

Chapter 2

Continuous Co-primary Endpoints

Abstract In this chapter, we provide an overview of the concepts and the technical fundamentals regarding power and sample size calculation when comparing two interventions with multiple co-primary continuous endpoints in a clinical trial. We provide numerical examples to illustrate the methods and introduce conservative sample sizing strategies for these clinical trials.

Keywords Conjunctive power • Conservative sample size • Intersection-union test • Multivariate normal

2.1 Introduction

Consider a randomized clinical trial comparing two interventions with n_T subjects in the test group and n_C subjects in the control group. There are $K (\geq 2)$ co-primary continuous endpoints with a K -variate normal distribution. Let the responses for the n_T subjects in the test group be denoted by Y_{Tjk} , $j = 1, \dots, n_T$, and those for the n_C subjects in the control group, by Y_{Cjk} , $j = 1, \dots, n_C$. Suppose that the vectors of responses $\mathbf{Y}_{Tj} = (Y_{Tj1}, \dots, Y_{TjK})^T$ and $\mathbf{Y}_{Cj} = (Y_{Cj1}, \dots, Y_{CjK})^T$ are independently distributed as K -variate normal distributions with mean vectors $E[\mathbf{Y}_{Tj}] = \boldsymbol{\mu}_T = (\mu_{T1}, \dots, \mu_{TK})^T$ and $E[\mathbf{Y}_{Cj}] = \boldsymbol{\mu}_C = (\mu_{C1}, \dots, \mu_{CK})^T$, respectively, and common covariance matrix $\boldsymbol{\Sigma}$, i.e.,

$$\mathbf{Y}_{Tj} \sim N_K(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}) \quad \text{and} \quad \mathbf{Y}_{Cj} \sim N_K(\boldsymbol{\mu}_C, \boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \cdots & \rho^{1K} \sigma_1 \sigma_K \\ \vdots & \ddots & \vdots \\ \rho^{1K} \sigma_1 \sigma_K & \cdots & \sigma_K^2 \end{pmatrix}$$

with $\text{var}[Y_{Tjk}] = \text{var}[Y_{Cjk}] = \sigma_k^2$, $\text{corr}[Y_{Tjk}, Y_{Tjk'}] = \text{corr}[Y_{Cjk}, Y_{Cjk'}] = \rho^{kk'}$ ($k \neq k' : 1 \leq k < k' \leq K$). In this setting, $\rho^{kk'}$ is the association measure among the endpoints.

We are interested in estimating the difference in the means $\mu_{Tk} - \mu_{Ck}$. A positive value of $\mu_{Tk} - \mu_{Ck}$ indicates an intervention benefit. We assert the superiority of the test intervention over the control in terms of all K primary endpoints if and only if $\mu_{Tk} - \mu_{Ck} > 0$ for all $k = 1, \dots, K$. Thus, the hypotheses for testing are

$$\begin{aligned} H_0 &: \mu_{Tk} - \mu_{Ck} \leq 0 \text{ for at least one } k, \\ H_1 &: \mu_{Tk} - \mu_{Ck} > 0 \text{ for all } k. \end{aligned}$$

In testing the preceding hypotheses, the null hypothesis H_0 is rejected if and only if all of the null hypotheses associated with each of the K primary endpoints are rejected at a significance level of α . The corresponding rejection region is the intersection of K regions associated with the K co-primary endpoints; therefore the test used in data analysis is an intersection-union test (IUT) (Berger 1982).

2.2 Test Statistics and Power

2.2.1 Known Variance

Assume that σ_k^2 is known. The following Z -statistic can be used to test the difference in the means for each endpoint:

$$Z_k = \frac{\bar{Y}_{Tk} - \bar{Y}_{Ck}}{\sigma_k \sqrt{\frac{1}{n_T} + \frac{1}{n_C}}}, \quad k = 1, \dots, K, \quad (2.1)$$

where \bar{Y}_{Tk} and \bar{Y}_{Ck} are the sample means given by

$$\bar{Y}_{Tk} = \frac{1}{n_T} \sum_{j=1}^{n_T} Y_{Tjk} \quad \text{and} \quad \bar{Y}_{Ck} = \frac{1}{n_C} \sum_{j=1}^{n_C} Y_{Cjk}.$$

The overall power function for the Z -statistics in (2.1) can be written as

$$1 - \beta = \Pr \left[\bigcap_{k=1}^K \{Z_k > z_\alpha\} \mid H_1 \right] = \Pr \left[\bigcap_{k=1}^K \{Z_k^* > c_k^*\} \mid H_1 \right], \quad (2.2)$$

where $Z_k^* = Z_k - \sqrt{\kappa n} \delta_k$, $c_k^* = z_\alpha - \sqrt{\kappa n} \delta_k$, $\delta_k = (\mu_{Tk} - \mu_{Ck})/\sigma_k$ (standardized effect size), $r = n_C/n_T$, $n = n_T$, and $\kappa = r/(1+r)$. Further, z_α is the $(1 - \alpha)$ quantile of the standard normal distribution. This overall power (2.2) is referred to as “complete power” (Westfall et al. 2011) or “conjunctive power” (Bretz et al. 2011; Senn and Bretz 2007). Since $E[Z_k^*] = 0$ and $\text{var}[Z_k^*] = 1$, the vector of $(Z_1^*, \dots, Z_K^*)^T$ is distributed as a K -variate normal distribution, $N_K(\mathbf{0}, \boldsymbol{\rho}_Z)$, where

the off-diagonal element of ρ_Z is given by $\rho^{kk'}$. The overall power function is calculated using $\Phi_K(-c_1^*, \dots, -c_K^*)$, where Φ_K is the cumulative distribution function of $N_K(\mathbf{0}, \rho_Z)$.

2.2.2 Unknown Variance

We assume that σ_k^2 is unknown as is realistic in practice. The following T -statistic can be used to test the difference in the means for each endpoint:

$$T_k = \frac{\bar{Y}_{Tk} - \bar{Y}_{Ck}}{s_k \sqrt{\frac{1}{n_T} + \frac{1}{n_C}}}, \quad k = 1, \dots, K, \quad (2.3)$$

where \bar{Y}_{Tk} and \bar{Y}_{Ck} are the sample means defined in the previous section, and s_k is the usual pooled standard deviation given by

$$s_k^2 = \frac{\sum_{j=1}^{n_T} (Y_{Tjk} - \bar{Y}_{Tk})^2 + \sum_{j=1}^{n_C} (Y_{Cjk} - \bar{Y}_{Ck})^2}{n_T + n_C - 2}.$$

Let $\mathbf{D} = (D_1, \dots, D_K)^T$ with $D_k = (\bar{Y}_{Tk} - \bar{Y}_{Ck})/\sigma_k$, then $\sqrt{\kappa n} \mathbf{D}$ is distributed as a K -variate normal distribution with mean vector $\sqrt{\kappa n} \boldsymbol{\delta}$ and covariance matrix ρ_Z , namely, $N_K(\sqrt{\kappa n} \boldsymbol{\delta}, \rho_Z)$. In addition, the pooled matrix of the sums of squares and cross products,

$$\mathbf{W} = \begin{pmatrix} w_{11} & \cdots & w_{1K} \\ \vdots & \ddots & \vdots \\ w_{1K} & \cdots & w_{KK} \end{pmatrix}$$

is distributed as a Wishart distribution with $n_T + n_C - 2$ degree of freedom and covariance matrix ρ_Z where

$$w_{kk'} = \begin{cases} \frac{1}{\sigma_k^2} \left(\sum_{j=1}^{n_T} (Y_{Tjk} - \bar{Y}_{Tk})^2 + \sum_{j=1}^{n_C} (Y_{Cjk} - \bar{Y}_{Ck})^2 \right), & k = k', \\ \frac{1}{\sigma_k \sigma_{k'}} \left(\sum_{j=1}^{n_T} (Y_{Tjk} - \bar{Y}_{Tk})(Y_{Tjk'} - \bar{Y}_{Tk'}) + \sum_{j=1}^{n_C} (Y_{Cjk} - \bar{Y}_{Ck})(Y_{Cjk'} - \bar{Y}_{Ck'}) \right), & k \neq k'. \end{cases}$$

Please see, e.g., Johnson and Kotz (1972) for the definition of the Wishart distribution. Subsequently statistic (2.3) can be rewritten as

$$T_k = \frac{\sqrt{\kappa n} D_k}{\sqrt{\frac{w_{kk}}{n_T + n_C - 2}}}$$

and the overall power function for statistic (2.3) is given by

$$1 - \beta = \Pr \left[\bigcap_{k=1}^K \{T_k > t_{\alpha, n_T + n_C - 2}\} \mid H_1 \right], \quad (2.4)$$

where $t_{\alpha, n_T + n_C - 2}$ is the $(1 - \alpha)$ quantile of the t -distribution with $n_T + n_C - 2$ degrees of freedom. If $K = 1$, then the overall power function (2.4) is based on a noncentral univariate t -distribution (e.g., Julious 2009). If $K \geq 2$, then the joint distribution of T_k is not a multivariate noncentral t -distribution because the joint distribution of $w_{kk'}$ is a Wishart distribution, which is not included in a multivariate gamma distribution. Hence, in order to calculate the overall power function of such a distribution, we consider rewriting (2.4) as

$$\begin{aligned} 1 - \beta &= \Pr \left[\bigcap_{k=1}^K \left\{ \sqrt{\kappa n} D_k > t_{\alpha, n_T + n_C - 2} \sqrt{\frac{w_{kk}}{n_T + n_C - 2}} \right\} \mid H_1 \right] \\ &= E \left[\Pr \left[\bigcap_{k=1}^K \{Z_k^* > c_k^*(w_{kk})\} \mid \mathbf{W} \right] \right] \\ &= E \left[\Phi_K(-c_1^*(w_{11}), \dots, -c_K^*(w_{KK})) \right], \end{aligned} \quad (2.5)$$

where

$$c_k^*(w_{kk}) = t_{\alpha, n_T + n_C - 2} \sqrt{\frac{w_{kk}}{n_T + n_C - 2}} - \sqrt{\kappa n} \delta_k \quad (\delta_k > 0 \text{ for all } k).$$

The equation (2.5) is calculated by a simulated average of $\Phi_K(-c_1^*(w_{11}), \dots, -c_K^*(w_{KK}))$ obtained by generating random numbers of \mathbf{W} . For additional details, please see Sozu et al. (2006).

2.3 Sample Size Calculation

In the sample size calculation, the means μ_{T_k} , μ_{C_k} , the variance σ_k^2 , and the correlation coefficient $\rho^{kk'}$ must be specified in advance. The sample size required to achieve the desired overall power of $1 - \beta$ at the significance level of α is the smallest integer not less than n satisfying $1 - \beta \leq \Phi_K(-c_1^*, \dots, -c_K^*)$ for the known variance and $1 - \beta \leq E[\Phi_K(-c_1^*(w_{11}), \dots, -c_K^*(w_{KK}))]$ for the unknown variance. An iterative procedure is required to find the required sample size. The easiest way is a grid search to increase n gradually until the power under n exceeds the desired overall power of $1 - \beta$, where the maximum value of the sample sizes separately calculated for each endpoint can be used as the initial value for sample size calculation. However, this often takes much computing time. To improve the convenience in the sample size calculation, Chap. 4 provides a more efficient and practical algorithm for

calculating the sample sizes and presents a useful sample size formula with numerical tables for multiple co-primary endpoints.

When the standardized effect size for one endpoint is relatively smaller than that for other endpoints, then the sample size is determined by the smallest standardized effect size and does not greatly depend on the correlation. In this situation, the sample size equation for co-primary continuous endpoints can be simplified, using the equation for the single continuous endpoint, as given by Eq. (2.8) in Sect. 2.5.

2.4 Behavior of the Type I Error Rate, Power and Sample Size

We focus on the behavior of the type I error rate, overall power and sample size calculated using the method based on the known variance in Sect. 2.2.1, because the method based on the unknown variance in Sect. 2.2.2 provides similar results. Sozu et al. (2011) show that the sample size per group calculated using the method based on the unknown variance is generally one participant larger than that using the method based on the known variance.

2.4.1 Type I Error Rate

There are alternative hypotheses in which the corresponding powers are lower than the nominal significance level in order to keep the maximum type I error rate below the nominal significance level as described in the ICH (1998). For more details, please see Chuang-Stein et al. (2007) and Eaton and Muirhead (2007).

Figure 2.1 illustrates the behavior of type I error rate for $\alpha = 0.025$ as a function of the correlation, where the off-diagonal elements of the correlation matrix are equal, i.e., $\rho = \rho^{12} = \dots = \rho^{K-1,K}$, and all of the standardized effect sizes are zero, i.e., $\delta_1 = \dots = \delta_K = 0$ ($K = 2, 3, 4, 5$, and 10).

Fig. 2.1 Behavior of the type I error rate as a function of the correlation, where the off-diagonal elements of the correlation matrix are equal, i.e., $\rho = \rho^{12} = \dots = \rho^{K-1,K}$, and all of the standardized effect sizes are zero, i.e., $\delta_1 = \dots = \delta_K = 0$ ($K = 2, 3, 4, 5$, and 10)

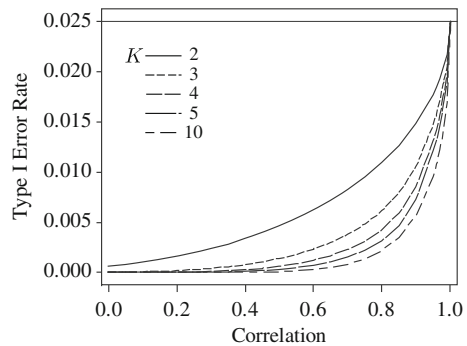


Fig. 2.2 Behavior of the type I error rate as a function of the correlation when there are two co-primary endpoints ($K = 2$)

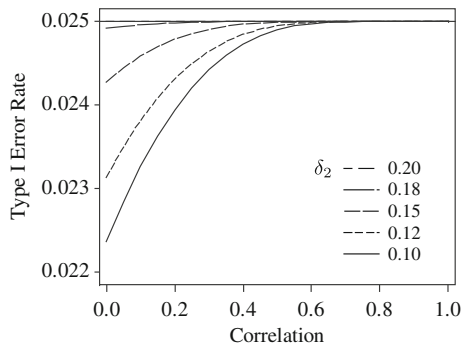


Figure 2.2 illustrates the behavior of type I error rate for $\alpha = 0.025$ as a function of the correlation when there are two co-primary endpoints ($K = 2$), where $\delta_1 = 0$ and $\delta_2 = 0.10, 0.12, 0.15, 0.18$, and 2.0 .

2.4.2 Overall Power

Figure 2.3 illustrates the behavior of overall power $1 - \beta$ as a function of the correlation for a given equal sample size per group $n = n_T = n_C$ (i.e., $r = 1.0$) so that the individual power for a single primary endpoint is at least 0.80 and 0.90 by a one-sided test at the significance level of $\alpha = 0.025$. Here, the off-diagonal elements of the correlation matrix are equal, i.e., $\rho = \rho^{12} = \dots = \rho^{K-1,K}$, and all of the standardized effect sizes are equal to 0.2, i.e., $\delta_1 = \dots = \delta_K = 0.2$ ($K = 2, 3, 4, 5$, and 10). The figure illustrates that the overall power increases as the correlation approaches one and decreases as the number of endpoints to be evaluated increases.

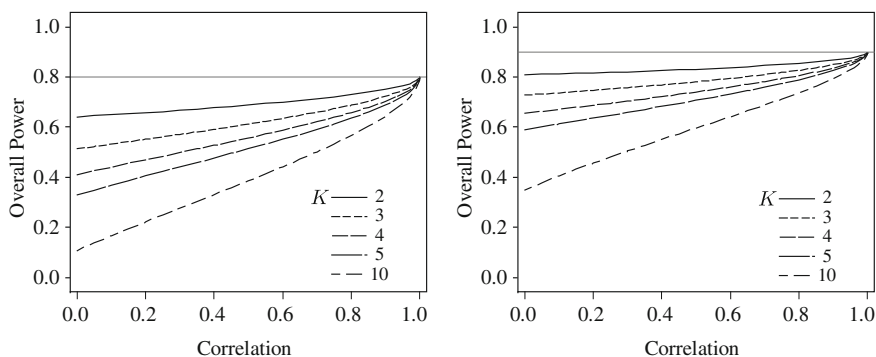


Fig. 2.3 Behavior of the overall power $1 - \beta$ as a function of the correlation for a given sample size so that the individual power for a single primary endpoint is at least 0.80 (the *left panel*) and 0.90 (the *right panel*) by a one-sided test at the significance level of $\alpha = 0.025$

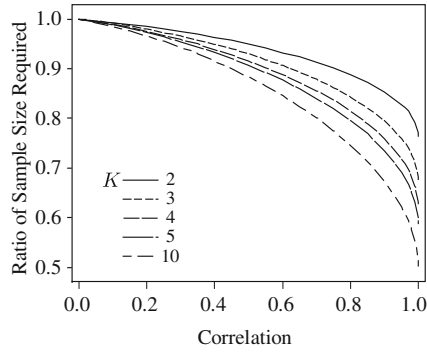


Fig. 2.4 Behavior of the ratio $(n(\rho)/n(0))$ as a function of the correlation, where the off-diagonal elements of the correlation matrix are equal, i.e., $\rho = \rho^{12} = \dots = \rho^{K-1,K}$, and all of the standardized effect sizes are equal to 0.2, i.e., $\delta_1 = \dots = \delta_K = 0.2$ ($K = 2, 3, 4, 5$, and 10). The sample size was calculated with the overall power of $1 - \beta = 0.80$ when each of the K endpoints is tested at the significance level of $\alpha = 0.025$ by a one-sided test

2.4.3 Sample Size

Figure 2.4 illustrates the behavior of the ratio of $n(\rho)$ to $n(0)$ as a function of the correlation when there are K co-primary endpoints ($K = 2, 3, 4, 5$, and 10), where the off-diagonal elements of the correlation matrix are equal $\rho = \rho^{12} = \dots = \rho^{K-1,K}$, and all of the standardized effect sizes are equal to 0.2, i.e., $\delta_1 = \dots = \delta_K = 0.2$. The equal sample sizes per group $n = n_T = n_C$ (i.e., $r = 1.0$) were calculated with the overall power of $1 - \beta = 0.80$ when each of the K endpoints is tested at the significance level of $\alpha = 0.025$ by a one-sided test. The figure illustrates that the ratio $n(\rho)/n(0)$ becomes smaller as the correlation approaches one and the degree of reduction is larger as the number of endpoints to be evaluated increases.

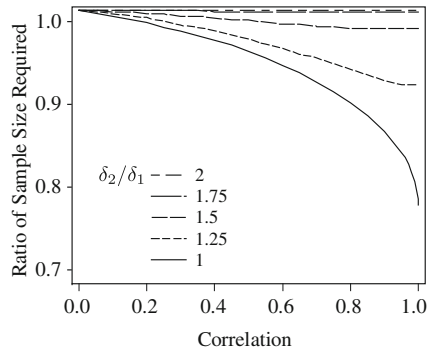


Fig. 2.5 Behavior of the ratio $(n(\rho)/n(0))$ as a function of the correlation for two co-primary endpoints ($K = 2$). The sample size was calculated with the overall power of $1 - \beta = 0.80$ when each of the two endpoints is tested at the significance level of $\alpha = 0.025$ by a one-sided test

Table 2.1 Sample size per group ($n = n_T = n_C$, $r = 1.0$) for two co-primary endpoints ($K = 2$) with the overall power of $1 - \beta = 0.80$ and 0.9 assuming that variance is known

Targeted power	Standardized effect size		Correlation ρ^{12}						
	δ_1	δ_2	0.0	0.3	0.5	0.8	1.0	E_1	E_2
0.80	0.20	0.20	516	503	490	458	393	393	393
	0.20	0.25	432	424	417	401	393	393	252
	0.20	0.30	402	399	397	393	393	393	175
	0.20	0.35	394	394	393	393	393	393	129
	0.20	0.40	393	393	393	393	393	393	99
	0.25	0.25	330	322	314	294	252	252	252
	0.25	0.30	284	278	272	260	252	252	175
	0.25	0.35	263	260	257	253	252	252	129
	0.25	0.40	254	253	253	252	252	252	99
	0.30	0.30	230	224	218	204	175	175	175
	0.30	0.35	201	197	192	183	175	175	129
	0.30	0.40	186	183	181	176	175	175	99
	0.35	0.35	169	165	160	150	129	129	129
	0.35	0.40	150	147	143	136	129	129	99
	0.40	0.40	129	126	123	115	99	99	99
0.90	0.20	0.20	646	637	626	597	526	526	526
	0.20	0.25	552	547	542	531	526	526	337
	0.20	0.30	529	528	527	526	526	526	234
	0.20	0.35	526	526	526	526	526	526	172
	0.20	0.40	526	526	526	526	526	526	132
	0.25	0.25	413	408	401	382	337	337	337
	0.25	0.30	360	356	352	343	337	337	234
	0.25	0.35	342	340	339	337	337	337	172
	0.25	0.40	337	337	337	337	337	337	132
	0.30	0.30	287	283	279	265	234	234	234
	0.30	0.35	254	251	248	240	234	234	172
	0.30	0.40	240	239	237	235	234	234	132
	0.35	0.35	211	208	205	195	172	172	172
	0.35	0.40	189	187	184	178	172	172	132
	0.40	0.40	162	160	157	150	132	132	132

E_1, E_2 : Sample size separately calculated for each endpoint 1 and 2 so that the individual power is at least 0.8 and 0.9

Figure 2.5 illustrates the behavior of the ratio of $n(\rho)$ to $n(0)$ as a function of the correlation when there are two co-primary endpoints ($K = 2$), where $\delta_2/\delta_1 = 1.0, 1.25, 1.50, 1.75$, and 2.0 . The equal sample sizes per group $n = n_T = n_C$ (i.e., $r = 1.0$) were calculated with the overall power of $1 - \beta = 0.80$ when each of two endpoints is tested at the significance level of $\alpha = 0.025$ by a one-sided test and the vertical axis is the ratio of $n(\rho^{12})$ to $n(0)$. When $\delta_2/\delta_1 = 1.0$, the ratio $(n(\rho)/n(0))$ decreases as the correlation approaches one. Even when $1.0 < \delta_2/\delta_1 < 1.5$, the ratio

Table 2.2 Sample size per group ($n = n_T = n_C, r = 1.0$) for three endpoints ($K = 3$) with the overall power of $1 - \beta = 0.80$ and 0.9 assuming that variance is known

Targeted power	Standardized effect size			Correlation ρ^{12}							
	δ_1	δ_2	δ_3	0.0	0.3	0.5	0.8	1.0	E_1	E_2	E_3
0.80	0.20	0.20	0.20	586	566	545	494	393	393	393	393
	0.20	0.20	0.30	517	504	490	458	393	393	393	175
	0.20	0.20	0.40	516	503	490	458	393	393	393	099
	0.20	0.30	0.30	410	404	400	394	393	393	175	175
	0.20	0.30	0.40	402	399	397	393	393	393	175	99
	0.20	0.40	0.40	393	393	393	393	393	393	99	99
	0.30	0.30	0.30	261	252	242	220	175	175	175	175
	0.30	0.30	0.40	233	226	220	204	175	175	175	99
	0.30	0.40	0.40	194	190	186	177	175	175	99	99
	0.40	0.40	0.40	147	142	137	124	99	99	99	99
0.90	0.20	0.20	0.20	714	700	683	635	526	526	526	526
	0.20	0.20	0.30	646	637	626	597	526	526	526	234
	0.20	0.20	0.40	646	637	626	597	526	526	526	132
	0.20	0.30	0.30	532	530	528	526	526	526	234	234
	0.20	0.30	0.40	529	528	527	526	526	526	234	132
	0.20	0.40	0.40	526	526	526	526	526	526	132	132
	0.30	0.30	0.30	318	311	304	283	234	234	234	234
	0.30	0.30	0.40	289	284	279	266	234	234	234	132
	0.30	0.40	0.40	245	243	240	235	234	234	132	132
	0.40	0.40	0.40	179	175	171	159	132	132	132	132

E_1, E_2, E_3 : Sample size separately calculated for each endpoint 1, 2, and 3 so that the individual power is at least 0.8 and 0.9

$(n(\rho)/n(0))$ still decreases as the correlation approaches one. However, when the ratio δ_2/δ_1 exceeds 1.5, the ratio $(n(\rho)/n(0))$ does not change considerably as the correlation varies.

Table 2.1 provides the equal sample sizes per group ($n = n_T = n_C, r = 1.0$) for two co-primary endpoints ($K = 2$) with correlation $\rho^{12} = 0.0$ (no correlation), 0.3 (low correlation), 0.5 (moderate correlation), 0.8 (high correlation), and 1.0 (perfect correlation). The sample size was calculated to detect standardized effect sizes of $0.2 \leq \delta_1, \delta_2 \leq 0.4$ with the overall power of $1 - \beta = 0.80$, when each of the two endpoints is tested at the significance level of $\alpha = 0.025$ by a one-sided test.

In the cases of equal effect sizes between the two endpoints, that is, $\delta_1 = \delta_2$, the sample size decreases as the correlation approaches one. Comparing the cases of $\rho^{12} = 0.0$ and $\rho^{12} = 0.8$, the decrease in the sample size is approximately 11 %. Even in the cases of unequal effect sizes, that is, $\delta_1 < \delta_2$, the sample sizes decrease as the correlation approaches one. However, when the ratio δ_2/δ_1 exceeds roughly 1.5, the sample size does not change considerably as the correlation varies. Consequently, the sample size is determined by the smaller effect size and is approximately equal to that calculated on the basis of the smaller effect size.

Similar to the case of two endpoints, Table 2.2 provides the equal sample sizes per group for three endpoints ($K = 3$) to detect standardized effect sizes $0.2 \leq \delta_1, \delta_2, \delta_3 \leq 0.4$ with overall power $1 - \beta = 0.8$ when each of the three endpoints is tested at the significance level of $\alpha = 0.025$ by a one-sided test, where the off-diagonal elements of the correlation matrix are equal, i.e., $\rho = \rho^{12} = \rho^{13} = \rho^{23} = 0.0, 0.3, 0.5, 0.8$, and 1.0 . In the cases of equal effect sizes among three endpoints, that is, $\delta_1 = \delta_2 = \delta_3$, the sample size decreases as the correlation approaches one. For example, comparing the cases of $\rho = 0.0$ and $\rho = 0.8$, the decrease in the sample size is approximately 16 %. Even in the cases of unequal effect sizes, that is, $\delta_1 < \delta_2 \leq \delta_3$, the sample size decreases as the correlation approaches one. However, when the ratios δ_2/δ_1 and δ_3/δ_1 exceed 1.5, the sample size does not change as the correlation varies. Consequently, the sample size is determined by the smallest effect size and is approximately equal to that calculated on the basis of the smallest effect size.

2.5 Conservative Sample Size Determination

When clinical trialists face the challenge of sizing clinical trials with multiple endpoints, one major concern is whether the correlations among the endpoints should be considered in the sample size calculation. The correlations may be estimated from external or internal pilot data, but they are usually unknown. When there are more than two endpoints, estimating the correlations is extremely difficult. If the correlations are over-estimated and are included into the sample size calculation for evaluating the joint effects on all of the endpoints, then the sample size is too small and important effects may not be detected. As a conservative approach, one could assume zero correlations among the endpoints as the overall power for detecting the joint statistical significance is lowest when the correlation is zero for $\rho^{kk'} \geq 0$.

Consider a conservative sample size strategy when evaluating superiority for ALL continuous endpoints by using a suggestion in Hung and Wang (2009). For illustration, first consider a situation where there are two continuous co-primary endpoints. As seen in Fig. 2.5, the overall power is lowest (because the corresponding sample size is highest) when there are equal standardized effect sizes and zero correlation among the endpoints. So that, with a common value of $c^* = c_1^* = c_2^*$ (i.e., $\delta = \delta_1 = \delta_2$) in the overall power function, we could set

$$1 - \beta = \Phi_2(-c^*, -c^* \mid \rho^{12} = 0) = \{\Phi(-c^*)\}^2 \quad (2.6)$$

where $c^* = z_\alpha - \sqrt{\kappa n} \delta$. Solving (2.6) for n provides the conservative sample size n_{CNSV} given by

$$n_{CNSV} \geq \frac{(z_\alpha + z_\gamma)^2}{\kappa \delta^2}$$

where $\gamma = 1 - (1 - \beta)^{1/2}$.

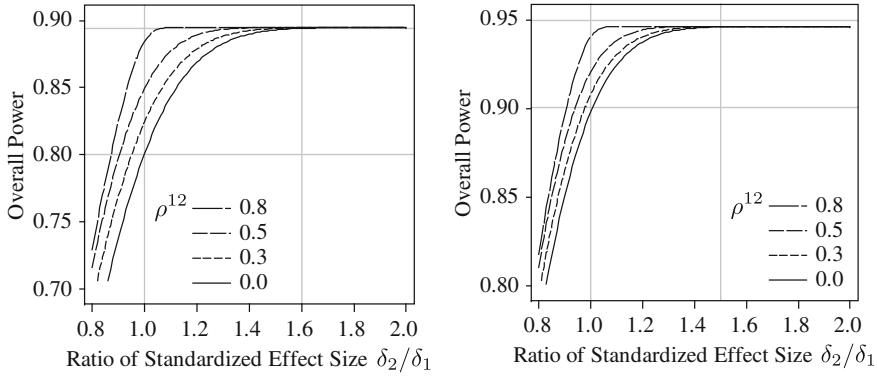


Fig. 2.6 Behavior of overall power $1 - \beta$ as a function of δ_2/δ_1 for a given equal sample size per group $n = n_T = n_C$ (i.e., $r = 1.0$) to detect superiority for endpoint 1 with the targeted individual power $1 - \gamma$ of $0.8^{1/3}$ (the left panel) and $0.9^{1/3}$ (the right panel) for a one-sided test at the significance level of $\alpha = 0.025$

In practice, one challenge is how to select a common value of c^* . The most conservative way is to choose a smaller value of either c_1^* or c_2^* . This may provide a sample size large enough to detect the joint superiority for both endpoints. Now calculate a sample size n required to detect superiority for endpoint 1, with the targeted individual power $1 - \gamma$ at the significance level of α assuming $\rho^{12} = 0$, i.e., $n_1 = (z_\alpha + z_\gamma)^2 / (\kappa \delta_1^2)$, where $\delta_1 \leq \delta_2$. The overall power $1 - \beta$ under n_1 is given as

$$1 - \beta = \Phi_2(-c_1^*, -c_2^* \mid \rho^{12} = 0) = \Phi(-c_1^*)\Phi(-c_2^*)$$

where $c_1^* = -z_\gamma$ and $c_2^* = z_\alpha - (\delta_2/\delta_1)(z_\alpha + z_\gamma)$. Therefore, the overall power can be expressed as a function of ratio of the standardized effect sizes.

Figure 2.6 illustrates the behavior of overall power $1 - \beta$ as a function of δ_2/δ_1 for a given equal sample size per group $n = n_T = n_C$ (i.e., $r = 1.0$) to detect superiority for endpoint 1 with the targeted individual power $1 - \gamma$ of $0.8^{1/3}$ and $0.9^{1/3}$ for a one-sided test at the significance level of $\alpha = 0.025$.

For the case of $1 - \gamma = 0.8^{1/2}$, the figure illustrates that the overall power increases toward $0.8^{1/2}$ as the ratio δ_2/δ_1 increases. In particular when the ratio δ_2/δ_1 is roughly greater than 1.6, the overall power almost reaches $0.8^{1/2}$. This is because the individual power for endpoint 2 is very close to one ($\Phi(-c_2^*) \rightarrow 1$) under the given sample size calculated for endpoint 1 and the overall power depends greatly on the smaller difference. For the case of $1 - \gamma = 0.9^{1/2}$, when the ratio δ_2/δ_1 is roughly greater than 1.4, then the overall power reaches $0.9^{1/2}$. From this result, if we observe a large difference in the values of δ_1 and δ_2 , roughly $\delta_2/\delta_1 > 1.5$, then we could calculate the conservative sample size given by

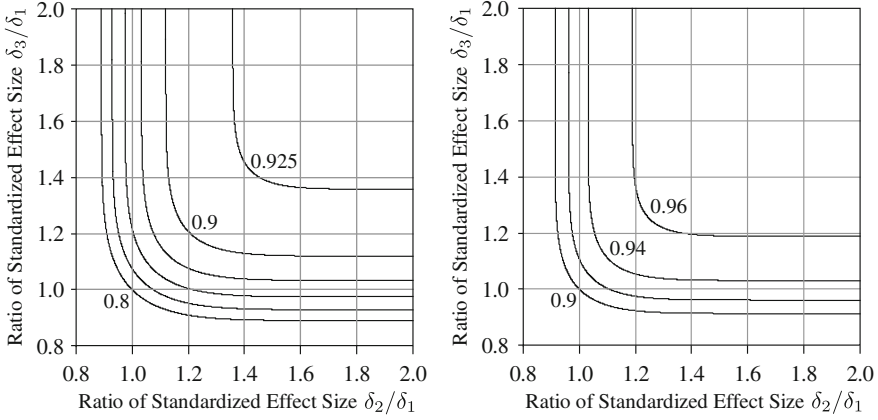


Fig. 2.7 Behavior of overall power $1 - \beta$ for three co-primary endpoints as a function of standardized effect sizes for a given equal sample size per group $n = n_T = n_C$ (i.e., $r = 1.0$) to detect superiority for endpoint 1 with the targeted individual power $1 - \gamma$ of $0.8^{1/3}$ (the left panel) and $0.9^{1/3}$ (the right panel) for a one-sided test at the significance level of $\alpha = 0.025$

$$n'_{CNSV} \geq \max \left(\frac{(z_\alpha + z_\beta)^2}{\kappa \delta_1^2}, \frac{(z_\alpha + z_\beta)^2}{\kappa \delta_2^2} \right).$$

Next we consider a more general situation where there are more than two endpoints. Similarly we calculate a sample size n to detect superiority for endpoint 1, with the targeted individual power $1 - \gamma = (1 - \beta)^{1/K}$ at the significance level of α assuming $\rho^{kk'} = 0$ i.e., $n_1 = (z_\alpha + z_\gamma)^2 / (\kappa \delta_1^2)$, where $\delta_1 \leq \dots \leq \delta_K$. Then the overall power $1 - \beta$ under n_1 is given as

$$1 - \beta = \Phi_K(-c_1^*, \dots, -c_K^* \mid \rho^{12} = \dots = \rho^{K-1,K} = 0) = \Phi(-c_1^*) \dots \Phi(-c_K^*) \quad (2.7)$$

where $c_1^* = -z_\gamma$ and $c_K^* = z_\alpha - (\delta_K / \delta_1)(z_\alpha + z_\gamma)$. Figure 2.7 illustrates the behavior of overall power for three co-primary endpoints as a function of δ_2 / δ_1 and δ_3 / δ_1 for a given equal sample size per group $n = n_T = n_C$ (i.e., $r = 1.0$) to detect superiority for endpoint 1 with the individual power $1 - \gamma$ of $0.8^{1/3}$ and $0.9^{1/3}$ for a one-sided test at the significance level of $\alpha = 0.025$.

For the case of $1 - \gamma = 0.8^{1/3}$, the figure illustrates that the overall power increases toward $0.8^{1/2}$ as the ratio δ_2 / δ_1 increases. In particular when both the ratio δ_2 / δ_1 and δ_3 / δ_1 are roughly greater than 1.5, the overall power almost reaches $0.8^{1/3}$. For the case of $1 - \gamma = 0.9^{1/3}$, when the both ratios are roughly greater than 1.4, the overall power almost reaches $0.9^{1/3}$. From this result, if we observe a large difference in the values of effect sizes, we could calculate the conservative sample size given by

$$n'_{CNSV} \geq \max \left(\frac{(z_\alpha + z_\beta)^2}{\kappa \delta_k^2} \right). \quad (2.8)$$

One question that arises is how large δ_k/δ_1 should be when the conservative sample size (2.8) is considered. To provide a reference value for δ_k/δ_1 , the overall power (2.7) is set to be at least $1 - \beta'$, i.e.,

$$(1 - r)\Phi(-z_\alpha + \delta_2/\delta_1(z_\alpha + z_\gamma)) \cdots \Phi(-z_\alpha + \delta_K/\delta_1(z_\alpha + z_\gamma)) > 1 - \beta' \quad (2.9)$$

and then the values of δ_k/δ_1 can be found satisfying the above inequality. For example, we consider a situation of $K = 2$. Solving (2.9) for δ_2/δ_1 gives

$$\frac{\delta_2}{\delta_1} > \frac{\Phi^{-1}\left(\frac{1 - \beta'}{1 - \gamma}\right) + z_\alpha}{z_\gamma + z_\alpha}. \quad (2.10)$$

If the target overall power $1 - \beta = 0.80$ and then the overall power is set to be at least greater than $1 - \beta' = 0.894$ as $1 - \gamma = \sqrt{0.8}$, by substituting these values into above inequality, we have $\delta_2/\delta_1 > 1.639$ with $\alpha = 0.025$. So that, when the one standardized effect size is large enough (or small enough) compared with the other, i.e., $\delta_2/\delta_1 > 1.639$, we may use the sample size equation (2.8). However, if $1 - \beta' = 0.8944$, $\delta_2/\delta_1 > 1.859$. Note that the ratio of standardized effect size will depend on a precision of decimal degree of $1 - \beta'$.

In addition, we discuss a more general situation with K endpoints. For simplicity, we assume $\delta_2 = \cdots = \delta_K$. Solving (2.9) for δ_k/δ_1 , we have

$$\frac{\delta_k}{\delta_1} > \frac{\Phi^{-1}\left(\left(\frac{1 - \beta'}{1 - \gamma}\right)^{\frac{1}{K-1}}\right) + z_\alpha}{z_\gamma + z_\alpha}.$$

Table 2.3 Reference values for ratio of standardized effect sizes for conservative sample sizing (2.8) with equal effect sizes $\delta_2 = \cdots = \delta_K$. $1 - \beta'$ is calculated by truncating the numbers beyond the fourth decimal point

Number of endpoints	Targeted overall $1 - \beta$			
	0.80	$(1 - \beta')$	0.90	$(1 - \beta')$
2	1.639	(0.894)	1.432	(0.948)
3	1.564	(0.928)	1.388	(0.965)
4	1.436	(0.945)	1.648	(0.974)
5	1.453	(0.956)	1.394	(0.979)
6	1.397	(0.963)	1.276	(0.982)
7	1.355	(0.968)	1.403	(0.985)
8	1.352	(0.972)	1.212	(0.986)
9	1.333	(0.975)	1.262	(0.988)
10	1.274	(0.977)	1.226	(0.989)

For example, consider a situation of $K = 3$. If the target overall power $1 - \beta = 0.80$ and the overall power is set to be at least greater than $1 - \beta' = 0.928$ as $1 - \gamma = 0.8^{1/3}$, we have $\delta_k/\delta_1 > 1.564$ with $\alpha = 0.025$. So that we may use the sample size equation (2.8) when both of the ratio of standardized effect sizes are larger than 1.564.

Table 2.3 shows typical reference values for ratio of standardized effect sizes given by (2.9) with equal effect sizes $\delta_2 = \dots = \delta_K$ when the conservative sample size (2.8) is considered.

2.6 Example

We illustrate the sample size calculations based on a clinical trial evaluating interventions for Alzheimer's disease. In Alzheimer's clinical trials, the change from the baseline in the ADAS-cog (the Alzheimer's Disease Assessment Scale-cognitive subscale) score and the CIBIC-plus (Clinician's Interview-Based Impression of Change, plus caregiver) at the last observed time point are commonly used as co-primary endpoints (e.g., Peskind et al. 2006; Rogers et al. 1998; Rösler et al. 1999; Tariot et al. 2000). In a 24-week, double-blind, placebo controlled trial of donepezil in patients with Alzheimer's disease in Rogers et al. (1998), the absolute values of the standardized effect size (with 95 % confidence interval) were estimated as 0.47 (0.24, 0.69) for ADAS-cog (δ_1) and 0.48 (0.25 0.70) for CIBIC-plus (δ_2). We use these estimates to define an alternative hypothesis to size a future trial. The sample sizes were calculated using the method based on the known variance to detect the standardized effect sizes $0.20 < \delta_1, \delta_2 < 0.70$ to achieve the overall power of $1 - \beta = 0.80$ at $\alpha = 0.025$, with $\rho^{12} = 0, 0.3, 0.5$, and 0.8 as the correlation between the two endpoints.

Figure 2.8 displays the contour plots of the sample sizes per group with two effect sizes δ_1 and δ_2 and correlation ρ^{12} . The figure displays how the sample size behaves as the two effect sizes and the correlations vary; when the effect sizes are approximately equal, the required sample size varies with the correlation. When one effect size is relatively smaller (or larger) than the other, the sample size is nearly determined by the smaller effect size, and does not depend greatly on the correlation. The correlation ρ^{12} is assumed to range between $-1 < \rho^{12} < 0.35$ by Offen et al. (2007) and $\rho^{12} = 0.5$ (as a trial value) by Xiong et al. 2005. As the baseline case of $(\delta_1, \delta_2) = (0.47, 0.48)$, the sample sizes per group for $\rho^{12} = 0, 0.3, 0.5$, and 0.8 were 92, 90, 87, and 82, respectively.

2.7 Summary

This chapter provides an overview of the concepts and technical fundamentals regarding power and sample size calculation for clinical trials with co-primary continuous endpoints when the alternative hypothesis is joint effects on all endpoints. The chapter also introduces conservative sample sizing strategies. Our major findings are as follows:

- There is an advantage of incorporating the correlation among endpoints into the power and sample size calculations with co-primary continuous endpoints. In general without design adjustments, the power is lower with additional endpoints, but can be improved by incorporating the correlation into the calculation (assuming a positive correlation). Thus incorporating the correlation into the sample size calculation may lead to a reduction in sample sizes. The reduction in sample size is greater with a greater number of endpoints, especially when the standardized effect sizes are approximately equal among the endpoints. For example, when the endpoints are positively correlated (correlation up to 0.8), with the power of 0.8

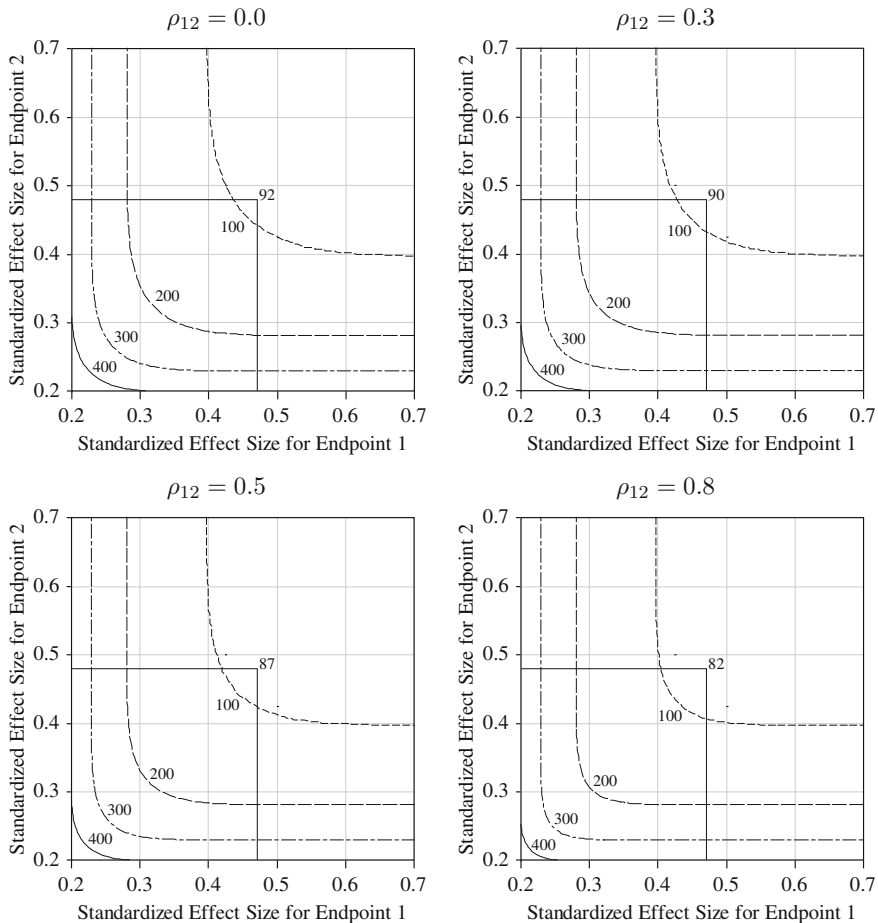


Fig. 2.8 Contour plot of the sample size (per group) for standardized effect sizes of endpoint 1 (SIB-J) and endpoint 2 (CIBIC plus-J) with $\rho^{12} = 0.0, 0.3, 0.5$, and 0.8 . The sample size was calculated to detect the superiority for all the endpoints with the overall power of $1 - \beta = 0.80$ for a one-sided test at the significance level of $\alpha = 0.025$

at the significant level of 0.025, there is approximately 11 % reduction in the case of two co-primary endpoints and 16 % reduction in the case of three co-primary endpoints, compared to the sample size calculated under the assumption of zero correlations among the endpoints.

- In most situations, the required sample size per group for co-primary continuous endpoints calculated using the method based on the assumption that variance is unknown is one participant larger than the method based on the assumption that the variance is known. This is very similar to results seen for a single continuous endpoint.
- When the standardized effect sizes for the endpoints are unequal, then the advantage of incorporating the correlation into sample size is less dramatic as the required sample size is primarily determined by the smaller standardized effect size and does not greatly depend on the correlation. In this situation, the sample size equation for co-primary continuous endpoints can be simplified using the equation for a single continuous endpoint, as given by Eq.(2.8). When the standardized effect sizes among endpoints are approximately equal, then the sample size method assuming zero correlation described in Hung and Wang (2009) may be used as the power is minimized with the equal standardized effect sizes.

References

- Berger RL (1982) Multiparameter hypothesis testing and acceptance sampling. *Technometrics* 24:295–300
- Bretz F, Hothorn T, Westfall P (2011) Multiple comparisons using R. Chapman & Hall/CRC, Boca Raton
- Chuang-Stein C, Stryczak P, Dmitrienko A, Offen W (2007) Challenge of multiple co-primary endpoints: a new approach. *Stat Med* 26:1181–1192
- Eaton ML, Muirhead RJ (2007) On a multiple endpoints testing problem. *J Stat Plann Infer* 137:3416–3429
- Hung HM, Wang SJ (2009) Some controversial multiple testing problems in regulatory applications. *J Biopharm Stat* 19:1–11
- International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use. ICH tripartite guideline. Statistical principles for clinical trials. 1998. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf Accessed 9 June 2014
- Johnson NL, Kotz S (1972) Distributions in statistics: continuous multivariate distributions. Wiley, New York
- Julious SA (2009) Sample sizes for clinical trials. Chapman & Hall, Boca Raton
- Offen W, Chuang-Stein C, Dmitrienko A, Littman G, Maca J, Meyerson L, Muirhead R, Stryczak P, Boddy A, Chen K, Copley-Merriman K, Dere W, Givens S, Hall D, Henry D, Jackson JD, Krishen A, Liu T, Ryder S, Sankoh AJ, Wang J, Yeh CH (2007) Multiple co-primary endpoints: medical and statistical solutions. *Drug Inf J* 41:31–46
- Peskind ER, Potkin SG, Pomara N, Ott BR, Graham SM, Olin JT, McDonald S (2006) Memantine treatment in mild to moderate Alzheimer disease: a 24-week randomized, controlled trial. *Am J Geriatr Psychiatry* 14:704–715

- Rogers SL, Farlow MR, Doody RS, Mohs R, Friedhoff LT (1998) The Donepezil Study Group. A 24-week, double-blind, placebo-controlled trial of donepezil in patients with Alzheimer's disease. *Neurology* 50:136–145
- Rösler M, Anand R, Cicin-Sain A, Gauthier S, Agid Y, Dal-Bianco P, Stähelin HB, Hartman R, Gharabawi M (1999) Efficacy and safety of rivastigmine in patients with Alzheimer's disease: international randomised controlled trial. *Br Med J* 318:633–640
- Senn S, Bretz F (2007) Power and sample size when multiple endpoints are considered
- Sozu T, Kanou T, Hamada C, Yoshimura I (2006) Power and sample size calculations in clinical trials with multiple primary variables. *Japan J Biometrics* 27:83–96
- Sozu T, Sugimoto T, Hamasaki T (2011) Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints. *J Biopharm Stat* 21:650–668
- Tariot PN, Solomon PR, Morris JC, Kershaw P, Lilienfeld S, Ding C (2000) The Galantamine USA Study Group. A 5-month, randomized, placebo-controlled trial of galantamine in AD. *Neurology* 54:2269–2276
- Westfall PH, Tobias RD, Wolfinger RD (2011) Multiple comparisons and multiple tests using SAS, 2nd edn. SAS Institute Inc, Cary
- Xiong C, Yu K, Gao F, Yan Y, Zhang Z (2005) Power and sample size for clinical trials when efficacy is required in multiple endpoints: application to an Alzheimer's treatment trial. *Clin Trials* 2:387–393

Sample Size Determination in Clinical Trials with Multiple
Endpoints

Sozu, T.; Sugimoto, T.; Hamasaki, T.; Evans, S.R.

2015, VI, 95 p. 17 illus., Softcover

ISBN: 978-3-319-22004-8