Chapter 4

Analysis of Multiple Endpoints in Clinical Trials

Ajit C. Tamhane
Northwestern University

Alex Dmitrienko
Eli Lilly and Company

4.1 Introduction

Most human diseases are characterized by multidimensional etiology and the efficacy of an experimental treatment frequently needs to be assessed on multiple outcome measures — commonly referred to as endpoints. There is a variety of ways in which the contribution of each endpoint can be accounted for in the primary analysis, for example, the trial's sponsor can treat endpoints as independent entities or as manifestations of a single underlying cause. The following examples illustrate the common approaches to analyzing multiple endpoints.

Example 1. If each endpoint independently provides a proof of efficacy, the trial's outcome is declared positive if at least one endpoint is associated with a significant improvement compared to the control. Gong, Pinheiro and DeMets (2000) gave several examples of cardiovascular trials in which the primary objective had two components (primary endpoint and principal secondary endpoint), e.g., mortality and mortality plus morbidity due to heart failure in the VEST trial (Cohn et al., 1998) and mortality plus cardiovascular morbidity and mortality in the PRAISE-I trial (Packer et al., 1996).

Example 2. Another commonly used approach to defining the primary objective in clinical trials with multiple endpoints is based on the development of *composite endpoints*. A composite endpoint can be based on a sum of multiple scores or combination of multiple events. In the case of multiple events, a patient achieves a composite endpoint if he or she experiences any of the pre-specified events related to morbidity or mor-

tality. For example, the primary objective of the Losartan Intervention For Endpoint reduction (LIFE) trial was to study the effect of losartan on the composite endpoint of cardiovascular death, myocardial infarction, and stroke (Dahlöf et al., 2002).

Example 3. When the multiple endpoints are biologically related to each other (for example, they measure different aspects of the same underlying cause), the primary objective can be defined in terms of a combination of individual effects across the endpoints. The mitoxantrone trial in patients with progressive multiple sclerosis evaluated the overall effect of five clinical measures: expanded disability status scale, ambulation index, number of treated relapses, time to first treated relapse and standardized neurological status (Hartung et al., 2002). The analysis was performed based on a multivariate approach that accounted for the correlations among the endpoints.

Example 4. In certain areas, a clinically meaningful effect is defined as the simultaneous improvement in multiple measures. In this case, the primary objective of a clinical trial is met if the test drug shows a significant effect with respect to all the endpoints. Offen et al. (2007) gave a list of more than 15 examples in which a positive trial is defined in terms of two or more significant endpoints. They included clinical trials for the treatment of migraine, Alzheimer's disease and osteoarthritis.

Due to an increasing number of studies dealing with conditions that exhibit a complex etiology, there has been much attention in the clinical trial literature on the development of statistical methods for analyzing multiple endpoints. O'Brien's (1984) paper ignited a flurry of research in this area and a number of procedures for testing multiple endpoints have been developed over the past two decades. O'Neill (2006) mentions multiple endpoints as one of the key areas for guidance development in FDA's critical path initiative and writes that "The many statistical methods and approaches that have been suggested in the literature in recent years now deserve to be digested and placed in the context of how they can be best used in clinical trials." The purpose of this chapter is to attempt to answer this need by giving a critical overview of the various procedures along with their assumptions, pros and cons, and domains of applicability. For some other reviews of the literature on multiple endpoints, the reader is referred to Chi (1998), Comelli and Klersy (1996), Geller (2004), Huque and Sankoh (1997), Kieser, Reitmeir and Wassmer (1995), Sankoh, Huque and Dubey (1997) and Zhang, Quan, Ng and Stepanavage (1997).

This chapter is organized as follows. Section 4.2 introduces inferential goals and main classes of multiple endpoint procedures. It is assumed in this chapter that all endpoints are primary or co-primary. Gatekeeping procedures that deal with endpoints classified into hierarchically ordered categories such as primary and secondary will be discussed in Chapter 5. We begin with an overview of multiple testing procedures for the assessment of the treatment's

effect on each individual endpoint (at-least-one procedures) in Section 4.3. Global procedures aimed at examining the overall efficacy of the treatment are reviewed in Section 4.4. Section 4.5 discusses all-or-none procedures arising in clinical trials when the treatment's efficacy needs to be demonstrated for all endpoints. Section 4.6 describes the superiority-noninferiority approach which requires demonstrating the treatment's superiority on at least one endpoint and noninferiority on others. Finally, Section 4.7 discusses software implementation of selected procedures for testing multiple endpoints in clinical trials.

4.2 Inferential goals

To define inferential goals of multiple endpoint procedures, we will consider a clinical trial with two treatment groups. The trial's objective is to assess the effect of the experimental treatment (hereafter referred to simply as the treatment) on m endpoints compared to that of the control, e.g., placebo. Let δ_i denote an appropriate measure of the true treatment effect for the ith endpoint (e.g., mean difference, odds ratio or log hazard ratio). The efficacy of the treatment on the ith endpoint is assessed by testing the hypothesis of no treatment difference. The multiple testing setting gives rise to the problem of Type I error rate inflation. Depending on the trial's objectives, this problem can be addressed by performing a multiplicity adjustment, combining evidence of the treatment's effect across the endpoints (e.g., by combining multiple endpoints into a single composite endpoint) or utilizing other approaches. This section gives a brief overview of four main approaches to the analysis of multiple endpoints.

4.2.1 At-least-one procedures

If each multiple endpoint is independently clinically relevant (and can potentially be associated with its own regulatory claim), the multiple endpoint problem can be formulated as a multiple testing problem. Cardiovascular clinical trials listed in Example 1 in the Introduction serve as an illustration of this approach. In these trials the endpoints are treated as clinically independent entities and the sponsor is interested in assessing the effect of the experimental treatment on each endpoint.

Given this structure of the primary objective, the trial is declared positive if at least one significant effect is detected. The global hypothesis is defined as the intersection of hypotheses for individual endpoints and the testing problem is stated as

$$H_I = \bigcap_{i=1}^{m} (\delta_i \leq 0) \text{ versus } K_U = \bigcup_{i=1}^{m} (\delta_i > 0). \tag{4.1}$$

The global hypothesis is rejected if one or more individual hypotheses of no treatment effect are demonstrated to be false. This is a prototypical multiple testing problem, known as the *union-intersection* (UI) problem (see Section 2.3.1), which requires a *multiplicity adjustment*. The objective of a multiplicity adjustment is to control the familywise error rate (FWER),

$$\text{FWER} = P\{\text{Reject at least one true hypothesis}\},$$

at a designated level α by adjusting the level of each test downward. As in other clinical trial applications, FWER needs to be controlled in the strong sense; i.e., it must not be greater than α regardless of which hypotheses are true and which are false (Hochberg and Tamhane, 1987). Procedures that can be used in this multiple endpoint problem (at-least-one procedures) are described in Section 4.3.

It is worth noting that the false discovery rate (FDR) proposed by Benjamini and Hochberg (1995) is generally not appropriate for the multiple endpoint problem for the following reasons.

- FDR is suitable for testing a large number of hypotheses whereas the number of endpoints is generally very small.
- FDR is suitable for exploratory studies in which a less stringent requirement of control of the proportion of false positives is acceptable. However, tests for multiple endpoints are generally confirmatory in nature required for drug approval and labeling.
- The problem of testing multiple endpoints often has additional complications such as ordered categories of endpoints, e.g., primary, co-primary and secondary with logical restrictions, and decision rules based on tests for superiority and noninferiority on different endpoints. The FDR approach is not designed to handle such complex decision rules.

4.2.2 Global procedures

In many clinical trial applications it is desired to show that the treatment has an *overall* effect across the endpoints without necessarily a large significant effect on any one endpoint (see Examples 2 and 3 in the Introduction).

To establish an overall treatment effect, usually a point null hypothesis of no difference between the treatment and control is tested against a one-sided alternative:

$$H_I^*: \ \delta_i = 0 \text{ for all } i \text{ versus } K_U^*: \ \delta_i \geq 0 \text{ for all } i \text{ and } \delta_i > 0 \text{ for some } i. \tag{4.2}$$

It is well-known (O'Brien, 1984) that Hotelling's T^2-test is inappropriate for this problem because it is a two-sided test, and hence lacks power for detecting one-sided alternatives necessary for showing the treatment efficacy.

In this case one-sided *global procedures* that have been proposed as alternatives to the T^2-test are more appropriate.

Global procedures are conceptually similar to *composite endpoints* (defined in Example 2) in that they also rely on reducing the number of dimensions in a multiple endpoint problem by combining multiple measures into a single one, e.g., combining tests for individual endpoints into a single test. However, unlike composite endpoints based on a sum or other simple function of individual scores, global testing procedures address the interdependence of multiple endpoints, i.e., they account for the correlations among them. For a detailed discussion of global procedures, see Section 4.4.

4.2.3 All-or-none procedures

Another formulation of the multiple endpoint problem pertains to the requirement that the treatment be effective on *all* endpoints (see Example 4 in the Introduction). This problem is referred to as the *reverse multiplicity problem* by Offen et al. (2007) and represents the most stringent inferential goal for multiple endpoints. In mathematical terms, this is an example of an *intersection-union* (IU) problem introduced in Section 2.3.2. Within the IU framework, the global hypothesis is defined as the union of hypotheses corresponding to individual endpoints and thus the testing problem is stated as

$$H_U = \bigcup_{i=1}^{m} (\delta_i \leq 0) \text{ versus } K_I = \bigcap_{i=1}^{m} (\delta_i > 0). \tag{4.3}$$

To reject H_U, one needs to show that all individual hypotheses are false (the treatment effects for all endpoints are significant). All-or-none testing procedures are discussed in Section 4.5.

4.2.4 Superiority-noninferiority procedures

The hybrid superiority-noninferiority testing approach to testing multiple endpoints provides a viable alternative to the stringent all-or-none testing approach described in Section 4.2.3. In this case, the goal is to demonstrate that the treatment is superior to the control on at least one endpoint and not inferior on all other endpoints.

To define the null and alternative hypotheses for this setting, let $\eta_k \geq 0$ denote the superiority threshold (commonly, $\eta_k = 0$) and $\varepsilon_k > 0$ denote the noninferiority threshold for the kth endpoint. In other words, the treatment is superior to the control on the kth endpoint if $\delta_k > \eta_k$ and noninferior if $\delta_k > -\varepsilon_k$, $k = 1, \ldots, m$.

For the kth endpoint, the superiority testing problem is stated as

$$H_k^{(S)} : \ \delta_k \leq \eta_k \text{ versus } K_k^{(S)} : \ \delta_k > \eta_k.$$

Similarly, the noninferiority testing problem for the kth endpoint is stated as

$$H_k^{(N)}: \ \delta_k \leq -\varepsilon_k \text{ versus } K_k^{(N)}: \ \delta_k > -\varepsilon_k.$$

Note that the difference between $H_k^{(S)}$ and $H_k^{(N)}$ is simply a shift. Showing superiority involves clearing a higher bar than showing noninferiority.

The overall superiority testing problem is given by

$$H_I^{(S)} = \bigcap_{k=1}^{m} H_k^{(S)} \text{ versus } K_U^{(S)} = \bigcup_{k=1}^{m} K_k^{(S)}.$$

The global superiority hypothesis $H_I^{(S)}$ is rejected if *at least one* $H_k^{(S)}$ is rejected. Similarly, the overall noninferiority testing problem is given by

$$H_U^{(N)} = \bigcup_{k=1}^{m} H_k^{(N)} \text{ versus } K_I^{(N)} = \bigcap_{k=1}^{m} K_k^{(N)}.$$

The global noninferiority hypothesis $H_U^{(N)}$ is rejected if *all* $H_k^{(N)}$ are rejected. To combine the two global hypotheses and formulate the superiority-noninferiority testing approach, we need to consider testing the union of the global superiority and global noninferiority hypotheses

$$H_U^{(SN)} = H_I^{(S)} \cup H_U^{(N)} \text{ versus } K_I^{(SN)} = K_U^{(S)} \cap K_I^{(N)}.$$

In other words, the trial's objective is met if there is superior efficacy for at least one endpoint ($K_U^{(S)}$) and noninferior efficacy for all endpoints ($K_I^{(N)}$). This superiority-noninferiority testing problem becomes equivalent to the all-or-none testing problem (Section 4.2.3) if the noninferiority margins are set to 0 for all endpoints, i.e., $\varepsilon_1 = \ldots = \varepsilon_m = 0$. Superiority-noninferiority procedures are discussed in Section 4.6.

4.3 At-least-one procedures

This section describes p-value-based procedures as well as parametric (normal theory) and resampling-based procedures for multiple endpoints when it is desired to demonstrate the treatment's superiority on at least one endpoint.

4.3.1 Procedures based on univariate p-values

Frequently, different test statistics (e.g., t-statistics, log-rank statistics, chi-square statistics, etc.) are used to compare the treatment and control groups on different endpoints because of the different scales on which the endpoints

are measured. A standard approach to put the results of these different tests on the same scale is via their p-values. Therefore p-value-based procedures are of interest in the analysis of multiple endpoints. Section 2.6 provides a detailed description of popular multiple testing procedures based on univariate p-values. Here we briefly review basic procedures from that section with emphasis on issues specific to multiple endpoints. In addition, we describe two p-value-based procedures introduced specifically for the multiple endpoint problem.

Bonferroni and related procedures

The simplest of the p-value-based procedures is the well-known Bonferroni procedure introduced in Section 2.6.1. This procedure is known to be conservative especially when there are many endpoints and/or they are highly correlated. More powerful stepwise versions of the Bonferroni procedure, e.g., the Holm and Hochberg procedures, are also described in Section 2.6.

The basic Bonferroni procedure allocates the Type I error rate α equally among the endpoints. A weighted version of this procedure allows unequal allocation which is useful for unequally important endpoints, e.g., a clinical trial with a principal secondary endpoint which may provide the basis for a new regulatory claim. The weighted Bonferroni procedure can also be employed in trials where some endpoints are adequately powered and the others are underpowered.

As an example, consider the Carvedilol cardiovascular trials (Fisher and Moyé, 1999) in which exercise capability plus quality of life served as the primary endpoint and mortality was a key secondary endpoint. The primary endpoint was not significant at the 0.05 level in the trials with the exercise capability endpoint while the secondary mortality endpoint was highly significant in the combined analysis across the trials. Although the problem of interpreting the results of such trials is a vexing one (see O'Neill (1997) and Davis (1997) for contrasting views), we will assume, for the sake of illustration, that the mortality endpoint was prospectively defined as a co-primary endpoint.

In this case a decision rule based on the weighted Bonferroni procedure can be set up. In the general case of m endpoints, the rule uses an additive alpha allocation scheme. Let $w_1, \ldots, w_m$ be positive weights representing the importance of the endpoints such that they sum to 1. The hypothesis of no treatment effect for the ith endpoint is tested at level α_i, where $\alpha_i = w_i\alpha$ and thus

$$\sum_{i=1}^{m} \alpha_i = \alpha.$$

In the Carvedilol example, a Bonferroni-based decision rule could have been constructed by assigning a large fraction of the Type I error rate α to the

exercise capability endpoint (i.e., by choosing α_1 close to α) and "spending" the rest on the mortality endpoint ($\alpha_2 = \alpha - \alpha_1$).

A slightly sharpened version of this rule, termed the *prospective alpha allocation scheme* (PAAS) method, was proposed by Moyé (2000). Assuming that the p-values for the individual endpoints are independent, we have

$$\prod_{i=1}^{m}(1-\alpha_i) = 1-\alpha.$$

Moyé's solution to the problem of two co-primary endpoints was to select the fraction of the Type I error rate α allocated to the primary endpoint and calculate the significance level for the other endpoint from the above identity. Specifically, let $0 < \alpha_1 < \alpha$ and

$$\alpha_2 = 1 - \frac{1-\alpha}{1-\alpha_1}.$$

In the context of the Carvedilol example, if $\alpha = 0.05$ and $\alpha_1 = 0.045$, then $\alpha_2 = 0.0052$, which is only marginally larger than $\alpha_2 = 0.005$ given by the weighted Bonferroni allocation.

Fixed-sequence procedure

Maurer, Hothorn and Lehmacher (1995) considered clinical trials in which the endpoints are *a priori* ordered (e.g., in terms of their importance). They applied a fixed-sequence method that tests the ordered endpoints sequentially at level α as long as the test is significant and stops testing when a non-significant result is encountered. Effectively, α is unused when the procedure rejects the hypothesis of no treatment effect for an endpoint and thus it can be carried over to the next endpoint. All α is used up when no treatment effect is detected with none left for the remaining endpoints. We refer to this as the "use it or lose it" principle.

The fixed-sequence testing approach is widely used in clinical trials and endorsed by regulatory agencies; see, for example, the CPMP guidance document on multiplicity issues (CPMP, 2002). As an example, this testing approach was adopted in the adalimumab trial in patients with rheumatoid arthritis (Keystone et al., 2004). The trial included three endpoints (American College of Rheumatology response rate, modified total Sharp score and Health Assessment Questionnaire score) that were prospectively ordered and tested sequentially. Since each test was carried out at an unadjusted α level, this approach helped the trial's sponsor maximize the power of each individual test. Note, however, that the overall power of the fixed-sequence procedure depends heavily on the true effect size of the earlier endpoints. The power of the procedure is increased if the likelihood of detecting a treatment effect for the endpoints at the beginning of the sequence is high. On the other hand, if at least one of the earlier endpoints is underpowered, the procedure is likely

to stop early and miss an opportunity to evaluate potentially useful endpoints later in the sequence. To improve the power of the fixed-sequence procedure, it is critical to order the endpoints based on the expected strength of evidence beginning with the endpoints associated with the largest effect size (Huque and Alosh, 2008).

Fallback procedure

A useful generalization of the fixed-sequence procedure was proposed by Wiens (2003). Wiens constructed the fallback procedure by allocating pre-specified fractions, $w_1, \ldots, w_m$, of α to the m *a priori* ordered endpoints subject to

$$\sum_{i=1}^{m} w_i = 1.$$

The procedure begins with the first endpoint in the sequence which is tested at level $\alpha_1 = \alpha w_1$. Further, the ith endpoint is tested at level $\alpha_i = \alpha_{i-1} + \alpha w_i$ if the previous endpoint is significant and level $\alpha_i = \alpha w_i$ otherwise. In other words, if a certain test is not significant, its significance level (α_i) is used up and, if it is significant, its level is carried over to the next endpoint, hence the name *fallback procedure*. Note that this procedure also uses the "use it or lose it" principle.

The fallback procedure is uniformly more powerful than the Bonferroni procedure and reduces to the fixed-sequence procedure if all Type I error rate is spent on the first endpoint in the sequence, i.e., $w_1 = 1$ and $w_2 = \cdots = w_m = 0$. The advantage of the fallback procedure is that one can continue testing even when the current test is not significant in contrast to the fixed-sequence procedure which stops testing as soon as it encounters a nonsignificant result.

As an illustration, consider a clinical trial with two endpoints, the first of which (functional capacity endpoint) is adequately powered and the other one (mortality endpoint) is not (Wiens, 2003). Wiens computed the power of the fallback procedure in this example assuming that $w_1 = 0.8$ and $w_2 = 0.2$ (80% of the Type I error rate is spent on the functional capacity endpoint and 20% on the mortality endpoint) and the two-sided $\alpha = 0.05$. Under this weight allocation scheme, the power for the mortality endpoint was substantially improved (from 50% to 88% compared to the Bonferroni procedure with the same set of weights) whereas the power for the functional capacity endpoint was reduced by a trivial amount (from 95% to 94%).

The overall power of the fallback procedure is heavily influenced by the effect sizes of the ordered endpoints and the significance levels for their tests (or, equivalently, the pre-specified weights). As shown in the next section, the power can be improved by arranging the endpoints in terms of the expected effect size, i.e., from the largest effect size to the smallest effect size. In addition, the expected effect size can help determine the significance levels. For

example, Huque and Alosh (2008) recommended defining the significance levels proportional to the reciprocals of the effect sizes. This choice helps increase the power of the early tests which will, in turn, raise the significance levels for the endpoints toward the end of the sequence.

Comparison of the fixed-sequence and fallback procedures

To assess the robustness of the fixed-sequence and fallback procedures with respect to the monotonicity assumption, a simulation study was conducted. A clinical trial with two arms (treatment and placebo) was simulated. The treatment-placebo comparison was performed for three ordered endpoints (Endpoints 1, 2 and 3). The endpoints were tested sequentially, beginning with Endpoint 1, using the fixed-sequence method at the one-sided 0.025 level. The endpoint outcomes were assumed to follow a multivariate normal distribution with a compound-symmetric correlation matrix (i.e., the outcomes were equicorrelated). The sample size per group ($n = 98$) was chosen to achieve 80% power for each univariate test when the true effect size is 0.4. The calculations were performed using 10,000 replications.

The power of the fixed-sequence and fallback procedures for the three endpoint tests is displayed in Table 4.1 for three values of the common correlation coefficient ($\rho = 0$, 0.2 and 0.5) and three sets of endpoint-specific effect sizes, e_i ($i = 1, 2, 3$). The following three scenarios were considered:

- Scenario 1. All tests are adequately powered, $e_1 = 0.4$, $e_2 = 0.4$, $e_3 = 0.4$.
- Scenario 2. The first test is underpowered but the other tests are adequately powered, $e_1 = 0.3$, $e_2 = 0.4$, $e_3 = 0.4$.
- Scenario 3. The first test is overpowered but the other tests are adequately powered, $e_1 = 0.5$, $e_2 = 0.4$, $e_3 = 0.4$.

Consider first the case of a constant effect size (Scenario 1 in Table 4.1). Since each test serves as a gatekeeper for the tests placed later in the sequence, the power of the individual tests in the fixed-sequence procedure declines fairly quickly as one progresses through the sequence. While the power of the first test is equal to its nominal value (80%), the power of the last test drops to 61% when the endpoints are moderately correlated ($\rho = 0.5$). A greater power loss is observed with the decreasing correlation among the endpoints. Furthermore, the fixed-sequence procedure is quite sensitive to the assumption that the true ordering of the endpoints (in terms of the effect sizes) is close to the actual ordering. If the first test is underpowered (Scenario 2), it creates a "domino effect" that supresses the power of the other tests. Comparing Scenario 2 to Scenario 1, the power of the last test decreases from 61% to 46% for moderately correlated endpoints and from 51% to 35% for uncorrelated endpoints. In general, the power of the fixed-sequence procedure is maximized if the outcome of the first test is very likely to be significant (see Westfall and Krishen, 2001; Huque and Alosh, 2008). This property of the fixed-sequence

TABLE 4.1: Power of the fixed-sequence and fallback procedures in a clinical trial with three endpoints as a function of the effect sizes and correlation. The fixed-sequence and fallback procedures are carried out at the one-sided 0.025 level. The weighting scheme for the fallback procedure is $w_1 = 0.5$, $w_2 = 0.25$ and $w_3 = 0.25$.

Correlation	Power of individual tests (%) (Endpoint 1, Endpoint 2, Endpoint 3)	
	Fixed-sequence procedure	Fallback procedure
Scenario 1 ($e_1 = 0.4$, $e_2 = 0.4$, $e_3 = 0.4$)		
0	(79.6, 63.4, 50.8)	(69.5, 72.3, 73.4)
0.2	(79.6, 65.2, 54.4)	(69.5, 71.7, 72.5)
0.5	(79.6, 68.1, 61.0)	(69.5, 70.7, 72.0)
Scenario 2 ($e_1 = 0.3, e_2 = 0.4, e_3 = 0.4$)		
0	(54.9, 43.9, 35.2)	(43.2, 68.3, 71.3)
0.2	(54.9, 46.1, 39.1)	(43.2, 67.5, 70.4)
0.5	(54.9, 49.4, 45.7)	(43.2, 66.1, 69.5)
Scenario 3 ($e_1 = 0.5, e_2 = 0.4, e_3 = 0.4$)		
0	(94.0, 75.0, 60.0)	(89.9, 75.0, 75.0)
0.2	(94.0, 75.7, 62.4)	(89.9, 74.8, 74.2)
0.5	(94.0, 77.1, 67.2)	(89.9, 74.7, 74.2)

procedure is illustrated in Scenario 3. It takes an overpowered test at the beginning of the sequence to bring the power of the other tests closer to its nominal level. For example, the power of the second test in the fixed-sequence procedure is only three to five percentage points lower than the nominal value (75-77% versus 80%) when the procedure successfully passes the first test 94% of the time.

Further, consider the properties of the fallback procedure based on the following weighting scheme for the three tests: $w_1 = 0.5$, $w_2 = 0.25$ and $w_3 = 0.25$. The power of the first test in the fallback procedure is uniformly lower across the three scenarios compared to the fixed-sequence procedure. This is due to the fact that the fallback procedure, unlike the fixed-sequence procedure, spends only half of the available Type I error rate on the first endpoint. The remaining fraction is distributed over the other two tests, which leads to a substantial improvement in their power in Scenarios 1 and 2. Specifically, the power of the second and third tests for the fallback procedure is much closer to the nominal level (80%) compared to the fixed-sequence procedure. Note also that, in the case of the fallback procedure, the power of individual tests stays at a constant level or increases toward the end of the sequence in all three scenarios (even when the monotonicity assumption is violated). Finally, while the power of tests placed later in the sequence improves with the increasing correlation for the fixed-sequence procedure, the fallback procedure exhibits an opposite trend. The power of the second and third tests declines slowly as the correlation among the endpoints increases

(the difference becomes very small when the first test is overpowered as in Scenario 3).

Adaptive alpha allocation approach

Li and Mehrotra (2008) proposed a multiple testing procedure, which they referred to as the *adaptive alpha allocation approach* or *4A procedure*. Consider a clinical trial with m endpoints and assume that the endpoints are grouped into two families. The first family includes m_1 endpoints that are adequately powered and the second family includes m_2 potentially underpowered endpoints ($m_1 + m_2 = m$). The endpoints in the first family are tested using any FWER controlling procedure at level $\alpha_1 = \alpha - \varepsilon$, where $\varepsilon > 0$ is small, e.g., $\alpha = 0.05$ and $\varepsilon = 0.005$. For example, the Hochberg procedure decides that all endpoints in the first family are significant if $p_{(m_1)} \leq \alpha_1$, where $p_{(m_1)}$ is the maximum p-value associated with those endpoints. The endpoints in the other family are tested using any FWER controlling procedure at level α_2, which is *adaptively* based on $p_{(m_1)}$ as follows:

$$\alpha_2(p_{(m_1)}) = \begin{cases} \alpha & \text{if} \quad p_{(m_1)} \leq \alpha_1, \\ \min\left(\alpha^*/p_{(m_1)}^2, \alpha_1\right) & \text{if} \quad p_{(m_1)} > \alpha_1, \end{cases}$$

where

$$\alpha^* = \begin{cases} \alpha_1\left(1 - \sqrt{2 - \alpha_1/m_1 - \alpha/\alpha_1}\right)^2 & \text{if} \quad \alpha_1 + \alpha_1^2/m_1 - \alpha_1^3/m_1^2 \leq \alpha, \\ \alpha_1(\alpha - \alpha_1)/(m_1 - \alpha_1) & \text{if} \quad \alpha_1 + \alpha_1^2/m_1 - \alpha_1^3/m_1^2 > \alpha. \end{cases}$$

It should be pointed out that this derivation assumes that all the p-values are independent. Li and Mehrotra also proposed an empirical adjustment to α^* if the endpoints follow a multivariate normal distribution.

An advantage of this method over the Bonferroni-based PAAS method is that the remaining endpoints are tested at a generally higher significance level, which improves their power. As an example, consider the case of two independent endpoints (Endpoints 1 and 2) and let $\alpha = 0.05$ and $\alpha_1 = 0.045$. The relationship between the significance level for Endpoint 2 (α_2) and p-value for Endpoint 1 (p_1) is depicted in Figure 4.1 (solid curve). The significance level for Endpoint 2 is 0.05 when $p_1 \leq 0.045$, 0.045 when $0.045 \leq p_1 \leq 0.081$ and remains higher than 0.0052 (the significance level for the PAAS method) when $p_1 \leq 0.244$.

Conceptually, the 4A procedure is similar to the fallback procedure except that in the latter α_2 takes only two values depending on whether $p_1 \leq \alpha_1$ or $> \alpha_1$. To compare the two procedures, consider again the clinical trial example with two independent endpoints. Figure 4.1 depicts the significance level for Endpoint 2 as a function of the p-value for Endpoint 1 for the fallback and 4A procedures. The two procedures test Endpoint 1 at the same level if $p_1 \leq 0.045$. The significance level for Endpoint 2 for the 4A method is less

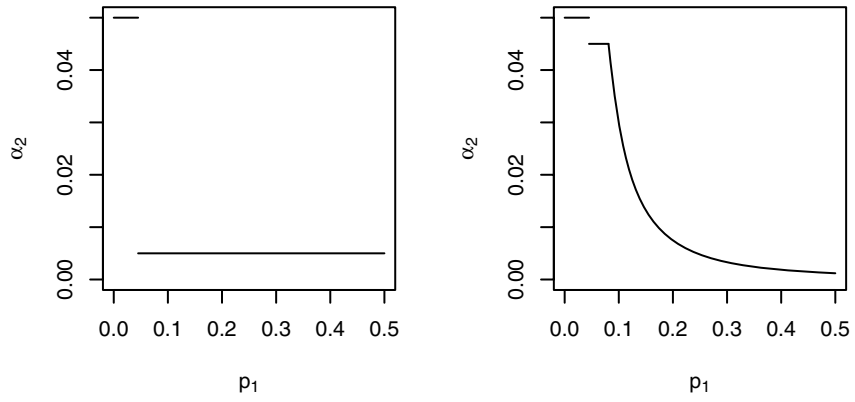

FIGURE 4.1: The significance level for Endpoint 2 (α_2) as a function of the p-value for Endpoint 1 (p_1) for the fallback procedure (left panel) and 4A procedure (right panel).

stringent than that for the fallback procedure when $0.045 \leq p_1 \leq 0.244$ and more stringent when $p_1 > 0.244$.

4.3.2 Parametric and resampling-based procedures

One of the limitations of procedures based on univariate p-values in the analysis of multiple endpoints is that they ignore the correlations among the endpoints. One can improve the power of these procedures by constructing parametric (normal theory) or resampling-based procedures that take correlations into account.

Bonferroni-type parametric procedure

Assume that the m endpoints follow a multivariate normal distribution and let t_i denote the t-statistic for testing the ith endpoint. The single-step parametric procedure is conceptually similar to the Bonferroni procedure in that the m hypotheses are tested simultaneously (i.e., in a single step). The global null hypothesis of no treatment effect is rejected if at least one test is significant, i.e., if $t_{\max} = \max(t_1, \ldots, t_m) \geq c$, where c is a critical value computed from $P\{t_{\max} < c\} = 1 - \alpha$. This calculation is performed under the global hypothesis (which is the least favorable configuration at which the

Type I error probability of this procedure is maximized over the null space). In other words, c is the $(1-\alpha)$-quantile of $t_{\max}$ when $\delta_1 = \ldots = \delta_m = 0$.

In general, this critical value is difficult to evaluate because the joint distribution of $t_1, \ldots, t_m$ (termed the *generalized multivariate t* distribution) is not known except for the case of two endpoints (Siddiqui, 1967; Gong, Pinheiro and DeMets, 2000). Note that this distribution is different from the standard multivariate t-distribution used in the Dunnett (1955) procedure because the denominator of each t_i uses a different error estimate, s_i. An additional complicating factor is that the joint distributions of both the numerators and denominators of the t_i statistics depend on the unknown correlation matrix.

Fallback-type parametric procedure

The main difficulty in computing critical values of the Bonferroni-type parametric procedure is that the m test statistics are evaluated simultaneously and one has to deal with a complicated null distribution. A stepwise method that examines the endpoints in a sequential manner considerably simplifies the process of calculating the null distributions of the test statistics and critical values. As an illustration, we will consider the stepwise procedure for multiple endpoints proposed by Huque and Alosh (2008). This procedure is a parametric extension of the fallback procedure introduced in Section 4.3.1.

Unlike the regular fallback procedure, the parametric fallback procedure takes into account the joint distribution of the test statistics associated with individual endpoints, which leads to improved power for endpoints placed later in the sequence. As before, let $t_1, \ldots, t_m$ denote the test statistics for the m endpoints and let $w_1, \ldots, w_m$ denote weights that represent the importance of the endpoints (the weights are positive and add up to 1). The test statistics are assumed to follow a standard multivariate normal distribution.

The first step involves computation of critical values $c_1, \ldots, c_m$ and significance levels $\gamma_1, \ldots, \gamma_m$ that are defined recursively using the following equations:

$$\begin{aligned} P(t_1 &\geq c_1) = \alpha w_1, \\ P(t_1 &< c_1, \ldots, t_{i-1} < c_{i-1}, t_i \geq c_i) = \alpha w_i, \quad i = 2, \ldots, m. \end{aligned}$$

The probabilities are computed under the global null hypothesis. The significance levels associated with these critical values are defined as $\gamma_i = 1 - \Phi(c_i)$, $i = 1, \ldots, m$, where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. Given these significance levels, the testing algorithm is set up as follows. The first endpoint is tested at the γ_1 level (note that $\gamma_1 = \alpha w_1$ and thus the parametric procedure uses the same level for the first endpoint as the regular procedure). At the ith step of the algorithm, the level is determined by the significance of the endpoints placed earlier in the sequence. For example, if s is the index of the last non-significant endpoint, the significance level for the ith endpoint is given by $\max(\alpha w_{s+1} + \ldots + \alpha w_i, \gamma_i)$. In

TABLE 4.2: Power of the parametric fallback procedure in a clinical trial with three endpoints as a function of the effect sizes and correlation. The procedure is carried out at the one-sided 0.025 level using the $(0.5, 0.25, 0.25)$ weighting scheme.

Correlation	Power of individual tests (%) (Endpoint 1, Endpoint 2, Endpoint 3)
Scenario 1 ($e_1 = 0.4$, $e_2 = 0.4$, $e_3 = 0.4$)	
0	(69.5, 72.4, 73.6)
0.2	(69.5, 71.9, 73.0)
0.5	(69.5, 71.5, 73.2)
Scenario 2 ($e_1 = 0.3, e_2 = 0.4, e_3 = 0.4$)	
0	(43.2, 68.6, 71.5)
0.2	(43.2, 67.9, 70.9)
0.5	(43.2, 67.6, 71.0)
Scenario 3 ($e_1 = 0.5, e_2 = 0.4, e_3 = 0.4$)	
0	(89.9, 74.9, 75.1)
0.2	(89.9, 74.9, 75.2)
0.5	(89.9, 74.9, 75.2)

other words, as in the regular procedure, the more consecutive endpoints are found significant, the higher the level for the current endpoint. However, if there is no evidence of a significant treatment effect for the $(i-1)$th endpoint, the ith endpoint is tested at the γ_i level.

To help the reader appreciate the benefits of the parametric approach, consider the example from Wiens (2003) used in Section 4.3.1. This example deals with a clinical trial with two unequally weighted endpoints ($w_1 = 0.8$ and $w_2 = 0.2$) tested at the overall two-sided $\alpha = 0.05$. Using the regular procedure, the first endpoint is tested at $\alpha w_1 = 0.04$ and, if the outcome is significant, the second endpoint is tested at the full 0.05 level. Otherwise, the significance level for the second endpoint is $\alpha w_2 = 0.01$. To apply the parametric procedure, assume that the test statistics for these two endpoints follow a standard bivariate normal distribution with the correlation coefficient ρ. The first endpoint is tested at $\gamma_1 = 0.04$ and, if a significant result is observed, the other endpoint is tested at the 0.05 level. Thus these two levels are identical to those used by the fallback procedure above. However, if the first endpoint is not significant, the level for the second endpoint can be higher than $\alpha w_2 = 0.01$. The parametric procedure tests the second endpoint at 0.0104 for $\rho = 0$, 0.0112 for $\rho = 0.3$ and 0.0146 for $\rho = 0.6$ (see Table 1 in Huque and Alosh, 2008).

Table 4.2 summarizes the power of the parametric fallback procedure in the setting described in Section 4.3.1. The weights assigned to the three tests are $w_1 = 0.5$, $w_2 = 0.25$ and $w_3 = 0.25$.

The parametric procedure is uniformly more powerful than the regular procedure in all three scenarios but, in general, the two power functions are

quite close to each other (the difference is less than two percentage points). The parametric procedure exhibits the same key features as the regular procedure, e.g.,

- The parametric procedure is robust with respect to the monotonicity assumption and performs well when the first test in the sequence is underpowered.
- When the effect sizes across the tests are comparable, the power of individual tests improves toward the end of the sequence.
- The power of tests later in the sequence declines with increasing correlation.

Additional simulations performed by Huque and Alosh (2008) for the case of two hypotheses demonstrated that the power of the parametric procedure is comparable to that of the regular procedure when the test statistics are uncorrelated or effect sizes are equal regardless of the weighting scheme. The power advantage of the parametric procedure for the second test increases with the increasing correlation when the effect size of the second test is greater than that of the first test.

Resampling-based procedures

Given the challenges associated with single-step parametric procedures for multiple endpoints (the joint distribution of the test statistics depends on an unknown correlation matrix), one can consider an alternative approach that uses the resampling-based methodology developed by Westfall and Young (1993). This alternative was explored by Reitmeir and Wassmer (1999) who introduced resampling-based versions of several single-step and stepwise procedures, e.g., Bonferroni and Hommel procedures, in the context of the multiple endpoint problem.

Along the lines of the general resampling-based method (see Section 2.8 for a detailed description of resampling-based procedures), Reitmeir and Wassmer proposed to estimate the joint distribution of the test statistics under the global null hypothesis using the bootstrap. Beginning with any multiple test, a resampling-based at-least-one procedure can be constructed using the following algorithm:

- Let p_i be the p-value for the ith endpoint, $i = 1, \ldots, m$ (this p-value is computed using a selected test).
- Generate K bootstrap samples (draw random samples with replacement of the same size as the original samples). Let $p_i(k)$ be the treatment comparison p-value for the ith endpoint, which is computed from the kth bootstrap run using the selected test, $i = 1, \ldots, m$ and $k = 1, \ldots, K$.

- Define the bootstrap multiplicity adjusted p-value for the ith endpoint as the proportion of bootstrap runs in which $p_i(k) \leq p_i$, $1, \ldots, m$.

The treatment effect for the ith endpoint is significant if the bootstrap multiplicity adjusted p-value is no greater than the pre-specified familywise error rate α. Reitmeir and Wassmer showed via simulations that the resampling-based tests for the multiple endpoint problem resulted in a consistent power gain compared to the original tests. The improvement in power was rather small for the Bonferroni test; however, a substantially larger gain was observed for some other tests, e.g., the Hommel test.

4.4 Global testing procedures

An important property of global testing procedures is that they combine evidence of treatment effect across several endpoints and thus they are more powerful than procedures for individual endpoints (provided the treatment effects are consistent across the endpoints). In this section we first consider procedures for the goal of demonstrating overall efficacy of the treatment and then describe inferences for individual endpoints when the global assessment produces a significant result.

4.4.1 Normal theory model

Global testing procedures considered in this section (with a few exceptions such as the global rank-sum procedure introduced by O'Brien, 1984) assume the normal theory model described below.

Consider a two-arm clinical trial with a parallel-group design in which the treatment group (Group 1) is tested versus the control group (Group 2). As in Lehmacher, Wassmer and Reitmeir (1991), the response of the jth patient in the ith group with respect to the kth endpoint is denoted by X_{ijk}, $i = 1, 2$, $j = 1, \ldots, n_i$, $k = 1, \ldots, m$. Let

$$\bar{X}_{i\cdot k} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ijk} \quad (i = 1, 2,\ 1 \leq k \leq m).$$

The vector of patient responses on the m endpoints, $(X_{ij1}, \ldots, X_{ijm})$, is assumed to follow a multivariate normal distribution with a mean vector $(\mu_{i1}, \ldots, \mu_{im})$ and a common covariance matrix Σ. The diagonal elements of the covariance matrix are $\sigma_k^2 = \text{var}(X_{ijk})$, $k = 1, \ldots, m$. The correlation matrix of the endpoints is denoted by R and its elements by ρ_{kl}. One may think that multiple endpoints are always highly correlated. In fact, independent endpoints are desirable because they are not proxies of each other and

thus contain more information. The correlations ρ_{kl} rarely exceed 0.6 in clinical trial applications; see Sankoh, D'Agostino and Huque (2003).

The mean treatment difference for the kth endpoint is defined as $\delta_k = \mu_{1k} - \mu_{2k}$ and it is assumed that large values of δ_k imply higher treatment efficacy. To define the test statistic for the kth endpoint, let $\bar{X}_{i\cdot k}$ denote the mean response in the ith group on the kth endpoint and let S denote the pooled sample covariance matrix. The diagonal elements of S are denoted by $s_1^2, \ldots, s_m^2$. The treatment effect for the kth endpoint is tested using the t-statistic

$$t_k = \frac{\bar{X}_{1\cdot k} - \bar{X}_{2\cdot k}}{s_k\sqrt{1/n_1 + 1/n_2}}.$$

4.4.2 OLS and GLS procedures

To motivate procedures described in this section, consider the Bonferroni procedure for multiple endpoints from Section 4.3.1. The global version of this procedure rejects the hypothesis H_I from (4.1) if

$$p_{\min} = \min(p_1, \ldots, p_m) \leq \alpha/m.$$

This procedure depends only on the smallest p-value and ignores all other p-values. Therefore it is not sensitive in the common scenario where small to modest effects are present in all endpoints.

To address this shortcoming of the Bonferroni global procedure, O'Brien (1984) considered the setup in which the multivariate testing problem is simplified by making an assumption of *a common standardized effect size*. Specifically, assume that the standardized effect sizes for the m endpoints, $\delta_1/\sigma_1, \ldots, \delta_m/\sigma_m$, are equal to, say, λ. In this case, the problem of testing the null hypothesis,

$$H_I^* : \ \delta_i = 0 \text{ for all } i,$$

reduces to a single parameter testing problem

$$H^* : \lambda = 0 \text{ versus } K^* : \lambda > 0.$$

OLS and GLS test statistics

O'Brien (1984) proposed two procedures for the hypothesis H^* based on standardized responses, $Y_{ijk} = X_{ijk}/\sigma_k$. Under the simplifying assumption of a common effect size, one can consider the following regression model for the standardized responses:

$$Y_{ijk} = \frac{\mu_k}{\sigma_k} + \frac{\lambda}{2} I_i + e_{ijk},$$

where $i = 1, 2$, $j = 1, \ldots, n_i$, $k = 1, \ldots, m$, $\mu_k = (\mu_{1k} + \mu_{2k})/2$, $I_i = +1$ if $i = 1$ and -1 if $i = 2$, and e_{ijk} is $N(0, 1)$ distributed error term with $\text{corr}(e_{ijk}, e_{i'j'k'}) = \rho_{kk'}$ if $i = i'$ and $j = j'$, and 0 otherwise.

The first procedure developed by O'Brien is based on the ordinary least squares (OLS) estimate of the common effect size λ while the second procedure is based on the generalized least squares (GLS) estimate. Let $\widehat{\lambda}_{\text{OLS}}$ and $\text{SE}(\widehat{\lambda}_{\text{OLS}})$ denote the OLS estimate of λ and its sample standard error, respectively. It can be shown that the OLS test statistic for H^* is given by

$$t_{\text{OLS}} = \frac{\widehat{\lambda}_{\text{OLS}}}{\text{SE}(\widehat{\lambda}_{\text{OLS}})} = \frac{J't}{\sqrt{J'\widehat{R}J}},$$

where J is an m-vector of all 1's and $t = (t_1, \ldots, t_m)'$ is the vector of t-statistics defined in Section 4.4.1.

Since the error terms e_{ijk} in the regression model for the standardized responses are correlated, it may be preferable to use the GLS estimate of λ, which leads to the following test statistic for H^*

$$t_{\text{GLS}} = \frac{\widehat{\lambda}_{\text{GLS}}}{\text{SE}(\widehat{\lambda}_{\text{GLS}})} = \frac{J'\widehat{R}^{-1}t}{\sqrt{J'\widehat{R}^{-1}J}}.$$

It is instructive to compare the OLS and GLS test statistics. Both testing procedures assess the composite effect of multiple endpoints by aggregating the t-statistics for the individual endpoints. In the case of the OLS test statistic, the t-statistics are equally weighted, while the GLS test statistic assigns unequal weights to the t-statistics. The weights are determined by the sample correlation matrix $\widehat{R}$. If a certain endpoint is highly correlated with the others, it is not very informative, so the GLS procedure gives its t-statistic a correspondingly low weight. A downside of this approach is that the weights can become negative. This leads to anomalous results, e.g., it becomes possible to reject H^* even if the treatment effect is negative on all the endpoints.

In order to compute critical values for the OLS and GLS procedures, one needs to derive the null distributions of their test statistics. For large sample sizes, the OLS and GLS statistics approach the standard normal distribution under H^*, but the approach of the GLS statistic is slower since it has the random matrix $\widehat{R}$ both in the numerator and denominator. For small sample sizes the standard normal distribution provides a liberal test of H^*. The exact small sample null distributions of t_{OLS} and t_{GLS} are not known. O'Brien (1984) proposed a t-distribution with $\nu = n_1 + n_2 - 2m$ df as an approximation. This approximation is exact for $m = 1$ but conservative for $m > 1$. Logan and Tamhane (2004) proposed the approximation, $\nu = 0.5(n_1 + n_2 - 2)(1 + 1/m^2)$, which is more accurate.

Logan and Tamhane (2004) also extended the OLS and GLS procedures to the heteroscedastic case (case of unequal Σs). Note that the heteroscedastic extension of the GLS test statistic given in Pocock, Geller and Tsiatis (1987) does not have the standard normal distribution under H^* as claimed there, and hence should not be used.

TABLE 4.3: Summary of the results of the osteoarthritis trial (SD, standard deviation).

Endpoint	Summary statistic	Treatment $n = 88$	Placebo $n = 90$
Pain subscale	Mean	59	35
	Pooled SD	96	96
Physical function subscale	Mean	202	111
	Pooled SD	278	278

Osteoarthritis trial example

To illustrate the use of global testing procedures, consider a clinical trial for the treatment of osteoarthritis. The study was conducted to evaluate the effects of a treatment on two endpoints, the pain and physical function subscales of the Western Ontario and McMaster Universities (WOMAC) Osteoarthritis Index (Bellamy, 2002), compared to placebo. The efficacy analysis was based on the mean changes in the two endpoints during a 6-week study period. The results of the study are shown in Table 4.3.

The OLS procedure was carried out to assess the overall efficacy of the treatment (note that the GLS procedure is equivalent to the OLS procedure in the case of two endpoints). The t-statistics for the pain and physical function endpoints and sample correlation matrix were

$$\begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = \begin{bmatrix} 1.67 \\ 2.18 \end{bmatrix}, \quad \widehat{R} = \begin{bmatrix} 1 & 0.36 \\ 0.36 & 1 \end{bmatrix}.$$

Given this information, it is easy to compute the OLS/GLS test statistic,

$$t_{\mathrm{OLS}} = t_{\mathrm{GLS}} = \frac{1.67 + 2.18}{\sqrt{1 + 0.36 + 0.36 + 1}} = 2.334.$$

Using the Logan-Tamhane formula, the one-sided p-value associated with this test statistic is 0.0107 (based on 110 df), which is significant at the one-sided 0.025 level.

It is worth noting that the OLS procedure becomes less powerful as the correlation between two endpoints increases. In fact, the OLS statistic would not be significant at the one-sided 0.025 level in the osteoarthritis trial if the sample correlation coefficient was greater than 0.91. The reason is that higher correlations imply correspondingly less independent information in the endpoints.

Power calculations for OLS and GLS procedures

From their construction, it is clear that the OLS and GLS procedures will be powerful when all endpoints have a similar positive effect, but in other

situations, they may lack power. Dallow, Leonov and Roger (2008) considered the problem of power and sample size calculations for the OLS and GLS procedures and introduced a simple measure, termed the *operational effect size*, that helps to quantify the composite effect of multiple endpoints. Let $\lambda_k = \delta_k/\sigma_k$ be the true standardized effect size for the kth endpoint and let $\lambda = (\lambda_1, \ldots, \lambda_m)'$. Under the alternative hypothesis, the distributions of t_{OLS} and t_{GLS} can be approximated by noncentral t-distributions with noncentrality parameters given by

$$\Delta_{\mathrm{OLS}} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{J'\lambda}{\sqrt{J'RJ}} \text{ and } \Delta_{\mathrm{GLS}} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{J'R^{-1}\lambda}{\sqrt{J'R^{-1}J}}.$$

Given this, Dallow et al. defined the operational effect sizes of the two procedures as

$$\Lambda_{\mathrm{OLS}} = \frac{J'\lambda}{\sqrt{J'RJ}} \text{ and } \Lambda_{\mathrm{GLS}} = \frac{J'R^{-1}\lambda}{\sqrt{J'R^{-1}J}}.$$

The two quantities serve the same role in the problem of testing multiple endpoints as the regular effect size in a single-endpoint problem. It is easy to show that Λ_{OLS} and Λ_{GLS} are in fact the standardized effect sizes for an equivalent single endpoint which would have the same overall power for the same sample sizes, n_1 and n_2.

Operational effect sizes help establish the relationship between key properties of multiple endpoints (e.g., λ and R) and the power of the two global procedures. For example, the numerator in Λ_{OLS} is the sum of the effect sizes for individual endpoints and thus the power of the OLS procedure is an increasing function of each λ_k. Further, the denominator depends on the correlations among the endpoints. It is clear that, as the endpoints become more correlated and thus less informative, the operational effect size decreases in magnitude. In addition, operational effect sizes facilitate power comparisons between the OLS and GLS procedures. Dallow et al. proved that the GLS procedure is more powerful than the OLS procedure when the effect sizes are equal across the endpoints ($\lambda_1 = \ldots = \lambda_m$) but, in general, the power of the GLS procedure can be lower than that of the OLS procedure. This happens, for example, when the effect sizes of strongly correlated endpoints are large and the effect sizes of weakly correlated endpoints are small.

Several authors have reported results of simulation studies to assess the power of the OLS and GLS procedures under various configurations of effect sizes and correlation values. The simulation study by Reitmeir and Wassmer (1996) showed that the power of the OLS procedure was comparable to that of the GLS procedure. Dallow et al. (2008) demonstrated that the GLS procedure is biased when the sample size is small to moderate which complicates the power comparison of the OLS and GLS procedures for $n \leq 40$. For larger sample sizes, the difference between the power functions of the two procedures is very small.

Power and sample size calculations can be performed using a normal approximation by specifying λ and R. Suppose, for example, that we are inter-

ested in computing the number of patients in each arm of a two-arm trial. The one-sided Type I error rate is given by α and the desired power of a global procedure (OLS or GLS) is set at $1-\beta$. As shown by Dallow et al., the sample size in each arm is given by the familiar formula,

$$n = \frac{2(z_\alpha + z_\beta)^2}{\Lambda^2},$$

where z_x is the $(1-x)$-quantile of the standard normal distribution and Λ is the operational effect size of the global procedure chosen in this trial.

As an illustration, we will return to the osteoarthritis clinical trial example and compute the sample size required to achieve 90% power at a one-sided 0.025 level ($\alpha = 0.025$ and $\beta = 0.1$). Based on the results displayed in Table 4.3, assume that the standardized effect sizes for the pain and physical function endpoints are 0.25 and 0.33, respectively, and the correlation coefficient is 0.36. Under these assumptions, the operational effect size for the OLS and GLS procedures is

$$\Lambda_{\text{OLS}} = \Lambda_{\text{GLS}} = \frac{0.25 + 0.33}{\sqrt{1 + 0.36 + 0.36 + 1}} = 0.35$$

and thus

$$n = \frac{2(1.96 + 1.28)^2}{0.35^2} \simeq 170$$

patients per arm need to be enrolled in the study.

Given that the GLS procedure does not generally dominate the OLS procedure in terms of power and because of the added difficulties caused by negative weights in the GLS statistic, we recommend the use of the OLS procedure in clinical trials.

Nonparametric global procedures

The OLS and GLS procedures can be formulated for non-normal responses as long as the test statistics for the m endpoints follow a multivariate normal distribution in large samples. Examples include binary and time-to-event variables (Pocock, Geller and Tsiatis, 1987). However, if the assumption of multivariate normality cannot be made, one can consider a nonparametric version of the OLS procedure proposed by O'Brien (1984). In this procedure, the data from the two groups are pooled and ranked on each endpoint separately as in the Wilcoxon rank-sum test. Let r_{ijk} be the rank of X_{ijk} in the pooled sample. Then a two-sample t-test is performed on the summed ranks,

$$r_{ij} = \sum_{k=1}^{m} r_{ijk}, \quad i = 1, 2,\ j = 1, \ldots, n_i.$$

This procedure offers a viable alternative to the OLS procedure particularly if the data are non-normal. For example, this global rank-sum procedure was

used in the azithromycin study in patients with coronary artery disease (Anderson et al., 1999). The procedure was chosen to evaluate the overall effect of the treatment on four inflammatory markers because the change scores were not expected to follow a normal distribution.

4.4.3 Likelihood ratio and other procedures

This section gives a review of a class of global procedures based on the likelihood ratio (LR) principle and two other global procedures (the Läuter and Follmann procedures). This review is rather brief because these procedures are not commonly used in clinical trial applications due to the limitations discussed below.

Exact likelihood ratio procedures

It will be assumed in this section that the point null and one-sided alternative hypotheses are given by (4.2). Kudô (1963) was the first to derive an exact one-sided LR procedure for the one-sample problem when the covariance matrix Σ is known. This procedure can be readily extended to the two-sample problem. Perlman (1969) extended the Kudô procedure to the case of an unknown covariance matrix but the null distribution of the resulting test statistic is not free of Σ and the procedure is biased. However, Perlman provided sharp lower and upper bounds on the null distribution that are free of Σ. Wang and McDermott (1998) solved the problem of dependence on unknown Σ by deriving an LR procedure conditional on the sample covariance matrix S.

These procedures are not commonly used because they are not easy to implement computationally. There is, however, a more basic problem with the LR procedures that they can reject H_I^* even when the vector of mean treatment differences has all negative elements (Silvapulle, 1997). These procedures are also nonmonotone in the sense that if the differences $\bar{X}_{1\cdot k} - \bar{X}_{2\cdot k}$ become more negative the test statistic can get larger.

Perlman and Wu (2002) showed that these difficulties are caused by the point null hypothesis. Basically, the LR procedure compares the *ratio* of the likelihood under K_U^* versus that under H_I^*. The apparent nonmonotonicity of the LR procedure results because, in some cases, as the sample outcomes move deeper into the part of the sample space corresponding to K_U^*, their likelihood under K_U^* increases, but so does their likelihood under H_I^*, and their ratio gets smaller. This is not a defect of the LR procedure, but rather that of the null hypothesis not being correctly specified. If the null hypothesis is defined as a full complement of K_U^* then the LR procedure no longer has these difficulties. However, computation of the test statistic under the complete null hypothesis and its null distribution are problematic.

Approximate LR procedures were proposed in the literature to circumvent the computational and analytical difficulties of the exact LR procedure; see, for example, Tang, Gnecco and Geller (1989) and Tamhane and Logan (2002).

These procedures, although easier to apply, suffer from the same anomalies that the exact LR procedures suffer because of the misspecification of the null hypothesis as a point null hypothesis.

Cohen and Sackrowitz (1998) proposed the cone-ordered monotone (COM) criterion to overcome the nonmonotonicity problem. However, their COM procedure is not entirely satisfactory either since, e.g., in the bivariate case, it can reject H_I^* if one mean difference is highly negative as long as the other mean difference is highly positive.

Läuter exact procedure

Läuter (1996) and Läuter, Glimm and Kropf (1996) proposed a class of test statistics having the property that they are exactly t-distributed under the point null hypothesis. To define the procedure, let $\bar{X}_{\cdot\cdot k}$ denote the overall sample mean for the kth endpoint, i.e.,

$$\bar{X}_{\cdot\cdot k} = \frac{1}{n_1 + n_2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} X_{ijk}.$$

Consider the total cross-products matrix V with elements

$$v_{kl} = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (X_{ijk} - \bar{X}_{\cdot\cdot k})(X_{ijl} - \bar{X}_{\cdot\cdot l}), \quad k, l = 1, \ldots, m.$$

Let $w = w(V)$ be any m-dimensional vector of weights depending only on V and $w \neq 0$ with probability 1. Läuter (1996) showed that

$$t_w = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\frac{w't}{\sqrt{w'Sw}} \right)$$

is t-distributed with $n_1 + n_2 - 2$ df under the point null hypothesis. Various choices of w were discussed by Läuter et al. (1996). The simplest among them is $w_k = 1/\sqrt{v_{kk}}$. The resulting statistic is called the standardized sum (SS) statistic (denoted by t_{SS}) which can be expressed as

$$t_{SS} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\frac{\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}}{s_y} \right),$$

where $\bar{Y}_{i\cdot}$ is the average of the standardized observations

$$Y_{ij} = \sum_{k=1}^{m} \frac{X_{ijk}}{\sqrt{v_{kk}}}, \quad i = 1, 2, \quad j = 1, \ldots, n_i.$$

Analytical and simulated power comparisons made in Logan and Tamhane (2004) showed that the OLS procedure is more powerful than the Läuter procedure when only a few endpoints have an effect. In fact, if only one endpoint

has an effect, which tends to infinity, the power of the Läuter procedure remains bounded strictly below 1, whereas the power of the OLS procedure tends to 1, as it should. Frick (1996) also noted this drawback of the Läuter procedures but argued that such a scenario is unlikely in practice. When all endpoints have roughly equal effects, the powers of the OLS and Läuter procedures are comparable.

The reason for the lack of power of the SS test statistic when only a few endpoints have an effect is that it standardizes the data on each endpoint by its pooled *total* group sample standard deviation and then computes an overall t-statistic. The pooled standard deviation overestimates the true standard deviation since it includes the differences between the means of the treatment and control groups which diminishes the power of the Läuter procedure. On the other hand, the OLS statistic is the sum of t-statistics obtained by standardizing the individual endpoints by their pooled *within* group sample standard deviations.

Follmann procedure

Follmann (1996) proposed an ad-hoc procedure which is simple to apply: Reject the point null hypothesis if the two-sided Hotelling's T^2-test is significant at the 2α-level and the average endpoint mean difference is positive,

$$\sum_{k=1}^{m}(\bar{X}_{1\cdot k} - \bar{X}_{2\cdot k}) > 0.$$

Unfortunately, the alternative for which this procedure is designed,

$$\sum_{k=1}^{m}(\mu_{1k} - \mu_{2k}) > 0,$$

is not very meaningful since it depends on the scaling used for the endpoints.

4.4.4 Procedures for individual endpoints

As was explained earlier in this section, global procedures are aimed at an overall evaluation of the treatment effect. However, if the overall treatment effect is positive, the trial's sponsor is likely to be interested in examining the treatment effect on individual endpoints/components. As pointed out in Chapter 1, the LIFE trial (Dahlöf et al., 2002) serves as an example of a trial in which the analysis of individual endpoints provided important insights into the nature of treatment benefit. This analysis revealed that the overall treatment effect was driven mainly by one component (stroke endpoint). Here we review extensions of global testing methods that can be used to perform multiplicity-adjusted univariate inferences for individual endpoints after a significant global result.

Lehmacher, Wassmer and Reitmeir (1991) applied the closure principle to make inferences on individual endpoints. As was explained in Section 2.3.3, this principle is a popular tool used in the construction of multiple testing procedures. For example, the Holm procedure is a closed procedure for testing individual hypotheses which is derived from the Bonferroni global procedure (see Section 4.4.2). A similar approach can be used to construct a closed testing procedure for testing individual endpoints based on any other global procedure. To accomplish this, one needs to consider all possible combinations of the m endpoints and test each combination using an α-level global procedure subject to the closure principle, i.e., if the procedure for any combination is not significant then all of its subset combinations are declared nonsignificant without testing them. The treatment effect on an endpoint is significant at level α if the global procedures for all the combinations including the selected endpoint are significant at this level.

This approach can be used with any α-level global procedure for testing different intersections. Lehmacher et al. constructed a procedure for testing individual endpoints based on the OLS and GLS procedures described in Section 4.4.2. Wang (1998) applied the Follmann procedure introduced in Section 4.4.3 as the global procedure and found the performance comparable to the Westfall-Young resampling procedures (Section 2.8). Logan and Tamhane (2001) proposed a hybrid approach that uses a combination of global procedures, each one of which is powerful against a different alternative. Specifically, the Logan-Tamhane hybrid procedure consists of the Bonferroni global procedure based on the smallest p-value and the OLS procedure. Here the former procedure is powerful against alternatives where only a few endpoints have large effects while the latter procedure is powerful against alternatives where all endpoints have small to modest effects. The hybrid test statistic for each intersection hypothesis is the minimum of the p-values for the Bonferroni and OLS procedures. A bootstrap method is used to estimate the adjusted p-value for this complex statistic. The resulting procedure has stable and high power against a range of alternatives (i.e., the procedure is robust), but is computationally more intensive.

Osteoarthritis trial example

A clinical study with two endpoints (pain and physical function) was considered in the osteoarthritis trial example (Section 4.4.2). The overall treatment effect of the two endpoints was evaluated using the OLS procedure. The global procedure was significant at the one-sided 0.025 level. We will apply the closure principle to assess the treatment effect on each endpoint and see whether the overall positive result was driven by both components or only one. First, we need to compute p-values for the endpoint-specific tests. From Section 4.4.2, the t-statistics for the pain and physical function endpoints are 1.67 and 2.18, respectively, with $n_1 + n_2 - 2 = 172$ df. The one-sided p-values associated with the t-statistics are 0.0486 and 0.0152 applying the

Logan-Tamhane formula for degrees of freedom. Using the closure principle, the multiplicity adjusted p-value for each endpoint is the larger of the p-value for the OLS procedure and the p-value for the endpoint-specific t-test. The treatment's effect on the pain endpoint is not significant at the one-sided 0.025 level ($p = 0.0486$) whereas the effect on the other endpoint is significant ($p = 0.0152$). It is worth remembering that it is possible for endpoint-specific tests to be non-significant even if the global procedure is highly significant.

4.5 All-or-none procedures

As was explained in Section 4.2.3, the goal of demonstrating the efficacy of the treatment on *all* endpoints requires an all-or-none or IU procedure (Berger, 1982) of the union of individual hypotheses, H_i. The all-or-none procedure has the following form:

$$\text{Reject all hypotheses if } t_{\min} = \min_{1 \le i \le m} t_i \ge t_\alpha(\nu),$$

where $t_\alpha(\nu)$ is the $(1-\alpha)$-quantile of the t-distribution with $\nu = n_1 + n_2 - 2$ df. This procedure is popularly known as the min test (Laska and Meisner, 1989).

Since this procedure does not use a multiplicity adjustment (each hypothesis H_i is tested at level α), it may appear at first that it must be highly powerful as a test of the global hypothesis H_U. In reality, the min test is very conservative because of the requirement that *all* hypotheses must be rejected at level α. The conservatism results from the least favorable configuration of the min test which can be shown to be of the following form:

- No treatment effect for any one endpoint ($\delta_i = 0$ for some i).
- Infinitely large treatment effects for all other endpoints ($\delta_j \to \infty$ for $j \ne i$).

This configuration leads to marginal α-level t-tests.

Figure 4.2, based on Offen et al. (2007), will help the reader appreciate how conservative the min test can be. This figure gives the multipliers to calculate the sample size required to guarantee 80% power for the min test if the base sample size guarantees 80% power for each endpoint. The calculation was done under the assumption of equicorrelated endpoints and a common standardized effect size for all endpoints. As can be seen from the table, the multiplier increases as the number of endpoints increases and the assumed common correlation between them decreases. Consider, for example, the clinical trial in patients with Alzheimer's disease described by Offen and Helterbrand (2003). It is commonly required that a new treatment should demonstrate a significant

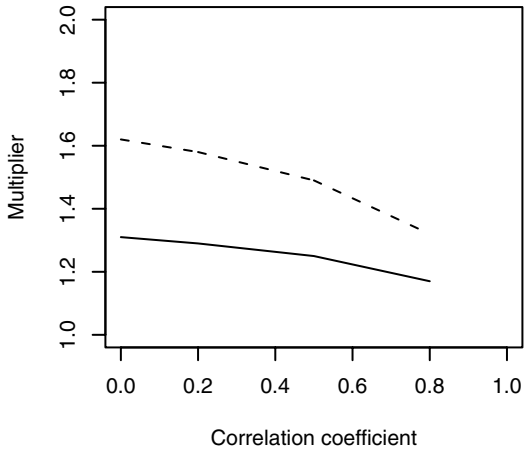

FIGURE 4.2: Sample size multipliers for the min test for two endpoints (solid curve) and four endpoints (dashed curve) as a function of the common correlation in all-or-none testing problems. Multiplier equals 1 for a single endpoint.

effect on at least two endpoints, e.g., a cognition endpoint (Alzheimer's Disease Assessment Scale-Cognitive Subpart) and a clinical global scale (Clinician's Interview-Based Impression of Change). The correlation between these two endpoints is usually around 0.2 and thus the sample size multiplier is 1.29 which corresponds to almost a 30% increase in the sample size. In other cases, e.g., when four weakly correlated endpoints are considered, the multiplier is 1.58 meaning that the sample size needs to be increased by almost 60% compared to the single-endpoint case.

It is important to note that the least favorable configuration for the min test is clinically not plausible. The global hypothesis H_U permits configurations with infinitely large positive effects on some endpoints and negative effects on others. However, it is uncommon for treatments to have substantially different effects on the endpoints. This has led researchers to put restrictions on the global hypothesis in order to develop more powerful versions of the min test.

Hochberg and Mosier (2001) suggested restricting H_U to the negative quadrant,

$$\bigcap_{k=1}^{m} (\delta_k \leq 0),$$

in which case the least favorable configuration is the overall null configuration, $\delta_1 = \ldots = \delta_m = 0$. Chuang-Stein et al. (2007) restricted the hypothesis to the subset of the global hypothesis which satisfies

$$\bigcap_{k=1}^{m} (-\varepsilon_k \leq \delta_k \leq \varepsilon_k),$$

where the thresholds ε_k, $k = 1, \ldots, m$, are prespecified based on clinical considerations. A similar approach, but based on estimated mean differences, $\bar{X}_{1\cdot k} - \bar{X}_{2\cdot k}$, was proposed by Snapinn (1987). Cappizi and Zhang (1996) suggested another alternative to the min test which requires that the treatment be shown effective at a more stringent significance level α_1 on say $m_1 < m$ endpoints and at a less stringent significance level $\alpha_2 > \alpha_1$ on the remaining $m_2 = m - m_1$ endpoints. For $m = 2$, they suggested this rule for $m_1 = m_2 = 1$, $\alpha_1 = 0.05$ and $\alpha_2 = 0.10$ or 0.20. However, as pointed out by Neuhäuser, Steinijans and Bretz (1999), this rule does not control the FWER at $\alpha = 0.05$.

Another approach to this formulation adopts a modified definition of the error rate to improve the power of the min test in clinical trials with several endpoints. Chuang-Stein et al. (2007) considered an error rate definition based on the average Type I error rate over the null space and developed a procedure that adjusts significance levels for the individual endpoints to control the average Type I error rate at a prespecified level.

This is a relatively new research area and further work is required to assess the utility of the methods described in this section and their applicability.

4.6 Superiority-noninferiority procedures

There are many situations in which the requirement that the treatment be superior to the control on all endpoints (all-or-none procedures in Section 4.5) is often too strong and the requirement that the treatment be superior to the control on at least one endpoint (at-least-one procedures in Section 4.3) is too weak. The superiority-noninferiority approach discussed in this section strengthens the latter requirement by augmenting it with the additional requirement that the treatment is not inferior to the control on all other endpoints.

Consider a clinical trial with m endpoints and suppose that its objective is to demonstrate the treatment is superior to the control on at least one endpoint and noninferior to the control on all other endpoints. Note that if superiority is established for both endpoints, the second part of this requirement (noninferiority) becomes redundant. This formulation of the superiority-noninferiority testing problem was considered by Bloch, Lai and Tubert-Bitter (2001) and Tamhane and Logan (2004). The null and alternative hypotheses

for the superiority-noninferiority problem are defined in Section 4.2.4:

$$H_U^{(SN)} = H_I^{(S)} \cup H_U^{(N)} \text{ versus } K_I^{(SN)} = K_U^{(S)} \cap K_I^{(N)}.$$

The trial's outcome is declared positive if there is evidence of superior efficacy for at least one endpoint ($K_U^{(S)}$) and noninferior efficacy for all endpoints ($K_I^{(N)}$).

As an illustration, consider a clinical trial with two endpoints. The region corresponding to the alternative hypothesis $K_I^{(SN)}$ with $\varepsilon_1 = 5$, $\varepsilon_2 = 7$, $\eta_1 = 3$ and $\eta_2 = 4$ is shown in Figure 4.3.

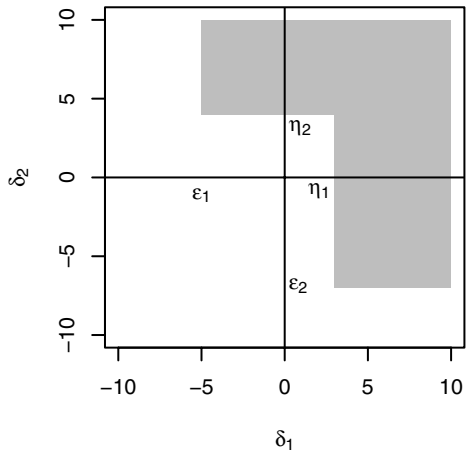

FIGURE 4.3: Region in the parameter space corresponding to the alternative hypothesis (shaded area) in a superiority-noninferiority testing problem with two endpoints.

Tamhane-Logan superiority-noninferiority procedure

Denote the t-statistics for superiority and noninferiority for the kth endpoint by

$$t_k^{(S)} = \frac{\bar{X}_{1\cdot k} - \bar{X}_{2\cdot k} - \eta_k}{s_k\sqrt{1/n_1 + 1/n_2}}, \quad t_k^{(N)} = \frac{\bar{X}_{1\cdot k} - \bar{X}_{2\cdot k} + \varepsilon_k}{s_k\sqrt{1/n_1 + 1/n_2}}.$$

Tamhane and Logan (2004) used the UI statistic

$$t_{\max}^{(S)} = \max(t_1^{(S)}, \ldots, t_m^{(S)})$$

for testing the superiority null hypothesis $H_I^{(S)}$ and the IU statistic

$$t_{\min}^{(N)} = \min(t_1^{(N)}, \ldots, t_m^{(N)})$$

for testing the noninferiority null hypothesis $H_U^{(N)}$. They proposed the following procedure of the global superiority-noninferiority hypothesis:

$$\text{Reject } H_U^{(SN)} \text{ if } t_{\max}^{(S)} \geq c^{(S)} \text{ and } t_{\min}^{(N)} \geq c^{(N)},$$

where the critical values $c^{(S)}$ and $c^{(N)}$ are chosen so that the procedure has level α. Bloch et al. (2001) used the Hotelling T^2-statistic for testing superiority in place of $t_{\max}^{(S)}$; however, as noted before, the T^2-statistic is not very powerful against one-sided superiority alternative. Perlman and Wu (2004) used Perlman's one-sided LR statistic instead of the T^2-statistic.

According to the intersection-union testing principle, the superiority and noninferiority tests must be of level α. Conservative values for $c^{(S)}$ and $c^{(N)}$ can be chosen to be the $(1-\alpha/m)$- and $(1-\alpha)$-quantiles of the t-distribution, respectively, with $\nu = n_1 + n_2 - 2$ df. The exact value of $c^{(S)}$ involves the generalized multivariate t distribution and thus, as was explained in Section 4.3.2, is difficult to evaluate. Also, a sharper critical constant $c^{(S)}$ can be evaluated by conditioning on the event that the noninferiority test is passed by all endpoints. However, the resulting value of $c^{(S)}$ needs to be evaluated by using bootstrap; see Bloch et al. (2001) and Tamhane and Logan (2004). Röhmel et al. (2006) objected to this conditioning arguing that it causes significance of the superiority test to be influenced by changes in the noninferiority margin for which there is no clinical justification. However, Logan and Tamhane (2008) showed that passing the noninferiority test at a more stringent margin adds more credence to the alternative hypothesis $K_U^{(S)}$ that the treatment is superior on at least one endpoint, and therefore it is retained more easily. If $H_U^{(SN)} = H_I^{(S)} \cup H_U^{(N)}$ is rejected then it is of interest to know which endpoints demonstrate superiority of the treatment over the control. Logan and Tamhane (2008) gave a closed procedure for this purpose which controls FWER for the family of $H_U^{(SN)}$ as well as the endpoint-specific superiority null hypotheses $H_k^{(S)}$. This closed procedure can be implemented in $m+1$ steps.

Alzheimer's disease trial example

The Alzheimer's disease example from Section 4.5 will be used to illustrate key properties of the Tamhane-Logan superiority-noninferiority procedure. Table 4.4 displays results of a 24-week study in patients with Alzheimer's disease that tested the efficacy and safety of an experimental treatment compared to placebo. The efficacy profile of the treatment was evaluated using two co-primary endpoints, Alzheimer's Disease Assessment Scale-Cognitive

TABLE 4.4: Summary of the results of Alzheimer's disease trial (SD, standard deviation).

Endpoint	Summary statistic	Treatment $n = 167$	Placebo $n = 161$
ADAS-Cog	Mean	0.5	2.5
	Pooled SD	7.4	7.4
CIBIC-Plus	Mean	4.2	4.4
	Pooled SD	1.1	1.1

Subpart (ADAS-Cog) and Clinician's Interview-Based Impression of Change (CIBIC-Plus).

Suppose that the superiority margins are set at 0 and the noninferiority margins for the ADAS-Cog and CIBIC-Plus endpoints at 0.8 and 0.1, respectively. Based on this information, the superiority and noninferiority test statistics are given by

$$\begin{bmatrix} t_1^{(S)} \\ t_2^{(S)} \end{bmatrix} = \begin{bmatrix} 2.45 \\ 1.65 \end{bmatrix}, \quad \begin{bmatrix} t_1^{(N)} \\ t_2^{(N)} \end{bmatrix} = \begin{bmatrix} 3.43 \\ 2.47 \end{bmatrix}.$$

Further, assuming a one-sided $\alpha = 0.025$, the critical values of the Tamhane-Logan procedure are

$$c^{(S)} = t_{0.0125}(326) = 2.25 \text{ and } c^{(N)} = t_{0.025}(326) = 1.97.$$

The rejection region of the superiority-noninferiority procedure is shown in Figure 4.4. The superiority and noninferiority test statistics are in their corresponding rejection regions, i.e.,

$$\max(t_1^{(S)}, t_2^{(S)}) \geq 2.25 \text{ and } \min(t_1^{(N)}, t_2^{(N)}) \geq 1.97,$$

and thus the procedure rejects the global superiority-noninferiority hypothesis. To determine which individual endpoints demonstrate superiority of the treatment, $t_1^{(S)}$ and $t_2^{(S)}$ are compared with $t_{0.025}(326) = 1.97$. Since only $t_1^{(S)}$ exceeds this critical constant, superiority is demonstrated only on ADAS-Cog; noninferiority is demonstrated on CIBIC-Plus.

Note that the Tamhane-Logan procedure is monotone in the sense that, if a certain set of test statistics leads to the rejection of the global superiority-noninferiority hypothesis, the hypothesis will be rejected for any set of more extreme test statistics.

It is instructive to compare the Tamhane-Logan superiority-noninferiority procedure to the min test used in the all-or-none testing problem described in Section 4.5. The min test rejects the global superiority hypothesis if

$$\min(t_1^{(S)}, t_2^{(S)}) \geq t_{0.025}(326) = 1.97.$$

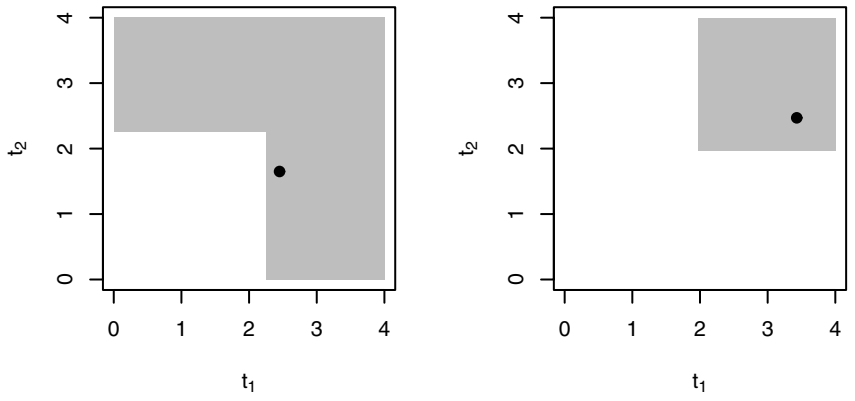

FIGURE 4.4: Left panel: Rejection region of the superiority component of the Tamhane-Logan procedure (shaded area) and superiority test statistics (dot). Right panel: Rejection region of the noninferiority component of the Tamhane-Logan procedure (shaded area) and noninferiority test statistics (dot).

The rejection region of the min test is displayed in Figure 4.5. It is clear that the global superiority hypothesis cannot be rejected since the treatment effect for CIBIC-Plus is not significant at the one-sided 0.025 level ($t_2^{(S)} < 1.97$). This serves as an illustration of the fact that the all-or-none testing approach is based on more stringent criteria compared to the superiority-noninferiority testing approach.

4.7 Software implementation

This section briefly describes software implementation of the multiple and global procedures discussed in this chapter. The following SAS programs were used in the examples included in this chapter. The programs can be downloaded from the book's Web site (http://www.multxpert.com).

- Program 2.1 can be used to implement at-least-one procedures for identifying the treatment effect on individual endpoints described in Section 4.3, including the Bonferroni, fixed-sequence and fallback proce-

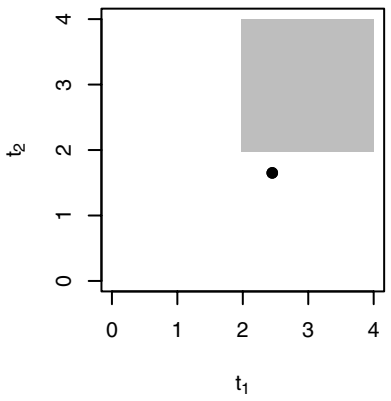

FIGURE 4.5: Rejection region of the min test (shaded area) and superiority test statistics (dot).

dures (note that this program also implements other p-value-based procedures introduced in Section 2.6). Software implementation of the parametric fallback and 4A procedures is not currently available. Critical values for these procedures are tabulated in Huque and Alosh (2008) and Li and Mehrotra (2008), respectively.

- Program 4.1 implements the OLS and GLS procedures in the osteoarthritis trial example introduced in Section 4.4.2. Program 4.2 performs sample size calculations for the OLS and GLS procedures in this clinical trial example.

- Program 4.3 implements the Tamhane-Logan superiority-noninferiority procedure in the Alzheimer's disease trial example (Section 4.6).

Acknowledgements

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