Chapter 7
Statistical Testing of Single and Multiple Endpoint Hypotheses in Group Sequential Clinical Trials

Mohammad Huque, Sirisha Mushti and Mohamed Alosh

7.1 Introduction

It is well recognized that a clinical trial of fixed-sample design planned without interim looks can falsely reject a hypothesis of no treatment effect on an endpoint by chance alone. This error commonly known as the false positive error or the Type I error can be excessive if the trial tests more than one hypothesis in the same study. This inflation of the Type I error is of concern as it can lead to false conclusions of treatment benefits in a trial. However, many statistical approaches for confirmatory clinical trials are now available for keeping the probability of falsely rejecting any hypothesis in testing a family of hypotheses (i.e., the familywise Type I error rate) controlled to a specified level; see, for example, a recently released FDA draft guidance "Multiple Endpoints in Clinical Trials," and Alosh et al. (2014).

However, many confirmatory clinical trials accrue patients over many months and enroll hundreds to thousands of patients; this is a widespread practice, for example, for some cardiovascular and oncology trials. Investigators, bound by ethical and economic constraints, usually design these large trials with interim looks, with the possibility of stopping the trial early at an interim stage if the study treatment has the desired efficacy that is clinically relevant, or if it is futile to continue the study, either for lack of efficacy of the study treatment or for safety concerns. These clinical trials are normally recognized as group sequential (GS) clinical trials. The Type I error rate for GS trials, even for the simplest case of testing a single hypothesis, can be inflated

M. Huque (✉)
Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA, USA
e-mail: huque.stat@gmail.com

S. Mushti
Division of Biometrics V, Office of Biostatistics, OTS, CDER, FDA, Silver Spring, MD, USA

M. Alosh
Division of Biometrics III, Office of Biostatistics, OTS, CDER, FDA, Silver Spring, MD, USA

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7820-0_7

if there are no adjustments for multiple looks, as compared to conventional non-GS trials, because of the repeated tests of the same hypothesis at interim looks. In GS trials, the same hypothesis is tested at different looks as the trial data accumulates over the time course of the trial, until the hypothesis is rejected or the trial reaches the final look for the last test of the hypothesis. Consequently, for assuring the credibility of a treatment benefit result even for a single-hypothesis GS trial, it is considered necessary to use a statistical adjustment method for controlling the probability of a Type I error at a pre-specified level through proper design and analysis methods that are prospectively planned.

There is an extensive literature for GS trials with plans to test a single primary hypothesis of a trial with repeated testing on accumulating data observed at different looks, and to stop the trial early at a look either for efficacy or for futility reasons. This literature covers in detail the technical and operational aspects of such trials, explaining how to plan, conduct, and analyze accumulating data of such trials. Emerson (2007) is an excellent review article on this topic. Also, there are useful books on this topic, including Whitehead (1997), Jennison and Turnbull (2000), and Proschan et al. (2006). Also, there are classical papers on this topic that are of historical importance, such as Armitage et al. (1969), Pocock (1977), O'Brien and Fleming (1979), and Lan and DeMets (1983). In addition, there are some extensions of the methods for multi-arm group sequential trials, e.g., comparison of multiple doses of the same treatment to a common control on a single primary endpoint with interim looks; see, for example, Follmann et al. (1994), Jennison and Turnbull (2000), Hellmich (2001), and Stallard and Friede (2008).

However, modern clinical trials are designed with multiple endpoints; some of these endpoints are given primary and secondary designations. The primary endpoint family along with their hypotheses holds a special position: If the study wins on one or more of its primary endpoint hypotheses then, depending on the level of evidence desired for this win, one can characterize a clinically relevant benefit of the study treatment. In this regard, O'Neill (1997), based on clinical and statistical considerations, made the case that secondary endpoint hypotheses need to be tested only when there is at least one rejection of the primary endpoint hypotheses leading to a clinically relevant benefit of the study treatment. Several innovative statistical procedures for confirmatory clinical trials were proposed that maximize the power for the tests of the primary hypotheses. In doing so, these approaches consider O'Neill's stipulation along with possibility of assigning weights to the different endpoint hypotheses and other logical restrictions. Further, these test procedures control the familywise Type I error rate (FWER) in the "strong sense" (see, e.g., Hochberg and Tamhane 1987), so that the conclusion of treatment efficacy can be made at the individual endpoints or hypotheses levels.

There is a fair amount of literature regarding these novel procedures for fixed-sample clinical trials but not so for GS clinical trials which are frequent for cardiovascular and oncology trials. Examples of such procedures for fixed-sample trial designs include the gatekeeping procedures (see, e.g., Dmitrienko et al. 2003, 2008; Dmitrienko and Tamhane 2009; and Huque et al. 2013 among others) and the graphical procedures (see, e.g., Bretz et al. 2009, 2011, 2014). The development of the

gatekeeping procedures and the graphical method have relied, either explicitly or implicitly, on shortcuts to the closed test procedure, as discussed by Hommel et al. (2007). These developments that utilize short-cut testing have been possible for weighted Bonferroni tests of the intersection hypotheses that satisfy "consonance" property (Hommel et al. 2007). Thereafter, the interest has been as to whether a similar approach for testing multiple hypotheses is possible for GS clinical trials. Recent publications, including Glimm et al. (2010), Tamhane et al. (2010), Maurer and Bretz (2013), Ye et al. (2013), Xi and Tamhane (2015), and Xi et al. (2016), have made this possible and have advanced multiple hypotheses testing methods for GS trials.

Tang and Geller (1999) proposed a general closed testing scheme for testing multiple hypotheses for GS clinical trials. This scheme, though conceptually simple to follow, seems complex to apply in practice, except for certain special situations. By taking advantage of the Hommel et al.'s findings and those of others, we make the case that that Tang and Geller's scheme can be simplified for application purposes by developing short-cut closed test procedures using, for example, the weighted Bonferroni tests. These short-cut procedures for testing multiple hypotheses in GS clinical trials also allow, indirectly, recycling the unused significance level of a rejected hypothesis to testing other hypotheses in a trial.

In this chapter, we first review the classical O'Brien-Fleming (OF) and Pocock (PK) approaches as well as the α-spending function methods, for setting the boundaries in a standard GS clinical trial for repeated testing of a single primary hypothesis. We will call herewith the α-spending function methods as spending function methods. As we will see later, these boundaries computed from the spending function approaches for testing a single hypothesis can still be used for testing multiple hypotheses in GS trials. Consequently, software developed for standard GS trials with a single-hypothesis test can also be used for multiple hypotheses tests. We also touch on the Tang and Geller (1999) closed testing approach as it is of historical importance and show that for testing two primary hypotheses of a trial, this approach simplifies when the weighted Bonferroni test is used for testing the intersection hypothesis. We then visit the graphical approach, for testing multiple primary and secondary hypotheses of GS trials, as discussed by Mauer and Bretz (2013), and present an illustrative example for testing two primary and two secondary endpoints of a trial. Thereafter, we consider the case that when the trial stops after the rejection of a primary hypothesis at a look say for ethical reasons, then other hypotheses need to be tested at the same look, as discussed by Tamhane et al. (2010). We close this chapter with some concluding remarks. Finally, we should point out that in all the discussions and methods presented for deriving boundaries of the GS trials and all tests considered are 1-sided comparing a study treatment to control.

7.2 Testing of a Single Hypothesis in a GS Trial

As in fixed-sample trials, the endpoints in a GS trial can be continuous, binary, or time-to-event. Although the associated test statistics for these endpoints may appear dissimilar, they share a common property: They can be expressed in terms of the standardized sums of independent observations of a random variable. Consequently, they span asymptotically the same joint distribution across time points of multiple looks of the data. Therefore, for the sake of simplicity in this chapter, we assume that the multiple endpoints considered are continuous, and the sample size for each arm of a 2-arm trial designed to compare the study treatment to control remains equal for each endpoint at each look. This case of equal sample size can be easily extended to the case when the sample size for the treated and control arms of the trial at a look can be of different sizes. Also, we consider the case that the total sample size for the final look is fixed in advance. In our discussion of GS trials, we do not consider them adaptive when the investigator continues to modifying the trial design based on the earlier results or what is known as adaptive study design. Adaptive study designs may allow for the possibility of adjusting the sample size of the trial, redefining the endpoint, or modifying the patient population based on the results of an interim look of the data of the trial. Methodological approaches for GS trials with such adaptations are more complex, and some of the assumptions and statements made here may not be valid. With these considerations, we first consider the case of testing a single endpoint hypothesis $H_0 : \delta \leq 0$ against the alternative hypothesis $H_a : \delta > 0$ for a trial with $K - 1$ interim looks and a final look, for a total of $K \geq 2$ looks. A positive value of δ indicates that the test treatment is better than the control.

7.2.1 Test Statistics and Their Distributions

Consider a 2-arm randomized trial designed to compare a treatment with a control on a single primary endpoint based on a total sample size of N subjects per arm. Let S_{n_1} be the sum statistic for the treatment difference at look 1 based on n_1 subjects per treatment arm. This sum statistics at look 1 is the sum of endpoint observations on n_1 subjects in the treatment arm minus the sum of endpoint observations on n_1 subjects in the control arm. Define the B-value at look 1 as

$$B(t_1) = S_{n_1}/\sqrt{V_N}, \text{ where } V_N = \text{Var}(S_N) = 2N\sigma^2. \tag{7.2.1}$$

In (7.2.1), S_N is the sum statistic for the final look yet to be observed and σ^2 is the known variance of individual observations which remains constant throughout the trial regardless of whether the subject observed is in the treatment arm or in the control arm. The value t_1 at look 1, usually known as the information fraction or the information time at look 1, is given by

$$\text{Var}\{B(t_1)\} = n_1/N = t_1. \tag{7.2.2}$$

Note that calling here $n_1/N = t_1$ as the information fraction or information time assumes that the sample sizes for the treatment and control groups are equal at each look and the variance of individual observations remains constant. In general, if d_1 and d_2 denote asymptotically normal estimates of a treatment group difference at interim and final looks, then the information fraction is defined as $I = Var(d_2)/Var(d_1)$. For normal outcomes, information time is the proportion of data available at the interim look, relative to the planned maximum if the trial is not stopped early. However, in presenting our results, for simplicity, we maintain our assumptions of equal sample sizes and constant variance. These results easily extend to the general case (Jennison and Turnbull 2000).

The standardized test statistic $Z(t_1)$ for testing H_0 at look 1 can then be expressed as

$$Z(t_1) = S_{n_1}/\sqrt{V_{n_1}} = (S_{n_1}/\sqrt{V_N})\sqrt{V_N/V_{n_1}} = B(t_1)/\sqrt{t_1}. \tag{7.2.3}$$

The relationship in (7.2.3) follows from $\text{Var}(S_{n_1}) = 2n_1\sigma^2$ and $V_N/V_{n_1} = 1/t_1$. Now consider the second look with the sample size of $n_2 = n_1 + r$ per treatment arm. Then $B(t_2) = (S_{n_1} + S_r)/\sqrt{V_N}$ where S_r is the sum statistic for the treatment difference based on the new data available at look 2. Consequently,

$$\text{Var}\{B(t_2)\} = n_2/N = t_2, \text{Cov}\{B(t_1), B(t_2)\} = t_1,$$

and

$$\text{Corr}\{B(t_1), B(t_2)\} = \text{Corr}\{Z(t_1), Z(t_2)\} = \sqrt{t_1/t_2} \text{ for } t_1 \le t_2. \tag{7.2.4}$$

Given $t_1 \le t_2 \le \cdots \le t_k \le \cdots \le t_K = 1$, we assume that $B(t_1), B(t_2), \ldots, B(t_K)$ follow a multivariate normal distribution with

$$\text{E}\{B(t_k)\} = 0 \text{ under } H_0 \text{ and } \text{Cov}\{B(t_k), B(t_l)\} = t_k \text{ for } t_k \le t_l \le t_K. \tag{7.2.5}$$

Therefore, the normal Z-statistics $\{Z(t_k) = B(t_k)/\sqrt{t_k}\}$ for $k = 1, \ldots, K$ follow a multivariate normal distribution with

$$\text{E}\{Z(t_k)\} = 0 \text{ under } H_0 \text{ and } \text{Cov}\{Z(t_k), Z(t_l)\} = \sqrt{t_k/t_l} \text{ for } t_k \le t_l \le t_K. \tag{7.2.6}$$

The non-central expected value of $B(t_k)$ in terms of the information fraction t_k is given by:

$$\text{E}\{B(t_k)\} = n_k\delta/\sqrt{2N\sigma^2} = (n_k/N)\sqrt{N/2}(\delta/\sigma) = t_k\theta, \tag{7.2.7}$$

where $\theta = \sqrt{N/2}(\delta/\sigma)$ is the "drift parameter." Consequently, the non-central expected value of $\mathrm{E}\{Z(t_k)\} = \sqrt{t_k}\theta$.

Note that $\theta = z_{1-\alpha} + z_{1-\beta}$ for a fixed-sample non-GS trial, where for such a trial, α is the probability of falsely rejecting the null hypothesis $H_0 : \delta \leq 0$ of no treatment effect in favor of the alternative hypothesis $H_a : \delta > 0$ of treatment effect, and power $1-\beta$ is the probability of rejecting H_0 when given the true treatment difference $\delta = \delta_0 > 0$. For example, when the trial $\alpha = 0.025$ and power $1-\beta = 0.90$, then $\theta = 3.2415$. Here the notation z_{1-x} stands for the deviate such that $\Pr(U \leq z_{1-x}) = 1-x$ with $0 \leq x \leq 1$, where U is the normal N(0, 1) random variable. More details about $B(t)$ values and $Z(t)$ normal scores can be found in Proschan et al. (2006) and Lan and Wittes (1988). In the following, we show how the well-known methods by Pocock (1977) and O'Brien and Fleming (1979) rely on these B-values and z-scores in finding their local significance levels, i.e., GS-boundary values, for the repeated testing of H_0. For convenience, we will call these historical methods as PK and OF methods and their boundaries as PK and OF classical boundaries.

7.2.2 Classical PK and OF Boundaries

When analyses of accumulating data of a GS trial occur at equally spaced information times, then the PK boundary is a constant boundary on the z-scale. That is, if $t_k = k/K$ for $k = 1, \ldots, K$, the constant PK boundary $c_{PK}(\alpha, K) = x$ for 1-sided tests can then be obtained by solving for x in the following equation:

$$\Pr[\bigcap_{k=1}^{K} \{Z(t_k) \leq x\} | H_0] = 1 - \alpha \text{ with } t_k = k/K \text{ for } k = 1, \ldots, K, \tag{7.2.8}$$

such that the Type I error rate is controlled at level α. This equation can be solved under the assumption that the joint distribution of the test statistics $\{Z(t_k); k = 1, \ldots, K\}$ is multivariate normal with zero mean vector and correlation matrix $(\rho_{kl}) = \left(\sqrt{t_k/t_l}\right)$ with $t_k \leq t_l$. For example, $c_{PK}(\alpha, K) = 2.28947$ for $K = 3$, ($t_1 = 1/3$, $t_2 = 2/3$, and $t_3 = 1$), and $\alpha = 0.025$. For solving for x in (7.2.8), we wrote SAS/IML codes that calculated the left-hand side of the equation using PROBBNRM and QUAD functions of SAS. PROBBNRM is a SAS function which gives values of the cumulative distribution functions of a standard bivariate normal distribution on specifying the value of the two variables and the correlation coefficient between them. QUAD is a SAS function which integrates numerically a function over an interval. This calculation expressed the joint distribution of $\{Z(t_k); k = 1, 2, 3\}$ as the product of the distribution of $Z(t_1)$ and the conditional bivariate distribution of $Z(t_2)$ and $Z(t_3)$ given $Z(t_1) = z(t_1)$.

Jennison and Turnbull (2000) and Proschan et al. (2006) include 2-sided PK boundary values for different values of K, and $\alpha = 0.01$, 0.05, and 0.10. These 2-sided boundary values at level α, if taken as 1-sided boundary values at level $\alpha/2$,

may not be identical to the actual 1-sided boundary values obtained from (7.2.8); see, for example, Sect. 2.4 in Wassmer and Brannath (2016). The PK boundary values for 2-sided tests are obtained by replacing $Z(t_k) \leq x$ by $|Z(t_k)| \leq x$ in (7.2.8). Thus, a GS trial, designed with PK boundary with looks at equally spaced information times with given α and K, would reject H_0 for efficacy and stop the trial at look k with the information fraction t_k when $Z(t_k) > c_{PK}(\alpha, K)$.

Likewise, the OF boundary is a constant boundary on the B-value scale when the trial looks occur at equally spaced information times. Therefore, when $t_k = k/K$, for $k = 1, \ldots, K$, the 1-sided OF boundary value can be obtained by solving for x in the following equation:

$$\Pr[\bigcap_{k=1}^{K} \{B(t_k) \leq x\} | H_0] = 1 - \alpha \text{ with } t_k = k/K \text{ for } k = 1, \ldots, K.$$

Using $Z(t_k) = B(t_k)/\sqrt{t_k}$ the above equation can be expressed as in (7.2.9) to solve for x using the joint distribution of the test statistics $\{Z(t_k); k = 1, \ldots, K\}$ as a multivariate normal with zero mean vector and correlation matrix $(\rho_{kl}) = \left(\sqrt{t_k/t_l}\right)$ for $t_k \leq t_l$:

$$\Pr[\bigcap_{k=1}^{K} \left\{Z(t_k) \leq x/\sqrt{t_k}\right\} | H_0] = 1 - \alpha \text{ with } t_k = k/K \text{ for } k = 1, \ldots, K. \quad (7.2.9)$$

For example, when $K = 2$, ($t_1 = 1/2$ and $t_2 = 1$), $\alpha = 0.025$, and the tests are 1-sided, then solving the equation PROBBNRM $(x\sqrt{2}, x, \sqrt{1/2}) = 0.975$ gives the value of $x = 1.97742$ which in turn gives the OF boundary values of $c_1(\alpha, K) = x\sqrt{2} = 2.796494$ for the first look at $t_1 = 1/2$ and $c_2(\alpha, K) = x = 1.97742$ for the final look on the z-score scale with the corresponding boundary values of $\alpha_1(\alpha, K) = 0.002583$ and $\alpha_2(\alpha, K) = 0.023997$ on the p-value scale. Thus, if a GS trial is designed with two looks with an interim look at $t_1 = 1/2$, and $\alpha = 0.025$, then H_0 will be rejected when the p-value at this look is less than $\alpha_1(\alpha, K) = 0.002583$ stopping the trial early; otherwise, the trial will continue to the next and final look, and H_0 will be rejected there when the p-value at this look is less than $\alpha_2(\alpha, K) = 0.023997$.

Jennison and Turnbull (2000) and Proschan et al. (2006) provide values of x for 2-sided tests for different values of K and $\alpha = 0.01$, 0.05, and 0.1. These 2-sided boundary values at level α, if read as 1-sided boundary values at level $\alpha/2$, may not agree with the actual 1-sided boundary values. Note that the methods described in this section are of historical importance and are not so frequently used; they lack flexibility because managing analysis at equally spaced information time can be challenging. A more flexible approach for GS trials is the spending function approach described in the next section.

7.2.3 Spending Function Approach

The classical PK and OF boundaries introduced above require specifying the total number of looks at equally spaced information times. This can be inconvenient for clinical trial applications as the Data Safety Monitoring Board (DSMB) or any other group charged with performing interim looks of the accumulating clinical trial data may have to postpone a look for logistical reasons, or may decide to have a look at an unspecified time because of certain concerns. Lan and DeMets (1983) proposed the spending function approach for this and showed that the construction of GS boundaries do not require pre-specification of the number or timings of looks.

Any non-decreasing function $f(\alpha, t)$ in the information time t, over the interval $0 \leq t \leq 1$ and parameterized by the overall significance level α for testing H_0, can be a spending function if it satisfies the following conditions: $f(\alpha, t) \leq f(\alpha, t')$ for $0 \leq t \leq t' \leq 1$; $f(\alpha, t = 0) = 0$; and $f(\alpha, t = 1) = \alpha$. A commonly used spending function for clinical trials is the OF-like:

$$f_1(\alpha, t) = 2\{1 - \Phi(z_{1-\alpha/2}/\sqrt{t})\},$$

where $\Phi(.)$ is the cumulative distribution function of the standard normal distribution.

Note that $f_1(\alpha, 0) = 0$ and $f_1(\alpha, 1) = \alpha$. If the trial had only 2 looks, one at $t = 1/2$ and the other at $t = 1$, and $\alpha = 0.025$, then $f_1(\alpha = 0.025, t = 1/2) = 2(1 - \Phi(2.241403/0.70711)) = 2\{1 - \Phi(3.1698)\} = 0.001525$ and $f_1(\alpha = 0.025, t = 1) = \alpha$. One can then find the significance level x for the final look by solving the equation $\Pr\{(P_1 < 0.001525) \cup (P_2 < x)\} = 0.025$. The next section shows how these equations are solved. The advantage of using the OF-like spending function for clinical trials is its shape which is convex. This allows spending very little of the total α for early looks and saves most of it for latter looks when the trial has sufficient number of patients exposed to the new treatment. The idea is to stop the trial early only when the treatment effect size is sufficiently large and clinically convincing.

Table 7.1 includes a few other spending functions. These and other spending functions give the cumulative Type I error rate spent at look k with the associated information fraction t_k. This cumulative value does not give directly the local significance level $\alpha_k(\alpha, t_k)$ (i.e., the boundary value) for testing H_0 at look k, except when $k = 1$ (the first look). Note that these boundary values are on the p-value scale and need to be converted for presentation on the z-scale. Finding $\alpha_k(\alpha, t_k)$ requires additional calculations which we describe in the following with an example. These calculations usually require solving equations in multiple integrals and are not easy when $K \geq 3$. Special computer software is normally used for this.

Table 7.1 Examples of spending functions

Linear	Pocock-like	Hwang-Shi-Decani (1990)
$f_2(\alpha, t) = \alpha t$	$f_3(\alpha, t) = \alpha \log_e\{1 + (e-1)t\}$	$f_4(\alpha, t) = \alpha\big[\{1 - \exp(-\lambda t)\}/\{1 - \exp(-\lambda)\}\big]$, for $\lambda \neq 0$

7.2.4 Calculations of Boundary Values Using Spending Functions

We illustrate the use of spending functions for finding the local significance level $\alpha_k(\alpha, t_k)$ at look k with the information fraction t_k, so that H_0 will be rejected when the 1-sided p-value p_k at this look is less than $\alpha_k(\alpha, t_k)$. Suppose a trial uses the OF-like spending function to control the Type I error rate at level $\alpha = 0.025$. Suppose that the first look occurs at $t_1 = 0.30$. Then at this look, we spend

$$f_1(\alpha = 0.025, t_1 = 0.30) = 2\left\{1 - \Phi\left(z_{1-\alpha/2}/\sqrt{0.30}\right)\right\}$$
$$= 2\left\{1 - \Phi\left(\frac{2.2414027}{\sqrt{0.30}}\right)\right\} = 0.0000427$$

Therefore, at this look, $\alpha_1(\alpha, t_1) = 0.0000427$ and the critical value $c_1(\alpha, t_1) = 3.9285725$ from $\Pr\{Z(t_1) > c_1(\alpha, t_1)\} = 0.0000427$; one will reject H_0 and stop the trial at the first look if $p_1 < 0.0000427$ or $Z(t_1) > 3.9285725$. Thus, at this look the investigator spends very little of the total $\alpha = 0.025$.

Suppose that the trial did not stop at the first look and the investigator decides to have the second look at $t_2 = 0.65$. Then the cumulative alpha spent at this look is

$$f_1(\alpha = 0.025, t = 0.65) = 2\left\{1 - \Phi\left(z_{1-\alpha/2}/\sqrt{0.65}\right)\right\}$$
$$= 2\left\{1 - \Phi\left(\frac{2.2414027}{\sqrt{0.65}}\right)\right\} = 0.0054339$$

Therefore, we determine the boundary critical values of $c_2(\alpha, t_2) = 2.5479$ or $\alpha_2(\alpha, t_2) = 0.0054187$ by solving the equation: $\Pr[\{(Z(t_1) > 3.9285725\} \cup \{(Z(t_2) > c_2(\alpha, t_2)\}] = 0.0054339$. Therefore, one can reject H_0 at the second look and stop the trial, if at this look, the observed p-value $p_2 < 0.005187$ or $Z(t_2) > 2.5479$.

Suppose the trial did not stop at this second look and the investigator moves to the final look at $t_3 = 1$. Then the cumulative alpha spent at the final look is $\alpha = 0.025$. One can then find $c_3(\alpha, t_3)$ by solving the equation:

$$\Pr[\{(Z(t_1) > 3.9285725\} \cup \{(Z(t_2) > 2.5479\} \cup \{(Z(t_3) > c_3(\alpha, t_3)\}] = 0.025$$

Table 7.2 Examples for the OF-like spending function with $\alpha = 0.025, 0.0125$, $K = 3$, and 1-sided tests

Look #	Information fraction	Cumulative α spent	Boundaries
$\alpha = 0.025$			
1	0.30	0.00004	0.00004
2	0.65	0.00543	0.00542
3	1.00	0.025	0.02331
$\alpha = 0.0125$			
1	0.30	0.00001	0.00001
2	0.65	0.00194	0.00194
3	1.00	0.0125	0.01188

Solving this equation gives $c_3(\alpha, t_3) = 1.9897$ and $\alpha_3(\alpha, t_3) = 0.023312$. Therefore, one can reject H_0 at the final look if at this look the p-value $p_3 < 0.023312$ or $Z(t_3) > 1.9897$.

A general recursive equation for finding $c_k(\alpha, t_k)$ and $\alpha_k(\alpha, t_k)$ for a spending function $f(\alpha, t)$ is given by $f(\alpha, t_k) = f(\alpha, t_{k-1}) + \Pr\left[\{\bigcap_{i=1}^{k-1} Z(t_i) \le c_i(\alpha, t_i)\} \cap \{Z(t_k) > c_k(\alpha, t_k)\}\right]$ for $k \ge 2$. There are software available that give values of $c_k(\alpha, t_k)$ and $\alpha_k(\alpha, t_k)$ for OF-like and other spending functions, see Zhu et al. (2011) for a review of these software. Table 7.2 shows the results from such a software. We show in Sect. 7.3 that such boundaries can also be used for testing multiple hypotheses of GS trials.

7.3 Testing of Multiple Hypotheses in GS Trials

Many GS trials are designed for testing multiple endpoint hypotheses, frequently, for testing two endpoint hypotheses. Two situations generally arise. Consider, for example, a GS trial for testing two endpoint hypotheses. The first case arises when after the rejection of one of the two hypotheses at an interim look the trial does not stop but continues to later looks for testing the other hypothesis. The second case arises when the two hypotheses are hierarchically ordered, e.g., one is primary and the other is secondary. The first hypothesis in the hierarchy (i.e., the primary hypothesis) is allocated first using the full trial α (e.g., $\alpha = 0.025$). If this hypothesis is rejected at an interim look, then the trial stops because of ethical considerations. For example, if the first hypothesis is associated with the mortality endpoint and the second hypothesis with a quality of life measure, then if the trial wins at a look for the mortality endpoint then the trial would generally discontinue for ethical reasons. In that case, the second hypothesis (i.e., the secondary hypothesis) is tested at the same look at which the first hypothesis was rejected. The remainder of this section considers the first case and Sect. 7.4 considers the second case. In the following, we first address methods based on the Bonferroni inequality and then move on to

α-recycling approaches based on the closed testing principle (CTP) of Marcus et al. (1976), and finally to the more recent graphical approach of Maurer and Bretz (2013).

7.3.1 Methods Based on the Bonferroni Inequality

Consider, for example, a trial which for the demonstration of superiority of a new treatment to control specifies two null hypotheses: H_1 and H_2. Rejection of either of the two hypotheses at a look can establish efficacy of the new treatment. However, if the trial rejects one of the two hypotheses at an interim look, the trial can continue to later looks for testing the other hypothesis. For such a trial, the use of the Bonferroni inequality leads to two approaches for a stronger claim. The first approach splits the significance level α as $\alpha_1 + \alpha_2 \leq \alpha$ for testing H_1 at level α_1 and H_2 at level α_2. For example, it may assign $\alpha_1 = 0.005$ for testing H_1 and $\alpha_2 = 0.02$ for testing H_2 for controlling the overall Type I error rate at $\alpha = 0.025$. Tests for H_1 and H_2 can then separately follow in a univariate GS testing framework for the separate control of the Type I error rates at levels α_1 and α_2, respectively, using the same or different spending functions for each. In Sect. 7.3.2, we show that this approach extends to an α-recycling approach, such that, if one of the multiple hypotheses is rejected at a look then the boundary value for testing other hypotheses is updated to larger values.

The second approach uses the Bonferroni inequality differently. It specifies the rejection boundary values as $\alpha'_k(t_k) > 0$ for looks $k = 1, \ldots, K$ such that $\sum_{k=1}^{K} \alpha'_k(t_k) = \alpha$. It then applies a conventional multiple hypothesis testing method at a look for the control of the Type I error rate at the local level $\alpha'_k(t_k)$ at that look. Suppose that $K = 2$, i.e., the trial is designed with two looks, and $\alpha'_1(t_1) = 0.005$ and $\alpha'_2(t_2 = 1) = 0.02$, for the first and second looks, respectively. One can then apply, for example, the conventional Hochberg procedure (1988) for testing H_1 and H_2 at level 0.005 at the first look, and similarly, can apply the same procedure for testing these hypotheses at the final look at level 0.02. The methods discussed in this section for testing two hypotheses generalize to testing more than two hypotheses.

7.3.2 Method Based on the Closed Testing Principle

The closed testing principle of Marcus et al. (1976) provides a general framework for constructing powerful closed test procedures (CTPs) for testing individual hypotheses based on tests of intersection hypotheses of different orders. One starts with a family of individual hypotheses $H_1, \ldots, H_h$ and constructs a closed set $\tilde{\boldsymbol{H}}$ of $2^h - 1$ non-empty intersection hypotheses as follows:

$$\tilde{\boldsymbol{H}} = \left\{ H_{\boldsymbol{J}} = \bigcap\nolimits_{j \in \boldsymbol{J}} H_j, \quad \boldsymbol{J} \subseteq \boldsymbol{I} = \{1, \ldots, h\} \right\}.$$

One then performs an α-level test for each hypothesis $H_{\boldsymbol{J}}$ in $\tilde{\boldsymbol{H}}$ by using, for example, the weighted Bonferroni test. One then rejects an individual hypothesis H_j when all $H_{\boldsymbol{J}}$ for $j \in \boldsymbol{J}$ are rejected by their corresponding α-level tests.

For example, when $h = 2$, the closed set $\tilde{\boldsymbol{H}} = \{H_{12}, H_1, H_2\}$. A CTP will reject the individual hypothesis H_1 only when H_1 and H_{12} are both rejected, each by an α-level test. If one uses, for example, the weighted Bonferroni test for H_{12}, then the procedure cuts down the extra step of testing H_1 after rejecting H_{12}. The weighted Bonferroni test rejects H_{12}, when $p_j < w_j\alpha$ for at least one $j \in \{1, 2\}$, where w_1 and w_2 are the nonnegative weights assigned to H_1 and H_2, respectively, such that $w_1 + w_2 \leq 1$, and p_j are the observed p-values associated with H_j for $j \in \{1, 2\}$. Suppose that this test rejects H_{12} for $j = 1$ on observing $p_1 < w_1\alpha$, then H_1 is automatically rejected, as the significance level α for the test of H_1 satisfies $\alpha \geq w_1\alpha$. This property in its general form, known as the *consonance* property, when satisfied for testing intersection hypotheses in a closed testing procedure, leads to short-cuts of closed test procedures and allows recycling of the significance level of a rejected hypothesis to other hypotheses (Hommel et al. 2007). This property basically means that the rejection of an intersection hypothesis $H_{\boldsymbol{J}}$ by an α-level test implies the rejection of at least one individual hypothesis H_j for $j \in \boldsymbol{J}$.

As a numerical example, consider testing the two hypotheses H_1 and H_2 with $\alpha = 0.025$, and suppose that weights assigned to H_1 and H_2 are $w_1 = 0.8$ and $w_2 = 0.2$, respectively, so that $w_1 + w_2 = 1$. Further, supposed that the associated observed p-values for the tests of H_1 and H_2 were $p_1 = 0.024$ for H_1 and $p_2 = 0.004$ for H_2. The simple weighted Bonferroni test would reject only H_2, as $p_1 > w_1\alpha = 0.020$ and $p_2 < w_2\alpha = 0.005$. However, the weighted Bonferroni based CTP with these weights would reject both hypotheses. This CTP, in its initial step, would reject the intersection hypothesis H_{12} as $p_j < w_j\alpha$ for $j = 2$. Consequently, as the procedure assigns the weights of one for testing each singleton hypotheses, satisfying consonance, it would then reject each of the two hypotheses as $p_j < 1.\alpha = 0.025$ for each $j \in \{1, 2\}$.

In the following, we first visit the GS closed test procedure by Tang and Geller (1999) for testing multiple hypotheses and show that this procedure leads to α-recycling procedures by using weighted Bonferroni tests of intersection hypotheses that satisfy consonance. The Tang and Geller procedure is of historical importance with respect to using the closed testing procedure for testing multiple hypotheses in group sequential trials. Although the procedure sounds complicated in its original form, it can be simplified if the weighted Bonferroni tests, with weights satisfying the consonance property, are used for testing its intersection hypotheses. However, selection of such weights can be cumbersome for testing more than three hypotheses. Section 7.3.3 toward the end illustrates how to find these weights when testing two primary hypotheses and a secondary hypothesis. In general, the graphical approach (Sect. 7.3.5) in this regard is easier to use when testing multiple hypotheses.

Consider testing $h \geq 2$ endpoint hypotheses in a GS trial designed to compare a new treatment to control. Consider, as before, the intersection hypotheses $H_{\boldsymbol{J}}$ for $\boldsymbol{J} \subseteq \boldsymbol{I} = \{1, \ldots, h\}$, i.e., the new treatment to control treatment difference $\delta_j \leq 0$ for all endpoints $j \in \boldsymbol{J} \subseteq \boldsymbol{I}$. Also, consider that multiple looks for the trial occur at

different information times $t \in \{t_1, t_2, \ldots, t_K\}$ such that $t_1 \leq t_2 \leq \cdots \leq t_K = 1$. Let $Z_{\boldsymbol{J}}$ be a test statistic for testing $H_{\boldsymbol{J}}$ (e.g., by a weighted Bonferroni test) and let $Z_{\boldsymbol{J}}(t)$ be the test statistic value of $Z_{\boldsymbol{J}}$ at a look with information fraction t. Further, let $c_{\boldsymbol{J}}(t)$ be the critical value for performing an α-level test of $H_{\boldsymbol{J}}$ at this look by using $Z_{\boldsymbol{J}}(t)$. That is, for each $\boldsymbol{J} \subseteq \boldsymbol{I}$, the $c_{\boldsymbol{J}}(t)$ values for different t (at which times repeated tests occur) satisfy $\Pr\{Z_{\boldsymbol{J}}(t) > c_{\boldsymbol{J}}(t) \text{ for some } t | H_{\boldsymbol{J}}\} \leq \alpha$. Then a closed test procedure for GS trials as proposed by Tang and Geller (1999) can be stated as follows:

Step 1: Start testing $H_{\boldsymbol{I}}$ as in a univariate case of a GS trial but using the group sequential boundary values $c_{\boldsymbol{I}}(t)$ for the test statistics $Z_{\boldsymbol{I}}(t)$, where $\boldsymbol{I} = \{1, \ldots, h\}$.

Step 2: Suppose that $H_{\boldsymbol{I}}$ is rejected first time at the look with $t = t^*$. Then, for rejecting at least one individual hypothesis at this look, apply a CTP to test $H_{\boldsymbol{J}}$ with $\boldsymbol{J} \subseteq \boldsymbol{I}$ using $Z_{\boldsymbol{J}}(t^*)$ and its critical value $c_{\boldsymbol{J}}(t^*)$. Note that $c_{\boldsymbol{J}}(t^*)$ can be different for different $H_{\boldsymbol{J}}$'s. In applying this CTP at $t = t^*$ either (a) none of the individual hypotheses will be rejected, or (b) at least one individual hypothesis H_j will be rejected for $j \in \boldsymbol{I}$.

Step 3(a): In *Step* 2, if none of the individual hypotheses are rejected at $t = t^*$ then continue to the next look; however, if $t^* = 1$ and none of the individual hypotheses are rejected, the trial will stop without the rejection of any hypothesis.

Step 3(b): In *Step* 2, if at least one hypothesis is rejected at $t = t^*$, then exclude the indices of the rejected hypotheses from the index set $\boldsymbol{I}$. With this updated index set $\boldsymbol{I}$, continue to the next look and repeat *Step* 1 and *Step* 2. Note that in this process, all previously rejected hypotheses are assumed rejected at later looks and are removed for further testing.

Step 4: Reiterate the above steps until all hypotheses are rejected or the trial reaches the final look.

Implementing the Tang and Geller (1999) approach for the general case can be complicated because of the computational difficulties in finding $c_{\boldsymbol{J}}(t)$ values for testing $H_{\boldsymbol{J}}$ for different $\boldsymbol{J}$ and different looks. However, this approach simplifies on using univariate tests for $H_{\boldsymbol{J}}$ that satisfy consonance. Examples, of such tests, are the max-T or min-p test, and the un-weighted Bonferroni test. Weighted Bonferroni test which is more useful for clinical trial applications also serves this purpose, but the weights for the weighted Bonferroni tests need to be pre-selected to satisfy consonance. This may be difficult when testing more than three hypotheses. An alternative to this which does not have this issue is the graphical approach addressed in Sect. 7.3.4. The following, however, addresses the weighted Bonferroni test approach and illustrates its application for testing two hypotheses in a GS trial.

In the weighted Bonferroni test approach, to satisfy consonance for the tests of $H_{\boldsymbol{J}}$ for $\boldsymbol{J} \subseteq \boldsymbol{I}$, one pre-selects weights $w_j(\boldsymbol{J})$ for $j \in \boldsymbol{J}$ with $\sum_{j \in \boldsymbol{J}} w_j(\boldsymbol{J}) \leq 1$ so that $w_j(\boldsymbol{J}^*) \geq w_j(\boldsymbol{J})$ for every $\boldsymbol{J}^* \subseteq \boldsymbol{J}$. For these cases, standard software developed for testing a single hypothesis with a spending function approach can still be used for testing multiple hypotheses. The following is an illustrative example for testing two hypotheses H_1 and H_2 in a GS trial.

In the case of testing two hypotheses, a CTP considers a single intersection hypothesis $H_{\boldsymbol{J}}$ with $\boldsymbol{J} = \{1, 2\}$, written as H_{12}, and two individual hypotheses H_1 and H_2. Suppose that for testing H_{12} one assigns weights $w_1\{1, 2\} = 0.8$ and $w_2\{1, 2\} = 0.2$ so that $w_1\{1, 2\}\alpha = 0.02$ and $w_2\{1, 2\}\alpha = 0.005$ with the trial $\alpha = 0.025$. Consonance is satisfied, because after H_{12} is rejected, the weights for testing each of the two individual hypotheses in the CTP is one. The following illustrates how one will test H_1 and H_2 in a GS trial with such initial weights.

Tests at the First Look

Suppose that the first look for the trial occurs at $t = t_1 = 0.30$, and suppose that at this look the unadjusted p-values associated with H_1 and H_2 are $p_1(t_1)$ and $p_2(t_1)$, respectively. The CTP will reject H_{12} by the weighted Bonferroni test if either $p_1(t_1) < \alpha_1(w_1\{1, 2\}\alpha = 0.02, t_1 = 0.30) = \alpha_1(0.02, t_1 = 0.30)$ or $p_2(t_1) < \alpha_2(0.005, t_1 = 0.30)$, where these boundary critical values can be obtained by specifying spending functions f_1 and f_2. If f_1 and f_2 are each OF-like, then

$$\alpha_1(0.020, t_1 = 0.30) = f_1(w_1\{1, 2\}\alpha = 0.02, t_1 = 0.30) = 0.00002$$
$$\alpha_2(0.005, t_1 = 0.30) = f_2(w_2\{1, 2\}\alpha = 0.005, t_1 = 0.30) = 2.977\text{E} - 07$$

Suppose that H_{12} is not rejected at this look with $t_1 = 0.30$ and the trial continues to the second look.

Tests at the Second Look

Suppose that the second look occurs at $t_2 = 0.65$. Further, suppose that at this look the unadjusted p-values associated with H_1 and H_2 are $p_1(t_2)$ and $p_2(t_2)$, respectively. Consequently, the CTP will reject H_{12} at this look if either $p_1(t_2) < \alpha_1(0.02, t_2 = 0.65)$ or $p_2(t_2) < \alpha_1(0.005, t_2 = 0.65)$. The use of the spending functions f_1 and f_2 as OF-like for this look gives the boundary values

$$\alpha_1(0.020, t_1 = 0.65) = 0.0039 \text{ and } \alpha_2(0.005, t_1 = 0.65) = 0.000498.$$

Section 7.2.4 has addressed how these boundary values are calculated. As indicated before, computer software is used to calculate such boundary values.

Now, suppose that $p_2(t_2) < 0.000498$, then H_{12} will be rejected leading to the automatic rejection of H_2 because of the consonance condition being satisfied. Therefore, as H_{12} and H_2 are rejected at $t^* = t_2 = 0.65$, the CTP will test the remaining hypothesis H_1 at the same look with ($t^* = t_2 = 0.65$) with the updated boundary value of $\alpha_1(0.025, t_2 = 0.65) = 0.00542$ by the same OF-like spending function. Thus, there is a recycling of alpha of 0.005 form the rejected H_2 to H_1, updating the alpha of 0.02 to $0.02 + 0.005 = 0.025$ which is incorporated in the first argument of $\alpha_1(0.025, t_2 = 0.65)$. Thus, a CTP with consonance allows recycling of alpha for GS trials, but here, this recycling updates the boundary values for testing H_1 starting from at $t^* = t_2 = 0.65$ using a spending function. Suppose that $p_1(t_2) = 0.015$ which is greater than 0.00542, then H_1 at this second look remains not rejected. The trial then continues to the final look with $t_3 = 1$ for testing H_1.

Test at the Final Look
The final look occurs with $t_3 = 1$ for testing H_1 with the assumption that H_2 (which was rejected at the second look) remain rejected at this look. Therefore, H_1 would be tested at this look at level $\alpha_1(0.025, t_3 = 1) = 0.02331$ by the same OF-like spending function.

7.3.3 Some Key Considerations and Comments

For applications, the spending functions to be used for testing different hypotheses need to be pre-specified, and for interpreting study findings, it is good practice to use the same spending functions for testing different hypotheses. It should be noted that although the total number of looks may not be pre-specified, however, specifying it may help reducing concerns about unnecessary looks of the data. In addition, in our previous discussion, including the illustrative example in Sect. 7.3.2, we assumed that information fractions for the two endpoints are equal at each look. This can be the case for continuous or binary endpoints; however, this may be not the general case. That is, if $t_k(E_1)$ and $t_k(E_2)$ are information fraction for two endpoints at looks $k = 1, .., K$ then it is possible that $t_k(E_1) \neq t_k(E_2)$ for at least one k. This can occur, for example, when E_1 or E_2 are time-to-event endpoints; it may also occur for other situations. Then the question may arise as how to adopt the above procedure for this general case.

In this regard, we note that the above procedure can be easily adopted to address this general case. To illustrate, suppose that in the above example, at the first look $t_1(E_1) = t_1(E_2) = 0.30$, but at the second look $t_2(E_1) = 0.40$ and $t_2(E_2) = 0.65$ and assume that H_{12} is not rejected at the first look; yet, it can be rejected at the second look if either $p_1(t_2) < \alpha_1(0.02, t_2(E_1) = 0.40)$ or $p_2(t_2) < \alpha_1(0.005, t_2(E_2) = 0.65)$. Now, suppose that at this stage H_{12} is rejected by observing that $p_2(t_2) < \alpha_1(0.005, t_2(E_2) = 0.65)$, leading to the rejection of H_2 as before. Therefore, the alpha of 0.005 for the rejected H_2 will now be recycled for testing H_1, that is by updating the old boundary value of $\alpha_1(0.02, t_2(E_1) = 0.40)$ to a new boundary value $\alpha_1(0.025, t_2(E_1) = 0.40)$ at this second look, and to $\alpha_1(0.025, t_3(E_1) = 1)$ at the final look.

Note that in above after rejecting H_2 at the second look, the significance level for testing for H_1 is $\alpha_1(0.025, t_2(E_1) = 0.40)$ which is not equal to $\alpha = 0.025$. Wrongfully, testing H_1 at $\alpha = 0.025$ instead of testing it at level $\alpha_1(0.025, t_2(E_1) = 0.40)$ after the rejection of H_2 can inflate the overall Type I error rate. Also, if the trial stops at a look after rejecting a hypothesis for ethical reasons, say after the rejection of H_2, then one cannot test a second hypothesis such as H_1 at the full significance level of $\alpha = 0.025$. Doing this can inflate the overall Type I error rate, except for the special case when the test statistics for the two hypotheses are independent. We consider this type of GS trials in Sect. 7.4.

The spending functions used to test each hypothesis needs to satisfy a monotonicity property. That is, the difference function $f(\lambda, t_k) - f(\lambda, t_{k-1})$ is monotonically

non-decreasing in λ for $k = 1, \ldots, K$. For example, the OF-like α-spending function satisfies this condition for $\lambda < 0.318$ (Maurer and Bretz 2013).

The above weighted Bonferroni-based CTP for testing two hypotheses can be extended to testing more than two hypotheses if weights assigned for testing intersection hypotheses in a CTP are such that consonance property is guaranteed, that is, weights assigned are such that rejection of an intersection hypothesis in the CTP leads to the rejection of at least one individual hypothesis in that intersection hypothesis. For example, for testing two primary hypotheses H_1 and H_2 and a secondary hypothesis H_3 of a trial, the CTP would consider four intersection hypotheses H_{123}, H_{12}, H_{13} and H_{23} and three individual hypotheses.

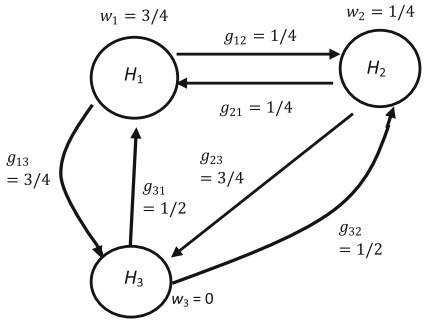
The following selection of weights for performing Bonferroni-based tests of intersection hypotheses in the CTP would then satisfy consonance property. Assign non-negative weights of w_1, w_2, and w_3 associated with indices (1, 2, and 3) of H_{123} to test this hypothesis with $w_1 + w_2 = 1$ and $w_3 = 0$; the selection of $w_3 = 0$ indicates that H_3 is tested only after at least one of the two primary hypotheses is first rejected. Assign weights of $\{w_1, w_2\}$ to H_1 and H_2, respectively to test H_{12}. Similarly, weights of $\{w_1 + \delta_2 w_2, (1 - \delta_2)w_2\}$ to test H_{13}, and weights of $\{w_2 + \delta_1 w_1, (1 - \delta_1)w_1\}$ to test H_{23}, where $0 \leq \delta_1 \leq 1$ and $0 \leq \delta_2 \leq 1$. The weights assigned to each of the individual hypotheses will be one. The selection of these weight and the recycling parameter δ_1 and δ_2, for example, can be based on the trial objectives. Once such weights for performing the weighted Bonferroni tests satisfy consonance, a CTP for testing the above three hypotheses in a GS trial can be proposed.

GS trials that are not properly conducted have the potential of unblinding the trial prematurely, and consequently, this may impact the integrity of the trial and its results. To address this important issue, usually an Independent Data Monitoring Committees (DMC) along with a charter is setup for GS trials. As our focus for this chapter is to overview the general multiple testing approaches for group sequential trials, we do not discuss this issue here. The interested reader may consult relevant literature in this regard, see, e.g., Ellenberg et al. (2017). The concerns about potential unblinding for testing single hypothesis over the course of GS trials remain the same for GS trials with testing multiple hypotheses related to multiple endpoints.

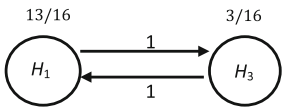
For a GS trial that include testing of multiple hypotheses, a Statistical Analysis Plan (SAP) that explains in sufficient details the design, the analyses method, and the DMC charter, is essential for proper interpretation of study findings. Such a SAP should in general be developed a priori and agreed upon by those involved before launching the trial.

7.3.4 Graphical Approach

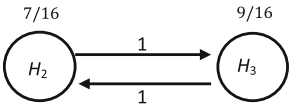
The above weighted Bonferroni-based CTP for testing multiple hypotheses of a GS trial, though possible, can be challenging in finding appropriate weights that guarantee consonance when the number of hypotheses tested are more than a few. The graphical approach of Bretz et al. (2009) which includes a special algorithm for



(a) Initial Graph for testing H_1, H_2, and H_3 along with their weights



(b) Graph after rejecting H_2 in (a), for testing H_1 and H_3 along with their weights



(c) Graph after rejecting H_1 in (a), for testing H_2 and H_3 along with their weights

Fig. 7.1 Graphical representation of testing with two primary hypotheses H_1 and H_2, and one secondary hypothesis H_3

doing this solves this problem. In this approach, one can graphically visualize the weighted Bonferroni tests for multiple hypotheses along with an α-propagation rule by which the procedure recycles the significance level of a rejected hypothesis to other remaining unrejected hypotheses. This graphical approach, originally developed for testing multiple hypotheses of non-GS trials, can also be conveniently used for testing multiple hypotheses of GS trials; see, for example, Maurer and Bretz (2013). The following explains the key concepts of this approach for testing multiple hypotheses.

In this graphical approach, the h individual hypotheses are represented initially by a set of h nodes with nonnegative weight of w_i at node $i (i = 1, \ldots, h)$ such that $\sum_{i=1}^{h} w_i \leq 1$. These weights when multiplied by α represent the local significance levels at those respective nodes. The weight g_{ij} (with $0 \leq g_{ij} \leq 1$) associated with a directed edge connecting the node i to the node j indicates the fraction of the local significance level at the tail node i that is added to the significance level at the terminal node j, if the hypothesis at the tail node i is rejected. For convenience, we will call these directed edges as "arrows" running from one node to the other, and the weight g_{ij} as the "transition weight" on the arrow running from node i to node j.

Figure 7.1 illustrates key concepts of this graphical approach for testing two primary hypotheses H_1 and H_2 and a secondary hypothesis H_3 of a trial. In this figure, the initial Graph (a) shows three nodes. Two nodes represent H_1 and H_2 with weights $w_1 = 3/4$ and $w_2 = (1 - w_1) = 1/4$, respectively. The node for H_3 shows a weight $w_3 = 0$, which can increase only after the rejection of a primary hypothesis. The nonnegative number $g_{12} = 1/4$ is the transition weight on the arrow going from H_1 to H_2; similarly, $g_{21} = 1/4$ is the transition weight on the arrow going from H_2 to H_1. The transition weight on the arrow going from H_1 to H_3 is 3/4 and that on the arrow going from H_2 to H_3 is also 3/4 satisfying the condition that sum of the transition weights of all outgoing arrows from a single node must be bounded above by 1.

Graph (b) of Fig. 7.1 represents the resulting graph after H_2 is rejected in Graph (a). The rejection of this hypothesis frees its weight w_2 which is then recycled to H_1 and H_3 according to an α-propagation rule addressed in the following for the general case. This rule also calculates new transition weights going from one node to the other for the new graph. Graph (c) of Fig. 7.1 similarly shows the resulting graph if H_1 is rejected in Graph (a). The following shows the general graphical procedure for testing h individual hypotheses $H_1, \ldots, H_h$ for a non-GS trial given their individual unadjusted p-values p_j for $j = 1, \ldots, h$.

(0) Set $\boldsymbol{I} = \{1, \ldots, h\}$. The set of weights $\{w_j(\boldsymbol{I}), j \in \boldsymbol{I}\}$ are such that $0 \leq w_j(\boldsymbol{I}) \leq 1$ with the sum $\sum_{j \in \boldsymbol{I}} w_j(\boldsymbol{I}) \leq 1$.

(i) Select a $j \in \boldsymbol{I}$ such that $p_j < \{w_j(\boldsymbol{I})\}\alpha$ and reject H_j; otherwise stop.

(ii) Update the graph as:

(a) $\boldsymbol{I} = \boldsymbol{I} \backslash \{j\}$, i.e., the index set $\boldsymbol{I}$ without the index j

(b)

$$w_l(\boldsymbol{I}) = w_l(\boldsymbol{I}) + w_j(\boldsymbol{I}) g_{jl}, \, l \in \boldsymbol{I}; \; 0, \text{ otherwise} \tag{7.3.1}$$

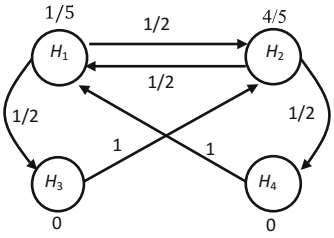
(c)

$$g_{lk} = \frac{g_{lk} + g_{lj} g_{jk}}{1 - g_{lj} g_{jl}}, \text{ where } (l, k) \in \boldsymbol{I}, \, l \neq k \text{ and } g_{lj} g_{jl} < 1; \; 0, \text{ otherwise} \tag{7.3.2}$$

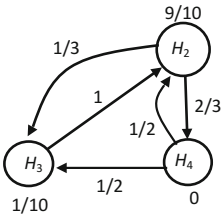
(iii) If $|\boldsymbol{I}| \geq 1$ then go to step (i); otherwise stop

After rejecting H_j, the Eq. (7.3.1) for a new graph updates the weight for H_l to a new weight which is its old weight $w_l(\boldsymbol{I})$ plus the weight $w_j(\boldsymbol{I})$ at H_j multiplied by the transition weight g_{jl} on the arrow connecting H_j to H_l. Also, the transition weights g_{lk} for the new graph are obtained by the algorithm (7.3.2) whose numerator $g_{lk} + g_{lj} g_{jk}$ is the transition weight on the arrow connecting H_l to H_k plus the product of the transition weights on arrows going from H_l to H_k through the rejected hypothesis H_j. The term $g_{lj} g_{jl}$ in (7.3.2) is the product of transition weights on arrows connecting H_l to H_j and then returning to H_l. The approach produces weights $w_l(\boldsymbol{I})$ which satisfy consonance.

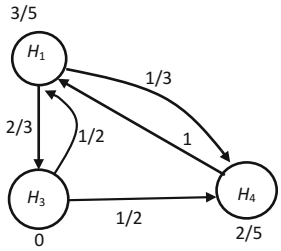
For explaining this procedure, consider a trial, which for demonstrating superiority of a new treatment A + Standard of Care (SOC) to placebo +SOC, plans to test two primary hypotheses H_1 and H_2 and two secondary hypotheses H_3 and H_4, where the pairs (H_1, H_3) and (H_2, H_4) being considered as parent–descendant (Maurer et al. 2011). That is, H_3 is tested only when H_1 is rejected, and similarly, H_4 is tested only when H_2 is rejected. Suppose that the trial specifies a graphical test strategy as in Fig. 7.2 for testing these four hypotheses. The initial Graph (a) in Fig. 7.2 gives a smaller weight of $w_1 = 1/5$ to H_1 as compared to a weight of $w_2 = 4/5$ to H_2 based on the prior experience that the trial may win easily for H_1 at the significance level of $w_1\alpha = 0.005$, but the trial may require a larger significance level of $w_2\alpha = 0.02$ for winning for H_2. As stated before, we assume that all tests in the procedure are



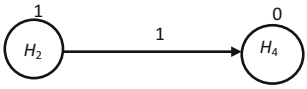
(a) Initial Graph for testing H_1, H_2, H_3, and H_4 along with their weights



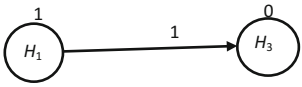
(b) Graph for testing for testing H_2, H_3, and H_4 along with their weights after the rejection of H_1 in (a)



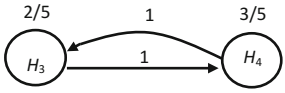
(c) Graph for testing H_1, H_3, and H_4 along with their weights after the rejection of H_2 in (a)



(d) Graph for testing H_2 and H_4 along with their B-weights after the rejection of H_3 in (b)



(e) Graph for testing H_1 and H_3 along with their weights after the rejection of H_4 in (c)



(f) Graph for testing H_3 and H_4 along with their weights after the rejection of H_2 in (b) or H_1 in (c)

Fig. 7.2 Graphical test procedure for two primary hypotheses H_1 and H_2, and two secondary hypotheses H_3 and H_4, where pairs (H_1, H_3) and (H_2, H_4) are parent–descendant

1-sided and the control of the overall Type I error rate is at level $\alpha = 0.025$. The Graph (a) assigns zero-weights to the two secondary hypotheses indicating that we do not want to reject a secondary hypothesis until its parent primary hypothesis is first rejected.

In Graph (a) of Fig. 7.2, $g_{12} = g_{21} = g_{13} = g_{24} = 1/2$ and $g_{32} = g_{41} = 1$. These settings mean that if H_1 was rejected in Graph (a) then a fraction 1/2 of w_1 would be recycled to H_2 so that the weight at H_2 would become $w_2 + (1/2)w_1 = 9/10$ and the remainder $(1/2)w_1 = 1/10$ would go to H_3; the weight at H_4 would remain 0 because there is no arrow going from H_1 to H_4 meaning that $g_{14} = 0$. The rejection of H_1 in Graph (a) would lead to Graph (b) with new transition weights obtained from (7.3.2) as: $g_{23} = 1/3$, $g_{24} = 2/3$, $g_{42} = g_{43} = 1/2$ and $g_{32} = 1$. Similarly, if H_2 was initially rejected in Graph (a), then a fraction 1/2 of w_2 would be recycled to

H_1 so that the weight at H_1 would become $w_1 + (1/2)w_2 = 3/5$ and the remainder $(1/2)w_2 = 2/5$ would go to H_4; the weight at H_3 would remain 0 as there is no arrow going from H_2 to H_3 giving $g_{23} = 0$. The rejection of H_2 in Graph (a) would lead to Graph (c) with transition weights obtained from (7.3.2) as: $g_{13} = 2/3$, $g_{14} = 1/3$, $g_{34} = g_{31} = 1/2$ and $g_{41} = 1$.

The value $g_{32} = 1$ in this Graph (b) indicates that if H_3 was rejected after the rejection of H_1 then the entire weight of $(1/2)w_1 = 1/10$ at H_3 would be recycled to H_2, so that the total weight at H_2 after the rejection of both H_1 and H_3 would be $(w_2+(1/2)w_1 = 9/10)+((1/2)w_1 = 1/10) = 1$; the weight at H_4 would remain zero as in this graph there is no arrow going from H_3 to H_4. Therefore, after the rejection of both H_1 and H_3, the Graph (b) would reduce to Graph (d). Similarly, $g_{41} = 1$ in Graph (c) indicates that if H_4 was rejected after the rejection of H_2 then the entire weight $(1/2)w_2 = 2/5$ at H_4 would be recycled to H_1, so that the total weight at H_1 after the rejection of both H_2 and H_4 would be $(w_1 + (1/2)w_2)) + ((1/2)w_2) = 1$; the weight at H_3 would remain zero. Therefore, after the rejection of both H_2 and H_4, the Graph (c) would reduce to Graph (e). However, if either H_2 was rejected in Graph (b) or H_1 was rejected in Graph (c), then these graphs would reduce to Graph (f).

7.3.5 *Illustrative Example of the Graphical Approach for GS Trials*

The above graphical approach originally developed for testing multiple hypotheses of non-GS trials also applies to GS trials. Recycling of alpha of a rejected hypothesis to other hypotheses occurs similarly, but boundary values for testing the unrejected hypotheses are calculated using spending functions. For example, consider the above trial for testing two primary hypotheses H_1 and H_2 and two secondary hypotheses H_3 and H_4, where pairs (H_1, H_3) and (H_2, H_4) are parent–descendant.

In the beginning, we start with Graph (a) of Fig. 7.2 with four hypotheses $\{H_j, j \in \boldsymbol{I}_1 = \{1, 2, 3, 4\}\}$ identified by four nodes and the associated weights $\{w_j(\boldsymbol{I}_1), j \in \boldsymbol{I}_1\} = \{1/5, 4/5, 0, 0\}$. These weights give the starting overall significance levels $\{w_j(\boldsymbol{I}_1)\alpha, j \in \boldsymbol{I}_1; \alpha = 0.025\} = \{0.005, 0.02, 0, 0\}$, and the j-th one for testing of H_j by using its spending function f_j for determining its boundary values for testing. That is, in the beginning, with Graph (a), we test each H_j $(j \in \boldsymbol{I}_1)$ in the univariate GS testing framework for the control of the overall Type I error rate at level $w_j(\boldsymbol{I}_1)\alpha$ so that the total overall Type I error rate control for the trial is at level $\sum_{j\in \boldsymbol{I}_1} w_j(\boldsymbol{I}_1)\alpha = \alpha$.

For this example, we assume that f_j's are all equal to $f(\gamma, t) = 2\{1 - \Phi(z_{1-\gamma/2}/\sqrt{t})\}$, which is OF-like, and γ is the overall significance level for the repeated testing of a hypothesis. The weights $w_3(\boldsymbol{I}_1) = w_4(\boldsymbol{I}_1) = 0$ indicate that H_3 and H_4 are not tested in Graph (a); if they were tested, they would remain unrejected. The following describes how the procedure performs tests of these hypotheses at

Table 7.3 Tests information at the first look at $t_1 = 1/2$ according to Graph (a)

Overall trial α	0.025			
$j \in \boldsymbol{I}_1$	1	2	3	4
$w_j(\boldsymbol{I}_1)$	1/5	4/5	0	0
$w_j(\boldsymbol{I}_1)\alpha$	0.005	0.02	0	0
$\alpha_j\big(w_j(\boldsymbol{I}_1)\alpha, t_1\big)$	0.00007	0.0010	0	0

Note As the p-values $\{p_j(t_1), j \in \boldsymbol{I}_1\}$ exceed their corresponding boundary values, there is no rejection of a hypothesis at this look

different looks and how it recycles the unused alpha of a rejected hypothesis to other unrejected hypotheses.

Tests at the First Interim Look:
Suppose that at the first look, the information fraction is $t_1 = 1/2$. For this example, we assume that the information fraction at a look remains the same for different hypotheses. If this is not the case, the procedure will proceed as discussed in Sect. 7.3.3. The univariate group sequential procedure for testing a hypothesis in a single-hypothesis trial calculates the boundary values for interim looks given the overall significance level α. However, in our case, there are more than one significance levels as $\big\{w_j(I)\alpha, j \in \boldsymbol{I}_1; \alpha = 0.025\big\} = \{0.005, 0.02, 0, 0\}$ assigned to $\{H_j, j \in \boldsymbol{I}_1\}$. These overall significance levels, and the use of the OF-like spending function at $t_1 = 0.5$, then give the boundary values $\{\alpha_j\big(w_j(\boldsymbol{I}_1)\alpha, t_1\big), j \in \boldsymbol{I}_1\} = \{0.00007, 0.0010, 0, 0\}$ for testing $\{H_j, j \in \boldsymbol{I}_1\}$ at the first look. Note that the subscript of t identifies the look number and the subscript j for the hypothesis H_j being tested. Also note that the boundary value of $\alpha_j\big(w_j(\boldsymbol{I}_1)\alpha, t_k\big)$ is a function of the overall significance level $w_j(\boldsymbol{I}_1)\alpha$ assigned to H_j and the information fraction t_k at look k; here $k = 1$.

Suppose that at the first look, the unadjusted p-values $\{p_j(t_1), j \in \boldsymbol{I}_1\}$ associate with $\{H_j, j \in \boldsymbol{I}_1\}$ are such that $p_j(t_1) \geq \alpha_j\big(w_j(\boldsymbol{I}_1)\alpha, t_1\big)$for $j \in \boldsymbol{I}_1$; consequently, the trial will continue to the second look without rejection of a hypothesis at the first look. For recording purposes, one can summarize the above testing information at the first look as in Table 7.3.

Tests at the Second Look:

Suppose that the trial conducts the second look when the information fraction is $t_2 = 3/4$. Since none of the hypotheses was rejected at the first look, we begin with Graph (a) at the second look, by using the same overall significance levels of $\{w_j(\boldsymbol{I}_1)\alpha, j \in \boldsymbol{I}_1\}) = \{0.005, 0.02, 0, 0\}$ that were used at the first look. However, as $t_2 = 3/4$ at the second look, the use OF-like spending function leads to the boundary values of $\{\alpha_j\big(w_j(\boldsymbol{I}_1)\alpha, t_2\big), j \in \boldsymbol{I}_1\} = \{0.00117, 0.0069, 0, 0\}$ for testing H_j for $j \in \boldsymbol{I}_1$. The boundary values for testing H_3 and H_4 remain zero, as there is no rejection of a primary hypothesis so far. Suppose that at this second look, the observed p-values associated with for H_1, H_3, H_2, and H_4 are $p_1(t_2) = 0.001$, $p_2(t_2) = 0.020$, $p_3(t_2) = 0.040$, and $p_4(t_2) = 0.091$, respectively. These results lead to the rejection

Table 7.4 **a** Tests information at the second look at $t_2 = 3/4$ according to Graph (a) after no rejection at the first look. **b** Tests information at the second look at $t_2 = 3/4$ according to Graph (b) after the rejection of H_1 at this look

Overall trial α	0.025 (Table 7.4a)			
$j \in \boldsymbol{I}_1$	1	2	3	4
$w_j(\boldsymbol{I}_1)$	1/5	4/5	0	0
$w_j(\boldsymbol{I}_1)\alpha$	0.005	0.02	0	0
$\alpha_j\big(w_j(\boldsymbol{I}_1)\alpha, t_2\big)$	0.00117	0.0069	0	0
p-values: $p_j(t_2)$	0.001	0.020	0.040	0.091
Overall trial α	0.025 (Table 7.4b)			
$j \in \boldsymbol{I}_2$	–	2	3	4
$w_j(\boldsymbol{I}_2)$	–	9/10	1/10	0
$w_j(\boldsymbol{I}_2)\alpha$	–	0.0225	0.00255	0
$\alpha_j\big(w_j(\boldsymbol{I}_2)\alpha, t_2\big)$	–	0.00802	0.00047	0
p-values: $p_j(t_2)$	0.001	0.020	0.040	0.091

Note H_1 is rejected as $p_1(t_2) = 0.001$ is less than its boundary value of 0.00117 (Table 7.4a)
Note As $p_2(t_2) = 0.020 > 0.00802$ and $p_3(t_2) = 0.040 > 0.00047$, there is no additional rejection at the second look (Table 7.4b)

of H_1 at the second look as $p_1(t_2) = 0.001$ is less than its boundary value of 0.00117; see Table 7.4a.

The above rejection of H_1 at the second look then frees its overall significance level of $w_1(\boldsymbol{I}_1)\alpha) = 0.005$ as unused alpha which is recycled to the remaining three hypotheses for their tests according to Graph (b). This revised graph, constructed after the rejection of H_1, allows retesting of the remaining hypotheses $\{H_j, j \in \boldsymbol{I}_2 = \{2, 3, 4\}\}$ at their corresponding overall significance levels of $\big\{w_j(\boldsymbol{I}_2)\alpha, j \in \boldsymbol{I}_2\big\} = \{-, (9/10)\alpha, (1/10)\alpha, (0)\alpha\} = \{-, 0.0225, 0.00255, 0\}$. Note that the overall significance levels for testing H_2, H_3 are now increased creating the possibility of additional rejections of hypotheses at the second look according to Graph (b). The use OF-like spending function with these updated overall significance levels and $t_2 = 3/4$, then produces the boundary values of $\{\alpha_j\big(w_j(\boldsymbol{I}_2)\alpha, t_2\big), j \in \boldsymbol{I}_2\} = \{-, 0.00802, 0.00047, 0\}$ for testing H_j for $j \in \boldsymbol{I}_2$; see Table 7.4b. However, in this table, as $p_2(t_2) = 0.020 > 0.00802$ and $p_3(t_2) = 0.040 > 0.00047$, there is no additional rejections at the second look. Therefore, the trial moves to the next look which is the final look.

Tests at the Final Look:
After the rejection of H_1 at the second look, the tests for the remaining three hypotheses $\{H_j, j \in \boldsymbol{I}_2\}$ at the final look start with the same Graph (b) and the same overall significance levels of $\big\{w_j(\boldsymbol{I}_2)\alpha, j \in \boldsymbol{I}_2\big\} = \{-, (9/10)\alpha, (1/10)\alpha, (0)\alpha\} = \{-, 0.0225, 0.00255, 0\}$ for testing $\{H_j, j \in \boldsymbol{I}_2 = \{2, 3, 4\}\}$. However, as $t_3 = 1$ at this look, the use of the same OF-like spending function produces the boundary val-

Table 7.5 **a** Tests information at the final look at $t_3 = 1$ according to Graph (b) after the rejection of H_1 at the second look. **b** Tests information at the final look at $t_3 = 1$ according to Graph (f) after the rejection of H_1 at the second look and the rejection of H_2 at the final look

Overall trial α	0.025 (Table 7.5a)			
$j \in \boldsymbol{I}_2$	–	2	3	4
$w_j(\boldsymbol{I}_2)$	–	9/10	1/10	0
$w_j(\boldsymbol{I}_2)\alpha$	–	0.0225	0.00255	0
$\alpha_j\big(w_j(\boldsymbol{I}_2)\alpha, t_3\big)$	–	0.01988	0.00234	0
p-values: $p_j(t_3)$	–	0.012	0.008	0.041
Overall trial α	0.025 (Table 7.5b)			
$j \in \boldsymbol{I}_2$	–	2	3	4
$w_j(\boldsymbol{I}_2)$	–	–	2/5	3/5
$w_j(\boldsymbol{I}_2)\alpha$	–	–	0.010	0.015
$\alpha_j\big(w_j(\boldsymbol{I}_2)\alpha, t_3\big)$	–	–	0.00907	0.013440
p-values: $p_j(t_3)$	–	0.012	0.008	0.041

Note As $p_2(t_3) = 0.0120 < 0.01988$ and $p_3(t_2) = 0.008 > 0.00234$, there is a rejection of H_2 at this look (Table 7.5a)
Note As $p_3(t_2) = 0.008 < 0.00907$, H_3 is also rejected at this look (Table 7.5b)

ues of $\{\alpha_j\big(w_j(\boldsymbol{I}_2)\alpha, t_3\big), j \in \boldsymbol{I}_2\} = \{-, 0.01988, 0.00234, 0\}$ for testing of $\{H_j, j \in \boldsymbol{I}_2\}$ at this look. Suppose that at this final look, the observed p-values associated with for H_3, H_2, and H_4 are $p_2(t_3) = 0.012, p_3(t_3) = 0.008$, and $p_4(t_3) = 0.041$, respectively. These results then lead to the rejection of H_2 at the final look as its $p_2(t_2) = 0.012$ is less than its corresponding boundary value of 0.01988; see Table 7.5a.

Now, as H_1 was rejected at the second look and as H_2 is rejected at the final look, the tests of hypotheses H_3 and H_4 at the final look will be at the increased overall significance levels of $\{w_j(\boldsymbol{I}_3)\alpha, j \in \boldsymbol{I}_3 = \{3, 4\}\} = \{(2/5)\alpha, (3/5)\alpha\} = \{0.010, 0.015\}$ according to Graph (f). These with the OF-like spending function give the boundary values of $\{-, -, 0.00907, 0.01344\}$ for testing $\{H_j, j \in \boldsymbol{I}_3\}$, rejecting also H_3 in this final look, as $p_3(t_2) = 0.008$ is less than 0.00907; see Table 7.5b. Consequently, the remaining H_4 can be tested at this look the at the full overall significance level of $\alpha = 0.025$ which gives the boundary value of 0.0220 for its testing. Therefore, as $p_4(t_2) = 0.041 > 0.0220$ for H_4, the trial stops without the rejection of this hypothesis.

7.4 Testing a Secondary Hypothesis When the Trial Stops After the Rejection of a Primary Hypothesis

Consider, for example, a trial with two looks for testing a primary hypothesis H_1 and a secondary hypothesis H_2 with one interim look and a final look at information fractions t_1 and $t_2 = 1 (0 < t_1 < t_2)$, respectively. The trial, if it rejects H_1 at the interim look, stops at that look for ethical reasons. This will in general be the case when H_1 is associated with an endpoint such as mortality. Therefore, H_2 must be tested at the same interim look when H_1 is rejected, and this test for H_2 must occur after the rejection of H_1.

A question often arises: Can the test of H_2 at the interim look, after the rejection of H_1 at that look, be at the full significance level α (e.g., $\alpha = 0.025$)? This question may arise based on the considerations that H_2 is not tested unless H_1 is first rejected and there is no repeated testing of H_2 after the rejection of H_1. Tamhane et al. (2010) (also Xi and Tamhane 2015) showed that the answer of this question is affirmative, only for the special case when the test statistics for testing H_1 and H_2 are independent. However, this can inflate the overall Type I error rate if the test statistics are correlated. They show that with certain distributional assumptions of the test statistics, the exact adjusted significance level for testing H_2 can be found if this correlation is known. However, if this correlation is unknown, then an upper bound of the adjusted significance levels can be set that covers all correlations. The following revisits this work in some detail because of its importance for clinical trial applications.

We assume that the trial is designed to demonstrate superiority of a new treatment to control such that $H_i : \delta_i \leq 0 (i = 1, 2)$, where δ is the treatment difference parameter. Also, X and Y are the test statistics for testing H_1 and H_2, respectively, which become $(X(t_k), Y(t_k))$ at information times $t_k (k = 1, 2)$. Also, following the results of Sect. 7.2, we assume that each pair $(X(t_1), X(t_2))$ and $(Y(t_1), Y(t_2))$ follows a standard bivariate normal distribution with the same correlation of $\sqrt{t_1}$. Further, we assume that each pair $(X(t_1), Y(t_1))$ and $(X(t_2), Y(t_2))$ follows a standard bivariate normal distribution with correlation coefficient of $\rho \geq 0$. Furthermore, we assume that (c_1, c_2) and (d_1, d_2) are boundary values for testing H_1 and H_2, respectively, so that d_1 is used only when H_1 is rejected at the first look; similarly, d_2 is used only when H_1 being retained at the first look is rejected at the final look. The test strategy for this 2-stage design can then be stated as follows:

Step 1:

$$\text{If } X(t_1) \leq c_1 \longrightarrow \text{Go to Step 2}$$

$$\text{If } X(t_1) > c_1 \longrightarrow \text{Reject } H_1 \text{ and test } H_2$$

$$\text{If } Y(t_1) > d_1 \longrightarrow \text{Reject } H_2\text{; else, retain it}$$

(In either case terminate the trial)

Step 2:

$$\text{If } X(t_2) \leq c_2 \longrightarrow \text{Terminate the trial without any rejection}$$

$$\text{If } X(t_2) > c_2 \longrightarrow \text{Reject } H_1 \text{ and test } H_2$$

$$\text{If } Y(t_2) > d_2 \longrightarrow \text{Reject } H_2\text{; else, retain it.}$$

Determining the Boundary Values of the Procedure
Tests for H_1 and H_2 for the above 2-stage design can be carried out by the method based on the closed testing for GS trials as addressed in Sect. 7.3.2. The intersection hypothesis H_{12} would be tested by the weighted Bonferroni tests with weights of $w_1 = 1$ and $w_2 = 0$ associated with the tests of H_1 and H_2, respectively; $w_2 = 0$ for H_2 implies that this weight can increase only after H_1 is rejected. Therefore, for this design, the rejection of H_1 at level α implies the rejection of H_{12} at level α. Consequently, H_2 can be tested at the full significance level α. But as the trial is a GS trial with one interim look, the boundary values c_1 and c_2 for testing H_1 can then be found from the following two equations:

$$\Pr\{X(t_1) > c_1 | H_1\} = f_1(\alpha, t_1(X))$$

and

$$f_1(\alpha, t_1(X)) + \Pr\{X(t_1) \leq c_1 \cap X(t_2) > c_2 | H_1\} = f_1(\alpha, t_2(X) = 1),$$

where $f_1(\alpha, t)$ is the spending function for testing H_1, and $t_1(X)$ and $t_2(X)$ are the information fractions for testing H_1 at the first and final looks, respectively. For example, when $f_1(\alpha, t)$ is OF-like, $\alpha = 0.025$, and $t_1(X) = 0.5$, then $c_1 = 2.95901$ and $c_2 = 1.96869$ on the normal z-scale which translates to $\alpha_1(0.025, t_1(X) = 0.5) = 0.00153$ and $\alpha_2(0.025, t_2(X) = 1) = 0.02449$ on the p-value scale.

Since the significance level α for the test of H_1 after its rejection recycles to test H_2, the boundary values (d_1, d_2) for H_2 need to be calculated also by a GS method but at the same level α. Reason for this is that, though H_2 is tested after the rejection of H_1, the rejection of H_2, similar to that for H_1, can occur either at the first look or at the final look. Thus, if one uses the Pocock (1977) method for calculating the

boundary values for testing H_2, then at $\alpha = 0.025$, $t_1(Y) = 0.5$ and $t_2(Y) = 1$, the value $d = d_1 = d_2 = 2.17828$ (on the z-scale) which is 0.01469 on the p-value scale. However, the test statistics X and Y in many applications will be positively correlated. Therefore, if this correlation is ρ, and remains the same for the two looks, then it is natural to ask a key question: Is it possible to take advantage of this correlation and find $d^* \leq d$ while maintaining the control of the overall Type I error rate at level $\alpha = 0.025$?

The following shows that this is possible. But the extent of the gain depends on the value of ρ. Larger is the value of ρ on the interval $0 \leq \rho \leq 1$, lesser is the gain, and as ρ approaches one, the value of d^* approaches d determined by the Pocock (1977) method.

Determining the Value of d^*

Testing of H_1 and H_2 gives rise to three null hypotheses configurations $H_{12} = H_1 \cap H_2$, $H_1 \cap K_2$, and $K_1 \cap H_2$, where K_1 and K_2 are alternatives to H_1 and H_2, respectively. The overall Type I error rate for testing H_1 and H_2 under the first two configurations is ≤α. That is, tests for H_1 control this error rate at level α regardless of whether H_2 is true or false. Therefore, we need to find $z_y = d^*$ by solving for z_y in the following equation under $K_1 \cap H_2$.

$$\Pr\{X(t_1) > c_1 \cap Y(t_1) > z_y\} + \Pr\{X(t_1) \leq c_1 \cap X(t_2) > c_2 \cap Y(t_2) > z_y\} = \alpha. \tag{7.4.1}$$

Now, Cov $\{X(t_1), X(t_2)\} = \sqrt{t_1}$, Cov $\{X(t_1), Y(t_2)\} = \sqrt{t_1}\,\rho$, and $\text{Cov}\{X(t_1), Y(t_1)\} = \text{Cov}\{X(t_2), Y(t_2)\} = \rho$. Also, $\text{E}\{X(t_1)\} = \theta\sqrt{t_1}$, $E\{X(t_2)\} = \theta$, and $E\{Y(t_i)\} = 0$ for $i = 1, 2$, because of $K_1 \cap H_2$ and θ being the drift parameter for X. Further, one can show that conditional on $X(t_2) = x(t_2)$, the test statistics $X(t_1)$ and $Y(t_2)$ are independently normally distributed as:

$X(t_1)$ is $N\left\{x(t_2)\sqrt{t_1}, 1 - t_1\right\}$ and $Y(t_2)$ is $N\left\{(x(t_2) - \theta)\rho, 1 - \rho^2\right\}$

Therefore, the Eq. (7.4.1) for finding $z_y = d^*$ can be written as:

$$\alpha = 1 - \Phi\left(c_1 - \theta\sqrt{t_1}\right) - \Phi\left(z_y\right) + \Phi_{12}\left(c_1 - \theta\sqrt{t_1}, z_y; \rho\right) + \int_{c_1-\theta}^{\infty} \Phi\left(\frac{c_1 - \theta\sqrt{t_1} - u\sqrt{t_1}}{\sqrt{1 - t_1}}\right) \Phi\left(\frac{-z_y - u\rho}{\sqrt{1 - \rho^2}}\right) \phi(u)du, \tag{7.4.2}$$

where Φ and ϕ are the density and the cumulative distribution functions of the $N(0,1)$ random variable, and Φ_{12} is the cumulative distribution function of the standard bivariate normal distribution with correlation coefficient of ρ.

Therefore, specifying values of ρ, t_1, c_1, and c_2, one can construct a graph $z_y = f(\theta)$ over the interval $\theta > 0$ that satisfy Eq. (7.4.2). Figure 7.3 shows such graphs for different values of ρ when $\alpha = 0.025$ (1-sided), $t_1 = 0.5$, and $c_1 = 2.95901$ and $c_2 = 1.96869$ on using the OF-like α-spending function. Constructing such a graph for a given ρ then gives $d^* = z_y$ where the maximum occurs for that ρ. Such a

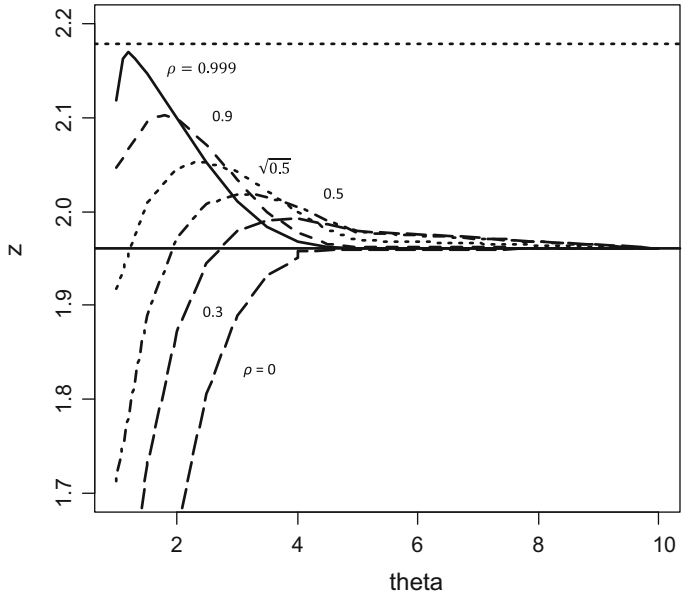


Fig. 7.3 Graph of $z_y = f(\theta)$ over the interval $\theta > 0$ satisfying Eq. (7.4.2). In this graph, theta $= \theta$ and $z = z_y$. The horizontal dashed line in the graph represents the Pocock boundary

selection of d^* assures that the right side of (7.4.2) is $\leq \alpha$ for all $\theta > 0$. Table 7.6, for the above values of α, t_1, c_1, and c_2, gives d^* values and the corresponding α_{d^*} values on the p-value scale for values of ρ shown in column 1 of this table. This table also includes values of θ^* where the d^* values occur. Results of this table show that if the test statistics for testing H_1 and H_2 are uncorrelated, then the test for H_2 at a look after the rejection of H_1 at that look can be at the full significance level α. However, if these test statistics are correlated, then this significance level for testing H_2 is correlation dependent. For positive correlations, this significance level for testing decreases with increasing correlation value and approaches to a value by the Pocock (1977) method.

7.5 Concluding Remarks

Confirmatory clinical trials have been gold standards for establishing efficacy of new treatments. However, such trials when designed with a single primary endpoint do not provide sufficient information when one must assess the effect of the new treatment on different but important multiple characteristics of the disease. For these situations, trials include multiple endpoints related to these disease characteristics and a statistical plan for testing multiple hypotheses on these endpoints for establishing

Table 7.6 Values of d^* for the 2-stage design for different correlations when $\alpha = 0.025$ (1-sided), $t_1 = 0.5$, and $c_1 = 2.95901$ and $c_2 = 1.96869$ on using the OF-like α-spending function

Correlation ρ	d^* (Z-scale)	α_{d^*} (p-value scale)	$\theta = \theta^*$
0.0	1.95996	0.02500	$\theta^* = \text{all}\,\theta > 6.5$
0.1	1.96958	0.02444	4.54
0.2	1.98063	0.02382	4.12
0.3	1.99160	0.02321	4.00
0.4	2.00497	0.02248	3.43
0.5	2.01872	0.02176	3.11
0.6	2.03407	0.02097	2.78
$\sqrt{0.5}$	2.05314	0.02003	2.45
0.8	2.07326	0.01907	2.15
0.9	2.10262	0.01775	1.79
0.99	2.15450	0.01560	1.31
0.999	2.17026	0.01499	1.20
PK value	$d = 2.17828$	$\alpha_d = 0.01469$	–
Conservative		$\alpha/2 = 0.0125$	–

Note $\theta = \theta^*$ is the value of θ where z_y is maximum on the graph $z_y = f(\theta)$ over the interval $\theta > 0$ satisfying Eq. (7.4.2)

efficacy findings of new treatments. However, testing multiple hypotheses in a trial can raise multiplicity issues causing inflation of the Type I error rate. Fortunately, many novel new statistical methods, such as gatekeeping and graphical methods, are now available in the literature for addressing all types of multiplicity issues of clinical trials. These novel methods have advanced the role of statistical methods in designing modern clinical trials with multiple endpoints or multiple objectives.

In clinical trials with serious endpoints, such as death, often a new treatment is added to an existing therapy for detecting a relatively small but clinically relevant improvement in the treatment effect beyond what the existing therapy provides. Designing and conducting such and other trials for serious diseases can be complex, as these trials may require thousands of patients to enroll and several years to complete. Ethical and economic reasons may necessitate that these trials be designed with interim looks for finding the effect of the treatment at an earlier time point allowing the possibility of stopping the trial early when it becomes clear that the study treatment has the desired efficacy or it is futile to continue the trial further. Such trials that allow analyses of the accumulated data at interim looks for the possibility of stopping the trial early for efficacy or futility reasons are commonly known as group sequential trials.

Obviously, interim analyses of the data in a group sequential trial amounts to repeated testing of one or more hypotheses and would result in Type I error rate inflation, so multiplicity adjustment would be required for drawing valid inference. As mentioned in this chapter, several approaches have been cited in the literature for

addressing the control of Type I error rate for repeated tests of a single hypothesis related to a single primary endpoint of the trial. However, approaches for addressing the multiplicity issues for testing multiple hypotheses related to multiple endpoints of group sequential trials are less frequent in the literature.

This chapter, in addition to providing a brief review of procedures and citing key references thereof for the repeated testing procedures of a single endpoint hypothesis in groups sequential trials, considers procedures for handling multiplicity issues for repeated testing of multiple endpoint hypotheses of trials. In this regard, we distinguish two cases of multiple endpoints which guide the approach for handling the multiplicity issue. The first case arises when after a hypothesis is rejected at an interim look, the trial can continue to test other hypotheses at subsequent looks for additional claims. A testing approach for this is to use the Bonferroni inequality which requires splitting the significance level either among the endpoints or among the different looks. This approach is now rarely used because of the low power of the tests.

A better approach (discussed in Sect. 7.3.2) is to consider the use of the closed testing with the weighted Bonferroni tests of the intersection hypotheses, when the weights satisfy the consonance property. This approach allows recycling of the significance level of a rejected hypothesis to the other hypotheses, thus increasing the power of the test procedure. However, as discussed, the recycling of the significance level from a rejected hypothesis to other hypotheses occurs through an α-spending function and is not simple as with non-group sequential trials.

The closed testing-based approach can be manageable when testing 2–3 hypotheses, but it may be difficult to set up for testing more than three hypotheses, for example, when testing two primary and two secondary hypotheses in a trial, as selecting weights for the weighted Bonferroni tests that satisfy the consonance property can be complicated. For these advanced cases, a graphical approach is recommended which is easier to plan, to use, and to communicate to non-statisticians. This chapter illustrates the application of these two approaches through illustrative examples, showing details of the derivations of the significance levels.

The second case arises (discussed in Sect. 7.4), for example, for a group sequential trial designed for testing a primary and a secondary endpoint hypotheses, and the trial stops at an interim look for ethical reasons when the primary hypothesis is rejected at that look in favor of the study treatment. The issue then arises as to what would be the significance level for testing the secondary hypothesis at that look, given that that the secondary hypothesis is tested only after the primary one is rejected first. This issue has been investigated in the literature in detail, but we have revisited it for increasing its awareness, as group sequential trials are frequently designed with a single primary hypothesis and multiple secondary hypotheses. A natural way to address this problem is to use the graphical procedure and recycle the significance level of the rejected primary hypothesis to secondary hypotheses using the Pocock-like α-spending function.

Glimm et al. (2010) illustrated that using the Pocock-like group sequential test to the secondary hypotheses has a power advantage over the O'Brien-Fleming boundary. Other approaches that consider correlation information between the test statistics

can also be used for simple cases, for example, for the case of testing a single primary and a single secondary hypothesis.

Power considerations in designing GS trials that tests multiple hypotheses are also important. However, this topic is beyond the scope of this paper. The power issue would generally be like those for testing multiple hypotheses in a non-GS trial.

Acknowledgements The authors are grateful to Drs. Frank Bretz, Dong Xi, and Estelle Russek-Cohen for providing detailed comments on this chapter which helped in improving the readability of the materials presented.

Disclaimer This paper reflects the views of the authors and must not be construed to represent FDA's views or policies.

References

Alosh, M., Bretz, F., & Huque, M. F. (2014). Advanced multiplicity adjustment methods in clinical trials. *Statistics in Medicine, 33*(4), 693–713.

Armitage, P., McPherson, C. K., Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society Series A, 132*, 235–244.

Bretz, F., Maurer, W., Brannath, W., & Posh, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine, 28,* 586–604.

Bretz, F., Maurer, W., & Maca, J. (2014). Graphical approaches to multiple testing. Chapter 14. In W. Young & D. G. Chen (Eds.), *Clinical trial biostatistics and biopharmaceutical applications* (pp. 349–394). Boca Raton: Chapman and Hall/CRC Press.

Bretz, F., Posch, M., Glimm, E., Klinglmueller, F., Maurer, W., & Rohmeyer, K. (2011). Graphical approaches for multiple comparison procedures using weighted Bonferroni Simes or parametric tests. *Biometrical Journal, 53*(6), 894–913.

SAS Online Doc. *Version 8*. Copyright 1999 by SAS Institute Inc., Cary, NC, U.S.A.

Dmitrienko, A., Offen, W. W., & Westfall, P. H. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine, 22,* 2387–2400.

Dmitrienko, A., & Tamhane, A. C. (2009). Gatekeeping procedures in clinical trials. In A. Dmitrienko, A. C. Tamhane, & F. Bretz (Eds.), *Multiple testing problems in pharmaceutical statistics* (Chap. 1). Boca Raton, FL: Chapman & Hall/CRC Biostatistics Series.

Dmitrienko, A., Tamhane, A. C., & Wiens, W. (2008). General multi-stage gatekeeping procedures. *Biometrical Journal, 50,* 667–677.

EAST software 6.3. (2014). By Cytel Software Corporation, Cambridge, MA, U.S.A.

Ellenberg, S. S., Fleming, T. R., & DeMets, D. L. (2017). *Data monitoring committees in clinical trials*. New York: Wiley.

Emerson, S. (2007). Frequentist evaluation of group sequential trial designs. *Statistics in Medicine, 26,* 5047–5080.

Follmann, D. A., Proschan, M. A., & Geller, N. L. (1994). Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics, 50,* 325–336.

Food and Drug Administration. (2017). Guidance for industry: Multiple endpoints in clinical trials. Retrieved May 3, 2017 from https://www.fda.gov/ucm/groups/fdagov-public/@fdagov-drugs-gen/documents/document/ucm536750.pdf.

Glimm, E., Maurer, W., & Bretz, F. (2010). Hierarchical testing of multiple endpoints in group sequential trials. *Statistics in Medicine, 29,* 219–228.

Hellmich, M. (2001). Monitoring clinical trials with multiple arms. *Biometrics, 57,* 892–898.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple significance testing. *Biometrika, 75,* 800–802.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York: Wiley.

Hommel, G., Bretz, F., & Maurer, W. (2007). Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies. *Statistics in Medicine, 26,* 4063–4073.

Huque, M. F., Dmitrienko, A., & D'Agostino, R. (2013). Multiplicity issues in clinical trials with multiple objectives. *Statistics in Biopharmaceutical Research, 5*(4), 321–337.

Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials.* Boca Raton, FL: Chapman and Hall/CRC.

Lan, K. K. G., & DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika, 70,* 659–663.

Lan, K. K. G., & Wittes, J. (1988). The B-value: A tool for monitoring data. *Biometrics, 44,* 579–585.

Marcus, R., Peritz, E., & Gabriel, K. R. (1976). On closed testing procedure with special reference to ordered analysis of variance. *Biometrika, 63,* 655–660.

Maurer, W., & Bretz, F. (2013). Multiple testing in group sequential trials using graphical approaches. *Statistics in Biopharmaceutical Research, 5*(4), 311–320.

Maurer, W., Glimm, E., & Bretz, F. (2011). Multiple and repeated testing of primary, co-primary and secondary hypotheses. *Statistics in Biopharmaceutical Research, 3,* 336–352.

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics, 5,* 549–556.

O'Neill, R. T. (1997). Secondary endpoints cannot be validly analyzed if primary endpoint does not demonstrate clear statistical significance. *Controlled Clinical Trials, 18,* 550–556.

Pocock, S. J. (1977). Group sequential methods in the design and analysis clinical trials. *Biometrika, 64,* 191–199.

Proschan, M. A., Lan, K. K. G., & Wittes, J. T. (2006). *Statistical monitoring of clinical trials. A unified approach*. New York: Springer.

Stallard, N., & Friede, T. (2008). A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine, 27,* 6209–6227.

Tamhane, A. C., Mehta, C. R., & Liu, L. (2010). Testing a primary and a secondary endpoint in a group sequential design. *Biometrics, 66,* 1174–1184.

Tang, D. L., & Geller, N. L. (1999). Closed testing procedures for group sequential clinical trials with multiple endpoints. *Biometrics, 55,* 1188–1192.

Wassmer, G., & Brannath, W. (2016). *Group sequential and confirmatory adaptive designs in clinical trials*. Switzerland: Springer International Publishing AG.

Whitehead, J. (1997). *The design and analysis of sequential clinical trials*. Chichester: Wiley.

Xi, D., Glimm, E., Bretz, F. (2016). *Multiplicity in chapter 3 of cancer clinical trials: Current and controversial issues in design and analysis*. In: S. L. George, X. Wang, & H. Pang. CRC Press, Taylor and Francis Group.

Xi, D., & Tamhane, A. C. (2015). Allocating recycled significance levels in group sequential procedures for multiple endpoints. *Biometrical Journal, 57*(1), 90–107.

Ye, Y., Li, A., Liu, L., & Yao, B. (2013). A group sequential Holm procedure with multiple primary endpoints. *Statistics in Medicine, 32,* 1112–1124.

Zhu, L., Ni, L., & Yao, B. (2011). Group sequential methods and software applications. *The American Statistician, 65*(2), 127–135.