Chapter 5

Gatekeeping Procedures in Clinical Trials

Alex Dmitrienko
Eli Lilly and Company

Ajit C. Tamhane
Northwestern University

5.1 Introduction

Multiple objectives pursued in clinical trials typically exhibit a hierarchical structure; e.g., they can be divided into primary and secondary objectives (for a detailed classification of primary and secondary endpoints, see D'Agostino, 2000). The primary objective is typically formulated in terms of the primary analysis which describes the most important features of the treatment. In most registration trials, the primary analysis determines the overall outcome of the trial, provides the basis for the regulatory claim and is included in the product label. Secondary analyses (including secondary endpoints and subgroup analyses) play a supportive role and provide additional information for prescribing physicians, patients, payers, etc.

Two approaches to the control of the false positive rate for the family of primary and secondary analyses can be considered in a clinical trial setting:

- Approach 1. The false positive rate is not protected. Regulatory agencies do not always require a strict control of the false positive rate. An example is the case of a small number of secondary analyses or secondary analyses that are highly correlated with the primary analysis.
- Approach 2. The familywise error rate (FWER) associated with the primary and secondary analyses is controlled (FWER is defined in Section 2.2.1).

Although multiplicity adjustments are not mandatory in registration studies to justify the inclusion of secondary endpoints or analyses in the product label, control of the Type I error probability (Approach 2) is becoming increasingly important. This approach is used by regulatory agencies to define the

acceptable statistical risk of false efficacy claims in registration trials. Gatekeeping methods described in this chapter offer a solution to this multiplicity problem. These methods enable the trial's sponsor to

- Control the risk of spurious conclusions (e.g., false efficacy claims) with respect to multiple ordered analyses.
- Take into account the hierarchical structure of the multiple testing problem and examine ordered analyses in a sequential manner beginning with the primary analyses. The gatekeeping methodology is consistent with a regulatory view that findings with respect to secondary/supportive objectives; e.g., secondary endpoints, are meaningful only if the primary objective is met (O'Neill, 1997).

For more information about the use of gatekeeping procedures in a clinical trial setting and literature review in this area of multiple comparison research, see Dmitrienko et al. (2005, Chapter 2) and Dmitrienko and Tamhane (2007).

This chapter begins with motivating examples and a review of gatekeeping procedures in Section 5.2. The next three sections provide a detailed description of three classes of gatekeeping procedures: serial (Section 5.3), parallel (Section 5.4) and tree-structured (Section 5.5). Each section includes a discussion of relevant methodology and clinical trial examples. The last section (Section 5.6) describes available software tools for implementing gatekeeping procedures in clinical trials.

5.2 Motivating examples

To construct a gatekeeping procedure, one first needs to define two or more families of analyses, for example, a family of primary endpoints and a family of secondary endpoints. Each family (except for the last one) serves as a gatekeeper in the sense that one must pass it to perform analyses in the next family.

In this section we will present clinical trial examples that motivate the use of gatekeeping methods in clinical trials and also set the stage for the review of main classes of gatekeeping procedures in Sections 5.3–5.5.

As a side note, gatekeeping procedures discussed in this section focus on multiplicity adjustments in a single trial. In the context of registration/marketing authorization packages that normally include two confirmatory trials with similar sets of primary and secondary analyses, gatekeeping methods can be applied independently to each trial. This approach will ensure Type I error rate control within each confirmatory trial. To justify the inclusion of secondary findings into the product label, the trial's sponsor can use consistency arguments and demonstrate that multiplicity-adjusted primary

and secondary analyses lead to similar conclusions in both trials. It is worth noting that regulatory guidelines do not currently discuss rules for combining secondary findings across several confirmatory trials.

5.2.1 Clinical trials with serial gatekeepers

We will begin with a two-family testing problem arising in clinical trials with noninferiority and superiority objectives (this example is based on Dmitrienko and Tamhane, 2007). Consider a trial in patients with Type II diabetes with three treatment groups, Group A (a new formulation of an insulin therapy), Group B (a standard formulation) and Group A+B (a combination of the formulations). The following two scenarios will be examined:

- Scenario 1. Noninferiority and superiority tests are carried out sequentially for the comparison of A versus B.
- Scenario 2. A noninferiority test for the comparison of A versus B is carried out first followed by a superiority test for the same comparison and a noninferiority test for the comparison of A+B versus B.

Let δ_1 and δ_2 denote the true treatment differences for the comparisons of A versus B and A+B versus B, respectively. The three sets of null and alternative hypotheses arising in this problem are defined as follows:

- A versus B (noninferiority), $H_1: \delta_1 \leq -\gamma_1$ versus $K_1: \delta_1 > -\gamma_1$, where γ_1 is a positive non-inferiority margin for the comparison of A versus B.
- A versus B (superiority), $H_2: \delta_1 \leq 0$ versus $K_2: \delta_1 > 0$.
- A+B versus B (noninferiority), $H_3: \delta_2 \leq -\gamma_2$ versus $K_3: \delta_2 > -\gamma_2$, where γ_2 is a positive non-inferiority margin for the comparison of A+B versus B.

The testing procedures used in the two scenarios are depicted in Figure 5.1. In both scenarios, testing begins with the first family that includes the test for H_1. This family serves as a serial gatekeeper in the sense that all hypotheses of no treatment effect must be rejected in the first family to proceed to the second family (in this case there is only one hypothesis in the first family). In Scenario 1, the second family includes the test for H_2; in Scenario 2, this family includes the tests for H_2 and H_3.

It is important to note that, even though two tests are performed in Scenario 1, no multiplicity adjustment is needed. Both tests can be carried out at the same α level, where α is the pre-specified FWER, e.g., $\alpha = 0.05$. This is due to the fact that this testing procedure is a special case of the fixed-sequence approach described in Section 2.6.3.

It is instructive to compare the straightforward multiple testing problem in Scenario 1 to the more complex one in Scenario 2. Although the two settings

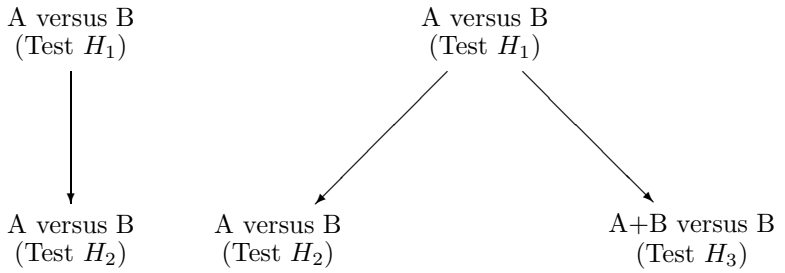

FIGURE 5.1: Decision trees in the combination-therapy clinical trial example (Scenario 1, left panel; Scenario 2, right panel).

look quite similar at first glance, one can no longer avoid multiplicity adjustments in Scenario 2. To see this, suppose that no multiplicity adjustment is performed; i.e., all three tests are carried out at the α level and assume that $\delta_1 = 0$ (the new formulation is equivalent to the standard formulation) and $\delta_2 \leq -\gamma_2$ (the combination is inferior to the standard formulation). Under this set of assumptions, if the noninferiority margin for the comparison of A versus B is very wide (γ_1 is large), one is virtually assured of observing a significant outcome for the first test and passing the gatekeeper. This means that the original multiple testing problem will simplify to the problem of testing H_2 and H_3. The two tests are carried at the unadjusted α level and thus the probability of at least one incorrect conclusion will be inflated. This example illustrates that multiplicity adjustments are needed in general for multi-family testing problems.

The multiple testing problem considered in this section is a two-family problem in which the first family serves as a serial gatekeeper. Serial gatekeepers are often found in clinical trials with multiple ordered endpoints. For example, in a clinical trial with a single primary endpoint and several key secondary endpoints, the endpoints may be arranged in a sequence. In this case, each endpoint defines a family and serves as a serial gatekeeper for the next family in the sequence. Multiple testing procedures that control the FWER in problems with serial gatekeepers are discussed in Section 5.3.

5.2.2 Clinical trials with parallel gatekeepers

To introduce parallel gatekeepers, consider an osteoporosis/breast cancer clinical trial in postmenopausal women that investigates the efficacy of a novel treatment compared to a placebo control (this example is based on Cummings et al., 1999; Ettinger et al., 1999). The treatment effect is evaluated using two primary endpoints, incidence of vertebral fractures (Endpoint 1) and incidence of breast cancer (Endpoint 2) and an important secondary endpoint, incidence

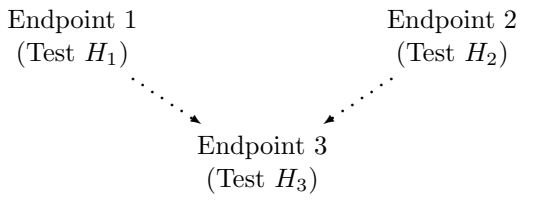

FIGURE 5.2: Decision tree in the osteoporosis/breast cancer clinical trial example. Dotted lines are used to indicate that only one test in the first family needs to be significant to perform the test in the second family.

of non-vertebral fractures (Endpoint 3). Let δ_i denote the true treatment difference for the ith endpoint. The associated null hypothesis, $H_i: \ \delta_i \leq 0$, is tested against a superiority hypothesis, $K_i: \ \delta_i > 0$, $i = 1, 2, 3$.

The first family includes the primary tests (tests for H_1 and H_2) and the second family includes the test for H_3. Each primary endpoint is associated with an independent regulatory claim and the trial will be declared positive if there is evidence of a beneficial treatment effect for at least one primary endpoint. Using mathematical terminology, the first family serves as a parallel gatekeeper; i.e., at least one hypothesis needs to be rejected in this family to pass the gatekeeper and carry out the test in the second family. The testing procedure is displayed in Figure 5.2.

As in the two-family problem described in Section 5.2.1, it is easy to show that an appropriate multiplicity adjustment strategy is required in this case to preserve the FWER. A naive strategy can be set up as follows:

- Since there are two tests in the first family, a multiple test is used to control the Type I error rate within this family, e.g., Bonferroni test (H_1 and H_2 are tested at the $\alpha/2$ level).

- If one or more tests in the first family are significant, H_3 is tested at the α level (after all, there is only one test in the second family).

To verify whether this approach prevents Type I error rate inflation, we can compute the probability of at least one erroneous conclusion when δ_1 is very large but $\delta_2 = 0$ and $\delta_3 = 0$.

Since the treatment is superior to placebo with a large margin for Endpoint 1, the testing procedure is virtually guaranteed to pass the gatekeeper and the test for H_3 will be carried out almost all of the time. As a result, the three-endpoint problem collapses to a two-endpoint problem in which H_2 is tested at the $\alpha/2$ level and H_3 is tested at the α level. It is clear that the probability of at least one incorrect conclusion will be greater than α (unless the test statistics associated with H_2 and H_3 are perfectly correlated). In other words, even though the naive strategy protects the Type I error rate within each family, the overall Type I error rate ends up being inflated. This clinical trial

example shows that a more sophisticated multiple comparison procedure (that goes beyond Type I error rate control within each family) may be required in trials with hierarchically ordered analyses.

In general, parallel gatekeepers could be utilized in clinical trials with several primary endpoints where each endpoint defines a successful trial outcome; e.g., each endpoint is associated with its own regulatory claim. In addition, parallel gatekeepers could be used in trials with multiple doses of a treatment tested against a control, e.g., a placebo or active control. In this case, the dose-control hypotheses corresponding to higher dose levels could be included in the first family that serves as a parallel gatekeeper for the family containing the other dose-control hypotheses. Section 5.4 introduces a general class of procedures that control the FWER in trials with parallel gatekeepers.

5.2.3 Clinical trials with tree-structured gatekeepers

Tree-structured gatekeeping procedures (or, simply, tree gatekeeping procedures) are used in clinical trials with multiple analyses that form a complex hierarchical structure. This includes structures with logical relationships among the analyses that go beyond more basic hierarchical structures associated with serial and parallel gatekeeping methods.

To illustrate, we will consider an extension of the combination-therapy clinical trial example given in Section 5.2.1. In this example, there are six tests that are carried out in four stages as shown below:

- Stage 1. A versus B (test of the noninferiority hypothesis H_1).
- Stage 2. A versus B (test of the superiority hypothesis H_2) and A+B versus B (test of the noninferiority hypothesis H_3).
- Stage 3. A+B versus B (test of the superiority hypothesis H_4) and A+B versus A (test of the noninferiority hypothesis H_5).
- Stage 4. A+B versus A (test of the superiority hypothesis H_6).

A decision tree associated with this testing strategy is displayed in Figure 5.3. The tree exhibits fairly complex logical relationships among the tests. In the parallel gatekeeping example in Section 5.2.2, the secondary test (test for H_3) was logically related to both primary tests (tests for H_1 and H_2). In this case, each test at Stages 3 and 4 is logically related to only one test carried out at the previous stage. For example, H_4 will be tested if and only if the test for H_3 is significant and the outcome of the test for H_2 is ignored. Gatekeeping procedures for problems with logical restrictions of this kind are known as tree gatekeeping procedures. Using a counterexample similar to the one given in Section 5.2.2, it is easy to show that Type I error rate control within each family does not, in general, guarantee control of the FWER in problems with logical restrictions.

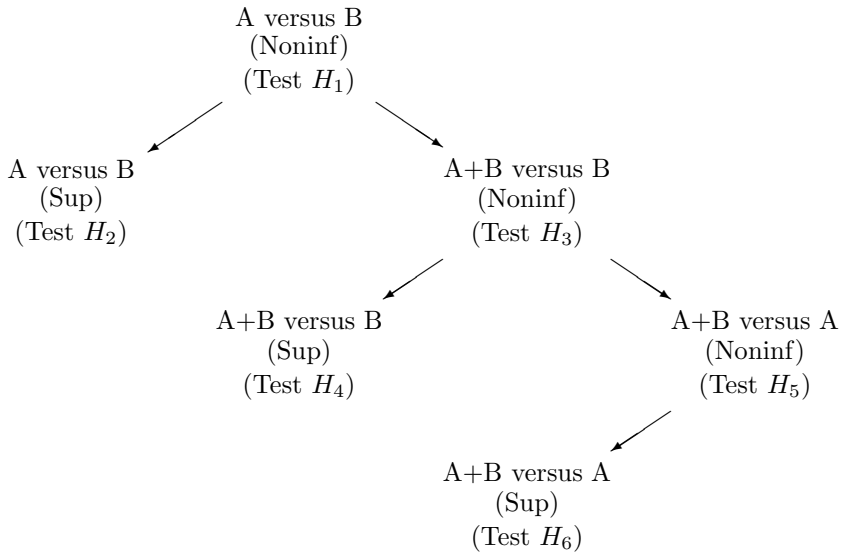

FIGURE 5.3: Decision tree in the combination-therapy clinical trial example (Noninf, Noninferiority; Sup, Superiority).

Tree gatekeepers arise in clinical trials with multiple objectives, e.g., multiple endpoints or multiple subgroups, when logical dependencies exist among the null hypotheses associated with these objectives. It is shown in Section 5.5 how to construct tree gatekeeping procedures that take these logical relationships into account and protect the FWER.

5.3 Serial gatekeeping procedures

Sections 5.3–5.5 give a comprehensive review of the three classes of gatekeeping procedures, including underlying theory, clinical trial examples and implementation details.

The following notation will be used in the three sections. Consider a clinical trial with multiple, hierarchically ordered objectives/analyses. To account for the hierarchical ordering, the analyses are grouped into m families denoted by $F_1, \ldots, F_m$. Each family includes null hypotheses corresponding to the analyses at the same level in the hierarchy; e.g., the hypotheses in F_1 may be related to a set of primary analyses and the hypotheses in F_2 may represent

TABLE 5.1: Families of null hypotheses corresponding to multiple, hierarchically ordered objectives.

Family	Null hypotheses	Hypothesis weights	Raw p-values
F_1	$H_{11}, \ldots, H_{1n_1}$	$w_{11}, \ldots, w_{1n_1}$	$p_{11}, \ldots, p_{1n_1}$
...	...	...	...
F_i	$H_{i1}, \ldots, H_{in_i}$	$w_{i1}, \ldots, w_{in_i}$	$p_{i1}, \ldots, p_{in_i}$
...	...	...	...
F_m	$H_{m1}, \ldots, H_{mn_m}$	$w_{m1}, \ldots, w_{mn_m}$	$p_{m1}, \ldots, p_{mn_m}$

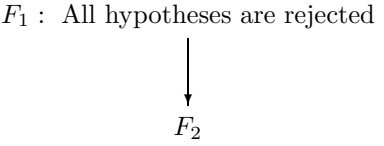

FIGURE 5.4: A problem with a serial gatekeeper (F_1 is a serial gatekeeper for F_2).

secondary analyses. As was stated in Section 5.2, each family (except for the last one) serves as a gatekeeper for the families placed later in the sequence.

The hypotheses included in the m families are shown in Table 5.1. Further, $w_{i1}, \ldots, w_{in_i}$ are the weights representing the importance of hypotheses within F_i (the weights are non-negative and $w_{i1} + \ldots + w_{in_i} = 1$) and $p_{i1}, \ldots, p_{in_i}$ are the associated raw p-values. Multiplicity adjusted p-values for the hypotheses in F_i are denoted by $\widetilde{p}_{i1}, \ldots, \widetilde{p}_{in_i}$ (note that the adjusted p-values are defined with respect to all m families rather than any individual family).

5.3.1 General serial gatekeeping framework

A family is termed a *serial gatekeeper* if all hypotheses must be rejected within that family in order to proceed to the next family in the sequence (see Figure 5.4). In other words, if F_i, $i = 1, \ldots, m-1$, is a serial gatekeeper, hypotheses in F_{i+1} are tested if and only if

$$\max_{j=1,\ldots,n_i} \widetilde{p}_{ij} \leq \alpha.$$

A clinical trial example with a serial gatekeeper was given in Section 5.2.1.

Serial gatekeeping procedures were studied by Maurer, Hothorn and Lehmacher (1995), Bauer et al. (1998) and Westfall and Krishen (2001). Most commonly, serial gatekeepers are encountered in trials where endpoints can be ordered from most important to least important:

- The adalimumab trial in patients with rheumatoid arthritis (Keystone et al., 2004) tested the effect of adalimumab on three endpoints that

were ordered and examined sequentially: symptomatic response, disease progression and physical function.

- Hierarchical arrangements of endpoints are often used in oncology trials, e.g., overall survival duration, progression-free survival duration, tumor response rate, time to treatment failure and duration of tumor response.

Serial gatekeeping procedures are widely used in clinical trials, mainly due to the fact that they do not require an adjustment for multiplicity. Note that serial gatekeeping procedures are closely related to the fixed-sequence test introduced in Section 2.6.3 (in fact, these procedures simplify to the fixed-sequence test if each family includes a single hypothesis). This approach to testing ordered endpoints is described in the CPMP guidance document on multiplicity issues in clinical trials (CPMP, 2002).

5.3.2 Serial gatekeeping procedures with a single decision-making branch

In their most basic form, serial gatekeeping procedures can be applied to problems in which multiple analyses define a single sequence of hypotheses. We refer to these serial gatekeeping procedures as single-branch procedures.

A single-branch procedure for multiple families of analysis is defined as follows. Within each family F_i, $i = 1, \ldots, m-1$, hypotheses are tested at the nominal α level. For example, the hypotheses in F_i can be tested using an intersection-union (IU) test (Section 2.3.2); i.e., all hypotheses are rejected in F_i if $p_{ij} \leq \alpha$, $j = 1, \ldots, n_i$, and all hypotheses are retained otherwise. Any FWER-controlling test can be used in F_m, including all popular multiple tests described in Sections 2.6–2.8.

Multiplicity adjustments are commonly summarized using adjusted p-values for hypotheses of interest. Adjusted p-values for single-branch procedures are easy to compute using the Westfall-Young definition discussed in Section 2.4.1. Assume that the IU test is used in $F_1, \ldots, F_{m-1}$. Let p_i^* denote the largest p-value in F_i, $i = 1, \ldots, m-1$, and p'_{mj} denote the adjusted p-value for H_{mj} produced by the test used in the last family, $j = 1, \ldots, n_m$. The adjusted p-value for H_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, is given by:

$$\widetilde{p}_{ij} = \begin{cases} \max(p_1^*, \ldots, p_i^*) & \text{if } i = 1, \ldots, m-1, \\ \max(p'_{ij}, p_1^*, \ldots, p_{i-1}^*) & \text{if } i = m. \end{cases}$$

Alzheimer's disease clinical trial example

The Alzheimer's disease clinical trial example from Dmitrienko and Tamhane (2007) serves as an example of a single-branch problem with a serial gatekeeper. In this example, the efficacy profile of an experimental treatment is compared to that of a placebo using four endpoints:

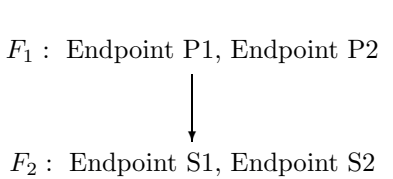

FIGURE 5.5: Single-branch serial gatekeeping procedure in the Alzheimer's disease clinical trial example.

TABLE 5.2: Serial gatekeeping procedure in the Alzheimer's disease clinical trial example. The asterisk identifies the adjusted p-values that are significant at the two-sided 0.05 level.

Family	Endpoint	Raw p-value	Adjusted p-value
F_1	P1	0.023	0.023*
F_1	P2	0.018	0.023*
F_2	S1	0.014	0.028*
F_2	S2	0.106	0.106

- The primary endpoints include a cognitive impairment endpoint, ADAS-Cog (Endpoint P1), and a clinical global performance endpoint, CIBIC (Endpoint P2).
- Two secondary endpoints are also examined in this trial, a biochemical endpoint (Endpoint S1) and an imaging endpoint (Endpoint S2).

The hypotheses for the primary and secondary endpoints are included in F_1 and F_2, respectively. Since a trial for the treatment of Alzheimer's disease is normally declared successful only if both primary endpoints are significant (Reisberg et al., 2003; Reines et al., 2004), F_1 serves as a serial gatekeeper (see Figure 5.5).

To illustrate the implementation of the serial gatekeeping procedure, Table 5.2 displays the two-sided raw p-values produced by the four tests in this clinical trial example as well as adjusted p-values. The hypotheses in F_1 are tested using the IU test and both of them are rejected at the 0.05 level. Because of this, the procedure can pass the gatekeeper and test the hypotheses in F_2. The Holm test is carried out in F_2 and the adjusted p-values for Endpoints S1 and S2 are given by 0.028 and 0.106, respectively. Endpoint S1 is significant at the 0.05 level, whereas Endpoint S2 is not. Since the serial gatekeeping procedure controls the FWER, the trial's sponsor can use these results to justify the inclusion of the two primary endpoints as well as one secondary endpoint (Endpoint S1) in the product label.

Serial gatekeeping procedures have a simple structure and are quite appealing in clinical trial applications. However, it is important to bear in mind that these procedures are based on the fixed-sequence approach and thus they

need to be considered only if there is sufficient clinical justification that can be used to prioritize the objectives of interest (pros and cons of the fixed-sequence approach are discussed in Section 4.3.1).

5.3.3 Serial gatekeeping procedures with multiple decision-making branches

In the previous section we considered a class of basic single-branch procedures. More complicated examples of serial gatekeeping procedures arise in clinical trials with multiple sequences of hypotheses or multiple decision-making branches, e.g., dose-finding studies with ordered endpoints. In this case, at each fixed dose level, dose-control comparisons for multiple endpoints form a branch within which hypotheses are tested sequentially.

Serial gatekeeping procedures with multiple branches can be constructed based on several multiple tests. Here we will focus on Bonferroni-based procedures (serial gatekeeping procedures based on other tests are briefly discussed in Section 5.3.4). Consider a multiple testing problem with m families and assume that each one contains n hypotheses, i.e., $n_1 = \ldots = n_m = n$. In this case there are n branches and the jth branch includes the hypotheses $H_{1j}, \ldots, H_{mj}$. Hypotheses within each branch are tested sequentially as follows:

- Consider the jth branch, $j = 1, \ldots, n$. The hypothesis H_{1j} is tested first at an α/n level. If H_{1j} is rejected, the next hypothesis in the sequence, i.e., H_{2j}, is tested, otherwise testing within this branch stops.

- In general, the hypothesis H_{ij} is rejected if $p_{kj} \leq \alpha/n$ for all $k = 1, \ldots, i$.

FWER control for the Bonferroni-based procedure is discussed in Quan, Luo and Capizzi (2005). Adjusted p-values for serial gatekeeping procedures with multiple branches can be found using the direct calculation algorithm defined in Section 5.4.2.

Type II diabetes clinical trial example

A multiple testing problem with three branches was described in Dmitrienko et al. (2006a) and Dmitrienko et al. (2007). The Type II diabetes clinical trial considered in these papers is conducted to compare three doses of an experimental treatment (labeled L, M and H) versus placebo (labeled P). Each dose-placebo test is carried out with respect to three ordered endpoints: hemoglobin A1c (Endpoint E1), fasting serum glucose (Endpoint E2) and HDL cholesterol (Endpoint E3). The E2 tests are restricted to the doses at which Endpoint E1 is significant and, similarly, the E3 tests are carried out only for the doses at which the E1 and E2 tests are both significant. Logical restrictions of this kind facilitate drug labeling and, in addition, improve the power of clinically relevant secondary dose-placebo tests. The resulting

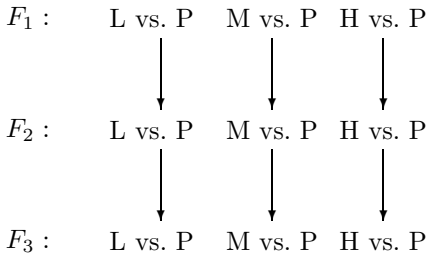

FIGURE 5.6: Three-branch serial gatekeeping procedure with three families of hypotheses in the Type II diabetes clinical trial example (F_1, Endpoint E1; F_2, Endpoint E2; F_3, Endpoint E3).

decision tree has three branches (see Figure 5.6) and the fixed-sequence approach is applied within each branch. The branches are "connected" using the Bonferroni test as described below.

To define the three-branch procedure, the hypotheses H_{i1} (H-P comparison), H_{i2} (M-P comparison) and H_{i3} (L-P comparison) for the ith endpoint are included in F_i, $i = 1, 2, 3$. The hypotheses are equally weighted within each family and the FWER is set at a two-sided $\alpha = 0.05$. The hypotheses within the three branches are tested sequentially using the Bonferroni-based procedure.

The two-sided raw and adjusted p-values in this clinical trial example are summarized in Table 5.3. The adjusted p-values are computed using the direct-calculation algorithm with $K = 100,000$. Note that only Doses M and H are significantly different from placebo for the primary endpoint (Endpoint E1) and thus the remaining branch corresponding to the L-P comparison is eliminated at the first stage of the procedure. At the second stage, the dose-placebo comparisons for Endpoint E2 are performed only for the dose levels at which Endpoint E1 is significant, i.e., Doses M and H. There is no evidence of a significant effect at Dose M compared to placebo for Endpoint E2 and thus testing within that branch stops. At the last stage, Dose H is tested against placebo for Endpoint E3. This test is significant and thus we conclude that Dose H is superior to placebo for all three endpoints whereas Dose M is superior to placebo only for Endpoint E1.

5.3.4 Other serial gatekeeping procedures

In general, sponsors of clinical trials may consider more complicated serial gatekeeping procedures, including multiple-branch with unequal length branches (this setting is encountered in trials that compare a treatment to multiple controls). Further, multiple-branch serial gatekeeping procedures can be constructed based on other multiple tests, e.g., the Hochberg test (Quan,

TABLE 5.3: Serial gatekeeping procedure in the Type II diabetes clinical trial example. The asterisk identifies the adjusted p-values that are significant at the two-sided 0.05 level.

Family	Endpoint	Comparison	Raw p-value	Adjusted p-value
F_1	E1	L vs. P	0.0176	0.0528
F_1	E1	M vs. P	0.0108	0.0324*
F_1	E1	H vs. P	0.0052	0.0156*
F_2	E2	L vs. P	0.0128	0.0528
F_2	E2	M vs. P	0.0259	0.0777
F_2	E2	H vs. P	0.0093	0.0279*
F_3	E3	L vs. P	0.0511	0.1533
F_3	E3	M vs. P	0.0058	0.0777
F_3	E3	H vs. P	0.0099	0.0297*

Luo and Capizzi, 2005) or Dunnett test (Dmitrienko et al., 2006a; Dmitrienko, Tamhane and Liu, 2008).

5.4 Parallel gatekeeping procedures

This section gives an overview of multiplicity adjustment methods used in parallel gatekeeping procedures.

5.4.1 General parallel gatekeeping framework

Family F_i is termed a *parallel gatekeeper* if at least one significant result must be observed in this family, i.e., one or more hypotheses must be rejected in $\{H_{i1}, \ldots, H_{in_i}\}$, to proceed to F_{i+1}, $i = 1, \ldots, m-1$ (see Figure 5.7). In other words, if testing is performed at the α level, the gatekeeper is passed if and only if

$$\min_{j=1,\ldots,n_i} \widetilde{p}_{ij} \leq \alpha.$$

As an illustration, a multiple testing problem with a parallel gatekeeper was discussed in Section 5.2.2. Other examples can be found in clinical trials with multiple primary endpoints when each endpoint provides independent proof of efficacy and can lead to a regulatory claim, e.g., the acute respiratory distress syndrome clinical trial (Dmitrienko, Offen and Westfall, 2003, Section 4) with two primary endpoints, number of ventilator-free days and 28-day all-cause mortality, or the EPHESUS trial (Pitt et al., 2003) that utilized two primary endpoints, all-cause mortality and cardiovascular mortality plus cardiovascular hospitalization.

The parallel gatekeeping methods were introduced in Dmitrienko, Offen

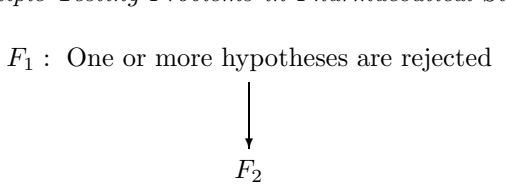

FIGURE 5.7: A problem with a parallel gatekeeper (F_1 is a parallel gatekeeper for F_2).

and Westfall (2003) who considered a Bonferroni-based procedure derived using the closure principle (see Section 2.3.3). Since this method relies on a complete enumeration of all intersection hypotheses in the closed family associated with $F_1, \ldots, F_m$, the resulting parallel gatekeeping procedures may lack transparency and their implementation can be computationally intensive since it takes order-2^n steps to test n hypotheses.

Further research in this area revealed that a broad class of parallel gatekeeping procedures have a stepwise form (Dmitrienko et al., 2006b; Hommel, Bretz and Maurer, 2007; Guilbaud, 2007; Dmitrienko, Tamhane and Wiens, 2008). This property streamlines their implementation and interpretation by clinical trial practitioners (US Food and Drug Administration statisticians have repeatedly emphasized the importance of multiple testing procedures that can be understood by clinicians). In this section we will focus on multistage parallel gatekeeping procedures developed in Dmitrienko, Tamhane and Wiens (2008).

5.4.2 Multistage parallel gatekeeping procedures

We will begin by introducing two concepts that play a key role in the framework for constructing multistage parallel gatekeeping procedures: the *error rate function* of a multiple test and *separable* multiple tests.

Consider the problem of testing a single family of n null hypotheses $H_1, \ldots, H_n$. For any subset I of the index set $N = \{1, 2, \ldots, n\}$, the error rate function $e(I)$ of a multiple test is the maximum probability of making at least one Type I error when testing the hypotheses H_i, $i \in I$, i.e.,

$$e(I) = \sup_{H_I} P\left\{ \bigcup_{i \in I} (\text{Reject } H_i) \,\middle|\, H_I \right\}.$$

Here the supremum of the probability is computed over the entire parameter space corresponding to the null hypothesis

$$H_I = \bigcap_{i \in I} H_i.$$

An exact expression for $e(I)$ is generally difficult to derive and an easily

computable upper bound on $e(I)$ can be used instead. For example, the upper bound for the error rate function of the Bonferroni test is given by $\alpha|I|/n$, where $|I|$ is the number of elements in the index set I. To simplify notation in this section, if an exact expression for $e(I)$ is available, we will use the original error rate function; otherwise $e(I)$ will denote an upper bound on the error rate function.

Note that, by definition, $e(\emptyset) = 0$ and $e(N) = \alpha$, where α is the FWER. In addition, it is natural to require that the error rate function be monotone, i.e., $e(I) \leq e(J)$ if $I \subseteq J$. If the monotonicity condition is not satisfied, one can easily enforce monotonicity by using the following upper bound in place of the original error rate function

$$e^*(I) = \max_{I' \subseteq I} e(I').$$

It is easy to see that $e^*(I)$ is a monotone error rate function.

A multiple test meets the *separability condition* (and is termed *separable*) if its error rate function is strictly less than (separates from) α unless all hypotheses are true, i.e.,

$$e(I) < \alpha \text{ for all } I \subset N.$$

The Bonferroni test clearly satisfies this condition since $e(I) < \alpha$ for any index set I with less than n elements.

Truncated multiple tests

It is easy to show that most popular multiple tests, with the exception of the Bonferroni test, do not meet the separability condition. To construct separable multiple tests, Dmitrienko, Tamhane and Wiens (2008) proposed *truncated* versions of popular tests by taking a convex combination of their critical values with the critical values of the Bonferroni test. As a result, a truncated test is uniformly more powerful than the Bonferroni test but uniformly less powerful than the original test. As an illustration, we will define the truncated Holm and Hochberg tests in this section. Truncated versions of other tests; e.g., the fallback and Dunnett tests, and their error rate functions are given in Dmitrienko, Tamhane and Wiens (2008).

To define the two truncated tests, consider the ordered p-values, $p_{(1)} \leq \ldots \leq p_{(n)}$ and let $H_{(1)}, \ldots, H_{(n)}$ denote the corresponding hypotheses. The truncated Holm test is a step-down test based on the following critical values:

$$c_i = \left[\frac{\gamma}{n-i+1} + \frac{1-\gamma}{n}\right]\alpha, \quad i = 1, \ldots, n,$$

where $0 \leq \gamma < 1$ is the truncation fraction. In other words, the truncated Holm test begins by testing $H_{(1)}$ at a c_1 level. If $p_{(1)} \leq c_1$, this hypothesis is rejected and the next hypothesis is examined. In general, the truncated test

rejects $H_{(i)}$ if $p_{(j)} \leq c_j$ for all $j \leq i$ and retains $H_{(i)}, \ldots, H_{(n)}$ otherwise. This test simplifies to the Bonferroni test if $\gamma = 0$ and to the regular Holm test if $\gamma = 1$. The error rate function of the truncated Holm test is given by $e(I) = [\gamma + (1-\gamma)|I|/n]\alpha$ if $|I| > 0$ and 0 otherwise.

The truncated Hochberg test utilizes the same set of critical values but it is set up as a step-up test. For $\gamma = 0$ and $\gamma = 1$, this truncated test reduces to the Bonferroni and regular Hochberg tests, respectively. The error rate function of the truncated Hochberg test is given by

$$e(I) = 1 - P\left\{p_{(i)}(I) > \left[\frac{\gamma}{|I| - i + 1} + \frac{1-\gamma}{n}\right]\alpha \text{ for all } i \in I\right\}$$

if $|I| > 0$ and 0 if $|I| = 0$. Here $p_{(i)}(I)$ denotes the ith ordered p-value associated with the index set I, $i = 1, \ldots, |I|$. The calculation of this error rate function is discussed in the Appendix. In the case of two hypotheses, the error rate function of the truncated Hochberg test is equivalent to that of the truncated Holm test, i.e., $e(I) = [\gamma + (1-\gamma)|I|/2]\alpha$ if $|I| > 0$ and 0 otherwise.

Multistage testing algorithm

Consider again families $F_1, \ldots, F_m$ corresponding to multiple analyses in a clinical trial and assume that F_i, $i = 1, \ldots, m-1$, is a parallel gatekeeper (as a side note, this framework also includes serial gatekeepers since any serial gatekeeper can be expressed as a series of single-hypothesis families). Let A_i denote the index set corresponding to the retained hypotheses in F_i and $e_i(I)$ denote the error rate function for the test used in F_i, $i = 1, \ldots, m-1$. The following algorithm defines a broad class of parallel gatekeeping procedures with a stepwise structure.

- Family F_1. The hypotheses are tested at an α_1 level using any FWER-controlling separable multiple test, where $\alpha_1 = \alpha$.
- Family F_i, $i = 2, \ldots, m-1$. The hypotheses are tested at an α_i level using any FWER-controlling separable multiple test, where
$$\alpha_i = \alpha_{i-1} - e_{i-1}(A_{i-1}).$$
- Family F_m. The hypotheses are tested at an α_m level using any FWER-controlling multiple test that controls the FWER within F_m, where
$$\alpha_m = \alpha_{m-1} - e_{m-1}(A_{m-1}).$$

Gatekeeping procedures constructed using this algorithm satisfy two important conditions:

- Parallel gatekeeping condition: A null hypothesis in F_i, $i = 2, \ldots, m$, cannot be rejected if all hypotheses in F_{i-1} are retained. This is a direct consequence of the fact that the "unused" Type I error rate, $\alpha_i = \alpha_{i-1} - e_{i-1}(A_{i-1}) = 0$ if $A_{i-1} = N_{i-1}$ (all hypotheses in F_{i-1} are retained).

- Independence condition: A decision to reject a null hypothesis in F_i, $i = 1, \ldots, m-1$, is independent of decisions made in $F_{i+1}, \ldots, F_m$ due to the stepwise form of gatekeeping procedures. This condition is consistent with the regulatory requirement that the primary analyses in a registration clinical trial be independent of secondary analyses. However, if the independence condition is not considered critical, one can construct gatekeeping procedures that have more power for tests in the first family. For a discussion of the independence condition, see Dmitrienko et al. (2005, Sections 2.7.2–2.7.3) and Hommel, Bretz and Maurer (2007, Section 4).

It follows from the multistage testing algorithm that the penalty paid for performing multiple inferences in F_i, $i = 2, \ldots, m$, depends on the number of the hypotheses rejected at earlier stages. Note that $\alpha_1, \ldots, \alpha_m$ is a nonincreasing sequence, which implies that one faces higher hurdles later in the sequence unless all hypotheses are rejected in previously examined families. The rate at which α_i decreases depends on the tests used at each stage of the procedure. As an illustration, assume that the hypotheses in $F_i, i = 1, \ldots, m-1$, are tested using the Bonferroni test. In this case,

$$\alpha_i = \frac{r_{i-1}\alpha_{i-1}}{n_{i-1}}, \quad i = 2, \ldots, m,$$

where r_{i-1} is the number of hypotheses rejected in F_{i-1}. In other words, the fraction of the FWER used in F_i is the product of the proportions of rejected hypotheses in F_1 through F_{i-1}. If the truncated Holm test is used in F_i, $i = 1, \ldots, m-1$,

$$\alpha_i = \begin{cases} (1-\gamma_{i-1})r_{i-1}\alpha_{i-1}/n_{i-1} & \text{if } r_{i-1} < n_{i-1}, \\ \alpha_{i-1} & \text{if } r_{i-1} = n_{i-1}, \end{cases} \qquad i = 2, \ldots, m,$$

where γ_{i-1} is the truncation fraction used in F_{i-1}. It follows from this formula that γ_{i-1} determines the fraction of α_{i-1} carried forward (unless all hypotheses are rejected in F_{i-1} in which case all of α_{i-1} is carried over to F_i). If the truncation fraction is close to 1 and some hypotheses are retained in F_{i-1}, an extremely small fraction of α_{i-1} will be carried over to F_i.

Computation of adjusted p-values

The Westfall-Young definition of an adjusted p-value given in Section 2.4.1 can be applied to calculate adjusted p-values for multistage gatekeeping procedures using the following direct calculation algorithm. This algorithm loops through a grid of significance levels between 0 and 1 to find the lowest level at which each hypothesis is rejected. The adjusted p-value for H_{ij} corresponds to the smallest k, $1 \leq k \leq K$, for which H_{ij} is rejected at the overall level $k\alpha/K$. The algorithm is quite fast since it takes only $K = 100,000$ iterations to compute adjusted p-values with four accurate decimal places. In addition,

multistage gatekeeping procedures have a stepwise form and thus each iteration requires order-n operations to test n hypotheses.

In special cases, a recursive approach can be applied to calculate adjusted p-values for multistage parallel gatekeeping procedures. For example, Guilbaud (2007) obtained a recursive formula for Bonferroni-based multistage parallel gatekeeping procedures. Consider, for simplicity, a multiple testing problem with two families, F_1 and F_2, and assume that the hypotheses are equally weighted within each family. The hypotheses in F_1 and F_2 are tested using the Bonferroni test and an arbitrary FWER-controlling test, respectively. Let p_{1j}, $j = 1, \ldots, n_1$, denote the raw p-values for the hypotheses in F_1. Further, let p'_{2j}, $j = 1, \ldots, n_2$, denote the adjusted p-values for the hypotheses in F_2 produced by the test used at the second stage of the procedure. The adjusted p-values in F_1 are given by

$$\widetilde{p}_{1j} = \min(1, n_1 p_{1j}), \quad j = 1, \ldots, n_1.$$

Now, consider the ordered adjusted p-values in F_1, i.e., $\widetilde{p}_{1(1)} < \ldots < \widetilde{p}_{1(n_1)}$. The adjusted p-values in F_2 are given by

$$\widetilde{p}_{2j} = \min_{k=1,\ldots,n_1} \max(\widetilde{p}_{1(k)},\ n_1 p'_{2j}/k), \quad j = 1, \ldots, n_2.$$

General parallel gatekeeping procedures

It is worth emphasizing that different multiple tests, can be used at different stages of the algorithm introduced earlier in this section. This includes truncated versions of all popular multiple tests introduced in Sections 2.6–2.8; e.g., p-value-based tests and tests that account for the correlation among the test statistics within each family (parametric and resampling multiple tests).

In addition, this parallel gatekeeping framework can be extended to procedures that account for the correlation across the families. Note that these gatekeeping procedures are constructed using the general closure method and may not have a stepwise form. For example, a closure-based parametric gatekeeping procedure derived from the Dunnett test was developed in Dmitrienko et al. (2006a). This procedure can be employed when the test statistics follow a multivariate normal distribution. Examples include dose-finding clinical trials with multiple normally distributed outcome variables, e.g., the Type II diabetes clinical trial example used in Section 5.3.3. Further, a Bonferroni-based resampling gatekeeping procedure was proposed in Dmitrienko, Offen and Westfall (2003). Unlike parametric procedures, this procedure does not make the normality assumption and can be applied to a broader class of multiple testing problems with a hierarchical structure.

5.4.3 Cardiovascular clinical trial example

The multistage parallel gatekeeping framework will be illustrated using a clinical trial example based on the EPHESUS trial (Pitt et al., 2003). This trial

TABLE 5.4: Two-sided p-values in the cardiovascular clinical trial example.

Family	Hypothesis	Endpoint	Raw p-value	
			Scenario 1	Scenario 2
F_1	H_{11}	P1	0.0121	0.0121
F_1	H_{12}	P2	0.0337	0.0872
F_2	H_{21}	S1	0.0084	0.0084
F_2	H_{22}	S2	0.0160	0.0160

was conducted to assess the effects of eplerenone on morbidity and mortality in patients with severe heart failure. In this clinical trial example, we will consider two families of endpoints:

- Two primary endpoints: all-cause mortality (Endpoint P1) and cardiovascular mortality plus cardiovascular hospitalization (Endpoint P2).
- Two major secondary endpoints: cardiovascular mortality (Endpoint S1) and all-cause mortality plus all-cause hospitalization (Endpoint S2).

The family of primary endpoints serves as a parallel gatekeeper for the family of secondary endpoints. The hypotheses of no treatment effect are defined as follows: The hypotheses H_{11} (Endpoint P1) and H_{12} (Endpoint P2) are included in F_1 and the hypotheses H_{21} (Endpoint S1) and H_{22} (Endpoint S2) are included in F_2. The hypotheses are equally weighted within each family and the pre-specified FWER is $\alpha = 0.05$. Table 5.4 displays two sets of two-sided p-values for the four endpoints that will be used in this example (note that these p-values are used here for illustration only). Under Scenario 1, the effect size is large for both primary endpoints and, under Scenario 2, there is evidence of treatment effect for only one primary endpoint (Endpoint P1).

A two-stage parallel gatekeeping procedure will be set up as follows. The hypotheses in F_1 and F_2 will be tested using the truncated and regular Holm tests, respectively. The truncated Holm test is carried out using four values of the truncation parameter ($\gamma = 0$, 0.25, 0.5 and 0.75) to evaluate the impact of this parameter on the outcomes of the four analyses.

To illustrate the process of applying the two-stage gatekeeping procedure, consider Scenario 1 and let $\gamma = 0.25$. The hypotheses H_{11} and H_{12} are tested using the truncated Holm test at $\alpha_1 = \alpha = 0.05$. The smaller p-value, $p_{11} = 0.0121$, is less than

$$[\gamma/2 + (1-\gamma)/2]\alpha = \alpha/2 = 0.025$$

and thus H_{11} is rejected. Further, the larger p-value, $p_{12} = 0.0337$, is compared to

$$[\gamma + (1-\gamma)/2]\alpha = 5\alpha/8 = 0.03125.$$

TABLE 5.5: Parallel gatekeeping procedure in the cardiovascular clinical trial example. The tests in F_1 are carried out using the truncated Holm test with $\gamma = 0$, 0.25, 0.5 and 0.75 and the tests in F_2 are carried out using the regular Holm test. The asterisk identifies the adjusted p-values that are significant at the two-sided 0.05 level.

Family	Endpoint	Adjusted p-value			
		$\gamma = 0$	$\gamma = 0.25$	$\gamma = 0.5$	$\gamma = 0.75$
Scenario 1					
F_1	P1	0.0242*	0.0242*	0.0242*	0.0242*
F_1	P2	0.0674	0.0539	0.0449*	0.0385*
F_2	S1	0.0336*	0.0448*	0.0449*	0.0385*
F_2	S2	0.0336*	0.0448*	0.0449*	0.0385*
Scenario 2					
F_1	P1	0.0242*	0.0242*	0.0242*	0.0242*
F_1	P2	0.1744	0.1395	0.1163	0.0997
F_2	S1	0.0336*	0.0448*	0.0672	0.0997
F_2	S2	0.0336*	0.0448*	0.0672	0.0997

The corresponding hypothesis cannot be rejected since $p_{12} > 0.03125$. To find the fraction of α that can be carried over to the hypotheses in F_2, note that the set of retained hypotheses in F_1 includes only one hypothesis. Thus, $|A_1| = 1$, $n = 2$ and

$$\alpha_2 = \alpha_1 - e_1(A_1) = \alpha - [\gamma + (1-\gamma)|A_1|/n]\alpha = 3\alpha/8 = 0.01875.$$

Applying the regular Holm test in F_2 at $\alpha_2 = 0.01875$, it is easy to verify that $p_{21} < \alpha_2/2$ and $p_{22} < \alpha_2$. This implies that the hypotheses H_{21} and H_{22} are rejected.

The adjusted p-values produced by the two-stage gatekeeping procedure are shown in Table 5.5. The adjusted p-values are computed using the direct-calculation algorithm with $K = 100,000$.

As we emphasized earlier in this section, the choice of the truncation parameter γ has a substantial impact on the outcomes of individual tests. It is clear from Table 5.5 that the adjusted p-values in the primary family (F_1) are non-increasing functions of γ (note that the adjusted p-value for Endpoint P1 is constant because the critical value of the truncated Holm test for the smallest p-value in F_1 does not actually depend on γ). However, the relationship between γ and the adjusted p-values for the secondary endpoints is more complicated. As γ increases, the fraction of α carried over to the secondary analyses may increase or decrease depending on effect sizes for false hypotheses and this directly influences the adjusted p-values in F_2.

In Scenario 1 a small increase in γ from 0 causes an increase in the adjusted p-values for Endpoints S1 and S2 (compare the columns for $\gamma = 0$ and $\gamma = 0.25$). Further, when $\gamma = 0.5$, these adjusted p-values stay at the same level and, when $\gamma = 0.75$, they drop to 0.037. This is due to the fact

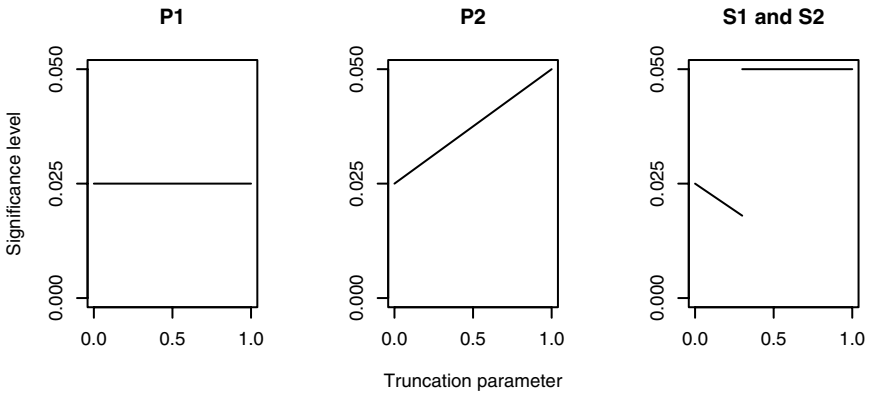

FIGURE 5.8: Significance level for Endpoint P1 (left panel), significance level for Endpoint P2 (middle panel) and overall significance level for Endpoints S1 and S2 (right panel) in Scenario 1 as a function of the truncation parameter γ.

that both primary tests achieve significance when $\gamma \geq 0.5$ and thus the entire α is propagated to the family of secondary endpoints. To illustrate this phenomenon, Figure 5.8 displays the individual significance levels for the two primary endpoints and overall α level for the secondary endpoints in Scenario 1 as a function of the truncation parameter.

Figure 5.8 shows that, as was pointed out above, the significance level for Endpoint P1 is constant, the significance level for Endpoint P2 increases with γ, and the overall significance level for Endpoints S1 and S2 is a nonlinear function of γ with a jump discontinuity at $\gamma = 0.3$. This discontinuity corresponds to the point when the p-value for Endpoint P2 becomes significant and thus, by the definition of the error rate function of the truncated Holm test, the α level for the secondary endpoints is set at 0.05. Table 5.5 and Figure 5.8 indicate that, when the effect sizes of both primary endpoints are large, the overall power is maximized by selecting a value of γ closer to 1.

Further, in Scenario 2 the adjusted p-values for Endpoints S1 and S2 steadily increase with γ because only one primary test is significant and, as a result, an increasingly smaller fraction of α is carried over to the secondary analyses. In this case, it will be desirable to choose a smaller value of the truncation parameter to improve the overall probability of success for the primary and secondary endpoints.

To summarize, the truncation parameter serves as a leverage that balances the power functions of the primary and secondary analyses. If the effect sizes of the primary endpoints are uniformly large, a truncation parameter near

1 will help improve the overall power. On the other hand, if the effect sizes are expected to vary across the endpoints, the overall power is likely to be maximized when the truncation parameter is small or in the middle of the $(0, 1)$ interval. In general, an optimal value of γ can be selected via simulations by maximizing an appropriately defined power function, e.g., the probability of rejecting all false hypotheses or at least one false hypothesis, under realistic assumptions about the effect sizes of individual endpoints.

5.5 Tree gatekeeping procedures

The tree gatekeeping methods serve as a unified framework that includes serial and parallel methods as well as a combination of serial and parallel methods with logical restrictions. This framework is quite general and can be used to address multiplicity issues in a wide variety of clinical trial applications.

5.5.1 General tree gatekeeping framework

Within the tree gatekeeping framework, gatekeepers are defined at the hypothesis rather than family level, i.e., a hypothesis in a certain family may be *testable* whereas another hypothesis in the same family may not. To give a formal definition, consider a hypothesis in F_i, say, H_{ij}, and define two sets of hypotheses associated with H_{ij} $(i = 2, \ldots, m, j = 1, \ldots, n_i)$. The sets are denoted by R_{ij}^S (serial rejection set) and R_{ij}^P (parallel rejection set). These sets include hypotheses from $F_1, \ldots, F_{i-1}$, at least one of them is non-empty and, without loss of generality, R_{ij}^S and R_{ij}^P do not overlap. The hypothesis H_{ij} is testable if all hypotheses are rejected in R_{ij}^S and at least one hypothesis is rejected in R_{ij}^P, i.e., if the following two conditions hold,

$$\max_{k,l \in R_{ij}^S} \widetilde{p}_{kl} \leq \alpha \text{ and } \min_{k,l \in R_{ij}^P} \widetilde{p}_{kl} \leq \alpha.$$

As an example, consider the two-family problem depicted in Figure 5.9. The first family, F_1, includes three hypotheses (H_{11}, H_{12}, H_{13}) and the second one, F_2, contains a single hypothesis (H_{21}). The serial and parallel rejection sets for H_{21} are defined as follows:

$$R_{21}^S = \{H_{11}\} \text{ and } R_{21}^P = \{H_{12}, H_{13}\}.$$

The hypothesis H_{21} can be tested only if there is a significant result in R_{21}^S and at least one significant result in R_{21}^P.

As was mentioned above, the tree gatekeeping framework includes the serial and parallel gatekeeping frameworks as special cases. Tree gatekeeping procedures simplifies to serial gatekeeping procedures if $R_{ij}^S = F_{i-1}$ and R_{ij}^P

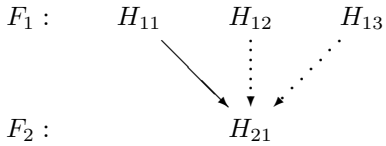

FIGURE 5.9: Tree gatekeeping procedure in a two-family problem. A solid line is used to define a "serial" connection and dotted lines are used for "parallel" connections.

is empty for all hypotheses H_{ij}, $i = 2, \ldots, m$, and to parallel gatekeeping procedures if R_{ij}^S is empty and $R_{ij}^P = F_{i-1}$ for all hypotheses H_{ij}, $i = 2, \ldots, m$.

The tree gatekeeping methodology was proposed in Dmitrienko, Wiens, Tamhane and Wang (2007) and was motivated by multiple testing problems that arise in trials when decision trees include multiple branches and/or logical restrictions, e.g.,

- Clinical trials with complex hierarchically ordered hypotheses, e.g., hypotheses associated with multiple endpoints (primary, secondary and tertiary) and multiple test types (noninferiority and superiority), e.g., a hypertension clinical trial with multiple endpoints and noninferiority/superiority tests (Dmitrienko et al., 2007, Section 5).

- Dose-finding studies with multiple endpoints and logical restrictions, e.g., a Type II diabetes clinical trial with a primary and two secondary endpoints (Dmitrienko et al., 2006a) and the metformin-rosiglitazone combination therapy trial (Fonseca et al., 2002) that included a comparison of two metformin-rosiglitazone regimens to metformin on several endpoints.

Multiple testing problems of this kind are quite complex and cannot be handled within the more basic serial or parallel gatekeeping frameworks.

Closure-based tree gatekeeping procedures

Dmitrienko, Wiens, Tamhane and Wang (2007) developed a framework for constructing tree gatekeeping procedures based on the Bonferroni test. Unlike parallel gatekeeping procedures introduced in Section 5.4.2, Bonferroni tree gatekeeping procedures do not, in general, have a straightforward stepwise form. To define a tree gatekeeping procedure, one needs to utilize the closure principle and use a weighted Bonferroni test for each intersection hypothesis in the closed family associated with the m families of interest. Dmitrienko, Wiens, Tamhane and Wang gave a general algorithm for assigning weights to individual hypotheses that takes into account logical relationships among multiple analyses in a clinical trial. Dmitrienko, Tamhane, Liu and Wiens (2008)

noted that Bonferroni tree gatekeeping procedures based on this algorithm may violate the tree gatekeeping property defined above, e.g., a hypothesis in F_i, $i = 2, \ldots, m$, may be rejected even though some hypotheses are retained in R_{ij}^S or all hypotheses are retained in R_{ij}^P. To address this problem, Dmitrienko, Tamhane, Liu and Wiens formulated a monotonicity condition which is sufficient to guarantee the tree gatekeeping property. Dmitrienko, Tamhane and Liu (2008) and Kordzakhia et al. (2008) derived a weight assignment algorithm that satisfies the monotonicity condition. This algorithm is given in the Appendix.

Dmitrienko, Tamhane and Liu (2008) defined a general approach to defining a broad family of tree gatekeeping procedures that includes Bonferroni tree gatekeeping procedures as a special case. This approach is based on combining multiple tests across families of hypotheses and enables clinical trial sponsors to set up powerful procedures that take into account complex logical restrictions. Examples include tree gatekeeping procedures based on the Hochberg or Dunnett tests.

5.5.2 Combination-therapy clinical trial example

To illustrate Bonferroni tree gatekeeping procedures, we will return to the clinical trial example described in Section 5.2.3. This example involves six hierarchically ordered null hypotheses grouped into four families.

To be consistent with the notation introduced earlier in this section, the hypotheses and families will be defined as follows:

- Family F_1 includes H_{11} (noninferiority hypothesis for A versus B).
- Family F_2 includes H_{21} (superiority hypothesis for A versus B) and H_{22} (noninferiority hypothesis for A+B versus B).
- Family F_3 includes H_{31} (superiority hypothesis for A+B versus B) and H_{32} (noninferiority hypothesis for A+B versus A).
- Family F_4 includes H_{41} (superiority hypothesis for A+B versus A).

Now, to account for the logical restrictions among the six hypotheses (the restrictions are displayed in Figure 5.3), the serial rejection sets are given by

$$\begin{aligned} R_{21}^S &= R_{22}^S = \{H_{11}\}, \\ R_{31}^S &= R_{32}^S = \{H_{22}\}, \\ R_{41}^S &= \{H_{32}\}. \end{aligned}$$

and the parallel rejections sets are empty.

A Bonferroni tree gatekeeping procedure based on the algorithm defined in the Appendix will be used to control the FWER at the two-sided 0.05 level. The adjusted p-values produced by this tree gatekeeping procedure are listed in Table 5.6. The table shows that the very first hypothesis, H_{11}, is rejected at

TABLE 5.6: Bonferroni tree gatekeeping procedure in the combination-therapy clinical trial example. The asterisk identifies the adjusted p-values that are significant at the two-sided 0.05 level.

Family	Hypothesis	Raw p-value	Adjusted p-value
F_1	H_{11}	0.011	0.011*
F_2	H_{21}	0.023	0.046*
F_2	H_{22}	0.006	0.012*
F_3	H_{31}	0.018	0.046*
F_3	H_{32}	0.042	0.084
F_4	H_{41}	0.088	0.088

the two-sided 0.05 level and thus the hypotheses H_{21} and H_{22} become testable. Both of them are also rejected and, since H_{22} is included in the serial rejection sets of the hypotheses in F_3, the tree gatekeeping procedure tests H_{31} and H_{32} at the next step. The adjusted p-value for H_{31} is significant but the adjusted p-value for H_{32} is not. Since the hypothesis H_{41} depends on H_{32}, the former is retained without testing. It can be seen from Table 5.6 that the adjusted p-value for H_{41} is greater than 0.05.

It is worth noting that the adjusted p-values displayed in Table 5.6 are equal to those computed in Dmitrienko and Tamhane (2007, Table IV) even though the latter set of adjusted p-values was obtained using another method (the method defined in Dmitrienko, Wiens, Tamhane and Wang, 2007). The two methods for implementing Bonferroni tree gatekeeping procedures are based on two different algorithms but they often produce identical sets of adjusted p-values.

5.6 Software implementation

This section describes the SAS programs that were used in this chapter to implement serial, parallel and tree gatekeeping procedures. These programs can be downloaded from the book's Web site (http://www.multxpert.com).

- Serial gatekeeping procedures. Program 5.1 implements the direct-calculation algorithm defined in Section 5.4.2 to compute adjusted p-values for the three-branch serial gatekeeping procedure in the Type II diabetes clinical trial example (Section 5.3.3).

- Parallel gatekeeping procedures. Program 5.2 computes adjusted p-values for the two-stage parallel gatekeeping procedure based on the truncated and regular Holm tests in the cardiovascular clinical trial ex-

ample (Section 5.4.3). This program also utilizes the direct-calculation algorithm.

- Tree gatekeeping procedures. Program 5.3 calculates adjusted p-values for the Bonferroni tree gatekeeping procedure in the combination-therapy clinical trial example (Section 5.5.2).

Acknowledgements

Ajit C. Tamhane's research was supported by grants from the National Heart, Lung and Blood Institute.

Appendix

Error rate function of the truncated Hochberg test

To compute the error rate function of the truncated Hochberg test for $|I| > 0$, note that $e(I) = 1 - P(a_1, \ldots, a_k)$, where $k = |I|$,

$$\begin{aligned} a_i &= \left(\frac{\gamma}{k-i+1} + \frac{1-\gamma}{n}\right)\alpha, \quad i = 1, \ldots, k, \\ P(a_1, \ldots, a_k) &= P(U_{(i)} > a_i \text{ for all } i = 1, \ldots, k) \end{aligned}$$

and $U_{(1)} < \ldots < U_{(k)}$ are the order statistics of i.i.d. observations from a uniform $(0, 1)$ distribution. Sen (1999) developed a recursive formula for computing $P(a_1, \ldots, a_k)$. Using this formula, it can be shown that

$$\begin{aligned} P(a_1) &= 1 - a_1, \\ P(a_1, a_2) &= (1 - a_2)(1 - 2a_1 + a_2) \\ P(a_1, a_2, a_3) &= (1 - a_3)(1 - 3a_1 + a_3 - 3a_2^2 + 6a_1a_2 - 3a_1a_3 + a_3^2). \end{aligned}$$

Weight assignment algorithm for Bonferroni tree gatekeeping procedures

Assuming the multiple testing problem formulated in Section 5.5.1, consider the closed family associated with the n null hypotheses in Families $F_1, \ldots, F_m$. For each intersection hypothesis H, define the indicator functions $\delta_{ij}(H)$ and $\xi_{ij}(H)$ as follows. Let $\delta_{ij}(H) = 1$ if H contains H_{ij} and 0 otherwise, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. Further, for $i = 2, \ldots, m$ and $j = 1, \ldots, n_i$, let $\xi_{ij}(H) = 0$ if H contains at least one hypothesis from R_{ij}^S or all hypotheses from R_{ij}^P. Otherwise, let $\xi_{ij}(H) = 1$. A Bonferroni tree gatekeeping procedure is defined by specifying a weighted Bonferroni test for each intersection hypothesis H. To accomplish this, it is sufficient to set up an n-dimensional

weight vector for H denoted by $v_{ij}(H)$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. The p-value for H is given by

$$p_H = \min_{i,j} \frac{p_{ij}}{v_{ij}(H)},$$

where p_{ij} is the p-value for H_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. Note that $p_{ij}/v_{ij}(H)$ can be set to 1 if $v_{ij}(H) = 0$. Based on the closure principle, the adjusted p-value for H_{ij} is found by computing the maximum p_H over all intersection hypotheses containing H_{ij}.

The weight vector for H is constructed sequentially by defining m subvectors

$$(v_{i1}, \ldots, v_{in_i}), \quad i = 1, \ldots, m,$$

using the algorithm described below (it is assumed in the algorithm that $0/0 = 0$).

Family F_1. Let

$$v_{1j}(H) = v_1^*(H) w_{1j} \delta_{1j}(H), \quad j = 1, \ldots, n_1,$$

where $v_1^*(H) = 1$, and let $v_2^*(H)$ denote the remaining weight, i.e.,

$$v_2^*(H) = v_1^*(H) \left(1 - \sum_{j=1}^{n_1} w_{1j} \delta_{1j}(H) \right).$$

Family F_k, $k = 2, \ldots, m-1$. Let

$$v_{kj}(H) = v_k^*(H) w_{kj} \delta_{kj}(H) \xi_{kj}(H), \quad j = 1, \ldots, n_k.$$

The remaining weight is given by

$$v_{k+1}^*(H) = v_k^*(H) \left(1 - \sum_{j=1}^{n_k} w_{kj} \delta_{kj}(H) \right).$$

Family F_m. Let

$$v_{mj}(H) = v_m^*(H) w_{mj} \delta_{mj}(H) \xi_{mj}(H) / \sum_{l=1}^{n_m} w_{ml} \delta_{ml}(H) \xi_{ml}(H),$$

where $j = 1, \ldots, n_m$.