HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2015 July 30.

Published in final edited form as:

Stat Med. 2014 July 30; 33(17): 2897–2913. doi:10.1002/sim.6154.

Sample size determination in group-sequential clinical trials with two co-primary endpoints

Koko Asakura[1], **Toshimitsu Hamasaki**[1,2], **Tomoyuki Sugimoto**[3], **Kenichi Hayashi**[1], **Scott R Evans**[4], and **Takashi Sozu**[5]

[1]Department of Biomedical Statistics, Osaka University Graduate School of Medicine, Osaka, Japan

[2]Office of Biostatistics and Data Management, Research & Development Initiative Center, National Cerebral and Cardiovascular Center, Osaka, Japan

[3]Department of Mathematical Sciences, Graduate School of Science & Technology, Hirosaki University, Aomori, Japan

[4]Department of Biostatistics and the Center for Biostatistics in AIDS Research, Harvard School of Public Health, Massachusetts, USA

[5]Department of Biostatistics, Kyoto University School of Public Health, Kyoto, Japan

Abstract

We discuss sample size determination in group-sequential designs with two endpoints as co-primary. We derive the power and sample size within two decision-making frameworks. One is to claim the test intervention's benefit relative to control when superiority is achieved for the two endpoints *at the same interim timepoint* of the trial. The other is when the superiority is achieved for the two endpoints *at any interim timepoint, not necessarily simultaneously*. We evaluate the behaviors of sample size and power with varying design elements and provide a real example to illustrate the proposed sample size methods. In addition, we discuss sample size recalculation based on observed data and evaluate the impact on the power and Type I error rate.

Keywords

Average sample number; Conditional Power; Cui–Hung–Wang statistics; Co-primary endpoints; Group-sequential methods; Maximum sample size; Sample size recalculation; Type I error

1 Introduction

Traditionally, in clinical trials, a single outcome is selected as a primary endpoint. This endpoint is then used as the basis for the trial design including sample size determination, interim monitoring, and final analyses. However, many recent clinical trials, especially in

Conflict of Interest

The authors have declared no conflict of interest.

medical product development, have utilized more than one endpoint as *co-primary*. "Co-primary" in this setting means that the trial is designed to evaluate if the new intervention is superior to the control on *all* endpoints, thus evaluating the intervention's multidimensional effects. Regulators have issued guidelines recommending co-primary endpoints in some disease areas. For example, the Committee for Medicinal Products for Human Use (CHMP) issued a guideline [1] recommending the use of cognitive, functional, and global endpoints to evaluate symptomatic improvement of dementia associated with in Alzheimer's disease, indicating that primary endpoints should be stipulated reflecting the cognitive and functional disease aspects. Offen et al. [2] provides other examples with co-primary endpoints for regulatory purposes.

The resulting need for new approaches to the design and analysis of clinical trials with co-primary endpoints has been noted [2–4]. Utilizing multiple endpoints may provide the opportunity for characterizing intervention's multidimensional effects, but also creates challenges. Specifically controlling the Type I and Type II error rates when the multiple co-primary endpoints are potentially correlated is non-trivial. When designing the trial to evaluate the joint effects on ALL of the endpoints, no adjustment is needed to control the Type I error rate. However, the Type II error rate increases as the number of endpoints to be evaluated increases. Thus adjustments in design (i.e., sample size) are needed to maintain the overall power. Methods for clinical trials with co-primary endpoints have been discussed in fixed sample size designs by many authors [5–16]. Even if the correlation among the endpoints is incorporated into the sample size calculation, existing methods often result in large and impractical sample sizes as the testing procedure for co-primary endpoints is conservative. Chuang-Stein et al. [7] and Kordzakhia et al [10] discuss the methods to adjust the significance levels that depend on the correlation among the endpoints in the fixed sample size designs. The methods may provide relatively smaller sample sizes, but also introduce the other challenges. For example, the sample size calculated to detect the joint effect may be smaller than the sample size calculated for each individual endpoint. The prespecified correlation incorporated into the significance level adjustment is usually unknown and may be incorrect. This calls into question whether or not the significance level should be updated based on the observed correlation.

In this paper, we extend previous work for the fixed sample size designs, considering sample size evaluation in the group-sequential setting with co-primary endpoints. As suggested in Hung and Wang [3], a group-sequential design may be a remedial, but practical approach because it offers the possibility to stop a trial early when evidence is overwhelming and thus offers efficiency (i.e., potentially fewer patients than the fixed sample size designs). We discuss the case of two positively correlated continuous outcomes. We consider a two-arm parallel-group trial designed to evaluate if an experimental intervention is superior to a control. The paper is structured as follows: in Section 2 we describe the statistical setting, decision-making frameworks for rejecting the null hypothesis, and definitions of power. In Section 3, we evaluate the behaviors of sample size and power with varying design elements and then provide a real example to illustrate the methods. In Section 4, we describe sample size recalculation and the resulting effect on power and Type I error rate. In Section 5 we summarize the findings and discuss the further developments.

2 Group-sequential designs with two co-primary endpoints

2.1 Statistical setting

Consider a randomized, group-sequential clinical trial of comparing the test intervention (T) with the control intervention (C). Two continuous outcomes are to be evaluated as co-primary endpoints. Suppose that a maximum of L analyses are planned, where the same number of analyses with the same information space are selected for both endpoints. Let n_l and $r_l n_l$ be the cumulative number of participants on the test and the control intervention groups at the l th analysis (l=1,…, L), respectively, where r_l is the sampling ratio. Hence, up to n_L and $r_L n_L$ participants are recruited and randomly assigned to the test and the control intervention groups, respectively. Then there are n_L paired outcomes (Y_{T1i},Y_{T2i}) (i = 1,…, n_L) for the test intervention group and $r_L n_L$ paired outcomes (Y_{C1j},Y_{C2j}) (j=1,…, $r_L n_L$) for the control intervention group. Assume that (Y_{T1i},Y_{T2i}) and (Y_{C1j},Y_{C2j}) are independently bivariate-normally distributed as $(Y_{\mathrm{T}1i}, Y_{\mathrm{T}2i}) \sim \mathrm{N}_2(\mu_{\mathrm{T}1}, \mu_{\mathrm{T}2}, \sigma^2_{\mathrm{T}1}, \sigma^2_{\mathrm{T}2}, \rho_{\mathrm{T}})$ and $(Y_{\mathrm{C}1j}, Y_{\mathrm{C}2j}) \sim \mathrm{N}_2(\mu_{\mathrm{C}1}, \mu_{\mathrm{C}2}, \sigma^2_{\mathrm{C}1}, \sigma^2_{\mathrm{C}2}, \rho_{\mathrm{C}})$, respectively. For simplicity, the variances are assumed to be known and common, i.e., $\sigma^2_{\mathrm{T}1}=\sigma^2_{\mathrm{C}1}=\sigma^2_1$ and $\sigma^2_{\mathrm{T}2}=\sigma^2_{\mathrm{C}2}=\sigma^2_2$. Note that the method can be applied to the case of unknown variances. For the fixed sample size 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 a one-sided significance level α = 0.025 and power $1-\beta$ = 0.8 or 0.9. By analogy from the fixed sample designs, there is 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 (δ_1, δ_2) denote the differences in the means for the test and the control intervention groups respectively, where $\delta_k = \mu_{\mathrm{T}k} - \mu_{\mathrm{C}k}$ (k = 1, 2). Suppose that positive values of (δ_1, δ_2) represent the test intervention's benefit. We are interested in conducting a hypothesis test to evaluate if the intervention is superior to the control intervention, i.e., the null hypothesis $\mathrm{H}_0 : \delta_1 \leq 0$ or $\delta_2 \leq 0$ versus the alternative hypothesis $\mathrm{H}_1 : \delta_1 > 0$ and $\delta_2 > 0$. Let (Z_{1l}, Z_{2l}) be the statistics for testing the hypotheses at the l th analysis, given by $Z_{kl}=(\overline{Y}_{\mathrm{T}kl}-\overline{Y}_{\mathrm{C}kl})/(\sigma_k\sqrt{\kappa_l/n_l})$, where $\kappa_l = (1+r_l)/r_l$, and $\overline{Y}_{\mathrm{T}kl}$ and $\overline{Y}_{\mathrm{C}kl}$ are the sample means given by $\overline{Y}_{\mathrm{T}kl}=n_l^{-1}\sum_{i=1}^{n_l} Y_{\mathrm{T}ki}$ and $\overline{Y}_{\mathrm{C}kl}=(r_l n_l)^{-1}\sum_{j=1}^{r_l n_l} Y_{\mathrm{C}kj}$. Z_{1l} and Z_{2l} are normally distributed as $\mathrm{N}(\sqrt{n_l/\kappa_l}\delta_1/\sigma_1, 1^2)$ and $\mathrm{N}(\sqrt{n_l/\kappa_l}\delta_2/\sigma_2, 1^2)$, respectively. Thus ($Z_{1l}$, Z_{2l}) is bivariate-normally distributed with the correlation $(r_l\rho_{\mathrm{T}} + \rho_{\mathrm{C}})/(1 + r_l)$. Furthermore, the joint distribution of (Z_{11}, Z_{21},…, Z_{1L}, Z_{2L}) is $2L$ multivariate normal with their correlations given by $\mathrm{corr}[Z_{kl}, Z_{k'l'}] = \sqrt{\kappa_l n_{l'}/\kappa_{l'} n_l}$ if $k = k'$; $\sqrt{\kappa_l n_{l'}}(r_l\rho_{\mathrm{T}}+\rho_{\mathrm{C}})/\{\sqrt{\kappa_{l'} n_l}(1+r_l)\}$ if $k \neq k'$.

2.2 Decision-making framework, stopping rules, and power

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 and only if superiority is achieved for the two

endpoints simultaneously (i.e., at the same interim timepoint of the trial) (DF-1). The other is to reject H_0 if superiority is achieved for the two endpoints at any interim timepoint (i.e., not necessarily simultaneously) (DF-2). We will discuss the two decision-making frameworks separately as the corresponding stopping rules and power definitions are unique.

DF-1—The DF-1 is relatively simple: if superiority is demonstrated on only one endpoint at an interim, then the trial continues and the hypothesis testing is repeated for both endpoints until the joint significance for the two endpoints is established simultaneously. The stopping rule for DF-1 is formally given as follows:

At the l th analysis (l=1,…, L−1)

$$\begin{array}{ll} \text{If } Z_{1l}>c_{1l} \text{ and } Z_{2l}>c_{2l}, & \text{then reject } \mathrm{H}_0 \text{ and stop the trial,} \\ \text{otherwise,} & \text{continue to the } (l+1) \text{ th analysis,} \end{array}$$

at the Lth analysis

$$\begin{array}{ll} \text{if } Z_{1L}>c_{1L} \text{ and } Z_{2L}>c_{2L}, & \text{then reject } \mathrm{H}_0, \\ \text{otherwise,} & \text{do not reject } \mathrm{H}_0, \end{array}$$

where c_{1l} and c_{2l} are the critical values, which are constant and selected separately, using any group-sequential method such as the Lan-DeMets (LD) alpha-spending method [17] to control the overall Type I error rate of α, as if they were a single primary endpoint, ignoring the other co-primary endpoint. The testing procedure for co-primary endpoints is conservative. For example, if a zero correlation between the two endpoints is assumed and each endpoint is tested at the one-sided significance level of 2.5%, then the Type I error rate is 0.0625 %. As shown in Section 4, the maximum Type I error rate associated with the rejection region of the null hypothesis increases as the correlation goes toward one, but it is not greater than the targeted significance level.

The power corresponding to DF-1 is

$$1-\beta=\Pr\left[\cup_{l=1}^{L}\{A_{1l}\cap A_{2l}\}\,|\mathrm{H}_1\right], \quad (1)$$

where $A_{kl}=\{Z_{kl}>c_{kl}\}$(k= 1,2; l = 1,…, L). The power (1) can be numerically assessed by using multivariate normal integrals. A detailed calculation is provided in Appendix A.1.

DF-2—DF-2 is more flexible than DF-1. If superiority is demonstrated on one endpoint at the interim, then the trial will continue but subsequent hypothesis testing is repeatedly conducted only for the previously non-significant endpoint until superiority is demonstrated. The stopping rule for DF-2 is formally given as follows:

At the l th analysis (l=1,…, L−1)

$$\begin{array}{ll} \text{If } Z_{1l}>c_{1l} \text{ and } Z_{2l'}>c_{2l'} \text{ for some } 1\le l'\le l, & \text{then reject } \mathrm{H}_0 \text{ and stop the trial,} \\ \text{if } Z_{2l}>c_{2l} \text{ and } Z_{1l'}>c_{1l'} \text{ for some } 1\le l'\le l, & \text{then reject } \mathrm{H}_0 \text{ and stop the trial,} \\ \text{otherwise,} & \text{continue to the } (l+1)\text{th analysis,} \end{array}$$

at the Lth analysis

$$\begin{array}{ll} \text{if } Z_{1L}>c_{1L} \text{ and } Z_{2l'}>c_{2l'} \text{ for some } 1\le l'\le L, & \text{then reject } \mathrm{H}_0, \\ \text{if } Z_{2L}>c_{2L} \text{ and } Z_{1l'}>c_{1l'} \text{ for some } 1\le l'\le L, & \text{then reject } \mathrm{H}_0, \\ \text{otherwise,} & \text{do not reject } \mathrm{H}_0. \end{array}$$

Therefore, following DF-2, the power is

$$1-\beta=\Pr\left[\left\{\cup_{l=1}^{L}A_{1l}\right\}\cap\left\{\cup_{l=1}^{L}A_{2l}\right\}|\mathrm{H}_1\right]. \quad (2)$$

Similarly as in the power (1), the power (2) can be calculated by using multivariate normal integrals. For the details, please refer to Appendix A.1.

For simplicity, consider a two-stage group-sequential design with one interim and one final analysis. The probability of rejecting the null hypothesis at the interim analysis is same for DF-1 and DF-2. The difference in power between DF-1 and DF-2 is due to whether or not the null hypothesis is rejected at the final analysis. The difference in decision-making for DF-1 and DF-2 comes from the following two situations where the interim analysis result is inconsistent with the final analysis result even the alternative hypothesis is true, i.e., (i) Endpoint 1 is statistically significant at the interim, but not at the final analysis and similarly and (ii) Endpoint 2 is statistically significant at the interim, but not at the final analysis. Thus DF-1 fails to reject the null hypothesis in both situations even if the alternative hypothesis is true, but DF-2 is able to reject the null hypothesis in both situations. However the likelihood of this scenario occurring is quite low. Thus there is little practical difference in the power and sample size determinations for DF-1 and DF-2. However, DF-2 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 although stopping measurement may also introduce an operational difficulty into the trial. This will be illustrated in Section 3.

2.3 Maximum sample size and average sample number

We discuss two sample size concepts, i.e., the maximum sample size (MSS) and the average sample number (ASN) based on DF-1 and DF-2, and the corresponding powers (1) and (2) discussed in the previous section.

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 (1) or (2) for a group-sequential design at the pre-specified δ_1, δ_2, ρ_T and ρ_C, with Fisher's information time for the interim analyses, n_l/n_L, $l = 1,\ldots, L$. To find a value of n_L, an iterative procedure is required to numerically solve for the power (1) or (2). This can be accomplished by using

a grid search to gradually increase n_L until the power under n_L exceeds the desired power, although this often requires considerable computing resources. To reduce the computational resources, the Newton–Raphson algorithm in Sugimoto et al. [14] or the basic linear interpolation algorithm in Hamasaki et al. [15] may be utilized.

The ASN is the expected sample size under a specific hypothetical reference. Given these pre-specifications, the ASN per intervention group for DF-1 is given by

$$\mathrm{ASN}=n_L\left(1+\sum_{l=1}^{L-1}\Pr\left[\left\{\overline{A}_{11}\cup\overline{A}_{21}\right\}\cap\cdots\cap\left\{\overline{A}_{1l}\cup\overline{A}_{2l}\right\}\right]\right)/L, \quad (3)$$

and for DF-2,

$$\mathrm{ASN}=n_L\left(1+\sum_{l=1}^{L-1}\Pr\left[\left\{\overline{A}_{11}\cap\cdots\cap\overline{A}_{1l}\right\}\cup\left\{\overline{A}_{21}\cap\cdots\cap\overline{A}_{2l}\right\}\right]\right)/L, \quad (4)$$

where $r_l = 1$ and $n_l = ln_1$, $l = 1,\ldots,L$. The representations for calculating ASN (3) and (4) are described in Appendix A.2.

The powers, MSS and ASN will depend on the design parameters including differences between means, the correlation structure between the endpoints, the testing procedure (e.g., O'Brien-Fleming (OF) boundary [20], Pocock (PC) boundary [21]), the number of analyses, and the information time.

3. Evaluation of the sample size

3.1 Behavior of the sample size

In this section, we evaluate the behavior of the power, MSS, and ASN as the design parameters vary. Here, without loss of generality, $\sigma_1^2=\sigma_2^2=1^2$ is chosen for simplicity, so that δ_1 and δ_2 are interpret as (standardized) effect sizes.

Figures 1 illustrates how the MSS and ASN per intervention group for DF-1 behave as a function of the number of analyses and the boundaries when effect sizes are equal and unequal, i.e., $\delta_1 = \delta_2$ and $\delta_1 \neq \delta_2$ between the two endpoints. The MSS and ASN for DF-1 and DF-2 (equally-sized groups: $r_l = 1$) were calculated to detect the joint difference in the two endpoints with the overall power of 80% at the one-sided significance level of 2.5%, where $\delta_1 = \delta_2 = 0.1$ for equal effect sizes and $\delta_1 = 0.1$ and $\delta_2 = 0.2$ for unequal effect sizes; $\sigma_1^2=\sigma_2^2=1^2$; $\rho_T = \rho_C = \rho = 0.0$, 0.3, 0.5 and 0.8. The critical values are determined by the three boundary combinations, i.e., (i) the OF for both endpoints (OF-OF), (ii) the PC for both endpoints (PC-PC) and (iii) the OF for δ_1 and the PC for δ_2 (OF-PC), with the LD alpha-spending method with equal information space.

When effect sizes are equal, the MSS for the three boundary combinations increases as the number of analyses increases and the correlation is smaller. In all of $\rho = 0$, 0.3, 0.5 and 0.8, the largest MSS is given by PC-PC, and the smallest MSS by OF-OF. On the other hand, the

ASN for the three boundary combinations decreases as the number of analyses increases and the correlation is larger. In all of $\rho = 0$, 0.3, 0.5 and 0.8, the largest ASN is given by OF-PC.

When effect sizes are unequal $\delta_1 < \delta_2$, in addition to the three boundary combinations, one more combination of (iv) the PC for δ_1 and the OF for δ_2 (PC-OF) is considered, $\delta_1 = 0.1$ and $\delta_2 = 0.2$. Similarly as seen with equal effect sizes, the MSS for the four boundary combinations increases as the number of analyses increases, but it does not change as with the correlation varies. The largest MSS is given by PC-PC and PC-OF, and the smallest MSS by OF-OF and OF-PC. On the other hand, the ASN for the four boundary combinations decreases as the number of analyses increases independently of the correlation. The largest ASN is given by OF-OF and OF-PC, and the smallest ASN by PC-PC and PC-OF. When one effect size is relatively smaller (or larger) than the other, the MSS and ASN will be driven by the smaller effect size. In this illustration, as the OF is selected for the smaller effect size and the PC for the larger, the MSS and ASN by OF-PC are approximately equal to those by OF-OF.

Figures 2 illustrates how the MSS and ASN per intervention group for DF-2 behave as a function of the number of analyses and the boundaries when effect sizes are equal $\delta_1 = \delta_2$ and unequal $\delta_1 \neq \delta_2$ between the two endpoints with the same parameter settings as in Figure 1. The MSS and ASN behaviors are similar to those observed for DF-1. The major difference between DF-1 and DF-2 is that the MSS and ASN for DF-2 are smaller than those for DF-1. They are notably smaller as the number of analyses increases, especially when the correlation is low.

If the trial was designed to detect effects on *at least one* endpoint with a prespecified ordering of endpoints, a choice 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 boundary for both endpoints [18, 19]. However, as shown in Figures 1 and 2, the selection of a different boundary has a minimal effect on the power.

3.2 Example

We provide an example to illustrate the sample size methods discussed in the previous sections. 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 disease (AD) [22]. Co-primary endpoints were cognition as assessed by the Alzheimer Disease Assessment Scale Cognitive Subscale (ADAS-Cog; 80-point scale) and functional ability as assessed by the Alzheimer 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 of 800 patients provided an overall power of 96% to detect the joint difference in the two primary endpoints between the tarenflurbil and placebo groups, by using a one-sided test at 2.5% significance level, with the standardized effect size of 0.2 for both endpoints. In addition, 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 [23]).

Table 1 displays the MSS and ASN per intervention group (equally-sized groups: $r_l = 1$) for the DF-1 and DF-2. The sample size was with an alternative hypothesis of a difference for both ADAS-Cog ($\delta_1 = 0.2$) and ADCS-ADL ($\delta_2 = 0.2$), with the overall power of 96% at the one-sided significance level of 2.5%, where $\rho = \rho_T = \rho_C = 0.0$, 0.3, 0.5, and 0.8; $L = 1$, 2, 3, 5, 8 and 10. The critical values are determined by the three boundary combinations, i.e., the OF for both endpoints (OF-OF), the PC for both endpoints (PC-PC), and OF for ADAS-Cog and the PC for ADCS-ADL (OF-PC).

Based on the selected parameters described in Green et al. [22], i.e., $L = 1$ and $\rho = 0.0$, the sample size per intervention group is calculated as 804. If four interims and one final analysis are planned (i.e., $L = 5$) with DF-1, and conservatively assuming a zero correlation between the endpoints, then the MSS is 825 for OF-OF, 945 for PC-PC and 895 for OF-PC, and the ASN is 604 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, then the MSS are 820, 810 and 785 for OF-OF; 940, 930 and 900 for PC-PC; 890, 885 and 860 for OF-PC. The ASN are 589, 574 and 543 for OF-OF; 525, 506 and 469 for PC-PC and 593, 582, and 556 for OF-PC. When comparing DF-2 to DF-1, there are no major differences in MSS and ASN for all of the boundary combinations, although DF-2 provides a slightly smaller MSS and ASN than DF-1, for PC-PC and OF-PC. However, if the endpoint is very invasive and thus stopping measurement may be ethically desirable, there is a benefit of using DF-2 as DF-2 offers the option of stopping measurement of an endpoint for which superiority has been demonstrated. For example, when four interims and one final analysis with DF-2 are planned (i.e., $L = 5$), the average total number of measurements for each intervention group are 1052, 1045, 1041 and 1021 for OF-OF; 846, 845, 841 and 831 for PC-PC; 966, 961, 958 and 944 for OF-PC, corresponding to $\rho = 0.0$, 0.3, 0.5, and 0.8. They are relatively smaller than those for DF-1 as the average total number of measurements for DF-1 are 1208, 1178, 1148 and 1086 for OF-OF; 1096, 1050, 1012 and 938 for PC-PC; 1216, 1186, 1164 and 1112 for OF-PC.

4. Sample size recalculation

Clinical trials are designed based on assumptions often constructed based on prior data. However prior data may be limited or an inaccurate indication of future data, resulting in trials that are over/under-powered. Interim analyses provides an opportunity to evaluate the accuracy of the design assumptions and potentially make design adjustments (i.e., to the sample size) if the assumptions were markedly inaccurate. The tarenflurbil trial mentioned in the previous section, failed to demonstrate a beneficial effect of tarenflurbil on both ADAS-Cog and ADCS-ADL. The observed treatment effects were smaller than the assumed effects. Group-sequential designs allow for early stopping when there is sufficient statistical evidence that the two treatments are different. However more modern adaptive designs may also allow for increases in the sample size if effects are smaller than assumed. Such adjustments must be conducted carefully for several reasons. Challenges include: (a) maintaining control of statistical error rates, (b) developing a plan to make sure that treatment effects cannot be inferred via back-calculation of a resulting change in the sample size, (c) consideration of the clinical relevance of the treatment effects, and (d) practical concerns such as an increase in cost and the challenge of accruing more trial participants. In

this section, we discuss sample size recalculation based on the observed intervention's effects at an interim analysis with a focus on control of statistical error rates.

4.1 Test statistics and conditional power

Consider that the maximum sample size is recalculated to $n_L^{'}$ based on the interim data at the R th analysis. Suppose that $n_L^{'}$ is subject to $n_R < n_L^{'} \leq \lambda n_L$, where λ is a pre-specified constant for the maximum allowable sample size. For simplicity, assume a common correlation between the treatment groups, i.e., $\rho_T = \rho_C = \rho$. Let $(\tilde{\delta_1}, \tilde{\delta_2})$ and let (δ_1^*, δ_2^*) be the mean differences used for planned sample size and for recalculated sample size, respectively.

Here we consider the Cui-Hung-Wang (CHW) statistics [24] for sample size recalculation in group-sequential designs with two co-primary endpoints to preserve the overall Type I error rate at a pre-specified alpha level even when the sample size is increased and conventional test statistics are used. The CHW statistics are,

$$Z_{km}^{'} = \sqrt{\frac{n_R}{n_m}} Z_{kR} + \sqrt{\frac{n_m - n_R}{n_m}} \frac{\sum_{i=n_R+1}^{n_m^{'}} Y_{Tki} - \sum_{j=n_R+1}^{n_m^{'}} Y_{Ckj}}{\sqrt{2(n_m^{'} - n_R)}},$$

where $n_m^{'} = (n_m - n_R)(n_L^{'} - n_R)/(n_L - n_R) + n_R$ and $r_R = r_m = 1$ (k=1,2; R=1,…, L−1; m=R +1,…, L). The same critical values utilized for the case without sample size recalculation are used.

The sample size is increased or decreased when the conditional power evaluated at the R th analysis is lower or higher than the desired power 1−β. Under the planned maximum sample size and a given observed value of (Z_{1R}, Z_{2R}), for DF-1, the conditional power is defined by

$$CP = \Pr\left[\cup_{m=R+1}^{L} \{A_{1m} \cap A_{2m}\} \,|a_{1R}, a_{2R}\right] \quad (5)$$

if $Z_{1l} \leq c_{1l}$ or $Z_{2l} \leq c_{2l}$ for all l = 1,…, R, where (a_{1R}, a_{2R}) is a given observed value of (Z_{1R}, Z_{2R}). On the other hand, the conditional power for DF-2 is given by

$$CP = \begin{cases} \Pr\left[\cup_{m=R+1}^{L} A_{1m} | a_{1R}, a_{2l'}\right] \\ \text{if } Z_{1l} \leq c_{1l} \text{ for all } l=1,\ldots,R \text{ and } Z_{2l'} > c_{2l'} \text{ for some } l'=1,\ldots,R, \\ \Pr\left[\cup_{m=R+1}^{L} A_{2m} | a_{2R}, a_{1l'}\right] \\ \text{if } Z_{2l} \leq c_{2l} \text{ for all } l=1,\ldots,R \text{ and } Z_{2l'} > c_{1l'} \text{ for some } l'=1,\ldots,R, \\ \Pr\left[\left\{\cup_{m=R+1}^{L} A_{1m}\right\} \cap \left\{\cup_{m=R+1}^{L} A_{2,m}\right\} | a_{1R}, a_{2R}\right] \\ \text{if } Z_{1l} \leq c_{1l} \text{ and } Z_{2l} \leq c_{2l} \text{ for all } l=1,\ldots,R. \end{cases} \quad (6)$$

The detailed calculation of the conditional powers for DF-1 and DF-2 are provided in the Appendix A.3. Since (δ_1, δ_2) is unknown, it is customary to substitute (δ_1^*, δ_2^*), the estimated

mean differences at the R th analysis ($\hat{\delta}_{1R}$, $\hat{\delta}_{2R}$) or the assumed mean differences during trial planning ($\tilde{\delta}_1$, $\tilde{\delta}_2$). We consider the conditional power based on $(\delta_1^*, \delta_2^*)=(\hat{\delta}_{1R}, \hat{\delta}_{2R})$, which allows evaluation of behavior of power independent of (δ_1, δ_2).

When recalculating the sample size, three options are possible: (i) only allowing an increase in the sample size, (ii) only allowing a decrease in the sample size, and (iii) allowing an increase or decrease in sample size. For all the cases, we assign Z'_{km} and n'_m instead of Z_{km} and n_m in the conditional powers (5) and (6) for the conditional power with sample size recalculation. Consider the rule for determining the recalculated sample size n'_L, when the sample size may be increased only, which is:

$$n'_L = \begin{cases} n_L, & \text{if } CP \geq 1-\beta \text{ or } \min(\hat{\delta}_{1R}, \hat{\delta}_{2R}) \leq 0, \\ \min\left(n''_L, \lambda n_L\right), & \text{otherwise,} \end{cases}$$

where n''_L is the smallest integer $n'_L (>n_R)$, where the conditional power achieves the desired power $1-\beta$. When the sample size may be decreased only, the recalculated sample size n'_L is:

$$n'_L = \begin{cases} n''_L, & \text{if } CP > 1-\beta, \\ n_L, & \text{otherwise.} \end{cases}$$

When the sample size may be increased or decreased, then the recalculated sample size n'_L is:

$$n'_L = \begin{cases} n''_L, & \text{if } CP > 1-\beta, \\ n_L, & \text{if } CP = 1-\beta \text{ or } \min(\hat{\delta}_{1R}, \hat{\delta}_{2R}) \leq 0, \\ \min\left(n''_L, \lambda n_L\right), & \text{otherwise.} \end{cases}$$

4.2 Simulation study

A simulation study was performed to evaluate the impact of sample size recalculation based on DF-1 and DF-2 on the power and Type I error rate. We consider group-sequential designs with a single interim, i.e., one interim and one final analyses, and with multiple interims, i.e., three interims and one final analyses. In addition, we discuss the three options of: (i) only decreasing the sample size, (ii) only increasing the sample size, and (iii) increasing or decreasing the sample size, based upon the observed intervention's effect. The planned MSS per intervention group is calculated to detect the joint difference for two endpoints with the overall power of 80% at the one-sided significance level of 2.5%, where ($\tilde{\delta}_1$, $\tilde{\delta}_2$) = (0.2, 0.2), $\sigma_1^2=\sigma_2^2=1^2$ and the correlation is assumed to be known correlation at the design stage, i.e., ρ = 0.0, 0.3, 0.5, and 0.8. For the evaluation of the Type I error rate, the two pairs of the mean differences (δ_1, δ_2) = (0.0, 0.0) and (0.0, 0.2) are considered under H_0. For the designs with a single interim, the timing of the interim analysis for sample size recalculation is evaluated at 0.25, 0.50 and 0.75 of information time. For designs with multiple interims, one sample size

recalculation is considered and the timing is evaluated at the 1st, 2nd and 3rd of interim analysis. The critical values are determined by the OF boundary for both endpoints with the LD alpha-spending method, with equal information space. The upper limit of the recalculated sample size is set to $n_2' = \lambda n_2$ with λ = 1.5. The number of replications for the simulation is set to 1,000,000 for the evaluation of the Type I error rate and 100,000 replications for the power. These number of replications for the simulation was determined based on the precision, where a sample size of 1,000,000 provides a two-sided 95% confidence interval with a width equal to 0.001 when the proportion is 0.025, and a total number of replications of 100,000 provides a two-sided 95% confidence interval with a width equal to 0.005 when the proportion is 0.80.

Suppose that the sample size recalculation is based on the interim estimates of (δ_1, δ_2). Note that the value of correlation assumed at the design stage is retained for the sample size recalculation, i.e., without updating based on observed correlation at the interim as the correlation is a nuisance parameter in hypothesis testing. All results are summarized in Tables S1 to S4 in the Supplemental Data. As there are no significant differences between DF-1 and DF-2 with respect to the Type I error rates and empirical powers, we limit discussion to the behavior of the Type I error rates and power for DF-1.

Figure 3 illustrates how the Type I error rates and powers behave as a function of the correlation, the timing of the interim analysis for sample size recalculation, and the sample size recalculation options for DF-1 in the single-interim case. In all three recalculation options, the Type I error rates increase as the correlation increases, but they are not exceed the targeted 2.5%. There is no practical difference in the behavior of the Type I error rates depending on the timing of the interim analysis for sample size recalculation. On the other hand, for the behavior of the power, when only allowing an increase in the sample size, the empirical powers are higher than the desired power of 80% in all of the three timings of sample size recalculation, although the power is slightly decreased with higher correlation. When allowing an increase or a decrease in the sample size, if the timing for sample size recalculation is at 25% information time, then the empirical power is lower than the desired power of 80%, especially with higher correlation. However, if the timing for sample size recalculation is 50% or 75%, then the empirical powers are higher than in all three timings of sample size recalculation. When only allowing a decrease in the sample size, if the timing for the sample size recalculation is at 25% or 50% information time, then the empirical powers are always lower than the desired power, especially with higher correlation. If the timing for sample size recalculation is at 75% information time, then the empirical power is almost achieved at the desired power of 80%.

Figure 4 illustrates how the Type I error rates and powers behave as a function of the correlation, the timing of the interim analysis for sample size recalculation, and the sample size recalculation options for DF-1, in the multiple-interim case. The results are similar to those in the single-interim case; when only allowing an increase in the sample size, compared with the desired power of 80%, the empirical powers are improved in all of the three timings for the sample size recalculation, but the empirical power is much lower than the desired power if the sample size recalculation is conducted early in the study, especially when allowing a decrease in the sample size.

These results suggest incorporating the uncertainty of the estimates at the interim into the sample size recalculation is important. The power is much lower than desired power if the sample size recalculation is conducted early in the study, especially when allowing for a decrease in the sample size.

5. Summary and discussion

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.

We discuss group-sequential designs with co-primary endpoints. We derive the power and sample size methods under two decision-making frameworks: (1) designing the trial to detect the test intervention's superiority for the two endpoints simultaneously (i.e., at the same interim timepoint of the trial) (DF-1), and (2) designing the trial to detect superiority for the two endpoints at any interim timepoint (i.e., not necessarily simultaneously) (DF-2). The former is simpler while the latter 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 evaluate the behavior of sample size with varying design elements and provide an example to illustrate the methods. We also discuss sample size recalculation using CHW statistics and evaluate the impact on the power and Type I error rate. Although DF-2 will provide a slightly smaller sample size than DF-1, there is modest difference between two. However, if the endpoint is very invasive and thus stopping measurement may be ethically desirable, there is a benefit of using DF-2 as DF-2 offers the option of stopping measurement of an endpoint for which superiority has been demonstrated. However stopping measurement on one endpoint could also create operational challenges in study conduct and patient monitoring. The timing of the sample size recalculation should also be carefully considered as the power does not reach desired levels if the sample size recalculation is done early in the study when considering a decrease in the sample size.

There are other practical issues and extensions to consider when designing a group-sequential clinical trial with co-primary endpoints. They include: how the value of correlation should be selected at the planning and interim, evaluating futility or efficacy and futility simultaneously, other endpoint scales, and other inferential goals. We discuss each of these issues.

There are two important questions regarding the choice of the correlation in sample size calculations. One is whether the observed correlation from external or pilot data should be utilized or whether correlation is assumed to be zero. The other is whether the sample size should be recalculated based on the observed correlation at the interim. Incorporating the observed correlation at the planning or interim may affect the Type I error rate and power. Our experience suggests that when standardized effect sizes are unequal between the endpoints, the power is not improved with higher correlation. With unequal standardized effect sizes, incorporating the correlation into the sample size calculation at planning or interim may have no advantage [25, 26]. Further investigation will be required to assess how the choice of the correlation impacts the operation characteristics of the design.

Since the main objective of the paper is to provide the fundamental foundation in group-sequential designs for co-primary endpoints, our discussion is restricted to a superiority clinical trial comparing two interventions based on two continuous endpoints. The study design allows for early stopping when larger intervention differences are observed, i.e., rejecting a null hypothesis only. However, this work provides a foundation for designing clinical trials with other design features. In addition to this fundamental situation, the method discussed here can be straightforwardly extended to other situations such as evaluating futility (rejecting the alternative hypothesis) or evaluating both efficacy and futility.

Time-to-event outcomes are common in oncology, cardiovascular and infectious disease clinical trials. The method for continuous endpoints described in the paper may not be directly extended to time-to-event endpoints. When considering a trial with two time-to-event outcomes as co-primary with a plan for using the logrank test to compare two interventions in a group-sequential design, information for the two endpoints may accumulate at different rates. This creates challenges when designing trials, i.e., the amount of information for the endpoints may be different at any particular interim timepoint of the trial. Further investigation is required to assess this issue.

Although our primary interest is *co-primary* endpoints, these results provide a fundamental foundation to other inferential goals, e.g., designing a trial to detect an effect on *at least one* endpoint. Many authors have proposed methods for the *at least one* endpoint goal in fixed sample size designs, e.g., a weighted Bonferroni procedure, the prospective alpha allocation scheme method, the adaptive alpha allocation approach, the Bonferroni-type parametric procedure, and the fallback-type parametric procedure (e.g., see Dmitrienko et al. [4], Moyé [27] and Moyé and Baranuik [28]). In addition, several authors have discussed an extension of methods to the group-sequential designs with an inferential goal of *at least one* endpoint [29–32]. For example, Tang and Geller [30] discuss a method based on closed testing procedures and Tamhane et al. [31, 32] discuss sample size methods in two-stage group-sequential designs based on the gatekeeping procedures with hierarchically ordered multiple endpoints.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful to the two anonymous referees, and the associate editor for their valuable suggestions and helpful comments that improved the content and presentation of the paper. The authors thank Dr. H.M. James Hung and Dr. Sue-Jane Wang for encouraging us in this research with their helpful advice. Research reported in this publication was supported by JSPS KAKENHI under Grant Number 23500348 and the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number UM1AI104681.

References

1. Committee for Medicinal Products for Human Use (CHMP). Guideline on Medicinal Products for the Treatment Alzheimer's Disease and Other Dementias (CPMP/EWP/553/95 Rev.1). EMEA; London: 2008.
2. Offen W, Chuang-Stein C, Dmitrienko A, Littman G, Maca J, Meyerson L, Muirhead R, Stryszak 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. Multiple co-primary endpoints: medical and statistical solutions. Drug Information Journal. 2007; 41:31–46.10.1177/009286150704100105
3. Hung HMJ, Wang SJ. Some controversial multiple testing problems in regulatory applications. Journal of Biopharmaceutical Statistics. 2009; 19:1–11.10.1080/10543400802541693 [PubMed: 19127460]
4. Dmitrienko, A.; Tamhane, AC.; Bretz, F. Multiple Testing Problems in Pharmaceutical Statistics. Chapman & Hall; Boca Raton, FL: 2010.
5. Xiong C, Yu K, Gao F, Yan Y, Zhang Z. Power and sample size for clinical trials when efficacy is required in multiple endpoints: application to an Alzheimer's treatment trial. Clinical Trials. 2005; 2:387–393.10.1191/1740774505cn112oa [PubMed: 16317808]
6. Sozu T, Kanou T, Hamada C, Yoshimura I. Power and sample size calculations in clinical trials with multiple primary variables. Japanese Journal of Biometrics. 2006; 27:83–96.10.5691/jjb.27.83
7. Chuang-Stein C, Stryszak P, Dmitrienko A, Offen W. Challenge of multiple co-primary endpoints: a new approach. Statistics in Medicine. 2007; 26:1181–1192.10.1002/sim.2604 [PubMed: 16927251]
8. Eaton ML, Muirhead RJ. On multiple endpoints testing problem. Journal of Statistical Planning & Inference. 2007; 137:3416–3429.10.1016/j.jspi.2007.03.021
9. Senn S, Bretz F. Power and sample size when multiple endpoints are considered. Pharmaceutical Statistics. 2007; 6:161–170.10.1002/pst.301 [PubMed: 17674404]
10. Kordzakhia G, Siddiqui O, Huque MF. Method of balanced adjustment in testing co-primary endpoints. Statistics in Medicine. 2010; 29:2055–2066.10.1002/sim.3950 [PubMed: 20683896]
11. Sozu T, Sugimoto T, Hamasaki T. Sample size determination in clinical trials with multiple co-primary binary endpoints. Statistics in Medicine. 2010; 29:2169–2179.10.1002/sim.3972 [PubMed: 20687162]
12. Sozu T, Sugimoto T, Hamasaki T. Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints. Journal of Biopharmaceutical Statistics. 2011; 21:1–19.10.1080/10543406.2011.551329 [PubMed: 21191850]
13. Julious S, Mclntyre NE. Sample sizes for trials involving multiple correlated must-win comparisons. Pharmaceutical Statistics. 2012; 11:177–185.10.1002/pst.515 [PubMed: 22383136]
14. Sugimoto T, Sozu T, Hamasaki T. A convenient formula for sample size calculations in clinical trials with multiple co-primary continuous endpoints. Pharmaceutical Statistics. 2012; 11:118–128.10.1002/pst.505 [PubMed: 22415870]
15. Hamasaki T, Sugimoto T, Evans SR, Sozu T. Sample size determination for clinical trials with co-primary outcomes: exponential event times. Pharmaceutical Statistics. 2013; 12:28–34.10.1002/pst.1545 [PubMed: 23081932]
16. Sugimoto T, Sozu T, Hamasaki T, Evans SR. A logrank test-based method for sizing clinical trials with two co-primary time-to-event endpoints. Biostatistics. 2013; 14:409–421.10.1093/biostatistics/kxs057 [PubMed: 23307913]
17. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika. 1983; 70:659–663.10.1093/biomet/70.3.659

18. Glimm E, Maurer W, Bretz F. Hierarchical testing of multiple endpoints in group-sequential trials. Statistics in Medicine. 2010; 29:219–228.10.1002/sim.3748 [PubMed: 19827011]
19. Tamhane AC, Mehta CR, Liu L. Testing a primary and secondary endpoint in a group sequential design. Biometrics. 2010; 66:1174–1184.10.1111/j.1541-0420.2010.01402.x [PubMed: 20337631]
20. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics. 1979; 35:549–556.10.2307/2530245 [PubMed: 497341]
21. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika. 1977; 64:191–199.10.1093/biomet/64.2.191
22. Green RC, Schneider LS, Amato DA, Beelen AP, Wilcock G, Swabb EA, Zavitz KH. for the Tarenflurbil Phase 3 Study Group. Effect of tarenflurbil on cognitive decline and activities of daily living in patients with mild Alzheimer disease: A randomized controlled trial. Journal of the American Medical Association. 2009; 302:2557–2564.10.1001/jama.2009.1866 [PubMed: 20009055]
23. Doraiswamy PM, Bieber F, Kaiser L, Krishnan KR, Reuning-Scherer J, Gulanski B. The Alzheimer's disease assessment scale: patterns and predictors of baseline cognitive performance in multicenter Alzheimer's disease trials. Neurology. 1997; 48:1511–1517.10.1212/WNL.48.6.1511 [PubMed: 9191757]
24. Cui L, Hung HMJ, Wang SJ. Modification of sample size in group sequential clinical trials. Biometrics. 1999; 55:853–857.10.1111/j.0006-341X.1999.00853.x [PubMed: 11315017]
25. Asakura, K.; Hayashi, K.; Sugimoto, T.; Sozu, T.; Hamasaki, T. Sample size evaluation in group sequential designs for clinical trials with two continuous endpoints as co-primary contrasts. Joint Statistical Meetings; Montreal, Quebec, Canada. August 3–8, 2013; 2013.
26. Hamasaki, T.; Asakura, K.; Sugimoto, T.; Evans, SR. Sample size modification in group-sequential clinical trials with two co-primary endpoints. Proceedings of Joint Meeting of the IASC Satellite Conference and 8th Conference of the Asian Regional Section of the IASC; Seoul, Korea. August 21–14, 2013; p. 311-317.
27. Moyé, LA. Multiple Analyses in Clinical Trials. Springer; New York, NY: 2003.
28. Moyé LA, Baraniuk S. Dependence, hyper-dependence and hypothesis testing in clinical trials. Contemporary Clinical Trials. 2013; 28:68–78.10.1016/j.cct.2006.05.010 [PubMed: 16857430]
29. Jennison C, Turnbull BW. Group sequential tests for bivariate response: interim analyses of clinical trials with both efficacy and safety. Biometrics. 1993; 49:741–752. [PubMed: 8241370]
30. Tang DI, Geller NL. Closed testing procedures for group sequential clinical trials with multiple endpoints. Biometrics. 1999; 55:1188–1192.10.1111/j.0006-341X.1999.01188.x [PubMed: 11315066]
31. Tamhane AC, Wu Y, Mehta C. Adaptive extensions of a two-stage group sequential procedure for testing primary and secondary endpoints (I): unknown correlation between the endpoints. Statistics in Medicine. 2012; 31:2027–2040.10.1002/sim.5372 [PubMed: 22729929]
32. Tamhane AC, Wu Y, Mehta C. Adaptive extensions of a two-stage group sequential procedure for testing primary and secondary endpoints (II): sample size re-estimation. Statistics in Medicine. 2012; 31:2041–2054.10.1002/sim.5377 [PubMed: 22733687]
33. Genz, A.; Bretz, F. Computation of Multivariate Normal and t Probabilities. Springer Verlag; Berlin: 2009.

Appendix

A.1 Power calculation

The power (1) for DF-1 can be calculated by partitioning the set in (1) into mutually exclusive subsets and taking the sum of their probabilities as follows:

$$1-\beta=\Pr\left[\cup_{l=1}^{L}\{A_{1l}\cap A_{2l}\}\,|\mathrm{H}_1\right]$$
$$=\Pr\left[A_{11}\cap A_{21}|\mathrm{H}_1\right]+\sum_{l=2}^{L}\Pr\left[\cap_{l'=1}^{l-1}\left\{\overline{A}_{1l'}\cup\overline{A}_{2l'}\right\}\cap\{A_{1l}\cap A_{2l}\}\,|\mathrm{H}_1\right], \quad \text{(A1)}$$

where $A_{kl}=\{Z_{kl}>c_{kl}\}$ and $\bar{A}_{kl}=\{Z_{kl}\le c_{kl}\}$ (k = 1,2; l = 1,…, L). The probability of $\{\bar{A}_{1l'}\cup\bar{A}_{2l'}\}$ can be written as $\Pr[\overline{A}_{1l'}\cup\overline{A}_{2l'}]=\Pr[\tilde{A}_{l'}^{1}]+\Pr[\tilde{A}_{l'}^{2}]+\Pr[\tilde{A}_{l'}^{3}]$, where $A_{l'}^{1}=\{\overline{A}_{1l'}\cap A_{2l'}\}$, $\tilde{A}_{l'}^{2}=\{A_{1l'}\cap\overline{A}_{2l'}\}$ and $\tilde{A}_{l'}^{3}=\{\overline{A}_{1l'}\cap\overline{A}_{2l'}\}$ $(l'=1,\ldots,L-1)$. Similarly, the probability of the union of $\{\bar{A}_{1l'}\cup\bar{A}_{2l'}\}$ can be written by the sum of the probabilities of the unions composed of $\tilde{A}_{l'}^{1}$, $\tilde{A}_{l'}^{2}$ and $\tilde{A}_{l'}^{3}$. Then, the second term of the right-hand side in (A1) can be rewritten as

$$\sum_{l=2}^{L}\Pr\left[\cap_{l'=1}^{l-1}\left\{\overline{A}_{1l'}\cup\overline{A}_{2l'}\right\}\cap\{A_{1l}\cap A_{2l}\}\,|\mathrm{H}_1\right]=\sum_{l=2}^{L}\left(\sum_{h_1=1}^{3}\cdots\sum_{h_{l-1}=1}^{3}\Pr\left[\left\{\cap_{l'=1}^{l-1}\tilde{A}_{l'}^{h_{l'}}\right\}\cap\{A_{1l}\cap A_{2l}\}\,|\mathrm{H}_1\right]\right).$$

The probability of $\tilde{A}_{l'}^{1}$ is calculated by a bivariate normal integral as follows:

$$\Pr\left[\tilde{A}_{l'}^{1}\right]=\int_{-\infty}^{c_{1l'}}\int_{c_{2l'}}^{\infty}f_2\left(z_{1l'},z_{2l'}\right)dz_{2l'}dz_{1l'},$$

where $f_2(z_{1l'}, z_{2l'})$ is the density function of the joint distribution of $(Z_{1l'}, Z_{2l'})$ with the means and the covariance matrix given in Section 2.1. The probabilities of $\tilde{A}_{l'}^{2}$, $\tilde{A}_{l'}^{3}$ and $\{A_{1l'}\cap A_{2l'}\}$ are calculated similarly. Then the probability of the union composed of $\tilde{A}_{l'}^{1}$, $\tilde{A}_{l'}^{2}$, $\tilde{A}_{l'}^{3}$ and $\{A_{1l'}\cap A_{2l'}\}$ is calculated by a multivariate normal integral and the power is the sum of $(3^L-1)/2$ multivariate normal integrals. For details of the computation related to multivariate normal, please see Genz and Bretz [24].

For illustration, we provide the case of L = 2 and $r = r_1 = r_2$. In this case, the power can be rewritten as

$$1-\beta=\Pr\left[A_{11}\cap A_{21}|\mathrm{H}_1\right]+\sum_{h_1=1}^{3}\Pr\left[\tilde{A}_{1}^{h_1}\cap\{A_{12}\cap A_{22}\}\,|\mathrm{H}_1\right]$$
$$=\int_{c_{11}}^{\infty}\int_{c_{21}}^{\infty}f_2\left(z_{11},z_{21}\right)dz_{21}dz_{11}+\int_{-\infty}^{c_{11}}\int_{c_{21}}^{\infty}\int_{c_{12}}^{\infty}\int_{c_{22}}^{\infty}f_4\left(z_{11},z_{21},z_{12},z_{22}\right)dz_{22}dz_{12}dz_{21}dz_{11}$$
$$+\int_{c_{11}}^{\infty}\int_{-\infty}^{c_{21}}\int_{c_{12}}^{\infty}\int_{c_{22}}^{\infty}f_4\left(z_{11},z_{21},z_{12},z_{22}\right)dz_{22}dz_{12}dz_{21}dz_{11}$$
$$+\int_{-\infty}^{c_{11}}\int_{-\infty}^{c_{21}}\int_{c_{12}}^{\infty}\int_{c_{22}}^{\infty}f_4\left(z_{11},z_{21},z_{12},z_{22}\right)dz_{22}dz_{12}dz_{21}dz_{11},$$

where $f_2(z_{11}, z_{21})$ is the density function of the bivariate normal distribution of $\boldsymbol{Z}_2=(Z_{11}, Z_{21})^{\mathrm{T}}$, which is given by

$$f_2(\boldsymbol{Z}_2)=\frac{1}{2\pi\left|\sum_2\right|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{Z}_2-\boldsymbol{\mu}_2)^{\mathrm{T}}\sum_2{}^{-1}(\boldsymbol{Z}_2-\boldsymbol{\mu}_2)\right],\ -\infty<z_{11},z_{21}<\infty$$

with mean vector $\boldsymbol{\mu}_2 = \sqrt{rn_1/(1+r)}(\delta_1/\sigma_1, \delta_2/\sigma_2)^{\mathrm{T}}$ and correlation matrix

$$\sum\nolimits_2 = \begin{pmatrix} 1^2 & (r\rho_{\mathrm{T}}+\rho_{\mathrm{C}})/(1+r) \\ (r\rho_{\mathrm{T}}+\rho_{\mathrm{C}})/(1+r) & 1^2 \end{pmatrix}.$$

and $f_4(z_{11}, z_{21}, z_{12}, z_{22})$ is the density function of the tetra-variate normal distribution of $\boldsymbol{Z}_4 = (Z_{11}, Z_{21}, Z_{12}, Z_{22})^{\mathrm{T}}$ given by

$$f_4(\boldsymbol{Z}_4) = \frac{1}{(2\pi)^2 |\sum_4|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{Z}_4-\boldsymbol{\mu}_4)^{\mathrm{T}} \sum\nolimits_4{}^{-1} (\boldsymbol{Z}_4-\boldsymbol{\mu}_4)\right], -\infty < z_{11}, z_{21}, z_{12}, z_{22} < \infty$$

with mean vector $\boldsymbol{\mu}_4 = \sqrt{(1+r)/r}(\sqrt{n_1}\delta_1/\sigma_1, \sqrt{n_1}\delta_2/\sigma_2, \sqrt{n_2}\delta_1/\sigma_1, \sqrt{n_2}\delta_2/\sigma_2)^{\mathrm{T}}$ and correlation matrix

$$\sum\nolimits_4 = \begin{pmatrix} \sum_2 & \sqrt{n_1/n_2}\sum_2 \\ \sqrt{n_1/n_2}\sum_2 & \sum_2 \end{pmatrix},$$

where $\boldsymbol{\Sigma}_4$ is positive definite matrix under $|\rho_{\mathrm{T}}|, |\rho_{\mathrm{C}}| < 1$ and $n_1 \neq n_2$ as $|\boldsymbol{\Sigma}_4| = |\boldsymbol{\Sigma}_2|^2 (1-n_1/n_2)^2$.

The power (2) for DF-2 can be calculated from two L–variate normal integrals and a $2L$–variate normal integral.

$$\begin{aligned} 1-\beta &= \Pr\left[\left\{\cup_{l=1}^{L} A_{1l}\right\} \cap \left\{\cup_{l=1}^{L} A_{2l}\right\} | \mathrm{H}_1\right] \\ &= 1 - \left(\Pr\left[\cap_{l=1}^{L} \overline{A}_{1l} | \mathrm{H}_1\right] + \Pr\left[\cap_{l=1}^{L} \overline{A}_{2l} | \mathrm{H}_1\right] - \Pr\left[\cap_{l=1}^{L} \left\{\overline{A}_{1l} \cap \overline{A}_{2l}\right\} | \mathrm{H}_1\right]\right). \end{aligned}$$

The power can be calculated similarly as discussed in the power (1) for DF-1.

A.2 ASN calculation

The ASN (3) for DF-1 can be calculated by the sum of multivariate normal integrals

$$\begin{aligned} \mathrm{ASN} &= n_L \left(1 + \sum_{l=1}^{L-1} \Pr\left[\left\{\overline{A}_{11} \cup \overline{A}_{21}\right\} \cap \cdots \cap \left\{\overline{A}_{1l} \cup \overline{A}_{2l}\right\}\right]\right)/L \\ &= n_L \left\{1 + \sum_{l=1}^{L-1} \left(\sum_{h_1=1}^{3} \cdots \sum_{h_l=1}^{3} \Pr\left[\cap_{l'=1}^{l} \tilde{A}_{l'}^{h_{l'}}\right]\right)\right\}/L. \end{aligned}$$

Similarly, the ASN (4) for DF-2 can be calculated by

$$\begin{aligned}\mathrm{ASN}&=n_L\left(1+\sum_{l=1}^{L-1}\Pr\left[\left\{\overline{A}_{11}\cap\cdots\cap\overline{A}_{1l}\right\}\cup\left\{\overline{A}_{21}\cap\cdots\cap\overline{A}_{2l}\right\}\right]\right)/L\\&=n_L\left\{1+\sum_{l=1}^{L-1}\left(\Pr\left[\cap_{l'=1}^{l}\overline{A}_{1l'}\right]+\Pr\left[\cap_{l'=1}^{l}\overline{A}_{2l'}\right]-\Pr\left[\cap_{l'=1}^{l}\left\{\overline{A}_{1l'}\cap\overline{A}_{2l'}\right\}\right]\right)\right\}/L.\end{aligned}$$

A.3 Conditional Power

The conditional power (5) for DF-1 is described by

$$\begin{aligned}CP&=\Pr\left[\cup_{m=R+1}^{L}\left\{A_{1m}\cap A_{2m}\right\}|a_{1R},a_{2R}\right]\\&=\Pr\left[A_{1,R+1}\cap A_{2,R+1}|a_{1R},a_{2R}\right]+\sum_{m=R+2}^{L}\Pr\left[\cap_{m'=R+1}^{m-1}\left\{\overline{A}_{1m'}\cup\overline{A}_{2m'}\right\}\cap\left\{A_{1m}\cap A_{2m}\right\}|a_{1R},a_{2R}\right],\end{aligned}\tag{A2}$$

if $Z_{1l}\le c_{1l}$ or $Z_{2l}\le c_{2l}$ for all $l=1,\ldots,R$, where $A_{km}=\{Z_{km}>c_{km}\}$, $\bar{A}_{km}=\{Z_{km}\le c_{km}\}$ ($k=1,2$; $m=R+1,\ldots,L$) and (a_{1R}, a_{2R}) is a given observed value of (Z_{1R}, Z_{2R}). The second term of the right-hand side in (A2) can be calculated in a similar way to that for the power calculation (Appendix A.1.). The conditional distribution of $(Z_{1,R+1}, Z_{2,R+1},\ldots, Z_{1L}, Z_{2L}|a_{1R}, a_{2R})$ is a multivariate normal with their means

$\mathrm{E}[Z_{km}|a_{1R},a_{2R}]=\sqrt{n_m/2}\delta_k+\sqrt{n_R/n_m}(a_{kR}-\sqrt{n_R/2}\delta_k)$ and covariance given by $\mathrm{cov}[Z_{km},Z_{k'm'}|a_{1R},a_{2R}]=(n_{m'}-n_R)/\sqrt{n_m n_{m'}}$ if $k=k'$; $(n_{m'}-n_R)\rho/\sqrt{n_m n_{m'}}$ if $k\ne k'$, where $m'\le m=R+1,\ldots,L$. For DF-2, the conditional power (6) can be described as

$$CP=\begin{cases}\Pr\left[\cup_{m=R+1}^{L}A_{1m}|a_{1R},a_{2l'}\right]=1-\Pr\left[\cap_{m=R+1}^{L}\overline{A}_{1m}|a_{1R},a_{2l'}\right]\\ \text{if } Z_{1l}\le c_{1l}\text{ for all } l=1,\ldots,R \text{ and } Z_{2l'}>c_{2l'}\text{ for some } l'=1,\ldots,R,\\ \Pr\left[\cup_{m=R+1}^{L}A_{2m}|a_{2R},a_{1l'}\right]=1-\Pr\left[\cap_{m=R+1}^{L}\overline{A}_{2m}|a_{2R},a_{1l'}\right]\\ \text{if } Z_{2l}\le c_{2l}\text{ for all } l=1,\ldots,R \text{ and } Z_{1l'}>c_{1l'}\text{ for some } l'=1,\ldots,R,\\ \Pr\left[\left\{\cup_{m=R+1}^{L}A_{1m}\right\}\cap\left\{\cup_{m=R+1}^{L}A_{2m}\right\}|a_{1R},a_{2R}\right]\\ =1-\Pr\left[\cap_{m=R+1}^{L}\overline{A}_{1m}|a_{1R},a_{2R}\right]-\Pr\left[\cap_{m=R+1}^{L}\overline{A}_{2m}|a_{1R},a_{2R}\right]\\ +\Pr\left[\cap_{m=R+1}^{L}\left\{\overline{A}_{1m}\cap\overline{A}_{2m}\right\}|a_{1R},a_{2R}\right]\\ \text{if } Z_{1l}\le c_{1l}\text{ and } Z_{2l}\le c_{2l}\text{ for all } l=1,\ldots,R,\end{cases}$$

and calculated similarly as discussed in the power for DF-2 (Appendix A.1.).

When $R=L-1$, the conditional power for DF-1 can be rewritten as

$$CP=\Pr\left[A_{1L}\cap A_{2L}|a_{1R},a_{2R}\right]=\Phi_2(-c_1^*,-c_2^*|\rho),$$

where $\Phi_2(\cdot,\cdot,|\rho)$ is the cumulative distribution function of the standard bivariate normal distribution with the correlation ρ, and $c_1^*=(c_{1L}-a_{1R}\sqrt{t})/\sqrt{1-t}-\delta_1\sqrt{n_L-n_R}/\sqrt{2}$ and $c_2^*=(c_{2L}-a_{2R}\sqrt{t})/\sqrt{1-t}-\delta_2\sqrt{n_L-n_R}/\sqrt{2}$ with $t=n_R/n_L$. For DF-2, the conditional power can be rewritten as

$$CP=\begin{cases} \Pr\left[A_{1L}|a_{1R},a_{2l'}\right]=1-\Phi(c_1^*) \\ \text{if } Z_{1l}\le c_{1l} \text{ for all } l=1,\ldots,R \text{ and } Z_{2l'}>c_{2l'} \text{ for some } l'=1,\ldots,R, \\ \Pr\left[A_{2L}|a_{2R},a_{1l'}\right]=1-\Phi(c_2^*) \\ \text{if } Z_{2l}\le c_{2l} \text{ for all } l=1,\ldots,R \text{ and } Z_{1l'}>c_{1l'} \text{ for some } l'=1,\ldots,R, \\ \Pr\left[A_{1L}\cap A_{2L}|a_{1R},a_{2R}\right]=\Phi_2(-c_1^*,-c_2^*|\rho) \\ \text{if } Z_{1l}\le c_{1l} \text{ and } Z_{2l}\le c_{2l} \text{ for all } l=1,\ldots,R, \end{cases}$$

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

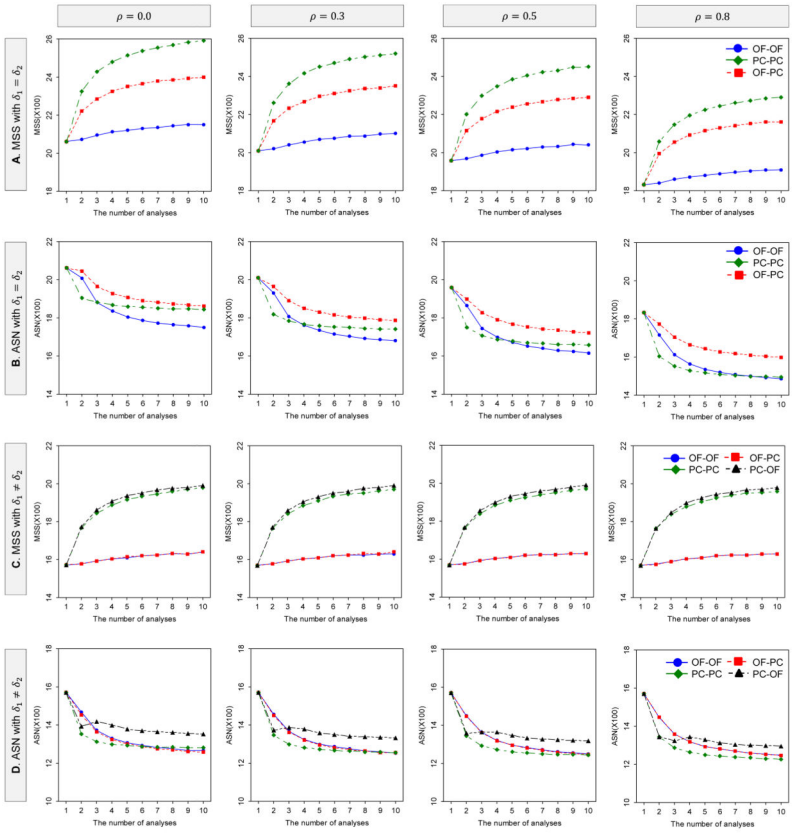

Figure 1.

Behavior of MSS and ASN for DF-1 as the number of analyses and boundaries vary. The MSS and ASN per intervention group (equally-sized groups: r_l=1) were calculated to detect the joint difference in the two endpoints with the overall power of 80% at the one-sided significance level of 2.5%, where $\delta_1 = \delta_2 = 0.1$ for A and B, and $\delta_1 = 0.1$ and $\delta_2 = 0.2$ for C and D; $\sigma_1^2 = \sigma_2^2 = 1^2$. When differences between means are equal, the and critical values are determined by the three boundary combinations, i.e, (i) the OF for both endpoints, (ii) the PC for both endpoints and (iii) the OF for δ_1 and the PC for δ_2, with the LD alpha-spending method with equal information space. When differences between means are unequal, in addtion to the three combinations, (iv) the PC for δ_1 and the OF for δ_2 is considered.

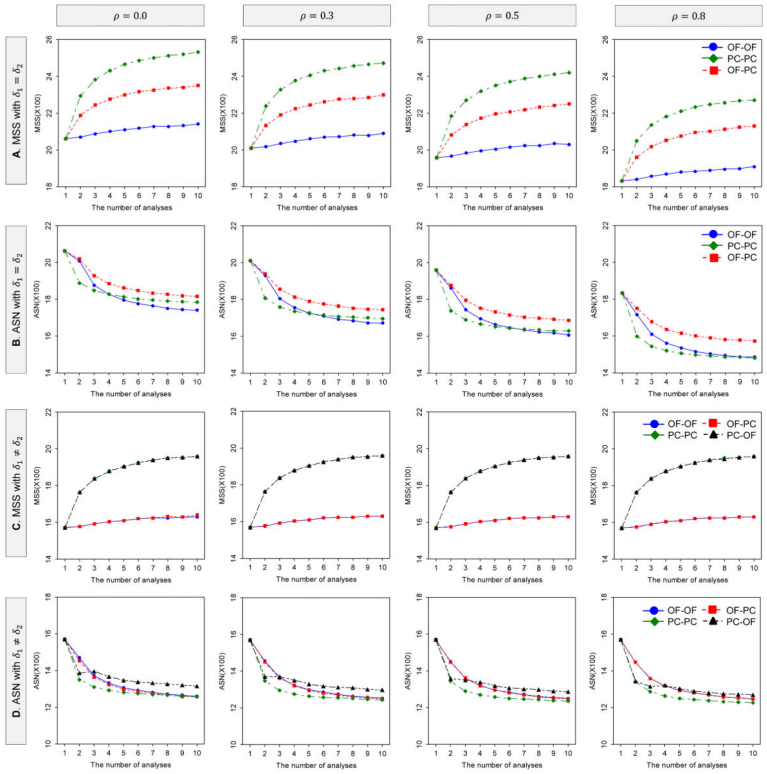

Figure 2.

Behavior of MSS and ASN for DF-2 as the number of analyses and boundaries vary. The MSS and ASN per intervention group (equally-sized groups: r_l=1) were calculated to detect the joint difference in the two endpoints with the overall power of 80% at the one-sided significance level of 2.5%, where $\delta_1 = \delta_2 = 0.1$ for A and B, and $\delta_1 = 0.1$ and $\delta_2 = 0.2$ for C and D; $\sigma_1^2 = \sigma_2^2 = 1^2$. When differences between means are equal, the critical values are determined by the three boundary combinations, i.e, (i) the OF for both endpoints, (ii) the PC for both endpoints and (iii) the OF for δ_1 and the PC for δ_2, with the LD alpha-spending method with equal information space. When differences between means are unequal, in addtion to the three combinations, (iv) the PC for δ_1 and the OF for δ_2 is considered.

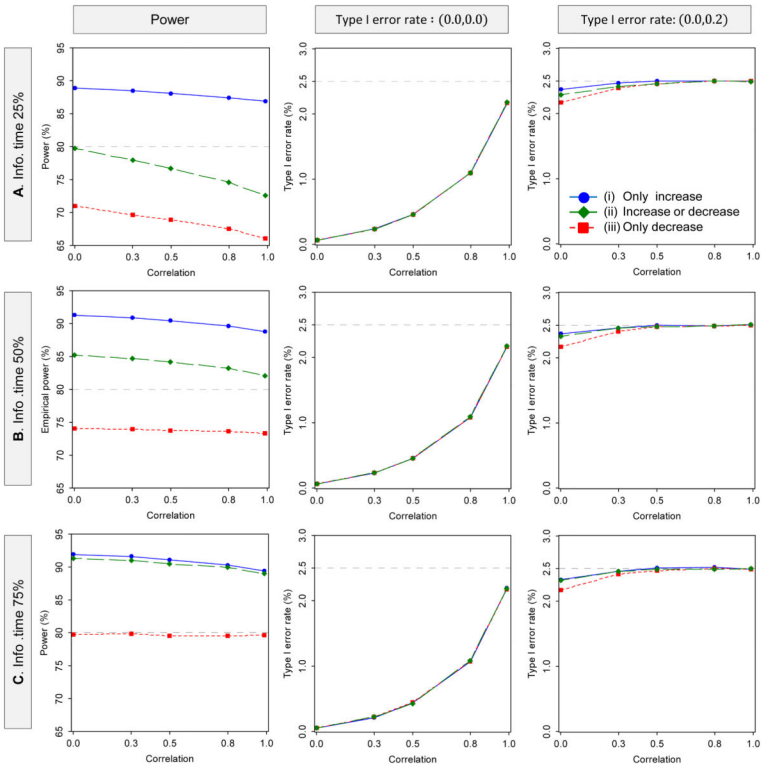

Figure 3.

Behavior of the power and Type I error rate as a function of the correlation with sample size recalculation in two-stage group-sequential designs, where the information times of 0.25, 0.50 and 0.75 were selected as the timing of the sample size recalculation. The planned MSS per intervention group is calculated to detect the joint difference for two endpoints with the overall power of 80% at the one-sided significance level of 2.5%, where one interim and one final analysis are to be performed. The critical values are determined by the OF boundary for both endpoints, with the LD alpha-spending method. The upper limit of recalculation sample size is $n_2^{'}=\lambda n_2$ with λ = 1.5. The number of replications for simulation is set to 1,000,000 for evaluation of the Type I error rate and 100,000 replications for the power (DF-1)

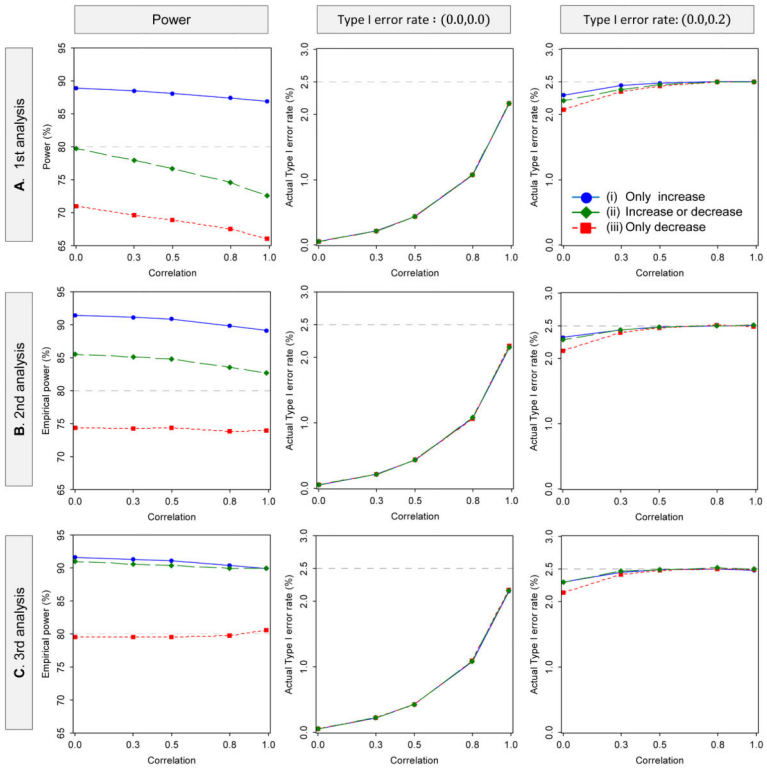

Figure 4.

Behavior of the power and Type I error rate as a function of the correlation with sample size recalculation in four-stage group-sequential designs, where the 1st, 2nd and 3rd interim point were selected as the timing of the sample size recalculation. The planned MSS per intervention group is calculated to detect the joint difference for two endpoints with the overall power of 80% at the one-sided significance level of 2.5%, where three interims and one final analysis are to be performed. The critical values are determined by the OF boundary for both endpoints, with the LD alpha-spending method. The upper limit of recalculation sample size is $n_2' = \lambda n_2$ with λ = 1.5. The number of replications for simulation is set to 1,000,000 for evaluation of the Type I error rate and 100,000 replications for the power (DF-1)

Table 1

MSS and ASN per intervention group (equally-sized groups) for detecting the joint difference for ADAS-Cog (0.2) and ADCS-ADL (0.2), with DF-1 and DF-2 and the overall power of 96% at the one-sided significance level of 2.5%

			(i) OF-OF		(ii) PC-PC		(iii) OF-PC	
Decision-making framework	**Correlation**	**Number of analyses**	**MSS**	**ASN**	**MSS**	**ASN**	**MSS**	**ASN**
DF-1	0.0	1	804	804	804	804	804	804
		2	808	725	886	607	854	693
		3	816	647	918	572	876	652
		5	825	604	945	548	895	608
		8	832	579	968	535	912	587
		10	840	573	970	530	920	581
	0.3	1	799	799	799	799	799	799
		2	802	702	880	593	850	676
		3	810	633	912	552	870	638
		5	820	589	940	525	890	593
		8	824	563	960	511	904	571
		10	830	556	970	507	910	564
	0.5	1	791	791	791	791	791	791
		2	794	684	872	580	842	662
		3	801	620	903	536	864	627
		5	810	574	930	506	885	582
		8	816	549	952	492	896	558
		10	820	542	960	488	900	551
	0.8	1	764	764	764	764	764	764
		2	768	644	842	549	818	635
		3	774	588	873	501	840	603
		5	785	543	900	469	860	556
		8	792	520	920	453	872	533
		10	800	514	920	447	880	527
DF-2	0.0	1	804	804	804	804	804	804
		2	808	725	882	605	848	690

Decision-making framework	Correlation	Number of analyses	(i) OF-OF		(ii) PC-PC		(iii) OF-PC	
			MSS	ASN	MSS	ASN	MSS	ASN
		3	813	645	912	569	867	646
		5	825	603	940	540	890	602
		8	832	578	960	524	904	579
		10	830	568	960	518	910	572
	0.3	1	799	799	799	799	799	799
		2	802	702	876	591	842	672
		3	807	632	906	549	861	632
		5	815	586	935	520	880	586
		8	824	562	952	503	896	564
		10	830	555	960	498	900	556
	0.5	1	791	791	791	791	791	791
		2	794	684	868	579	834	658
		3	801	620	897	533	855	621
		5	810	574	925	502	875	575
		8	816	549	944	486	888	552
		10	820	541	950	481	890	544
	0.8	1	764	764	764	764	764	764
		2	768	644	840	549	810	631
		3	774	588	870	499	831	597
		5	785	543	895	467	850	550
		8	792	520	912	450	864	528
		10	790	510	920	445	870	521