

Chapter 2

Interim Evaluation of Efficacy in Clinical Trials with Two Co-primary Endpoints

Abstract We discuss group-sequential designs for early efficacy stopping in clinical trials with two outcomes as co-primary endpoints, i.e., trials designed to evaluate whether the test intervention is superior to the control on *all* primary endpoints. We discuss two outcome scale situations: (i) when both outcomes are continuous, and (ii) when both outcomes are binary. We derive the power and sample size formulae within two decision-making frameworks: (A) evaluation of superiority not necessarily simultaneously and (B) evaluation of superiority for the two primary endpoints simultaneously. We evaluate the behaviors of sample size and power with varying design characteristics and provide an example to illustrate the methods.

Keywords Average sample number • Binary outcomes • Continuous outcomes • Efficacy stopping • Lan–DeMets error-spending method • Maximum sample size • O’Brien–Fleming-type boundary • Pocock-type boundary • Type I error • Type II error • Intersection–union test

2.1 Introduction

In this chapter, we describe the methods for designing group-sequential clinical trials with two outcomes as co-primary endpoints, where a trial is designed to evaluate whether the test intervention is superior to the control on *all* primary endpoints, and to be terminated early when evidence is overwhelming (early stopping for efficacy). Group-sequential designs for multiple co-primary endpoints are a more attractive design feature rather than the fixed-sample designs because they offer the possibility of stopping a trial when evidence is overwhelming, thus providing efficiency (Hung and Wang 2009) as the sample size in fixed-sample clinical trials with multiple co-primary endpoints is often unnecessarily large and impractical.

Recently, Asakura et al. (2014, 2015) discussed two decision-making frameworks associated with interim evaluation of efficacy in clinical trials with two

co-primary endpoints in a group-sequential setting. One framework is to reject the null hypothesis if and only if statistical significance is achieved for the two endpoints simultaneously (i.e., at the same interim time-point of the trial). The other is a generalization of this, i.e., to reject the null hypothesis if superiority is demonstrated for the two endpoints at any interim time-point (i.e., not necessarily simultaneously). The former framework is independently discussed by Chang et al. (2014) and evaluated in clinical trials with two co-primary endpoints. Hamasaki et al. (2015) discussed more flexible decision-making frameworks, allowing the different time-points of analyses among the endpoints. In addition, Jennison and Turnbull (1993) and Cook and Farewell (1994) discussed the decision-making frameworks associated with interim evaluation of efficacy and futility to monitor the efficacy and safety responses and considered a simple method for determining the boundaries as if the responses are not correlated (i.e., assuming zero correlations between the responses). The methods for the interim evaluation of efficacy and futility will be discussed in Chap. 4.

We discuss two outcome scale situations: (i) when both outcomes are continuous (in Sect. 2.2) and (ii) when both outcomes are binary (Sect. 2.3). We derive the power and sample size formulae within two decision-making frameworks for early efficacy stopping: (A) evaluation of superiority not necessarily simultaneously and (B) evaluation of superiority for the two primary endpoints simultaneously. We evaluate the behaviors of sample size and power with varying design characteristics and provide an example to illustrate the methods. For more than two endpoints, see Hamasaki et al. (2015).

2.2 Continuous Outcomes

2.2.1 Notation and Statistical Setting

Consider a randomized, group-sequential clinical trial of comparing the test intervention (T) with the control intervention (C). Two continuous outcomes (i.e., $K = 2$), EP1 and EP2, are to be evaluated as co-primary endpoints. Suppose that a maximum of L analyses is planned, where the same number of planned analyses with the same information space is selected for both endpoints. Let n_l and $r_C n_l$ be the cumulative number of participants on the T and the C at the l th analysis ($l = 1, \dots, L$), respectively, where r_C is the allocation ratio of the C to the T. Hence, up to n_L and $r_C n_L$ participants are recruited and randomly assigned to the T and the C, respectively. Then, there are n_L paired outcomes (Y_{T1i}, Y_{T2i}) ($i = 1, \dots, n_L$) for the T and $r_C n_L$ paired outcomes (Y_{C1j}, Y_{C2j}) ($j = 1, \dots, r_C n_L$) for the C. Assume that (Y_{T1i}, Y_{T2i}) and (Y_{C1j}, Y_{C2j}) are independently bivariate distributed with means $E[Y_{Tki}] = \mu_{Tk}$ and $E[Y_{Ckj}] = \mu_{Ck}$, variances $\text{var}[Y_{Tki}] = \sigma_{Tk}^2$ and $\text{var}[Y_{Ckj}] = \sigma_{Ck}^2$, and correlation $\text{corr}[Y_{T1i}, Y_{T2i}] = \rho_T$ and $\text{corr}[Y_{C1j}, Y_{C2j}] = \rho_C$, respectively ($k = 1, 2$). For simplicity, the variances are assumed to be known and common,

i.e., $\sigma_{Tk}^2 = \sigma_{Ck}^2 = \sigma_k^2$. Note that the method can be applied to the case of unknown variances. For the fixed-sample designs, Sozu et al. (2011) discuss a method for the unknown variance case and show that the calculated sample size is nearly equivalent to that for the known variance in the setting of 80 % or 90 % power at 2.5 % significance level for one-sided test. By analogy from the fixed-sample designs, there may be no practical difference in the group-sequential setting and the methodology for a known variance provides a reasonable approximation for the unknown variances case.

Let $\delta_k = \mu_{Tk} - \mu_{Ck}$ and $\Delta_k = \delta_k / \sigma_k$ denote the mean differences and standardized mean differences for the T and the C, respectively ($k = 1, 2$). Suppose that positive values of δ_k represent the test intervention's benefit. There is an interest in conducting a one-sided hypothesis test at the significance level of α to evaluate whether the T is superior to the C on both endpoints. The hypothesis for each endpoint is tested at significance level of α : The hypotheses are $H_{0k}: \delta_k \leq 0$ versus $H_{1k}: \delta_k > 0$. For multiple co-primary endpoints, "success" can be declared if the superiority is achieved on both endpoints. The hypotheses for co-primary endpoints are the null hypothesis $H_0: H_{01} \cup H_{02}$ versus the alternative hypothesis (the union H_0 of both individual nulls is tested against the intersection alternative $H_1: H_{11} \cap H_{12}$). This is referred to as the intersection-union test (Berger 1982).

Let (Z_{1l}, Z_{2l}) be the statistics for testing the hypotheses at the l th analysis, given by

$$Z_{kl} = \frac{\bar{Y}_{Tkl} - \bar{Y}_{Ckl}}{\sigma_k \sqrt{(1 + 1/r_C)/n_l}},$$

where \bar{Y}_{Tkl} and \bar{Y}_{Ckl} are the sample means given by $\bar{Y}_{Tkl} = \sum_{i=1}^{n_l} Y_{Tki}/n_l$ and $\bar{Y}_{Ckl} = \sum_{j=1}^{r_C n_l} Y_{Ckj}/(r_C n_l)$. For large samples, under the alternative hypothesis H_1 , each Z_{kl} is approximately normally distributed as $Z_{kl} \sim N(\sqrt{r_C n_l / (1 + r_C)} \delta_k / \sigma_k, 1^2)$. Thus, (Z_{1l}, Z_{2l}) is approximately bivariate normally distributed with the correlation $\text{corr}[Z_{1l}, Z_{2l}] = (r_C \rho_T + \rho_C) / (1 + r_C) = \rho_Z$ at the l th interim analysis. Furthermore, the joint distribution of $(Z_{11}, Z_{21}, \dots, Z_{1l}, Z_{2l}, \dots, Z_{1L}, Z_{2L})$ is $2L$ multivariate normal with their correlations given by $\text{corr}[Z_{1l'}, Z_{1l}] = \text{corr}[Z_{2l'}, Z_{2l}] = \sqrt{n_{l'}/n_l}$, and $\text{corr}[Z_{1l'}, Z_{2l}] = \text{corr}[Z_{1l}, Z_{2l'}] = \rho_Z \sqrt{n_{l'}/n_l}$, where $1 \leq l' \leq l \leq L$ and $k' \leq k$. If the correlation between the two endpoints is assumed be common between the two intervention groups, i.e., $\rho_T = \rho_C = \rho$, then the correlation among test statistics across the interim analyses is simply given by $\text{corr}[Z_{1l'}, Z_{2l}] = \text{corr}[Z_{1l}, Z_{2l'}] = \rho \sqrt{n_{l'}/n_l}$ as $\rho_Z = \rho$.

2.2.2 Decision-Making Frameworks and Stopping Rules

When evaluating the joint effects on both of the endpoints within the context of group-sequential designs, there are the two decision-making frameworks associated

with hypothesis testing. One is to reject H_0 if statistical significance of T relative to C is achieved for both endpoints at any interim analysis until the final analysis (i.e., not necessarily simultaneously at the same interim analysis) (DF-A) (Asakura et al. 2014), and the other is the special case of DF-A and is to reject H_0 if and only if superiority is achieved for the two endpoints simultaneously (i.e., at the same interim analysis of the trial) (DF-B) (Asakura et al. 2014; Cheng et al. 2014). We will discuss the two decision-making frameworks separately as the corresponding stopping rules and power definitions are unique.

DF-A is flexible. If only the hypothesis for one endpoint is rejected at an interim analysis, then the trial will continue but in subsequent interim analyses the not-yet-rejected hypothesis for other endpoint is repeatedly tested until it is rejected or the trial is completed. The stopping rule based on DF-A is formally given as follows:

At the l th analysis ($l = 1, \dots, L - 1$)

if $Z_{1l} > c_{1l}^E(\alpha)$ and $Z_{2l'} > c_{2l'}^E(\alpha)$ for some $1 \leq l' \leq l$, or if $Z_{1l'} > c_{1l'}^E(\alpha)$ for some $1 \leq l' \leq l$ and $Z_{2l} > c_{2l}^E(\alpha)$, then reject H_0 and stop the trial, otherwise, continue the $(l + 1)$ th analysis,

at the L th analysis

if $Z_{1L} > c_{1L}^E(\alpha)$ and $Z_{2l'} > c_{2l'}^E(\alpha)$ for some $1 \leq l' \leq L$, or if $Z_{1l'} > c_{1l'}^E(\alpha)$ for some $1 \leq l' \leq L$ and $Z_{2L} > c_{2L}^E(\alpha)$, then reject H_0 and stop the trial, otherwise, then do not reject H_0 ,

where $c_{1l}^E(\alpha)$ and $c_{2l}^E(\alpha)$ are the critical boundaries, which are constant and selected separately, using any group-sequential method such as the Lan–DeMets error-spending method (Lan and DeMets 1983) to control the overall Type I error rate, as if they were a single primary endpoint, ignoring the other co-primary endpoint.

For example, consider a group-sequential clinical trial with the five planned analyses ($L = 5$). The hypothesis for the joint effect on both endpoints is tested at 2.5 % significance level. If the critical boundaries for both endpoints are commonly determined by the O’Brien–Fleming-type boundary (OF) (O’Brien and Fleming 1979), using the Lan–DeMets error-spending method with equally spaced increments of information, then critical boundaries for each analysis are 4.8769, 3.3569, 2.6803, 2.2898, and 2.0310. Figure 2.1 illustrates the region for rejecting each H_{0k} ($k = 1, 2$). For example, if we observe the test statistics $Z_{14} = 3.5073$ for EP1 and $Z_{24} = 2.2294$ for EP2 at the fourth analysis, then H_0 is not rejected as Z_{14} is larger than the corresponding critical boundary of $c_{14}^E(2.5) = 2.2898$ but Z_{24} is not. In the subsequent analysis, i.e., the final analysis, the hypothesis testing is repeatedly conducted only for EP2. At the final analysis, if we observe $Z_{25} = 2.9732$ for EP2, then H_0 is rejected as Z_{25} is larger than the corresponding critical boundary of $c_{25}^E(2.5) = 2.0310$.

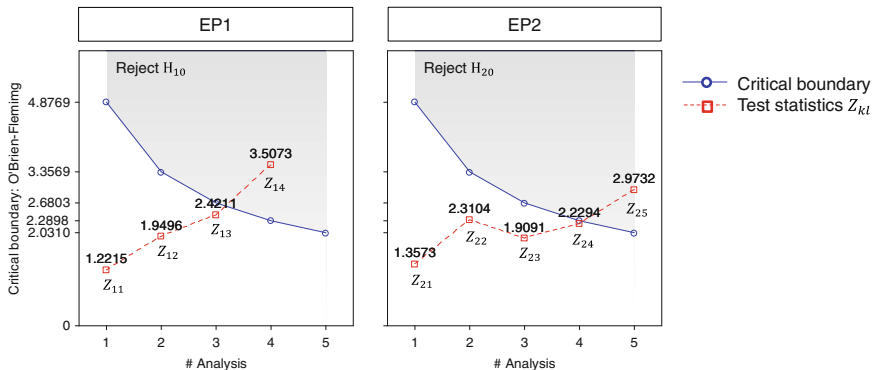


Fig. 2.1 The region for rejecting H_0 based on DF-A in a group-sequential clinical trial with the five planned analyses ($L = 5$), where the decision-making is based on DF-A. The hypothesis for the joint effect on both endpoints is tested at 2.5 % significance level. The critical boundaries for both endpoints are commonly determined by the OF, using the Lan–DeMets error-spending method with equally spaced increments of information

The power for the joint effect on both endpoints, corresponding to DF-A, is

$$1 - \beta = \Pr \left[\left\{ \bigcup_{l=1}^L A_{1l} \right\} \cap \left\{ \bigcup_{l=1}^L A_{2l} \right\} \middle| H_1 \right], \quad (2.1)$$

where $A_{kl} = \{Z_{kl} > c_{kl}^E\}$. The power based on DF-A (2.1) can be numerically assessed by using multivariate normal integrals. A detailed calculation is provided in Appendix A.

DF-B is relatively simple. If only the hypothesis for one endpoint is rejected at an interim analysis, then the trial continues and the hypotheses for both endpoints are repeatedly tested until they are rejected simultaneously, i.e., during the same interim analysis. The stopping rule based on DF-B is formally given as follows:

At the l th analysis ($l = 1, \dots, L - 1$)

if $Z_{1l} > c_{1l}^E(\alpha)$ and $Z_{2l} > c_{2l}^E(\alpha)$, then reject H_0 and stop the trial,
otherwise, continue to the $(l + 1)$ th analysis,

at the L th analysis

if $Z_{1L} > c_{1L}^E(\alpha)$ and $Z_{2L} > c_{2L}^E(\alpha)$, then reject H_0 ,
otherwise, do not reject H_0 .

Figure 2.2 illustrates the region for rejecting each H_{k0} with the number of planned analyses similarly as in Fig. 2.1. For example, if we observe the test

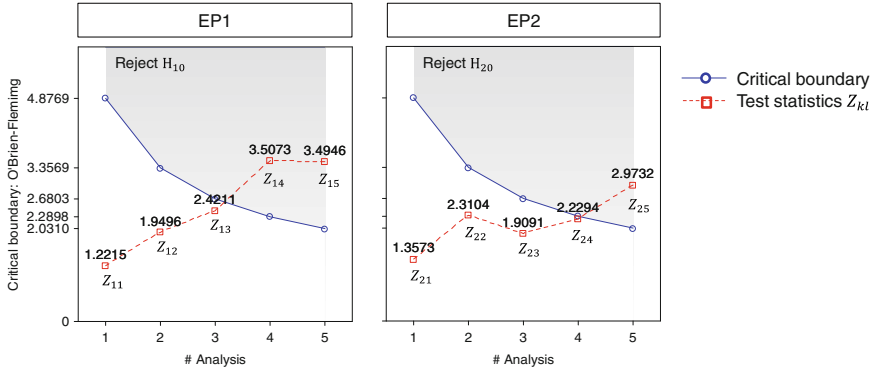


Fig. 2.2 The region for rejecting H_0 in a group-sequential clinical trial with the five planned analyses ($L = 5$), where the decision-making is based on DF-B. The hypothesis for the joint effect on both endpoints is tested at 2.5 % significance level. The critical boundaries for both endpoints are commonly determined by the OF, using the Lan-DeMets error-spending method, with equally spaced increments of information

statistics $Z_{14} = 3.5073$ for EP1 and $Z_{24} = 2.2294$ for EP2 at the fourth analysis, then H_0 is not rejected as Z_{14} is larger than the corresponding critical boundary of $c_{14}^E(2.5) = 2.2898$ but Z_{24} is not. At the final analysis, both H_{01} and H_{02} is tested again. If we observe $Z_{15} = 3.4946$ and $Z_{25} = 2.9732$, then H_0 is rejected as both Z_{15} and Z_{25} are larger than the corresponding critical boundary of $c_{15}^E(2.5) = c_{25}^E(2.5) = 2.0310$ simultaneously.

The power for the joint effect on both endpoints, corresponding to DF-B, is

$$1 - \beta = \Pr \left[\bigcup_{l=1}^L \{A_{1l} \cap A_{2l}\} \middle| H_1 \right]. \quad (2.2)$$

Similarly as in the power based on DF-A, the power based on DF-B can be numerically assessed by using multivariate normal integrals. A detailed calculation is provided in Appendix A.

To illustrate the difference in the power for the joint effect on both endpoints between the DF-A and DF-B, Fig. 2.3 summarizes how the powers based on DF-A and DF-B behave with correlation ($\rho_T = \rho_C = \rho$), critical boundary combinations, and the number of planned analyses under a given sample size, in a group-sequential clinical trial with the two or four planned analyses ($L = 2$ or 4), assuming equal standardized mean differences $\Delta_1 = \Delta_2 = 0.2$. The hypothesis for the joint effect on both endpoints is tested at 2.5 % significance level. The given sample size (equally sized groups: $r_C = 1$) is 393 per intervention group has 80 % power to detect a standardized mean difference for each endpoint at 2.5 % significance level for a one-sided test. The three critical boundary combinations are considered: OF for both endpoints (OF-OF), Pocock-type boundary (PC) (Pocock 1977) for both endpoints (PC-PC), and OF for EP1 and PC for EP2 (OF-PC).

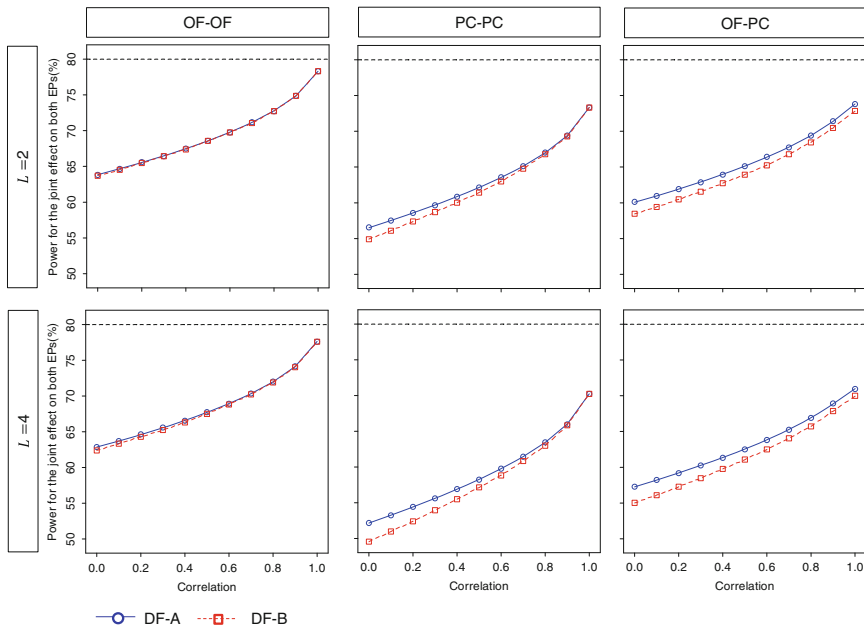


Fig. 2.3 Behavior of power for detecting a joint effect on both endpoints with correlation, critical boundary combinations, and the number of planned analyses under a given sample size, in a group-sequential clinical trial with the two or four planned analyses ($L = 2, 4$), assuming equal standardized mean differences $\Delta_1 = \Delta_2 = 0.2$, where the decision-making is based on DF-A or DF-B. The given sample size (equally sized groups) is 393 per intervention group has 80 % power to detect a standardized mean difference for each endpoint at 2.5 % significance level for a one-sided test. The hypothesis for the joint effect on both endpoints is tested at 2.5 % significance level. The critical boundary combinations are OF for both endpoints (OF-OF), and PC are for both endpoints (PC-PC) and OF for EP1 and PC for EP2 (OF-PC)

A range of correlation between the two endpoints considered in the evaluation is $\rho \geq 0$ since the correlation between the endpoints are usually non-negative as suggested in Offen et al. (2007).

The figure shows that the powers based on both DF-A and DF-B increase as the correlation approaches one in all of the three critical boundary combinations and the numbers of analyses. DF-A provides a slightly higher power than DF-B. In both of $L = 2$ and 4, the largest difference in the power between DF-A and DF-B is observed in PC-PC, and the smallest in OF-OF. However, the difference between DF-A and DF-B is smaller with higher correlation or smaller number of planned analyses in all of the three critical boundary combinations.

The testing procedure for co-primary endpoints is conservative. For example, in fixed-sample designs, if a zero correlation between the two endpoints is assumed and each endpoint is tested at 2.5 % significance level for a one-sided test, then the Type I error rate is 0.0625 % ($= 2.5 \% \times 2.5 \%$) (DF-A). As shown in Asakura

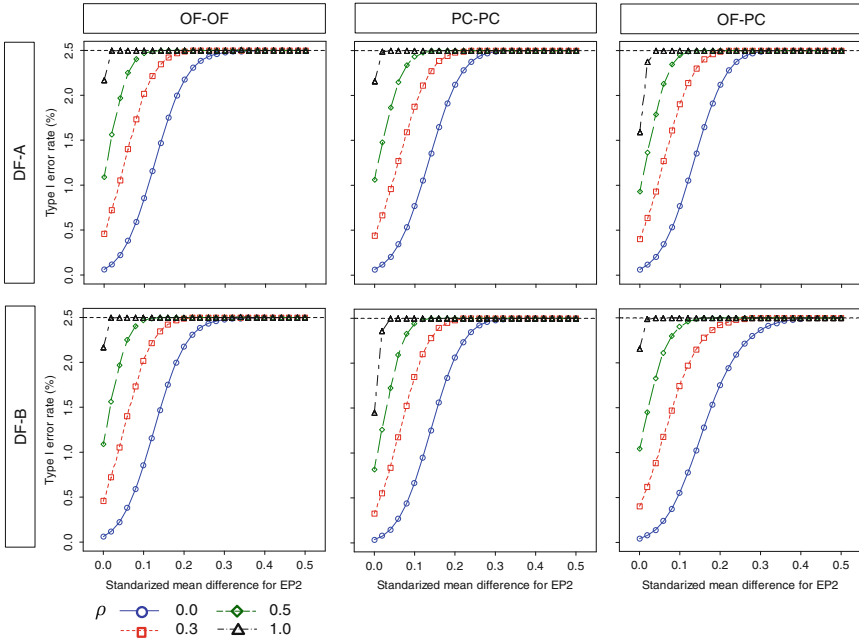


Fig. 2.4 Behavior of Type I error rate with correlation, critical boundary combinations, and standardized mean difference for EP2 in a group-sequential clinical trial with the two planned analyses ($L = 2$), assuming zero standardized mean difference for EP1 $\Delta_1 = 0$, where the decision-making is based on DF-A or DF-B. The hypothesis for the joint effect on both endpoints is tested at 2.5 % significance level. The three critical boundary combinations are OF for both endpoints (OF-OF), PC for both endpoints (PC-PC), OF for EP1 and PC for EP2 (OF-PC)

et al. (2014), the maximum overall Type I error rate associated with the rejection region of the null hypothesis increases as the correlation approaches one, but it is not larger than the prespecified significance level. Figure 2.4 summarizes how the overall Type I error rates based on DF-A and DF-B behave with correlation ($\rho_T = \rho_C = \rho$), critical boundary combinations, and standardized mean difference for EP2, in a group-sequential clinical trial with the two planned analyses ($L = 2$) and zero standardized mean difference for EP1, $\Delta_1 = 0.0$. The hypothesis for the joint effect on both endpoints is tested at 2.5 % significance level. The correlations are $\rho = 0.0, 0.3, 0.5$, and 1.0 . The three critical boundary combinations are considered: OF for both endpoints (OF-OF), PC for both endpoints (PC-PC), and OF for EP1 and PC for EP2 (OF-PC). The figure shows that the Type I error rate for both decision-making frameworks increases as the correlation approaches one, but they are not larger than the prespecified significance level of 2.5 %, in all of the three critical boundary combinations, and DF-B is always slightly conservative than DF-A.

The above differences in power and the Type I error between DF-A and DF-B can be illustrated from the following two situations where the interim analysis result

is inconsistent with the final analysis result even when the alternative hypothesis is true; that is, (i) EP1 is statistically significant at the interim, but not at the final analysis and similarly, and (ii) EP2 is statistically significant at the interim, but not at the final analysis. Thus, DF-B fails to reject the null hypothesis in both situations even if the alternative hypothesis is true, but DF-A is able to reject the null hypothesis in both situations. However, the likelihood of this scenario occurring is low and hence little practical difference in the power and sample size determinations based on DF-A and DF-B. However, DF-A offers the option of stopping measurement of an endpoint for which superiority has been demonstrated. Stopping measurement may be desirable if the endpoint is very invasive or expensive but may also introduce an operational challenge into the trial. For more details, see Asakura et al. (2014) and Hamasaki et al. (2015).

2.2.3 Sample Sizes

We describe two sample size concepts, i.e., the maximum sample size (MSS) and the average sample number (ASN) (i.e., expected sample size) based on the power (2.1) or (2.2). The MSS is the sample size required for the final analysis to achieve the desired power $1 - \beta$. The MSS is given by the smallest integer not less than n_L satisfying the power for a group-sequential strategy at the prespecified δ_k , σ_k , and ρ_T and ρ_C with Fisher's information time for the interim analyses, n_l/n_L ($l = 1, \dots, L$).

To identify the value of n_L , an easy strategy is a grid search to gradually increase (or decrease) n_L until the power under n_L exceeds (or falls below) the desired power. The grid search often requires considerable computing time, especially with a larger number of endpoints, a larger number of planned analyses, or a small mean difference. To reduce the computing time, the Newton–Raphson algorithm in Sugimoto et al. (2012) or the basic linear interpolation algorithm in Hamasaki et al. (2013) may be utilized.

The ASN is the expected sample size under hypothetical reference values and provides information regarding the number of participants anticipated in a group-sequential clinical trial in order to reach a decision point. The ASN per intervention group is given by

$$\text{ASN} = \sum_{l=1}^{L-1} n_l P_l + n_L \left(1 - \sum_{l=1}^{L-1} P_l \right),$$

where $P_l = P_l(\delta_1, \delta_2, \sigma_1, \sigma_2, \rho_T, \rho_C)$ is the stopping probability (or exit probability) as defined by the likelihood of crossing the critical boundaries at the l th interim analysis assuming that the true values of the intervention's effect are (δ_1, δ_2) .

Both MSS and ASN depend on the design parameters including the differences in means, the correlation structure among the endpoints, the selected critical

boundary based on Lan–DeMets error-spending method, the number of planned analyses, and whether there are equally or unequally spaced increments of information. As shown in Hamasaki et al. (2015), our experience suggests that when considering more than two endpoints as co-primary in a group-sequential setting with more than five analyses, calculating the multivariate normal integrals often requires considerable computing time. A Monte Carlo simulation-based method provides an alternative but the number of replications for simulations should be carefully chosen to control simulation error in calculating the empirical power.

Figures 2.5 and 2.6 display how the reduction in MSS and ASN varies with the ratio of the two standardized mean differences (Δ_2/Δ_1), correlation ($\rho_T = \rho_C = \rho$),

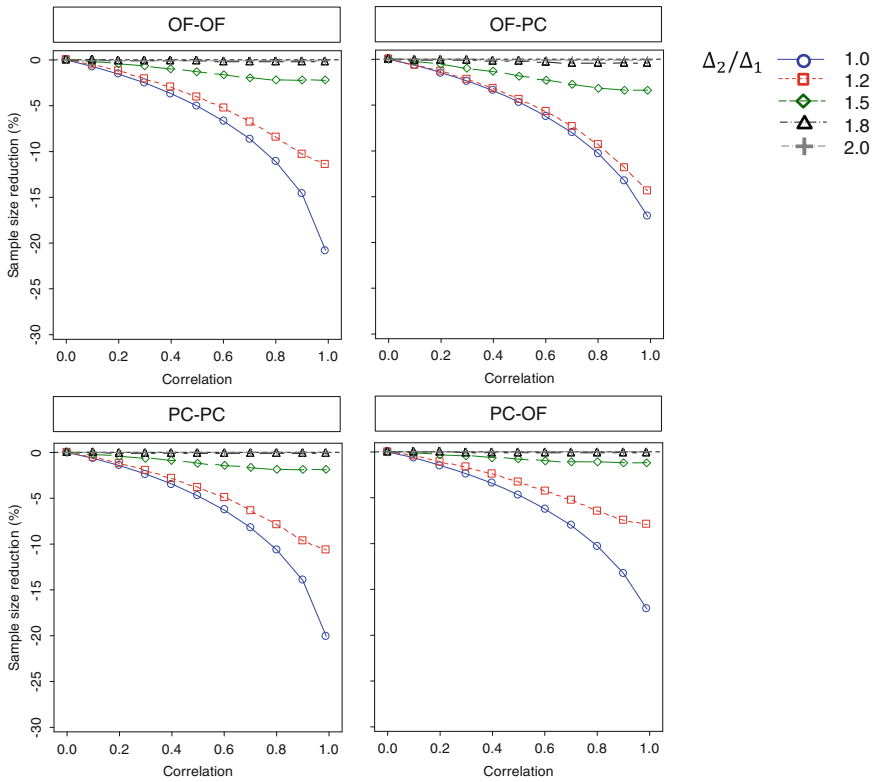


Fig. 2.5 Behavior of reduction in MSS with standardized mean difference, correlation, and critical boundary combination in a group-sequential clinical trial with the two planned analyses ($L = 2$), where the decision-making is based on DF-A. The sample size reduction is calculated as $[MSS(\rho) - MSS(0)]/MSS(0)$, where $MSS(\rho)$ is MSS calculated using ρ and $MSS(0)$ is calculated using zero correlation. The sample size (equally sized groups) per intervention group is calculated to detect the joint effect on both endpoints with 80 % power at 2.5 % significance level for a one-sided test. The four critical boundary combinations are OF for both endpoints (OF-OF), PC for both endpoints (PC-PC), OF for EP1 and PC for EP2 (OF-PC), and PC for EP1 and OF for EP2 (PC-OF)

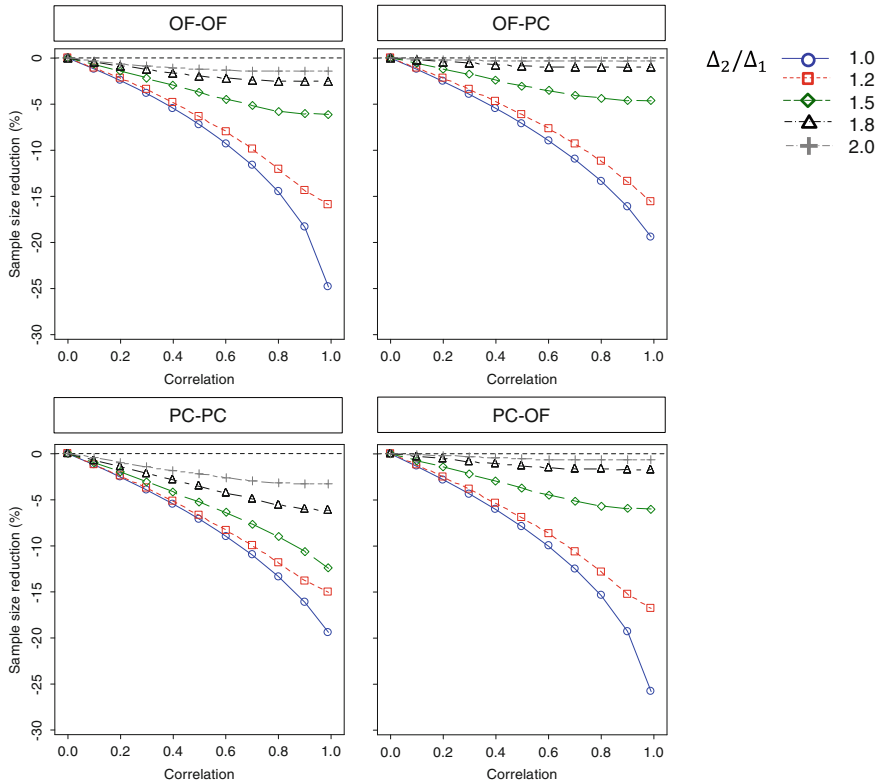


Fig. 2.6 Behavior of reduction in ASN with standardized mean difference, correlation, and critical boundary combination in a group-sequential clinical trial with the two planned analyses ($L = 2$), where the decision-making is based on DF-A. The reduction is calculated as $[MSS(\rho) - MSS(0)]/MSS(0)$, where $MSS(\rho)$ is MSS calculated using ρ and $MSS(0)$ is calculated using zero correlation. The sample size per intervention group (equally sized groups) is calculated to detect the joint effect on both endpoints with 80 % power at 2.5 % significance level for a one-sided test. The four critical boundary combinations are OF for both endpoints (OF-OF), PC for both endpoints (PC-PC), OF for EP1 and PC for EP2 (OF-PC), and PC for EP1 and OF for EP2 (PC-OF)

and critical boundary combinations in a group-sequential clinical trial with the two planned analyses ($L = 2$), where the decision-making is based on DF-A. The reduction is calculated as $[MSS(\rho) - MSS(0)]/MSS(0)$, where $MSS(\rho)$ is MSS calculated using ρ and $MSS(0)$ is calculated using zero correlation. The sample size per intervention group (equally sized groups: $r_C = 1$) is calculated to detect the joint effect on both endpoints with 80 % power at 2.5 % significance level for a one-sided test. The four critical boundary combinations are considered: OF for both endpoints (OF-OF), PC for both endpoints (PC-PC), OF for EP1 and PC for EP2 (OF-PC), and PC for EP1 and OF for EP2 (PC-OF).

Similarly as in fixed-sample designs shown in Sozu et al. (2015), the figures show that the absolute reduction in both MSS and ASN decreases as the correlation approaches one in all of critical boundary combinations when $\Delta_2/\Delta_1 = 1.0$. OF-OF and PC-PC provide a larger reduction than OF-PC and PC-OF. When $1.0 < \Delta_2/\Delta_1 < 1.5$, they still decreases as the correlation approaches one. However, when Δ_2/Δ_1 exceeds 1.5, especially larger than 1.8, the reduction does not change considerably as the correlation varies. Thus, incorporating the correlation into the sample size calculation may lead to a reduction in sample sizes when the standardized mean differences between the two endpoints are approximately equal. However, it is less dramatic as it does not greatly depend on the correlation when the standardized mean differences between the two endpoints are unequal.

2.2.4 Illustration

We provide an example to illustrate these sample size methods. Consider the clinical trial, “Effect of Tarenflurbil on Cognitive Decline and Activities of Daily Living in Patients With Mild Alzheimer Disease,” a multicenter, randomized, double-blind, placebo-controlled trial in patients with mild Alzheimer’s disease (Green et al. 2009). Co-primary endpoints were cognitive as assessed by the Alzheimer’s Disease Assessment Scale Cognitive Subscale (ADAS-Cog: 80-point scale) and functional ability as assessed by the Alzheimer’s Disease Cooperative Study Activities of Daily Living (ADCS-ADL: 78-point scale). A negative change score from baseline on the ADAS-Cog indicates improvement while a positive change score on the ADCS-ADL indicates improvement. The original sample size per intervention group (equally sized groups) of 800 patients provided 96 % power to detect the joint effect on the two primary endpoints, by using a one-sided test at 2.5 % significance level, with the standardized mean differences for both endpoints of $\Delta_1 = \Delta_2 = 0.2$. The correlation between the two endpoints was assumed to be zero in the calculation of the sample size although the two endpoints were expected to be correlated [for example, see Doraiswamy et al. (1997)].

Based on the selected parameters described in Green et al. (2009), i.e., $L = 1$ and $\rho_T = \rho_C = \rho = 0.0$, the sample size per intervention group is calculated as 804. As shown in Table 2.2, if four interims and one final analysis are planned (i.e., $L = 5$) based on DF-B, and conservatively assuming a zero correlation between the endpoints, then the MSS is 822 for OF-OF, 945 for PC-PC and 895 for OF-PC, and the ASN is 602 for OF-OF, 548 for PC-PC, and 608 for OF-PC. If the correlation is incorporated into the calculation when $\rho = 0.3, 0.5$, and 0.8 , the MSS are 817, 809, and 782 for OF-OF; 939, 929, and 898 for PC-PC; and 890, 883, and 859 for OF-PC.

Table 2.1 MSS and ASN per intervention group (equally sized groups) for detecting the joint difference for ADAS-Cog ($\Delta_1 = 0.2$) and ADCS-ADL ($\Delta_2 = 0.2$), with the power of $1 - \beta = 96\%$ for detect the joint effect on both endpoints at 2.5 % significance level for one-sided test, based on DF-A

Correlation ρ	# of analyses L	OF-OF		PC-PC		OF-PC	
		MSS	ASN (H_1)	MSS	ASN (H_1)	MSS	ASN (H_1)
0.0	1	804		804		804	
	2	807	725	881	605	847	690
	3	813	645	911	570	867	647
	4	817	618	927	551	878	615
	5	821	601	937	540	886	600
0.3	1	799		799		799	
	2	801	702	875	591	841	672
	3	807	632	905	550	861	633
	4	812	602	921	530	873	602
	5	815	586	931	519	880	586
0.5	1	791		791		791	
	2	793	683	867	578	833	658
	3	799	619	896	534	854	622
	4	804	589	912	513	865	590
	5	807	572	922	502	873	574
0.8	1	764		764		764	
	2	767	643	839	548	809	631
	3	773	589	869	500	830	599
	4	777	557	884	478	841	566
	5	781	542	894	466	849	550

The three critical boundary combinations are OF for both endpoints (OF-OF), PC for both endpoints (PC-PC), and OF for EP1 and PC for EP2 (OF-PC). The ASN is calculated under H_1 ($\Delta_1 = \Delta_2 = 0.2$)

The ASN are 587, 574, and 542 for OF-OF; 525, 506, and 468 for PC-PC; and 593, 581, and 556 for OF-PC. When comparing DF-A (Table 2.1) to DF-B (Table 2.2), there are no major differences in MSS and ASN for all of the critical boundary combinations, although DF-A provides a slightly smaller MSS and ASN than DF-B, for PC-PC and OF-PC. The advantage and disadvantage of the decision-making frameworks are given in Sect. 2.5.

Figure 2.7 illustrates the probability of rejecting/not rejecting the null hypothesis under H_1 in a group-sequential clinical trial with the five planned analyses ($L = 5$), assuming the correlation $\rho = 0.0$ or 0.8, where the decision-making is based on DF-A. The figure shows that the method offers the possibility to stop a trial early if

Table 2.2 MSS and ASN per intervention group (equally sized groups) for detecting the joint difference for ADAS-Cog ($\Delta_1 = 0.2$) and ADCS-ADL ($\Delta_2 = 0.2$), with the power of $1 - \beta = 96\%$ to detect the joint effect on both endpoints at 2.5 % significance level for one-sided test, based on DF-B

Correlation ρ	# of analyses L	OF-OF		PC-PC		OF-PC	
		MSS	ASN (H_1)	MSS	ASN (H_1)	MSS	ASN (H_1)
0.0	1	804		804		804	
	2	807	725	885	607	854	693
	3	814	646	917	574	875	653
	4	819	619	934	557	887	622
	5	822	602	945	548	895	608
0.3	1	799		799		799	
	2	802	702	880	593	849	676
	3	808	632	911	553	870	639
	4	813	603	928	535	882	608
	5	817	587	939	525	890	593
0.5	1	791		791		791	
	2	794	684	871	580	841	661
	3	800	620	902	537	863	628
	4	805	589	919	517	875	597
	5	809	574	929	506	883	581
0.8	1	764		764		764	
	2	767	643	841	549	818	635
	3	773	589	871	501	839	604
	4	778	558	887	480	851	571
	5	782	542	898	468	859	556

The three critical boundary combinations are considered: OF for both endpoints (OF-OF), PC for both endpoints (PC-PC), and OF for ADAS-Cog and PC for ADCS-ADL (OF-PC). The ASN is calculated under H_1 ($\Delta_1 = \Delta_2 = 0.2$)

evidence is overwhelming and thus offers potentially fewer patients than the fixed-sample designs. In the OF-OF and PC-OF testing procedure combinations, it is more difficult to reject the null hypothesis at the earliest analyses, but easier later on. On the other hand, in the PC-PC and OF-PC testing procedure combination, it is easier to reject the null hypothesis at the earliest analysis.

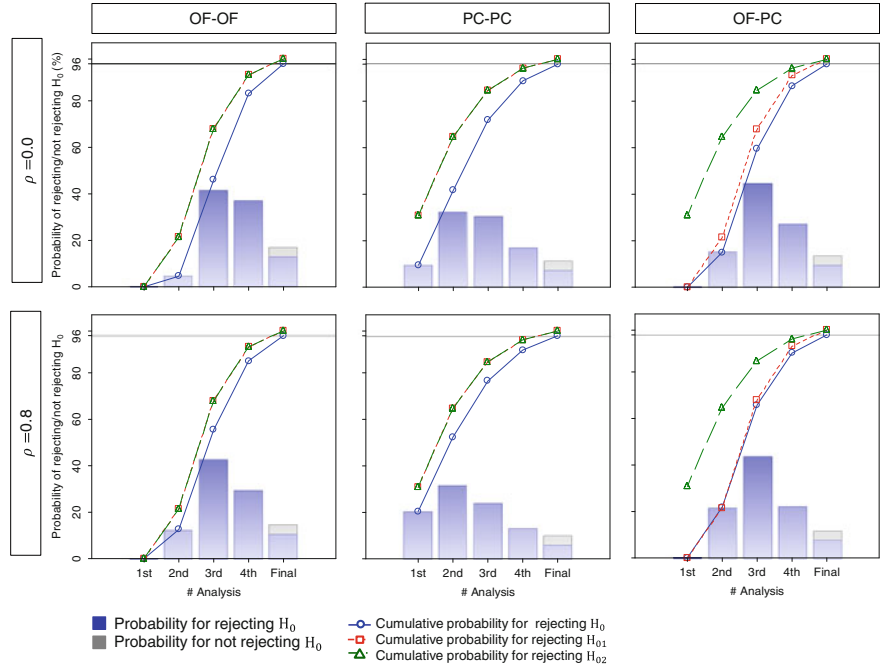


Fig. 2.7 The probability of rejecting/not rejecting the null hypothesis under H_1 in a group-sequential clinical trial with the five planned analyses ($L = 5$), where the decision-making is based on DF-A. The MSS are calculated to detect the joint effect for both endpoints with 96 % power at 2.5 % significance level for one-sided test, based on the assumption $\Delta_1 = \Delta_2 = 0.2$ from the tarenflurbil study. The critical boundaries are determined using the Lan-DeMets error-spending method with equally spaced increments of information. The three critical boundary combinations are OF for both endpoints (OF-OF), PC for both endpoints (PC-PC), and OF for ADAS-Cog and PC for ADCS-ADL (OF-PC)

2.3 Binary Outcomes

Clinical trials are often conducted with the objective of comparing a test intervention with that of a standard intervention based on several binary outcomes. For example, irritable bowel syndrome (IBS) is one of the most common gastrointestinal disorders and is characterized by symptoms of abdominal pain, discomfort, and altered bowel function (Grundmann and Yoon 2010; American College of Gastroenterology 2013). The comparison of the interventions to treat IBS is based on the proportions of participants with adequate relief of abdominal pain and discomfort, and improvements in urgency, stool frequency, and stool consistency. As described in Chap. 1, Food and Drug Administration (FDA) recommends the use of two endpoints for assessing IBS signs and symptoms: (1) pain intensity and (2) stool frequency (FDA 2013). The Committee for Medicinal Products for Human Use (CHMP) (2008) recommends the use of two endpoints for assessing IBS signs

and symptoms: (1) global assessment of symptoms and (2) assessment of symptoms of abdominal discomfort/pain.

In this section, we discuss group-sequential designs in clinical trials with two binary outcomes as co-primary. Similar to the previous section, we consider a two-arm parallel-group trial designed to evaluate whether the T is superior to the C based on two binary endpoints.

2.3.1 Notation and Statistical Setting

Consider a randomized, group-sequential clinical trial of comparing the T with the C. Two binary outcomes are to be evaluated as co-primary endpoints. As a measure the effect, we consider the difference in the proportions between two interventions as it is the most commonly used measure in many clinical trials. The risk ratio and odds ratio are also frequently used in clinical trials to measure a risk reduction. The methods discussed here can be straightforwardly extended to these measures. For details, see Ando et al. (2015).

Assume that Y_{Tki} and Y_{Ckj} are independently binomial distributed with probabilities of success p_{Tk} and p_{Ck} , i.e., $Y_{Tki} \sim B(1, p_{Tk})$ and $Y_{Ckj} \sim B(1, p_{Ck})$, but the observations within pairs for the two interventions are correlated with a common correlation $\text{corr}[Y_{T1i}, Y_{T2i}] = \rho_T$ and $\text{corr}[Y_{C1j}, Y_{C2j}] = \rho_C$. The range of the correlations ρ_T and ρ_C are restricted, depending on the marginal probabilities (Prentice 1988; Le Cessie and van Houwelingen 1994). Let (δ_1, δ_2) denote the differences in proportions for the T and the C, respectively, where $\delta_k = p_{Tk} - p_{Ck}$ ($k = 1, 2$). Suppose that positive values of (δ_1, δ_2) represent the test intervention's benefit. We now have the two observed differences in proportions at the l th analysis, i.e., $(\hat{\delta}_1, \hat{\delta}_2)$, where $\hat{\delta}_{kl} = \hat{p}_{Tkl} - \hat{p}_{Ckl}$ with $\hat{p}_{Tkl} = Y_{Tkl}/n_l$ and $\hat{p}_{Ckl} = Y_{Ckl}/r_{Cn_l}$, and $Y_{Tkl} = \sum_{i=1}^{n_l} Y_{Tli}$ and $Y_{Ckl} = \sum_{j=1}^{r_{Cn_l}} Y_{Ckj}$ denote the number of success under the T and the C. It follows that $Y_{Tkl} \sim B(n_l, p_{Tk})$ and $Y_{Ckl} \sim B(r_{Cn_l}, p_{Ck})$.

We are interested in conducting a hypothesis test to evaluate whether the T is superior to the C, i.e., the null hypothesis $H_0: \delta_1 \leq 0$ or $\delta_2 \leq 0$ versus the alternative hypothesis $H_1: \delta_1 > 0$ and $\delta_2 > 0$. Let (Z_{1l}, Z_{2l}) be the Z-score statistics for testing the hypotheses at the l th analysis, given by

$$Z_{kl} = \frac{\hat{p}_{Tkl} - \hat{p}_{Ckl}}{\sqrt{\hat{p}_{kl}\hat{q}_{kl}(r_C + 1/r_C)/n_l}},$$

where $\hat{p}_{kl} = (\hat{p}_{Tkl} + r_C\hat{p}_{Ckl})/(1 + r_C)$ and $\hat{q}_{kl} = 1 - \hat{p}_{kl}$. For large samples, each Z_{kl} is approximately normally distributed [e.g., see Fleiss et al. (2003)]. Thus, the two test statistics at l th analysis (Z_{1l}, Z_{2l}) is approximately bivariate normally distributed with the correlation

$$\text{corr}[Z_{1l}, Z_{2l}] = \rho_Z = \frac{r_C \rho_T \sqrt{p_{T1} q_{T1} p_{T2} q_{T2}} + \rho_C \sqrt{p_{C1} q_{C1} p_{C2} q_{C2}}}{\sqrt{r_C p_{T1} q_{T1} + p_{C1} q_{C1}} \sqrt{r_C p_{T2} q_{T2} + p_{C2} q_{C2}}},$$

$q_{Tk} = 1 - p_{Tk}$ and $q_{Ck} = 1 - p_{Ck}$. Furthermore, the joint distribution of $(Z_{11}, Z_{21}, \dots, Z_{1l}, Z_{2l}, \dots, Z_{1L}, Z_{2L})$ is approximately $2L$ multivariate normal with their correlations given by $\text{corr}[Z_{1l'}, Z_{1l}] = \text{corr}[Z_{2l'}, Z_{2l}] = \sqrt{n_{l'}/n_l}$ and $\text{corr}[Z_{1l'}, Z_{2l}] = \text{corr}[Z_{1l}, Z_{2l'}] = \rho_Z \sqrt{n_{l'}/n_l}$, where $1 \leq l' \leq l \leq L$. Similarly as discussed in Sect. 2.2, we can calculate the power, Type I error rate, and sample sizes based on the two decision-making frameworks associated with hypothesis testing, i.e., DF-A and DF-B.

The method is based on the normal approximation which works well in most situations (Asakura et al. 2015). However, it may not work well in the occurrence of extremely small event rates or small sample sizes as the joint distribution is not fully specified in the first- and second-order moments. In such situations, Monte Carlo simulation-based method or more direct methods may be more appropriate although this occurs at the expense of considerable computational resources. For more direct methods for multiple co-primary endpoints in fixed-sample designs including Fisher's exact test, see Sozu et al. (2010, 2015).

2.3.2 Illustration

We provide an example to illustrate these sample size methods. Consider the double-blind, randomized, parallel-group, placebo-controlled trial evaluating lactobacilli and bifidobacteria in the prevention of antibiotic-associated diarrhea in older people admitted to hospital (the PLACIDE study) (Allen et al. 2012, 2013). The study was designed to demonstrate that the administration of a probiotic comprising two strains of lactobacilli and two strains of bifidobacteria alongside antibiotic treatment prevents antibiotic-associated diarrhea. The co-primary outcomes were: (EP1) the occurrence of antibiotic-associated diarrhea (AAD) within 8 weeks and (EP2) the occurrence of *C difficile* diarrhea (CDD) within 12 weeks of recruitment.

The original sample size per intervention group (equally sized groups) of 1239 participants provided 80 % power to detect a 50 % reduction in CDD in the probiotic group compared with the placebo group, by using a two-sided Fisher's exact test at 5 % significance level, assuming CDD frequencies of 4 % in placebo group and 2 % in probiotic group. Although Cochran's condition seems to be hold for this setting, the normal approximation method was not used for the sample size calculation and the sample size was conservatively calculated. This sample size would provide a power of more than 99 % to detect a 50 % reduction in AAD, by using a

two-sided Fisher’s exact test at 5 % significance level, assuming AAD frequencies of 20 % in placebo group and 10 % in probiotic group as the normal approximation. The correlation between the two outcomes was not incorporated into the original sample size calculation.

Tables 2.3 and 2.4 display the MSS and ASN per intervention group (equally sized groups: $r_C = 1$) based on DF-A and DF-B. The sample size was derived using an alternative hypothesis of differences in proportions for AAD ($p_{T1} = 0.2$ and $p_{C1} = 0.4$) and CDD ($p_{T2} = 0.02$ and $p_{C2} = 0.04$) with 80 % power at 2.5 % significance level for one-sided test, using the normal approximation method where $\rho = \rho_T = \rho_C = 0.0, 0.3, 0.5$, and 0.8 ; $L = 2, 3, 4$, and 5 . The critical boundaries are determined by using the Lan–DeMets error-spending method, with equally spaced increments of information. The critical boundary combinations are OF for both endpoints (OF-OF), PC for both endpoints (PC-PC), OF for AAD and PC for CDD (OF-PC), and PC for AAD and OF for CDD (PC-OF).

Based on the selected parameters described in Allen et al. (2012), i.e., $L = 1$ and $\rho = 0.0$, the sample size per intervention group (equally sized groups) is calculated as 1141. If four interims and one final analysis are planned (i.e., $L = 5$) based on DF-A, and conservatively assuming a zero correlation between the endpoints, then

Table 2.3 MSS and ASN per intervention group (equally sized groups) for detecting the joint difference for AAD ($p_{T1} = 0.2$ and $p_{C1} = 0.4$) and CDD ($p_{T2} = 0.02$ and $p_{C2} = 0.04$), with 80 % power at 2.5 % significance level for a one-sided test, where the decision-making is based on DF-A

Correlation ρ	# of analyses L	OF-OF		PC-PC		OF-PC		PC-OF	
		MSS	ASN (H_1)	MSS	ASN (H_1)	MSS	ASN (H_1)	MSS	ASN (H_1)
0.0	2	1146	1056	1282	977	1282	982	1146	1053
	3	1156	991	1337	941	1337	981	1156	989
	4	1164	960	1366	925	1366	972	1164	958
	5	1170	943	1385	918	1385	956	1170	941
0.3	2	1146	1056	1282	977	1282	982	1146	1053
	3	1156	991	1337	941	1337	981	1156	989
	4	1164	960	1366	925	1366	972	1164	958
	5	1170	943	1385	918	1385	956	1170	941
0.5	2	1146	1056	1282	977	1282	982	1146	1053
	3	1156	991	1337	941	1337	981	1156	989
	4	1164	960	1366	925	1366	972	1164	958
	5	1170	943	1385	918	1385	956	1170	941
0.8	2	1146	1056	1282	977	1282	982	1146	1053
	3	1156	991	1337	941	1337	981	1156	989
	4	1164	960	1366	925	1366	972	1164	958
	5	1170	943	1385	918	1385	956	1170	941

The critical boundaries are determined by using the Lan–DeMets error-spending method, with equally spaced increments of information. The critical boundary combinations are OF for both endpoints (OF-OF), PC for both endpoints (PC-PC), OF for AAD and PC for CDD (OF-PC), and PC for AAD and OF for CDD (PC-OF). The ASN is calculated under H_1 ($p_{T1} = 0.2$ and $p_{C1} = 0.4$, and $p_{T2} = 0.02$ and $p_{C2} = 0.04$)

Table 2.4 MSS and ASN per intervention group (equally sized groups) for detecting the joint difference for AAD ($p_{T1} = 0.2$ and $p_{C1} = 0.4$) and CDD ($p_{T2} = 0.02$ and $p_{C2} = 0.04$), with 80 % at power at 2.5 % significance level for a one-sided test, where the decision-making is based on DF-B

Correlation ρ	# of analyses L	OF-OFF		PC-PC		OF-PC		PC-OFF	
		MSS	ASN (H_1)	MSS	ASN (H_1)	MSS	ASN (H_1)	MSS	ASN (H_1)
0.0	2	1146	1056	1282	977	1283	983	1146	1053
	3	1156	991	1337	941	1346	989	1156	989
	4	1164	960	1368	928	1380	989	1164	958
	5	1170	944	1387	921	1399	976	1170	941
0.3	2	1146	1055	1282	977	1283	982	1146	1053
	3	1156	991	1337	940	1346	985	1156	989
	4	1164	959	1367	926	1380	987	1164	958
	5	1170	943	1387	920	1398	973	1170	941
0.5	2	1146	1055	1282	977	1283	982	1146	1053
	3	1156	991	1337	940	1346	985	1156	989
	4	1164	959	1367	926	1380	987	1164	958
	5	1170	943	1387	920	1398	973	1170	941
0.8	2	1146	1055	1282	977	1283	982	1146	1053
	3	1156	991	1337	940	1346	985	1156	989
	4	1164	959	1367	926	1380	987	1164	958
	5	1170	943	1387	920	1398	973	1170	941

The critical boundaries are determined by using the Lan-DeMets error-spending method, with equally spaced increments of information. The critical boundary combinations are OF for both endpoints (OF-OFF), PC for both endpoints (PC-PC), OF for AAD and PC for CDD (OF-PC), and PC for AAD and OF for CDD (PC-OFF). The ASN is calculated under H_1 ($p_{T1} = 0.2$ and $p_{C1} = 0.4$, and $p_{T2} = 0.02$ and $p_{C2} = 0.04$)

the MSS is 1170 for OF-OFF, 1387 for PC-PC, 1399 for OF-PC, and 1170 for PC-OFF, and the ASN is 944 for OF-OFF, 921 for PC-PC, 976 for OF-PC, and 941 for PC-OFF. On the other hand, even if the correlation is incorporated into the calculation, the MSS and ASN do not change as the correlation varies. This means that when one standardized difference in proportions is relatively larger than the other, i.e., $\delta_1 > \delta_2$ (or $\delta_1 < \delta_2$) with $p_{C1} \neq p_{C2}$, then there is little benefit in incorporating the correlation into sample size calculation.

When comparing DF-A (Table 2.3) to DF-B (Table 2.4), there are no major differences in MSS and ASN for all of the testing procedure combinations, although DF-A provides a slightly smaller MSS and ASN than DF-B: for DF-B, the MSS is 1170 for OF-OFF, 1387 for PC-PC, 1399 for OF-PC, and 1170 for PC-OFF, and the ASN is 944 for OF-OFF, 921 for PC-PC, 976 for OF-PC, and 941 for PC-OFF when assuming a zero correlation.

2.4 Practical Issues

Two important decisions must be made when constructing efficient group-sequential strategies in clinical trials with multiple co-primary endpoints. The first decision is the choice of the critical boundary based on an error-spending method for each endpoint. If the trial was designed to detect effects on at least one endpoint with a prespecified ordering of endpoints, then the selection of different boundaries for each endpoint (i.e., the OF for the primary endpoint and the PC for the secondary endpoint) can provide a higher power than using the same critical boundary for both endpoints (Glimm et al. 2010; Tamhane et al. 2010). However, as shown in Asakura et al. (2014), the selection of a different critical boundary has a minimal effect on the overall power and ASN. In both decision-making frameworks, regardless of equal or unequal standardized mean difference among the endpoints, the largest power is obtained from the OF for all of the endpoints, and the lowest is the PC for all of the endpoints. Regarding the ASN, the smallest is provided by the PC for all of the endpoints while the largest is provided by the OF. One possible scenario for selecting a different boundary is when one endpoint is invasive and stopping measurement of the endpoint is desirable as soon as possible, i.e., once the superiority for the endpoint has been demonstrated.

Table 2.5 illustrates the average observation number (AON) per intervention group (equally sized groups) for each endpoint based on the decision-making frameworks DF-A under a given MSS in clinical trials with two co-primary endpoints, EP1 and EP2, when their standardized mean differences are $(\Delta_1, \Delta_2) = (0.2, 0.2)$ and $(0.2, 0.3)$. The AON is the expected sample size for each endpoint

Table 2.5 The AON per intervention group for each endpoint based on the decision-making framework DF-A under a given MSS in clinical trials with two co-primary endpoints, EP1 and EP2, when their standardized mean differences are $(\Delta_1, \Delta_2) = (0.2, 0.2)$ and $(0.2, 0.3)$

Standardized mean difference	Sample sizes		Critical boundary combination			
			OF-OF	PC-PC	OF-PC	PC-OF
(0.2, 0.2)	AON (H_{1k})	EP1	403	454	474	490
		EP2	403	454	390	474
		MSS	574	518	547	547
		ASN (H_1)	472	502	505	505
(0.2, 0.3)	AON (H_{1k})	EP1	259	298	316	243
		EP2	341	368	339	373
		MSS	450	403	446	408
		ASN (H_1)	357	385	384	380

The MSS per intervention group (equally sized groups) is calculated to detect the joint effect for two endpoints with 80 % power at 2.5 % significance level for a one-sided test, where one interim and one final analysis are to be performed. The critical boundaries are determined by using the Lan–DeMets error-spending method, with equally spaced increments of information. The critical boundary combinations are OF for both endpoints (OF-OF), PC for both endpoints (PC-PC), OF for AAD and PC for CDD (OF-PC), and PC for AAD and OF for CDD (PC-OF). The ASN is calculated under H_1 ($\Delta_1 = \Delta_2 = 0.2$). AON is calculated under H_{1k} with the calculated MSS

and it is calculated under the hypothetical reference values and provides information on the number of observations anticipated in a group-sequential clinical trial in order to reach a decision point for each endpoint. The AON is calculated under $H_{1,k}$ with the calculated MSS. The MSS per intervention group (equally sized groups) is calculated to detect the joint effect for two endpoints with 80 % power at 2.5 % significance level for a one-sided test, where one interim and one final analysis are to be performed. The critical boundaries are determined by the combinations of the OF and the PC, using the Lan–DeMets error-spending method with equally spaced increments of information; if EP1 is an invasive endpoint, then the critical boundary combination of the PC for EP1 and the OF for EP2 provides the smallest AON for EP1 in all of the standardized mean difference combinations.

Another practical decision is the selection of the correlations in the power evaluation and sample size calculation, i.e., whether the observed correlation from external or pilot data should be utilized. As shown in Sect. 2.2.3, when the standardized mean differences for the endpoints are unequal, the advantage of incorporating the correlation into sample size calculation is less dramatic as the required sample size is primarily determined by the smaller standardized mean difference and does not greatly depend on the correlation. In this situation, the sample size equation for multiple co-primary continuous endpoints can be simplified using the equation for a single endpoint. When the standardized mean differences among endpoints are approximately equal, one conservative approach is to assume that the correlations are zero even if nonzero correlations are expected. Group-sequential designs discussed in this chapter offer the possibility of reducing the sample size compared to fixed-sample designs even if zero correlation is assumed at the design stage.

Table 2.6 summarizes MSS and ASN per intervention group in clinical trials with two co-primary endpoints. The MSS per intervention group (equally sized

Table 2.6 MSS and ASN per intervention group in clinical trials with two co-primary endpoints

Decision-making framework	# of analyses L	MSS	ASN (H_1)			
			0.0	0.3	0.5	0.8
DF-A	2	518	502	494	488	475
	3	522	470	461	455	442
	4	525	457	447	440	426
	5	528	449	440	432	418
DF-B	2	518	502	494	488	475
	3	523	471	462	455	443
	4	528	459	449	442	428
	5	530	451	441	434	419

The MSS per intervention group (equally sized groups) is calculated to detect the joint effect for two endpoints ($\Delta_1 = \Delta_2 = 0.2$) with 80 % power at 2.5 % significance level for a one-sided test, where the correlation between the two endpoints is assumed to be zero, i.e., $\rho_T = \rho_C = \rho = 0.0$ and the critical boundaries are determined by OF, using the Lan–DeMets error-spending method with equally spaced increments of information. The ASN is calculated under H_1 ($\Delta_1 = \Delta_2 = 0.2$) with $\rho = 0.0, 0.3, 0.5$, and 0.8

groups) is calculated to detect the joint effect for two endpoints with 80 % power at 2.5 % significance level for a one-sided test, where the correlation between the two endpoints is assumed to be zero, i.e., $\rho_T = \rho_C = \rho = 0.0$ and the critical boundaries are determined by the OF, using the Lan–DeMets error-spending method with equally spaced increments of information. The ASN is calculated under H_1 ($\Delta_1 = \Delta_2 = 0.2$) and $\rho = 0.0, 0.3, 0.5$, and 0.8 . For example, when considering a clinical trial with two co-primary endpoints, 516, 503, 490, 458 participants per intervention group are required to detect a joint effect of equal standardized mean difference $\Delta_1 = \Delta_2 = 0.2$ with 80 % power at 2.5 % significance level for a one-sided test in a fixed-sample design, if the correlation between two endpoints is $\rho = 0.0, 0.3, 0.5$, and 0.8 . In a group-sequential design based on DF-B, assuming zero correlation between the two endpoints, the MSS are 518, 523, 528, and 530 corresponding to the number of planned analyses $L = 2, 3, 4$, and 5 . The critical boundaries for both endpoints are determined by OF, using the Lan–DeMets error-spending method with equally spaced increments of information. Under these MSS, the ASN are 488, 455, 442, and 434. The ASN are approximately equal or smaller than the fixed-sample designs, depending on the number of planned analyses. Our experience suggests that when standardized mean differences are unequal among the endpoints, the power is not increased with higher correlations. With unequal standardized mean differences, incorporating the correlation into the sample size calculation at the planning stage may have less of an advantage because the sample size is determined by the smaller standardized mean difference.

2.5 Summary

The determination of sample size and the evaluation of power are fundamental and critical elements in the design of a clinical trial. If a sample size is too small, then important effects may not be detected, while a sample size that is too large is wasteful of resources and unethically puts more participants at risk than necessary. Recently, many clinical trials are designed with more than one endpoint considered as co-primary. As with trials involving a single primary endpoint, designing such trials to include interim analyses (i.e., with repeated testing) may provide efficiencies by detecting trends prior to planned completion of the trial. It may also be prudent to evaluate design assumptions at the interim and potentially make design adjustments (i.e., sample size recalculation) if design assumptions were dramatically inaccurate. However, such design complexities create challenges in the evaluation of power and the calculation of sample size during trial design.

In this chapter, we discuss group-sequential designs with two co-primary endpoints, where both endpoints are continuous or both are binary. We derive the power and sample size methods under two decision-making frameworks: (1) designing the trial to detect superiority for the two endpoints at any interim time-point (i.e., not necessarily simultaneously) (DF-A) and (2) designing the trial to detect the test intervention's superiority for the two endpoints simultaneously (i.e., at the same

Table 2.7 Advantages and disadvantages of the two decision-making frameworks in clinical trials with multiple co-primary endpoints

Decision-making framework	Advantages	Disadvantages
DF-A	<ul style="list-style-type: none">• Controls the Type I error rate adequately• Flexible to allow the option of selecting different timings for interim analyses among the endpoints; this is useful when designing clinical trials with the endpoints requiring different information times such as progression-free survival and overall survival• Possible to stop measuring an endpoint for which superiority has been demonstrated—this is desirable if the endpoint is very invasive or expensive (e.g., data from a liver biopsy or gastro-fiberscope, or data from expensive imaging)	<ul style="list-style-type: none">• Conservative as the rejection region of the null hypothesis becomes more restricted as the number of endpoints increases• Difficult to maintain the integrity and validity of clinical trial if stopping measurement of an endpoint for which superiority has been demonstrated
DF-B	<ul style="list-style-type: none">• Controls the Type I error rate adequately• Makes the decision-making simple and easy to use in practice	<ul style="list-style-type: none">• Conservative as the rejection region of the null hypothesis becomes more restricted as the number of endpoints increases• Cannot stop measuring an endpoint for which superiority has been demonstrated• Provides the lowest power and largest sample sizes among the decision-making frameworks

interim time-point of the trial) (DF-B). The latter is simpler while the former is more flexible and may be useful when the endpoint is very invasive or expensive, as it allows for stopping the measurement of any endpoint upon which superiority has been demonstrated. We summarize advantages and disadvantages of the two decision-making frameworks in clinical trials with multiple co-primary endpoints in Table 2.7. For other decision-making frameworks, see Hamasaki et al. (2015).

References

Allen SJ, Wareham K, Bradley C, Harris W, Dhar A, Brown H, Foden A, Cheung WY, Gravenor MB, Plummer S, Phillips CJ, Mack D (2012) A multicentre randomised controlled trial evaluating lactobacilli and bifidobacteria in the prevention of antibiotic-associated diarrhoea in older people admitted to hospital: the PLACIDE study protocol. *BMC Infect Dis* 12:108

- Allen SJ, Wareham K, Wang D, Bradley C, Hutchings H, Harris W, Dhar A, Brown H, Foden A, Gravenor MB, Mack D (2013) Lactobacilli and bifidobacteria in the prevention of antibiotic-associated diarrhoea and *Clostridium difficile* diarrhoea in older inpatients (PLACIDE): a randomised, double-blind, placebo-controlled, multicentre trial. *Lancet* 382:1249–1257
- American College of Gastroenterology Website (2013) Understanding irritable bowel syndrome. www.patients.gi.org/gi-health-and-disease/understanding-irritable-bowel-syndromeleavingsiteicon. Accessed 25 Nov 2015
- Ando Y, Hamasaki T, Evans SR, Asakura K, Sugimoto T, Sozu T, Ohno Y (2015) Sample size considerations in clinical trials when comparing two interventions using multiple co-primary binary relative risk contrasts. *Stat Biopharm Res* 7:81–89
- Asakura K, Hamasaki T, Sugimoto T, Hayashi K, Evans SR, Sozu T (2014) Sample size determination in group-sequential clinical trials with two co-primary endpoints. *Stat Med* 33:2897–2913
- Asakura K, Hamasaki T, Evans SR, Sugimoto T, Sozu T (2015) Group-sequential designs when considering two binary outcomes as co-primary endpoints. In: Chen Z, Liu A, Qu Y, Tang L, Ting N, Tsong Y (eds) *Applied statistics in biomedicine and clinical trials design* (Chap. 14). Springer International Publishing, Cham, pp 235–262
- Berger RL (1982) Multiparameter hypothesis testing and acceptance sampling. *Technometrics* 24:295–300
- Cheng Y, Ray S, Chang M, Menon S (2014) Statistical monitoring of clinical trials with multiple co-primary endpoints using multivariate B-value. *Stat Biopharm Res* 6:241–250
- Committee for Medicinal Products for Human Use (2008) Guideline on medicinal products for the treatment Alzheimer's disease and Other dementias (CPMP/EWP/553/95 Rev.1). European Medicines Agency, London, UK. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003562.pdf. Accessed 25 Nov 2015
- Cook RJ, Farewell VT (1994) Guideline for monitoring efficacy and toxicity responses in clinical trials. *Biometrics* 50:1146–1162
- Doraiswamy PM, Bieber F, Kaiser L, Krishnan KR, Reuning-Scherer J, Gulanski B (1997) The Alzheimer's disease assessment scale: patterns and predictors of baseline cognitive performance in multicenter Alzheimer's disease trials. *Neurol* 48:1511–1517
- Fleiss JL, Levin B, Paik MC (2003) *Statistical methods for rates and proportions*, 3rd edn. Wiley, Hoboken
- Food and Drug Administration (2013) Guidance for industry. Alzheimer's disease: developing drugs for the treatment of early stage disease. Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, MD, USA. Available at: <http://www.fda.gov/ucm/groups/fdagov-public/@fdagov-drugs-gen/documents/document/ucm338287.pdf>. Accessed 25 Nov 2015
- Glimm E, Maurer W, Bretz F (2010) Hierarchical testing of multiple endpoints in group-sequential trials. *Stat Med* 29:219–228
- Green R et al (2009) Effect of tarenflurbil on cognitive decline and activities of daily living in patients with mild Alzheimer disease: a randomized controlled trial. *J Am Med Assoc* 302:2557–2564
- Grundmann O, Yoon SL (2010) Irritable bowel syndrome: epidemiology, diagnosis, and treatment: an update for health-care practitioners. *J Gastroenterol and Hepatol* 25:691–699
- Hamasaki T, Sugimoto T, Evans SR, Sozu T (2013) Sample size determination for clinical trials with co-primary outcomes: exponential event times. *Pharm Stat* 12:28–34
- Hamasaki T, Asakura K, Evans SR, Sugimoto T, Sozu T (2015) Group sequential strategies for clinical trials with multiple co-primary endpoints. *Stat Biopharm Res* 7:36–54
- Hung HMJ, Wang SJ (2009) Some controversial multiple testing problems in regulatory applications. *J Biopharm Stat* 19:1–11
- Jennison C, Turnbull BW (1993) Group sequential tests for bivariate response: interim analyses of clinical trials with both efficacy and safety. *Biometrics* 49:741–752

- Lan KKG, DeMets DL (1983) Discrete sequential boundaries for clinical trials. *Biometrika* 70:659–663
- Le Cessie S, van Houwelingen JC (1994) Logistic regression for correlated binary data. *Appl Stat* 43:95–108
- O’Brien PC, Fleming TR (1979) A multiple testing procedure for clinical trials. *Biometrics* 35:549–556
- Offen W et al (2007) Multiple co-primary endpoints: medical and statistical solutions. *Drug Inf J* 41:31–46
- Pocock SJ (1977) Group sequential methods in the design and analysis of clinical trials. *Biometrika* 64:191–199
- Prentice RL (1988) Correlated binary regression with covariates specific to each binary observation. *Biometrics* 44:1033–1048
- Sozu T, Sugimoto T, Hamasaki T (2010) Sample size determination in clinical trials with multiple co-primary binary endpoints. *Stat Med* 29:2169–2179
- Sozu T, Sugimoto T, Hamasaki T (2011) Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints. *J Biopharm Stat* 21:1–19
- Sozu T, Sugimoto T, Hamasaki T, Evans SR (2015) Sample size determination in clinical trials with multiple primary endpoints. Springer International Press, Cham
- Sugimoto T, Sozu T, Hamasaki T (2012) A convenient formula for sample size calculations in clinical trials with multiple co-primary continuous endpoints. *Pharm Stat* 11:118–128
- Tamhane AC, Mehta CR, Liu L (2010) Testing a primary and secondary endpoint in a group sequential design. *Biometrics* 66:1174–1184

Group-Sequential Clinical Trials with Multiple
Co-Objectives

Hamasaki, T.; Asakura, K.; Evans, S.R.; Ochiai, T.

2016, IX, 113 p. 14 illus., Softcover

ISBN: 978-4-431-55898-9