accepted, $\widetilde{L}_i = 0$.
- Case 2. If all null hypotheses are rejected, $\widetilde{L}_i = \max\left(0, \widehat{\theta}_i - z_{\alpha/m} s_i\right)$.
- Case 3. If H_i is accepted, $\widetilde{L}_i = \widehat{\theta}_i - z_{\alpha/(m-r)} s_i$, where r is the number of rejected null hypotheses.

It follows from this algorithm that the lower limit tends to be noninformative, for example, the lower limit equals 0, if the Holm procedure rejects the corresponding null hypothesis. In this case, the lower limit may be strictly positive only if the procedure rejects all hypotheses. Thus, even though the Holm procedure is more attractive than the Bonferroni procedure in terms of power for each individual hypothesis, the Holm-adjusted confidence limits are likely to be less meaningful than those produced by the Bonferroni procedure.

Key properties of adjusted p-values and simultaneous confidence intervals are illustrated using the multiplicity problem in Case study 2. The primary endpoint in this dose-finding trial is normally distributed, and a larger value of the endpoint indicates a beneficial effect. The top section of Table I displays the sample mean treatment differences for the three dose–placebo contrasts accompanied by a pooled standard deviation. The table also shows the raw two-sided p-value for each comparison computed from the two-sample t test.

The middle section of Table I presents the adjusted two-sided p-values produced by the Bonferroni and Holm procedures. By using a two-sided $\alpha = 0.05$, it is easy to see that the Bonferroni-adjusted p-value for the null hypothesis H_2 is significant, whereas the other two Bonferroni-adjusted p-values

Table I. Summary statistics, adjusted p-values, and simultaneous lower one-sided 97.5% confidence limits in Case study 2.

	Comparison versus placebo		
	High dose (H_1)	Medium dose (H_2)	Low dose (H_3)
Mean	2.3	2.5	1.9
Pooled standard deviation	9.5	9.5	9.5
Raw p-value	0.0216	0.0125	0.0578
	Adjusted p-values		
Bonferroni	0.0649	0.0376*	0.1733
Holm	0.0433*	0.0376*	0.0578
	Simultaneous confidence limits		
Bonferroni	−0.09	0.10	−0.50
Holm	0	0	−0.06

The asterisk identifies the adjusted p-values that are significant at a two-sided $\alpha = 0.05$.

are not significant. This means that only the medium dose provides a significant improvement over placebo after the Bonferroni adjustment. The Holm procedure detects a significant treatment effect at two doses (high and medium) but not at the low dose. This is to be expected because the Holm procedure is a stepwise procedure that is uniformly more powerful than the basic Bonferroni procedure.

The bottom section of Table I displays the lower limits of one-sided 97.5% simultaneous confidence intervals for the true mean differences. The Bonferroni-adjusted limits are consistent with the decision rules based on the Bonferroni-adjusted p-values. Because the Bonferroni procedure rejects H_2, the associated lower confidence limit is greater than 0. The lower confidence limits for H_1 and H_3 are negative, which implies that there is not enough evidence to reject these null hypotheses. The simultaneous confidence limits produced by the Holm procedure are also formally consistent with the adjusted p-values, but in most cases, they are not informative. Specifically, because the Holm procedure rejects H_1 and H_2, the lower limits for the corresponding true mean differences are set to 0. This means that the simultaneous confidence intervals are given by $(0, \infty)$ and thus cover the entire range of possible values under the alternative hypothesis. These confidence intervals do not provide meaningful information on the likely range of the true treatment effect at the high and medium doses. The advantage of using the more powerful Holm procedure compared with the Bonferroni procedure is clear only in the context of testing the null hypothesis H_3, which is not rejected by either procedure. In this case, the Holm procedure produces a sharper lower limit compared with the Bonferroni procedure.

Because simultaneous confidence intervals for more powerful stepwise procedures tend to be noninformative, trial sponsors can potentially consider using slightly less powerful procedures to obtain more meaningful confidence intervals. For example, the fallback procedure is generally comparable with the Holm procedure in terms of the overall power in multiplicity problems with three or more null hypotheses, and the fallback-adjusted confidence intervals are more informative than the Holm-adjusted confidence intervals.

9. Software implementation

Because of increasing interest in the general field of multiplicity research, a large number of software packages have been developed to implement various multiple testing procedures. Multiplicity adjustments are performed on the basis of adjusted p-values for all procedures, and simultaneous confidence intervals are computed for selected procedures. Additionally, some packages support related tasks, for example, perform sample size calculations in clinical trials with multiple objectives. A brief summary of commercially available and open-source packages is provided in the following:

- SAS software. PROC MULTTEST supports a number of nonparametric and resampling-based procedures. In addition, multiple SAS macros have been developed to implement popular procedures.
- R software. Several R packages are available to support implementation of multiple testing procedures. This includes the *multcomp* package, which supports nonparametric procedures as well as parametric procedures based on linear and other models, and the *multxpert* package,

Statistics in Medicine

which supports procedures for multiplicity problems with a single family and multiple families of null hypotheses (gatekeeping procedures).
- Other software. Excel-based packages have been developed to support popular multiple testing procedures and enable simulation-based sample size calculations based on these procedures.

For more information about software implementation of multiple testing procedures widely used in clinical trials, see http://multxpert.com/wiki/Software.

Glossary

Important terms introduced in this paper are defined as follows:

α allocation rule is used in nonparameteric and parameteric chain procedures. This rule defines the initial distribution of the overall error rate among the null hypotheses. See Section 4 and [22] for more information.

α propagation rule is used in nonparameteric and parameteric chain procedures. This rule defines the process of redistributing the available error rate among the nonrejected null hypotheses after each rejection according to the prespecified logical relationships among the null hypotheses. See Section 4 and [22] for more information.

Bonferroni procedure is a basic nonparametric multiple testing procedure that provides the most conservative multiplicity adjustment. A variety of more powerful stepwise procedures derived from the Bonferroni procedure have been constructed, including the Holm procedure, nonparametric fallback, and chain procedures. See Section 4 and [3, Section 2.6] for more information.

Closure principle (also known as the closed testing principle) is a fundamental principle that defines algorithms for constructing multiple testing procedures. Virtually, all multiple testing methods used in clinical trials are based on the closure principle. See [3, Section 2.3] for more information.

Chain procedures are stepwise multiple testing procedures that support very general α propagation rules. A wide variety of multiple testing procedures are members of the class of chain procedures, including the Holm, nonparametric (Bonferroni-based), and parametric fallback procedures. See Sections 4 and 6 as well as [3, 22, 49] for more information.

Fallback procedures are stepwise multiple testing procedures that enable more straightforward α propagation rules, for example, the error rate can be carried over along the prespecified sequence. Fallback procedures can be constructed within the nonparametric and parametric frameworks and are members of the class of chain procedures. See [3, 18, 19] for more information.

False discovery rate (FDR) is a definition of error rate control used mostly in a preclinical setting. FDR is defined as the expected ratio of the number of rejected true null hypotheses to the number of rejected hypotheses. FDR control does not imply FWER control. See [3, Section 2.2] for more information.

Familywise error rate (FWER) is a definition of error rate control required to be used in confirmatory clinical trials. FWER is defined as the probability of incorrectly rejecting at least one true null hypothesis. See [3, Section 2.2] for more information.

Fixed-sequence procedure is a simple nonparametric procedure that relies on a sequentially rejective algorithm. The fixed-sequence procedure is a special case of the fallback procedures. See Section 4 and [3, Section 2.6] for more information.

Gatekeepers are families of null hypotheses encountered in complex multiplicity problems. See Section 7 and [4] for more information.

Gatekeeping procedures are multiple testing procedures designed for addressing complex multiplicity issues in problems with several families of null hypotheses. See Section 7 and [4] for more information.

Generalized familywise error rate (gFWER) is a definition of error rate control used mostly in a preclinical setting. gFWER is defined as the probability of incorrectly rejecting at least k true null hypotheses, where $k \geqslant 2$ is a prespecified parameter. gFWER control does not imply FWER control. See [3, Section 2.2] for more information.

Hochberg procedure is a semiparametric procedure derived from the global Simes procedure. The Hochberg procedure is based on a straightforward step-up testing algorithm. This procedure is more powerful than the Holm but less powerful than the Hommel procedure. See Section 5 and [3, Section 2.6] for more information.

Holm procedure is a stepwise Bonferroni-based procedure. The Holm procedure relies on a step-down testing algorithm and is less powerful than either the Hochberg or Hommel procedures. See Section 4 and [3, Section 2.6] for more information.

Hommel procedure is a semiparametric procedure derived from the global Simes procedure. The Hommel procedure is based on a step-up testing algorithm and is more powerful than the Holm and Hochberg procedures. See Section 5 and [3, Section 2.6] for more information.

Intersection-union method is a general method used in multiplicity problems where the global null hypothesis of no effect is rejected if all individual null hypotheses are rejected, for example, if all individual objectives are met. See [3, Section 2.3] for more information.

Nonparametric procedures are model-free multiple testing procedures that rely on probabilistic inequalities to establish FWER control and thus do not make any assumptions about the joint distribution of the hypothesis test statistics. All Bonferroni-based procedures are nonparametric. See Section 4 and [3, Section 2.6] for more information.

O'Brien global procedures are examples of global testing procedures that can be used in clinical trials with global win criteria. In clinical trials with multiple endpoints, global testing procedures pool the treatment effects across multiple endpoints. See [14] for more information.

Parallel gatekeeper is a type of gatekeeper encountered in multiplicity problems with several families of null hypotheses. A parallel gatekeeper is passed if one or more null hypotheses are rejected within the gatekeeper.

Parametric procedures are model-based multiple testing procedures that make explicit assumptions about the joint distribution of the hypothesis test statistics in a multiplicity problem. Dunnett-based procedures serve as examples of parametric procedures. See Section 6 and [3, Section 2.7] for more information.

Semiparametric procedures are multiple testing procedures that utilize decision rules that do not explicitly depend on the joint distribution of the hypothesis test statistics, but additional distributional assumptions are required to establish FWER control. Examples include the Hochberg and Hommel procedures. See Section 5 and [3, Section 2.6] for more information.

Serial gatekeeper is a type of gatekeeper encountered in multiplicity problems with several families of null hypotheses. A serial gatekeeper is passed if all null hypotheses are rejected within the gatekeeper.

Šidák procedure is a simple semiparametric procedure. More powerful multiple testing procedures can be derived from the Šidák procedure, for example, the step-down Šidák procedure. See Section 5 and [3, Section 2.6] for more information.

Simes procedure is a global testing procedure. This procedure cannot be used for testing the individual null hypotheses, but powerful stepwise procedures can be derived from the Simes procedure, including the Hochberg and Hommel procedures. See Section 5 and [3, Section 2.6] for more information.

Step-down procedures are stepwise multiple testing procedures that test ordered null hypotheses sequentially beginning with the null hypothesis corresponding to the most significant p-value.

Step-up procedures are stepwise multiple testing procedures that test ordered null hypotheses sequentially beginning with the null hypothesis corresponding to the least significant p-value.

Strong FWER control refers to control of FWER under all possible configurations of true and false null hypotheses or, in other words, control of the maximum probability of incorrectly rejecting at least one true null hypothesis.

Superchain procedures are flexible gatekeeping procedures that rely on simultaneous testing of several families of null hypotheses. See Section 7.4 and [57] for more information.

Truncated procedure is a modified multiple testing procedure, which is defined as a combination of a given procedure and the Bonferroni procedure. Truncated procedures play an important role in the construction of multistep gatekeeping procedures. See Section 7.2 and [4] for more information.

Union–intersection method is a general method used in multiplicity problems where the global null hypothesis of no effect is rejected if at least one individual null hypothesis is rejected, that is, if there is evidence of a positive effect with respect to at least one individual objective. See [3, Section 2.3] for more information.

Weak FWER control refers to control of FWER under a single configuration of true and false null hypotheses, that is, when all null hypotheses are true.

Win criterion is a set of clinical decision rules used in clinical trials with multiple objectives. Examples include at-least-one win criteria (it is sufficient to meet only one objective for the overall outcome of

Statistics in Medicine

a trial to be declared positive), all-or-none win criteria (all objectives must be met to declare success in a trial), and global win criteria (the overall outcome of a trial is determined on the basis of a simultaneous analysis of all objectives). See Section 2 for more information.

Appendix A

The following example shows that weak control of the FWER does not provide adequate protection of the probability of an incorrect conclusion in a clinical trial setting. Consider the multiplicity problem arising in Case study 4. Let H_1 denote the primary null hypothesis and H_2, H_3, and H_4 denote the secondary null hypotheses. The null hypotheses are tested in a sequential manner as follows: the primary endpoint is tested first, and the other three endpoints are tested simultaneously provided the primary objective is met. Each endpoint is tested at the full α.

We first evaluate the error rate for this testing strategy when all null hypotheses are true. Because the secondary null hypotheses are tested if and only if H_1 is rejected, the probability of incorrectly rejecting at least one true null hypothesis in this problem is simply equal to the probability of rejecting H_1. Recall that this null hypothesis is tested at the full α level and thus the error rate does not exceed α. Even though the secondary endpoints are each tested at α, this simple testing strategy guarantees weak FWER control.

However, focusing on only one configuration of true and false null hypotheses can lead to a substantial underestimation of the true error rate. If other configurations are considered, the error rate will be inflated. To see this, assume that the secondary null hypotheses are true and the secondary test statistics are independent. Further, the primary null hypothesis is false with an extremely large effect size, which causes H_1 to be rejected virtually all the time. In this case, the probability of incorrectly rejecting at least one true null hypothesis will be very close to

$$P(\text{Reject } H_2 \text{ or } H_3 \text{ or } H_4).$$

Because each null hypothesis is tested at the full α level and the test statistics are independent, this probability is equal to

$$1 - P(\text{Reject } H_2)\,P(\text{Reject } H_3)\,P(\text{Reject } H_4) = 1 - (1-\alpha)^3.$$

The probability reaches 0.14 if $\alpha = 0.05$. This means that although the simple testing strategy controls the FWER in the weak sense, it clearly fails to control the error rate in the strong sense.

Acknowledgements

The authors would like to thank Drs George Kordzakhia, Ilya Lipkovich, and Guochen Song for helpful comments.

Disclaimer

The views and opinions presented in this paper are solely those of the authors and should not be taken to represent policies or regulations associated with the U.S. Food and Drug Administration in any meaningful way.

References

1. Committee for Proprietary Medicinal Products. Points to consider on multiplicity issues in clinical trials, 2002.
2. Hochberg Y, Tamhane AC. *Multiple Comparison Procedures*. Wiley: New York, 1987.
3. Dmitrienko A, Bretz F, Westfall PH, Troendle J, Wiens BL, Tamhane AC, Hsu JC. Multiple testing methodology. In *Multiple Testing Problems in Pharmaceutical Statistics*, Dmitrienko A, Tamhane AC, Bretz F (eds). Chapman and Hall/CRC Press: New York, 2009.
4. Dmitrienko A, Tamhane AC. Gatekeeping procedures in clinical trials. In *Multiple Testing Problems in Pharmaceutical Statistics*, Dmitrienko A, Tamhane AC, Bretz F (eds). Chapman and Hall/CRC Press: New York, 2009.
5. D'Agostino R, Massaro J, Kwan H, Cabral H. Strategies for dealing with multiple treatment comparisons in confirmatory clinical trials. *Drug Information Journal* 1993; **27**:625–641.
6. Sankoh AJ, Huque MF, Dubey SD. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statistics Medicine* 1997; **16**:2529–2542.
7. Sankoh AJ, D'Agostino RB, Huque MF. Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. *Statistics in Medicine* 2003; **22**:3133–3150.

8. Wang S, O'Neill R, Hung H. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics* 2007; **6**:227–244.
9. Millen B, Dmitrienko A, Ruberg S, Shen L. A statistical framework for decision making in confirmatory multipopulation tailoring clinical trials. *Drug Information Journal* 2012. DOI: 10.1177/0092861512454116. In press.
10. Offen W, Chuang-Stein C, Dmitrienko A, Littman G, Maca J, Meyerson L, Muirhead R, Stryszak P, Boddy A, Chen K, Copley-Merriman K, Dere W, Givens S, Hall D, Henry D, Jackson JD, Krishen A, Liu T, Ryder S, Sankoh AJ, Wang J, Yeh CH. Multiple co-primary endpoints: medical and statistical solutions. A report from the Multiple Endpoints Expert Team of the Pharmaceutical Research and Manufacturers of America. *Drug Information Journal* 2007; **41**:31–46.
11. Chuang-Stein C, Dmitrienko A, Offen W. Discussion of "Some controversial multiple testing problems in regulatory applications" by H.M. James Hung and Sue-Jane Wang. *Journal of Biopharmaceutical Statistics* 2009; **19**:14–21.
12. 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.
13. Sankoh AJ, Huque MF, Russell HK, D'Agostino RD. Global two groups multiple endpoint adjustment methods applied to clinical trials. *Drug Information Journal* 1999; **33**:119–140.
14. Tamhane AC, Dmitrienko A. Analysis of multiple endpoints in clinical trials. In *Multiple Testing Problems in Pharmaceutical Statistics*, Dmitrienko A, Tamhane AC, Bretz F (eds). Chapman and Hall/CRC Press: New York, 2009.
15. Mehrotra DV, Heyse JF. Use of false discovery rate for evaluating clinical safety data. *Statistical Methods in Medical Research* 2004; **13**:227–238.
16. Westfall PH, Young SS. *Resampling-based Multiple Testing: Examples and Methods for* P*-Value Adjustment*. John Wiley: New York, 1993.
17. Burman CF, Sonesson C, Guilbaud O. A recycling framework for the construction of Bonferroni-based multiple tests. *Statistics in Medicine* 2009; **28**:739–761.
18. Wiens B. A fixed-sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics* 2003; **2**:211–215.
19. Wiens B, Dmitrienko A. The fallback procedure for evaluating a single family of hypotheses. *Journal of Biopharmaceutical Statistics* 2005; **15**:929–942.
20. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979; **6**:65–70.
21. Marcus R, Peritz E, Gabriel KR. On closed testing procedure with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
22. Millen B, Dmitrienko A. Chain procedures: a class of flexible closed testing procedures with clinical trial applications. *Statistics in Biopharmaceutical Research* 2011; **3**:14–30.
23. Goeman JJ, Solari A. The sequential rejection principle of familywise error control. *The Annals of Statistics* 2010; **38**:3782–3810.
24. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **63**:655–660.
25. Šidák Z. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* 1967; **62**:626–633.
26. Holland BS, Copenhaver MD. An improved sequentially rejective Bonferroni test procedure. *Biometrics* 1987; **43**: 417–423. Correction in *Biometrics* **43**:737.
27. Hochberg Y. A sharper Bonferroni procedure for multiple significance testing. *Biometrika* 1988; **75**:800–802.
28. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 1988; **75**:383–386.
29. Brechenmacher T, Xu J, Dmitrienko A, Tamhane AC. A mixture gatekeeping procedure based on the Hommel test for clinical trial applications. *Journal of Biopharmaceutical Statistics* 2011; **21**:748–767.
30. Sarkar S, Chang CK. Simes' method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* 1997; **92**:1601–1608.
31. Sarkar SK. Some probability inequalities for censored MTP2 random variables: a proof of the Simes conjecture. *The Annals of Statistics* 1998; **26**:494–504.
32. Tamhane AC, Liu L. On weighted Hochberg procedures. *Biometrika* 2008; **95**:279–294.
33. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* 1955; **50**:1096–1121.
34. Dunnett CW, Tamhane AC. Step-down multiple tests for comparing treatments with acontrol in unbalanced one-way layouts. *Statistics in Medicine* 1991; **10**:939–947.
35. Dunnett CW, Tamhane AC. A step-up multiple test procedure. *Journal of the American Statistical Association* 1992; **87**:162–170.
36. Hothorn T, Bretz F, Westfall P. Simultaneous inference in general parametric models. *Biometrical Journal* 2008; **50**:346–363.
37. Voss DT, Hsu JC. Multiple comparisons for an unbalanced $a \times b$ design under mixed models with interaction. *Journal of Statistical Planning and Inference* 1998; **67**:297–309.
38. 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.
39. Zhao YD, Dmitrienko A, Tamura R. On optimal designs of clinical trials with a sensitive subgroup. *Statistics in Biopharmaceutical Research* 2010; **2**:72–83.
40. Li D, Mehrotra DV. An efficient method for accommodating potentially underpowered primary endpoints. *Statistics in Medicine* 2008; **27**:5377–5391.
41. Alosh M, Huque M. A flexible strategy for testing subgroups and overall population. *Statistics in Medicine* 2009; **28**:3–23.
42. Alosh M, Huque MF. A consistency-adjusted alpha-adaptive strategy for sequential testing. *Statistics in Medicine* 2010; **29**:1559–1571.

43. D'Agostino R, Heeren TC. Multiple comparisons in over-the-counter drug clinical trials with both positive and placebo controls. *Statistics in Medicine* 1991; **10**:1–6.
44. 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.
45. D'Agostino RB. Controlling alpha in clinical trials: the case for secondary endpoints. *Statistics in Medicine* 2000; **19**:763–766.
46. Huque MF, Röhmel J. Multiplicity problems in clinical trials, a regulatory perspective. In *Multiple Testing Problems in Pharmaceutical Statistics*, Dmitrienko A, Tamhane AC, Bretz F (eds). Chapman and Hall/CRC Press: New York, 2009.
47. Hung J, Wang SJ. Challenges to multiple testing in clinical trials. *Biometrical Journal* 2010; **52**:747–756.
48. Bauer P, Röhmel J, Maurer W, Hothorn L. Testing strategies in multi-dose experiments including active control. *Statistics in Medicine* 1998; **17**:2133–2146.
49. Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* 2009; **28**:586–604.
50. Dmitrienko A, Tamhane AC. Mixtures of multiple testing procedures for gatekeeping applications in clinical trials. *Statistics in Medicine* 2011; **30**:1473–1488.
51. Hung J, Wang SJ. Some controversial multiple testing problems in regulatory applications. *Journal of Biopharmaceutical Statistics* 2009; **19**:1–11.
52. Dmitrienko A, Tamhane AC, Wiens B. General multistage gatekeeping procedures. *Biometrical Journal* 2008; **50**:667–677.
53. Dmitrienko A, Kordzakhia G, Tamhane AC. Multistage and mixture gatekeeping procedures in clinical trials. *Journal of Biopharmaceutical Statistics* 2011; **53**:726–747.
54. Dmitrienko A, Millen B, Brechenmacher T, Paux G. Development of gatekeeping strategies in confirmatory clinical trials. *Biometrical Journal* 2011; **53**:875–893.
55. Meltzer HY, Cucchiaro J, Silva R, Ogasa M, Phillips D, Xu J, Kalali AH, Schweizer E, Pikalov A, Loebel A. Lurasidone in the treatment of schizophrenia: a randomized, double-blind, placebo- and olanzapine-controlled study. *American Journal of Psychiatry* 2011; **168**:957–967.
56. Dmitrienko A, Soulakova JN, Millen B. Three methods for constructing parallel gatekeeping procedures in clinical trials. *Journal of Biopharmaceutical Statistics* 2011; **53**:768–786.
57. Kordzakhia G, Dmitrienko A. Superchain procedures in clinical trials with multiple objectives. *Statistics in Medicine* 2012. DOI: 10.1002/sim.5537. In press.
58. Guilbaud O. Simultaneous confidence regions corresponding to Holm's stepdown procedure and other closed-testing procedures. *Biometrical Journal* 2008; **50**:678–692.
59. 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.
60. Guilbaud O. Simultaneous confidence regions for closed tests, including Holm, Hochberg and Hommel related procedures. *Biometrical Journal* 2012; **54**:317–342.
61. Stefansson G, Kim WC, Hsu JC. On confidence sets in multiple comparisons. In *Statistical, Decision Theory and Related Topics IV*, Gupta SS, Berger JO (eds). Academic Press: New York, 1988; 89–104.
62. 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.