Published in final edited form as:
J Biopharm Stat. 2018 ; 28(1): 28–51. doi:10.1080/10543406.2017.1378668.

Design, data monitoring and analysis of clinical trials with co-primary endpoints: a review

Toshimitsu Hamasaki[1,2,*], **Scott R. Evans**[3], and **Koko Asakura**[1,2]

[1]Department of Data Science, National Cerebral and Cardiovascular Center, Osaka, Japan.

[2]Department of Innovative Clinical Trials and Data Science, Osaka University Graduate School of Medicine, Osaka, Japan

[3]Department of Biostatistics and the Center for Biostatistics in AIDS Research, Harvard T.H. Chan School of Public Heath, MA, USA.

Abstract

We review the design, data monitoring, and analyses of clinical trials with co-primary endpoints. Recently developed methods for fixed-sample and group-sequential settings are described. Practical considerations are discussed and guidance for the application of these methods is provided.

Keywords

Co-primary endpoints (CPE); Fixed-sample designs; Group-sequential designs; Intersection–union test (IUT); Multiple primary endpoints (MPE); Type II error adjustment.

1. Introduction

Typically in clinical trials, a single outcome is selected as a primary endpoint. This endpoint serves as the basis for the trial design including sample size determination, interim data monitoring, final analyses, and the reporting of the trial result. The primary endpoint should be an outcome which can provide the most clinically relevant measure to address the primary objective of a trial (e.g., see ICH E9 guideline (1998)).

However, the effects of interventions are multidimensional. Thus a single primary endpoint may not provide a comprehensive picture of the important effects of the intervention. For this reason, many recent clinical trials have been designed with more than one primary outcome. Multiple primary endpoints offer an attractive design feature as they could capture a more complete characterization of the effect of an intervention.

But multiple primary endpoints also create challenges. In December 2016, the European Medical Agency (EMA) released draft guidelines on multiplicity issues in clinical trials (CHMP, 2017), and in January 2017, the US Food and Drug Administration (FDA) issued guidance on multiple endpoints in clinical trials (FDA, 2017). The documents describe the

[*]5-7-1 Fujishirodai, Suita, Osaka 565-8565, Japan. toshi.hamasaki@ncvc.go.jp.

challenges raised by multiple endpoints and provide a regulatory perspective on how to deal with the issues, especially multiple comparisons and Type I and Type II error control during the planning and analysis of clinical trials. The guidelines distinguish two decision-making frameworks based upon whether it is desirable to evaluate if there are effects on AT LEAST ONE of the endpoints or whether there are effects on ALL of the endpoints. This decision defines the alternative hypothesis to be tested and provides a framework for approaching trial design. When designing the trial to evaluate an effect on AT LEAST ONE of the endpoints, then an adjustment is needed to control the Type I error rate. This is referred to as “multiple primary endpoints” (MPE) or “alternative primary endpoints” (Offen et al. 2007; Hung and Wang 2009; Dmitrienko et al. 2010) and is related to the union-intersection problem (Roy, 1953). In contrast, 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 being evaluated increases. This is referred to as “co-primary endpoints” (CPE) (Offen et al., 2007; Hung and Wang, 2009; Dmitrienko et al., 2010) and is related to the intersection-union problem (Berger 1982). In CPE, failure to demonstrate an effect on any single endpoint implies that effects cannot be concluded. Table 1 summarizes the issues in MPE and CPE. Although CPE is a special case of MPE, it is important to recognize their differences. This paper will focus on statistical challenges created by CPE. We integrate recent methodological developments for design and analysis of CPE clinical trials.

In complex diseases, CPE may be preferable to a single primary endpoint as they offer the opportunity of characterizing intervention’s multidimensional effects. The use of CPE is increasingly common especially in medical product development. Regulators have issued guidelines recommending use of CPE in e.g., acute heart failure (Committee for Medicinal Products for Human Use, CHMP 2012a), Alzheimer’s disease (CHMP 2008; FDA 2013), diabetes mellitus (CHMP 2012b), Duchenne and Becker muscular dystrophy (CHMP 2013a), and irritable bowel syndrome (IBS) (FDA 2012; CHMP 2013b). For example, CHMP 2008 and FDA (2013) recommend a co-primary endpoint approach using cognitive and functional or global endpoints to evaluate symptomatic improvement of dementia associated with Alzheimer’s disease. In some CPE trials, the sample size is often unnecessarily large and impractical. For example, Green et al. (2009) reported the results of a multicenter, randomized, double-blind, placebo-controlled trial in patients with mild Alzheimer disease (Tarenflurbil study), where co-primary endpoints were cognition as assessed by the Alzheimer Disease Assessment Scale Cognitive Subscale (ADAS-Cog) and functional ability as assessed by the Alzheimer Disease Cooperative Study activities of daily living (ADCS-ADL). The study was sized for 1600 participants in total (equally sized groups) based on a power of 96% to detect the between-group joint difference in the two primary endpoints (using a one-sided test at 2.5% significance level, with the standardized mean differences between the two groups of 0.2 for both endpoints, assuming zero correlation between the two endpoints). To overcome these issues, approaches to the design and analysis of CPE clinical trials in fixed-sample and group-sequential settings have been discussed (e.g., extensive references found in Offen et al. (2007), Alosh et al. (2014), Dmitrienko et al. (2013, 2014), Sozu et al. (2015) and Hamasaki et al. (2016)).

We provide an overview of the design, data monitoring, and analyses of CPE clinical trials, summarizing recent developments. The paper is structured as follows: In Section 2, we describe the intersection-union principle and review recently developed approaches for testing hypotheses associated with CPE. We describe sample size determinations for fixed-sample and group-sequential settings in Section 3 and Section 4, respectively. In Section 5, we discuss practical considerations and provide guidance for the design, data monitoring, and analyses of CPE clinical trials. In Section 6, we discuss developments for designing CPE clinical trials with other design characteristics including endpoints with other measurement scales, multiple intervention arms, enrichment designs and subgroup analyses, and multi-regional clinical trials.

2. Methods for evaluating CPE

2.1 Preliminaries and definition

Consider a randomized, fixed-sample clinical trial comparing the test intervention (T) with the control intervention (C), where $K(\geq 2)$ continuous outcomes are to be evaluated as CPE, and n and rn participants are recruited and randomly assigned to the T and the C, respectively, where r is the allocation ratio of the C to the T. Then, there are n sets of K-paired outcomes $(Y_{\mathrm{T}1i}, \ldots, Y_{\mathrm{T}Ki})$ $(i = 1, \ldots., n)$ for the T and rn sets of K-paired outcomes $(Y_{\mathrm{C}1j}, \ldots, Y\mathrm{C}_{Kj})$ $(j = 1, \ldots., rn)$ for the C. For $k = 1, \ldots, K$, consider K mean difference and standardized mean difference for the T and C,, i.e., $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)$ and $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_K)$, wher $\delta_k = \mu_{Tk} - \mu_{\mathrm{C}k}$ and $\Delta_k = \delta_k/\sigma_k$. Suppose that a positive value of δ_k represent the test intervention’s advantage.

Assume that $(Y_{\mathrm{T}1i}, \ldots, Y_{\mathrm{T}Ki})$ and $(Y_{\mathrm{C}1j}, \ldots, Y_{\mathrm{C}Kj})$ are independently multivariate distributed with means $\mathrm{E}[Y_{\mathrm{T}ki}] = \mu_{\mathrm{T}k}$ and $\mathrm{E}[Y_{Ckj}] = \mu_{\mathrm{C}k}$, and common known variance-covariance matrix $\boldsymbol{\Sigma}$ with diagonal elements $\mathrm{var}[Y_{\mathrm{T}ki}] = \mathrm{var}[Y_{\mathrm{C}ki}] = \sigma_k^2$, i.e.,

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

where $\mathrm{corr}[Y_{\mathrm{T}ki}, Y_{\mathrm{T}k'i}] = \mathrm{corr}[Y_{\mathrm{C}kj}, Y_{\mathrm{T}k'j}] = \rho_{kk'}$ $(k \neq k'; 1 \leq k < k' \leq K)$. Let Z_k be the statistic for testing the hypotheses, given by

$$Z_k = \frac{\hat{\delta}_k}{\sigma_k\sqrt{(1 + 1/r)/n}}$$

were $\hat{\delta}_k = \overline{Y}_{\mathrm{T}k} - \overline{Y}_{\mathrm{C}k}$, and $\overline{Y}_{\mathrm{T}k}$ and $\overline{Y}_{\mathrm{T}k}$ the sample means given by $\overline{Y}_{\mathrm{T}k} = n^{-1}\Sigma_{i=1}^{n} Y_{\mathrm{T}ki}$ and $\overline{Y}_{\mathrm{C}k} = (rn)^{-1}\Sigma_{j=1}^{n} Y_{\mathrm{C}ki}$. For large sample, each Z_k is approximately normally distrusted as $Z_k \sim N(\sqrt{rn(1 + r)}\Delta_k, 1^2)$ and $(Z_1, \ldots, Z_K)$ is approximately multivariate normally distributed with the correlation $\mathrm{corr}[Z_k, Z_k{}^{\prime}] = \rho_{kk'}$.

2.2 Intersection–union principle

Suppose that there is an interest in evaluating whether T is superior to the C on all of the endpoints using a one-sided test. For CPE, “success” is declared if the superiority is achieved on all of the endpoints. The hypotheses for each endpoint are H_{0k}: $\delta_k \le 0$ versus H_{1k}: $\delta_k > 0$, and the each hypothesis is tested at significance level α_k. The hypotheses for CPE are H_0: $\cup_{k=1}^{K} H_{0k}$ versus H_1: $\cap_{k=1}^{K} H_{1k}$, and the null hypothesis H_0 is rejected if and only if each null hypothesis H_{0k} is rejected. In this procedure, the union H_0 of all individual nulls is tested against the intersection of alternatives and this is referred to as the intersection-union test (IUT) (Berger, 1982). If α_k is the size of test of H_{0k} with the rejection region R_k, then the IUT with rejection region $R = \cap_{k=1}^{K} R_k$ is a test of level α and the size of the test is most α with $\alpha = \max_{k=1,\ldots,K} \alpha_k$. Therefore, if each endpoint is tested at level α then the size of the test for CPE is α. The IUT was described in Lehmann (1952) and was first called IUT by Gleser (1973). Since then, The IUT methods have been discussed in a variety of problems, for example, sampling problems by Berger (1982) and contingency table problems by Cohen et al. (1983).

When each endpoint is tested at α using IUT, then the Type I error is not inflated as the maximum Type I error remains bounded above by α. However, the rejection region of the null hypothesis defined as the intersection of K regions associated with the K endpoints, is considerably restricted and thus the hypothesis test is conservative, especially when the number of endpoints being evaluated is large and the correlations among the endpoints are small. Figure 1 illustrates the behavior of Type I error rate for $\alpha = 2.5\%$ as a function of standardized mean difference Δ_1 and correlation ρ_{12} for H_0 when $\Delta_2 = 0$, where $K = 2$. The figure shows that the Type I error rate is the smallest when $\Delta_1 = \Delta_2 = 0$ and its maximum is not larger than the prespecified significance level of 2.5 % although the Type I error rate increases as Δ_1 or ρ_{12} increases. When $\rho_{12} = 0$, the Type I error rate is 0.0625%. On the other hand, the Type II error rate increases as the number of endpoints being evaluated increases or the correlations among the endpoints are smaller. Here sample size adjustment is required to maintain the power of the test. This may result in a sample size that is too large and impractical to conduct the clinical trial. In order to provide a more reasonable and practical sample size, methods for CPE clinical trials have been discussed in fixed-sample designs (Chuang-Stein et al. 2007, 2017; Offen et al. 2007; Hamasaki et al., 2013; Hung and Wang, 2007, 2009; Kordzakhia et al., 2010; Julious and Mclntyre, 2012; Li, 2009; Sozu et al. 2010, 2011, 2012; 2016; Ristl et al, 2016; Senn and Bretz 2007; Sugimoto et al., 2012, 2013, 2017; Xiong et al. 2005). Most methods consider incorporating the correlations among the endpoints into the calculations. We discuss methods for Type I error adjustment in Section 2.3 and for sample size determination with Type II error adjustment in Section 3.

2.3 Type I error adjustment methods

Several methods are available to improve the power for CPE with an adjusted Type I error. Based on examination of the false positive rate over a restricted null space discussed in Patel (1991), Chuang-Stein et al. (2007) proposed the “average Type I error” method. Instead of controlling the Type I error over the entire null space for CPE, their method takes the average Type I error rate over all possible null hypothesis configurations with an equal weight. The significance level for each endpoint is adjusted to α^* to ensure that the average Type I error rate is equal to the prespecified significant level of α. The value of α^* depends on the correlation among the endpoints. Figure 2 illustrates the adjusted significance level α^* for $\alpha = 2.5\%$ and $\alpha = 5\%$ as a function of the correlation ρ_{12}, where $K = 2$. The figure shows that the adjusted significance level is largest when $\rho_{12} = 0$ ($\alpha^* = 3.6\%$ for $\alpha = 2.5\%$ and $\alpha^* = 7\%$ for $\alpha = 5\%$), and is closer to α with increasing correlation toward one.

The method is attractive but introduces complexities. Within the IUT framework, the maximum of the Type I error rate is larger than the prespecified significance level of α if each endpoint is tested at α^*. During trial planning, the adjusted significance level is prespecified by using the method with assumed correlations among the endpoints. But assumptions regarding the correlation may be incorrect. This calls into question of how such assumptions regarding the correlation affect the decision-making in clinical trials. Figure 3 illustrates the behavior of Type I error for α^* as observed the correlation $\hat{\rho}_{12}$ for H_0 when $\Delta_1 = \Delta_2 = 0$, where $K=2$ and $\alpha = 2.5\%$. The adjusted significance levels using the average Type I error method are $\alpha^* = 3.6\%$, 3.5%, 3.3%, 3.0% and 2.6%, corresponding to $\rho_{12} = 0.0$, 0.3, 0.5, 0.8 and 0.99. The figure illustrates that the Type I error rate is larger than $\alpha = 2.5\%$ when $\hat{\rho}_{12}$ is close to one.

Kordzakhia et al. (2010) introduced an interesting approach, called the “balanced adjustment method” for the Type I error. The method consists of adjustment of the significance level on other endpoints only if the intervention shows the highest significant difference on one endpoint (or more than one endpoint). Kordzakhia et al. (2010) illustrated a situation where a p-value from one of two endpoints is sufficiently small (e.g., p <0.001), but other p-value is slightly larger than the prespecified level, and the adjusted significance level α_L ($\alpha < \alpha_L < K_a$), which is slightly larger than α, is selected to testing a hypothesis. The method compensates a smaller intervention effect in one endpoint by a stronger intervention effect in other endpoint (Chuang-Stein and Li., 2017). For example, when $K = 2$, once α_L has been allocated for the one endpoint with higher significant difference, the significance level for the other endpoint is given by $(\alpha_L - p_1)/(1 + cp_1)$, where p_1 is the p-value from the test for the endpoint with the higher significant difference and $c = (\alpha_L - 2\alpha)/\alpha^2$. In the average Type I error method, the rejection regions of H_0 include a region that all Z_k are greater than $c(\alpha)$ and smaller than $c(\alpha^*)$, where $c(\alpha)$ and $c(\alpha^*)$ are the $(1 - \alpha)$th and $(1 - \alpha^*)$th percentiles of the standardized normal distribution, respectively (Kordzakhia et al (2010) calls this region the “bad region”). The balanced adjustment method can exclude this region. Kordzakhia et al. (2010) discussed the balanced adjustment method under the average Type I error method and provide recommended values for α_L via simulation. Similarly as in the average Type I error method, the adjustment depends on the correlation among the endpoints and a larger adjustment is needed when correlations are lower. One major concern is, similarly as in the average Type I error method, that at the planning, the adjusted significance levels for all of the endpoints can be prespecified by using the method with assumed standardized mean differences and correlations among the endpoints, but the assumptions may be incorrect. In such situations, an important question is whether the prespecified significance levels can be modified based on the observed data.

The two procedures discussed above are motivated by the fact that the Type I error rate under the IUT near the origin in the null space is lower than the significance level. Recently Chuang-Stein and Li (2017) proposed a new test for which the information on the relative size of the mean difference and their distance from the origin in the null space are used to provide a more liberal critical value. Their procedure tests the endpoints sequentially according to their ordering by the size of mean difference even although they are equally important. This procedure also has associated concerns. Existing data may provide information on the size of the mean difference, but the assumptions may be incorrect. For example, $\Delta_1^* > \Delta_2^*$ were assumed and significance levels for each endpoint were determined by this order. But if the observed order is reversed, i.e., $\hat{\Delta}_1 < \hat{\Delta}_2$, it is unclear how the two endpoints should be tested.

The problem in restricted null space associated with IUT has been also discussed in a different setting, where a clinical trial is designed to evaluate whether a combination or simultaneous administration of the interventions has a better benefit rather than monotherapy when two monotherapies are available for the treatment of a disease. For more details, please see Laska and Meisner(1989) and Sarkar et al. (1995).

Furthermore, Ristl et al. (2016) discussed a use of fallback procedure (Wiens, 2003; Wiens and Dmitrienko, 2005) for a special case of CPE, where a joint statistical significance have been demonstrated on not all, but a subset of the endpoints, and it is still of interest to make best use of the collected data by making at least partial claims on the efficacy in such a subset of endpoints. The procedure has the same rejection region as the conventional CPE test for simultaneous rejection of all null hypotheses, but allows one to reject elementary or intersection null hypotheses if this objective is not achieved. This is related to a problem of at least s endpoints must-win out of K endpoints. Delorme et al. (2016) defined the generalized Type II error rate and methods for sample size calculation in such a problem.

3. Sample size determination for CPE clinical trials

Methods for sample size determination in CPE clinical trials in a fixed-sample setting have been discussed by several authors. Jennison and Turnbull (1993) formulated one-sided testing and the decision-making framework, and discussed the behavior of sample size with varying correlation between the endpoints in clinical trials with efficacy and safety endpoints as co-primary. Xiong et al. (2005) discussed methods for power assessment and sample size calculation in clinical trials with two co-primary endpoints with application to Alzheimer’s disease, where two continuous endpoints are assumed to be bivariate-normally distributed with known variance-covariance matrix. Sozu et al. (2006) extended their method under unknown variance-covariance matrix using the Wishart distribution. Sozu et al. (2011) extended the methods to more than two endpoints under both known and unknown variance-covariance matrices, and showed that the sample size using the method based on the known variance could be a good approximation to that using the unknown variance. Eaton and Muirhead (2007) provided a simple expression for calculating the p-value and computable bounds for the power function for CPE. Sugimoto et al. (2012) discussed a convenient and practical formula with accompanying numerical tables for sample size calculation in CPE clinical trials. In this section, we outline the methods for sample size determinations for fixed-sample designs.

3.1 Type II adjustment and sample size calculation

When more than one endpoint is viewed as important in a clinical trial and the trial is designed to evaluate a joint effect on all of the endpoints, then the Type II error rate increases as the number of endpoints being evaluated increases. For example, if there are two co-primary endpoints with an equivalent standardized mean difference, then the sample size required for 80% power (i.e., the Type II error is 20%) for each endpoint provides the power of 80% × 80% =64% to evaluate the joint effect on both endpoints, assuming zero correlation between the endpoints. Therefore, to maintain 80% power for evaluating the joint effect on both endpoints, the sample size should be increased to provide the power of (100% – 20%)$^{1/2}$ =89.4% for each endpoint. Similarly to the case with the Type I error, the Type II error changes with varying correlation among the endpoints (Sen and Bretz, 2007; Sozu et al., 2015). Therefore, it is important to evaluate the impact of the correlation on the power and sample size when designing CPE trials.

For CPE, the hypotheses for testing H_0 versus H_1 are tested by the statistics $(Z_1,\ldots, Z_K)$. The null hypothesis for each endpoint, H_{0k} is rejected if the statistic Z_k is larger than $c(\alpha)$ and the rejection regions of H_0 are $[\{Z_1 > c(\alpha)\} \cap \ldots \cap \{Z_K > c(\alpha)\}]$, where $c(\alpha)$ is the $(1 - \alpha)$the percentile of the standardized normal distribution. Defining $\boldsymbol{\delta}^* = (\delta_1^*, \ldots, \delta_K^*)$ to be the clinically meaningful difference in means between the two interventions to be detected with high probability. For large samples, we have the power for CPE at $\boldsymbol{\delta} = \boldsymbol{\delta}^*$ given as

$$1 - \beta = \Pr_{\delta = \delta^*}\left[\bigcap_{k=1}^{K} \{Z_k > c(\alpha)\} \middle| \mathrm{H}_1 \right] \approx \Pr_{\delta = \delta^*}\left[\bigcap_{k=1}^{K} \{Z_k^* > c_k^*(\alpha)\} \middle| \mathrm{H}_1 \right],$$

where $Z_k^* = Z_k - \sqrt{r/(1+r)n}\Delta_k$ and $c_k^*(\alpha) = c(\alpha) - \sqrt{r/(1+r)n}\Delta_k$ with $\mathrm{E}[Z_k^*] = 0$ and $\mathrm{var}[Z_k^*] = 1$. This power is referred to as “complete power” (Westfall et al., 2011) or “conjunctive power” (Senn and Bretz, 2011). The joint distribution of $(Z_1^*, \ldots, Z_K^*)$ is K-variate standardized normal $\mathrm{N}_K(\mathbf{0}, \boldsymbol{\rho}_Z)$, where the off-diagonal element of $\boldsymbol{\rho}_Z$ is given by $\mathrm{corr}[Z_k, Z_{k'}] = \rho_{kk'}$. Once the design parameters δ_k, σ_k, and $\rho_{kk'}$ has been given, the power can be numerically evaluated by the cumulative distribution function of K-variate standardized normal, i.e., $\Phi_K(-c_1^*(\alpha), \cdots, -c_K^*(\alpha) \mid \boldsymbol{\rho}_Z)$.

The sample size n required for achieving the desired power $1 - \beta$ at the significance level α is given by

$$n = \begin{cases} n^*, & \text{if } n \text{ is an interger,} \\ [n^*] + 1, & \text{otherwise,} \end{cases}$$

where n* is the smallest value satisfying $1 - \beta \le \Phi_K(-c_1^*(\alpha), \ldots, -c_K^*(\alpha) \mid \boldsymbol{\rho}_Z)$, and [$n$*] is the greatest integer less than n. An iterative procedure is required to calculate n. The easiest way is a grid search to increase n gradually until the power under n exceeds the desired power of $1 - \beta$, where a maximum of sample sizes separately calculated for each endpoint with the power 1 ′ β at the significance level α can be used as an initial value for the calculation. This often requires considerable computing resources. The Newton–Raphson algorithm in Sugimoto et al. (2012) or the basic linear interpolation algorithm in Hamasaki eat al. (2013) may be utilized to reduce the necessary computational resources. For R and SAS® codes for implementing the calculations, please see Sozu et al. (2015).

Table 2 provides the sample size per intervention group (equally-sized groups: $r = 1$) in clinical trials with two co-primary endpoints ($K = 2$), using the method described above (hereafter we refer to this as the “conventional method”), with varying clinical meaningful standardized mean differences (Δ_1^*, Δ_2^*), and assumed correlation ρ_{12}^* The sample size is calcualed to detect a joint effect on two endpoints with a power of $1 - \beta$ =80% or 90% at the significance level α = 2.5% by a one-sided test. The table also includes the sample size using the average Type I error method and the balanced adjustment method.

For the conventional method, when (Δ_1^*, Δ_2^*) = (0.2, 0.2), the sample size decreases as the correlation approaches one. Comparing the sample size for $\rho_{12}^* = 0.0$ to that for for $\rho_{12}^* \ge 0.8$, the decrease in the sample size is more than 10%. However, when correlation is smaller, i.e., $\rho_{12}^* \le 0.5$, the decrease is less than 5%. When (Δ_1^*, Δ_2^*) = (0.3, 0.2), the sample size still decreases, but does not change considerably as the correlation varies. When (Δ_1^*, Δ_2^*) = (0.4, 0.2), there is no decrease with varying correlation and the sample size is equivalent to that required for the endpoint with the smaller standardized mean difference (i.e., Δ_2^*) under the non-adjusted Type II error rate. Both the average Type I method and the balanced adjusted method provide relatively smaller sample size compared with the conventional method. The sample size increases toward that required from the conventional method as the correlation approaches one. When (Δ_1^*, Δ_2^*) = (0.3, 0.2) and (0.4, 0.2), then the sample size calculated for evaluating a joint effect is smaller than that required for the endpoint with the smaller standardized mean difference under the non-adjusted Type II error rate.

3.2 A simpler approach for sample size calculation

One major consideration in sizing CPE clinical trials is whether the correlations among the endpoints should be incorporated into the power assessment and sample size calculation (Sozu et al., 2015). The correlations may be estimated from external or internal pilot data, though such data are usually limited or unreliable. If correlations are overestimated at the planning stage when including them into the sample size calculation, then the sample size is too small and important effects may not be detected.

Using the conventional method, there is no practical advantage to incorporating correlations into the power assessment and sample size calculation (i) when correlations are low or moderate, and/or (ii) when the standardized mean difference for one endpoint(s) is smaller than the other endpoints as the reduction in the sample size is relatively small (less than 5%). For Situation (i), if the mean effect sizes are assumed to be the same on all K primary endpoints $\Delta = \Delta_1 = \cdots = \Delta_K$, and their correlations are ignored, then the sample size can be calculated simply using an equation for a single endpoint with adjusted power $1 - \gamma$, given by

$$n = \frac{(c(\alpha) + c(\gamma))^2}{r \Big/ (1 + r)\Delta^2}.$$

where $\gamma = 1 - (1\ 0 - \beta)^{1/K}$ and $c(\gamma)$ is the $(1 - \gamma)$th percentile of the standardized normal distribution. If there are differences in mean effect sizes among the endpoints, Varga et al. (2017) discussed a simple procedure to calculate the sample size required to achieve the desired power using SAS® code. However, when the standardized mean difference for one endpoint(s) is smaller than the other endpoints (e.g., $\Delta_1/\Delta_2 > 1.5$ or $\Delta_2/\Delta_1 > 1.5$ for $K = 2$), then the sample size can be determined based on the one endpoint with the small standardized mean difference using the equation for the singe endpoint, without the adjusted power. Sozu et al. (2015) and Ando et al. (2015) provides reference values for when the sample size equation can be simplified using the equation for a single endpoint.

When the standardized mean difference for one endpoint(s) is smaller than the other endpoint(s), the sample size required may be too large to evaluate an effect on the endpoint(s) with the larger standardized mean difference. If these endpoints are very invasive or expensive to obtain (e.g., data from a liver biopsy or gastro-fiberscope, or data from expensive imaging), then one may consider stopping measurement of this endpoint as soon as possible, that is, once the number of participants required for that endpoint are accrued. However, the trial will continue until the number of participants required for CPE without further examination of the invasive endpoint. Sozu et al. (2015) discussed methods for allowing the option of selecting different sample sizes among the endpoints in CPE and MPE settings. The method does not reduce the sample size required for CPE, but reduces the participant's burden.

3.3 Binary and time-to-event outcomes

We have reviewed the methods to address continuous endpoints in the previous sections. However clinical trials may be conducted with the objective of comparing a test intervention with that of a standard intervention based on several binary outcomes. For example, irritable bowel syndrome (IBS) is one of the most common gastrointestinal disorders and is characterized by symptoms of abdominal pain, discomfort, and altered bowel function (American College of Gastroenterology, 2013; Grundmann and Yoon, 2010). The comparison of the interventions to treat IBS is based on the proportions of participants with adequate relief of abdominal pain and discomfort, and improvements in urgency, stool frequency, and stool consistency.

Sozu et al. (2010) discussed sample size determination in clinical trials with multiple binary endpoints when risk differences are evaluated as co-primary. They introduced three different measures to define the associations among multiple binary endpoints and discussed a formula for power for the five common testing strategies frequently used in the analysis of binary data for a two-group comparison. Based on simulations, they conclude that the normal approximation method works well in most situations except for extremely small event rates or small sample sizes. In these situations, they recommend more direct ways of calculating the sample size without using a normal approximation. Ando et al. (2015) and Sozu et al. (2015) described methods for power assessment and sample size calculation for clinical trials with multiple binary endpoints when relative risks or odds ratios are evaluated as co-primary. Song (2009) discussed sample size calculations with multiple co-primary binary endpoints in the case of noninferiority clinical trials.

Methods for time-to-event outcomes are more complex. Considerable care is needed to design event-time trials. As discussed in Sugimoto et al. (2013) and Hamasaki et al. (2013), the magnitude of the association among the time-to-event outcomes may depend on time. For example, the outcomes may be less correlated in earlier stages but more highly correlated in later stages. The censoring mechanism further complicates the design of these trials. For example, coinfection/comorbidity trials may utilize primary endpoints to evaluate multiple comorbidities; e.g., a trial evaluating therapies to treat Kaposi's sarcoma (KS) in HIV-infected individuals may have the time to KS progression and the time to HIV virologic failure, as primary endpoints. Both events are non-fatal and neither event-time is censored by the other event. In new anticancer drug trials, the most commonly used primary endpoint is overall survival (OS) defined as the time from randomization until death from any cause. OS often requires long follow-up periods after disease progression leading to long and expensive trials. Therefore, in addition to OS, as a primary endpoint, many trials evaluate the time from randomization to the first of tumor progression (TTP) or progression-free survival (PFS) which is composite of tumor progression and death. In this example, a death event censors TTP: Death is a competing risk for TTP but not vice versa. This is referred to as "semi-competing risks" (Fine et al. 2001).

Hamasaki et al. (2013) and Sugimoto et al. (2013) developed methods for sizing clinical trials with two time-to-event outcomes under a time-dependent correlation structure of three bivariate exponential distributions, where both events are non-fatal. Sugimoto et al. (2013) discussed the log-rank test based method using the normal approximation for calculating both sample size and the number of events. They evaluate how the sample size varies as a function of the correlation between the endpoints for CPE and MPC. Hamasaki et al. (2013) discussed a simpler normal approximation method for calculating the sample size based on the log-transformed hazard ratio. Sugimoto et al. (2017) extended this methodology to two additional situations, i.e., when one event is fatal and other is non-fatal, and when both are fatal.

Furthermore, there may be instances where CPE are of mixed scales of measurement. For example, a trial evaluating interventions for pain may have pain evaluated on a continuous scale (e.g., Gracely pain scale) but have a binary safety endpoint (occurrence of an adverse event). Sozu et al. (2012) discussed sample size methodology assuming that the endpoints are distributed as a multivariate normal distribution, where binary variables are observed in a dichotomized normal distribution with a certain point of dichotomy. For mixed time-to-event and binary endpoints, Sugimoto et al. (2013) defined the relationship between the endpoints under the limited distributions of copulas. They evaluate how the correlation is restricted depending on the marginal probabilities of binary endpoints, and discussed how the sample size varies as a function of the correlation.

4. Group-sequential designs for CPE clinical trials

The Tarenflurbil trial, mentioned in Introduction, failed to demonstrate a beneficial effect of tarenflurbil as the observed ADCS-ADL scores in the tarenflurbil group were smaller than for the placebo group. If the design had included an interim efficacy or futility assessment, the trial may have been stopped earlier, saving resources and time, and preventing trial participants from being exposed to an ineffective intervention unnecessarily. Standard methods for sizing trials with CPE often results in large sample sizes due to the conservative nature of the testing procedure even when the correlations among the endpoints is incorporated into the calculation. Therefore, researchers may naturally consider interim analyses to evaluate if the research questions can be answered with fewer trial participants or shorter follow-up.

In this section, we review methods for group-sequential designs in CPE clinical trials. The improvement in power by incorporating the correlations among endpoints into Type I or Type II error adjustments is limited and the calculated sample size may still be large in practical situations. As suggested by Hung and Wang (2009), use of group-sequential designs may be a remedial but practical approach to improve efficiency. But it also creates operational challenges in study conduct and data monitoring.

4.1 Preliminaries and definitions

Consider a randomized, group-sequential clinical trial of comparing T with C, where the same L maximum planned analyses and the same information space are planned for all endpoints. Let n_l and rn_l be the cumulative number of participants on the T and the C at the lth analysis ($l = 1, \ldots, L$). Hence, up to n_L and rn_L participants are recruited and randomly assigned to the T and the C. Then, there are n sets of n_L-paired continuous outcomes ($Y_{\mathrm{T}1i}, \ldots, Y_{\mathrm{T}Ki}$) ($i = 1, \ldots., n_L$) for the T and rn sets of rn_L-paired continuous outcomes ($Y_{\mathrm{C}1j}, \ldots, Y_{\mathrm{C}Kj}$)($j = 1, \ldots., rn_L$) for the C. Let ($Z_{1l}, \ldots Z_{\mathrm{Kl}}$) be the statistics for testing the hypotheses at the lth analysis $Z_{kl} = \hat{\delta}_{kl} \big/ (\sigma_k\sqrt{(1+1/r)/n_l})$, were $\hat{\delta}_{kl} = \bar{Y}_{\mathrm{T}kl} - \bar{Y}_{\mathrm{C}kl}$, and $\bar{Y}_{\mathrm{T}kl}$ and $\bar{Y}_{\mathrm{C}kl}$ the sample means given by $\bar{Y}_{\mathrm{T}kl} = n_l^{-1}\Sigma_{i=1}^{n_l}Y_{\mathrm{T}ki}$ and $\bar{Y}_{\mathrm{C}kl} = (rn_l)^{-1}\Sigma_{j=1}^{rn_l}Y_{\mathrm{C}kj}$. For large samples, each Z_{kl} is approximately normally distributed as $Z_{kl} \sim \mathrm{N}(\sqrt{rn(1+r)}\Delta_k, 1^2)$. Furthermore, as the joint distribution of ($Z_{1l}, \ldots, Z_{Kl}$) is approximately K-variate normally distributed with the correlation $p_{kk}{}'$ and the joint distribut on of ($Z_{k1}, \ldots, Z_{kL}$) is approximately K-variate normally distributed with the correlation $\sqrt{n_l/n_{l'}}(1 \le l \le l' \le L)$, the joint distribution of the joint distribution of ($Z_{11}, \ldots, Z_{K1}, \ldots, Z_{1L}, \ldots, Z_{KL}$) is KL-variate normal with their correlations given by $\rho_{kk'}\sqrt{n_l/n_{l'}}(k \ne k'; 1 \ne l')$.

In group-sequential designs for CPE clinical trials, several decision-making frameworks associated with interim evaluation of efficacy in two endpoints (Ando et al., 2015; Asakura et al., 2014, 2015; Cheng et al., 2014) and two or more endpoints (Hamasaki et al., 2015) or efficacy or futility in two endpoints (Cook and Farewell, 1994; Jennison and Turnbull, 1993; Schüler et al., 2017) and two or more endpoints (Asakura et al., 2017) have been discussed. In the subsequence sections, we briefly review these methods.

4.2 Interim evaluation of efficacy only

When evaluating the joint effects on all K endpoints within the context of group-sequential designs, a general decision-making framework associated with hypothesis testing is to reject H_0 if statistical significance of the T relative to the C is achieved for all endpoints at any interim analysis until the final analysis (i.e., not necessarily simultaneously) (Asakura et al., 2014; Hamasaki et al., 2015). If statistical significance is achieved on some but not all of the endpoints at the interim, then the trial will continue but subsequent hypothesis testing is repeatedly conducted only for the previously nonsignificant endpoint(s). This decision-making framework offers the opportunity of stopping measurement of an endpoint for which superiority has already been demonstrated. This may be desirable if the endpoint is very invasive or expensive (e.g., data from a liver biopsy or gastro-fiberscope, or data from expensive imaging). The stopping rule is formally described as follows;

Until the lth analysis ($l = 1, \ldots, L - 1$),

If $Z_{kl} > c^{\mathrm{E}}_{kl}(\alpha)$ for each endpoint, for some $1 \le l' \le l$, then reject H_0 and stop the trial, otherwise, continue to the ($l + 1$)th analysis

at the Lth analysis, for the endpoints that statistics have not yet crossed the efficacy boundary until ($L - 1$)th analysis,

If $Z_{kL} > c^{\mathrm{E}}_{kL}(\alpha)$ for nonsignificant endpoint(s) until the ($L - 1$)th analysis, then reject H_0, otherwise, do not reject H_0,

where $c^{\mathrm{E}}_{kl}(\alpha)(k = 1, \ldots, K; l = 1, \ldots, L)$ are the efficacy boundaries. The efficacy boundaries for each endpoint can be simply prespecified using any group-sequential method such as the Lan–DeMets error-spending method (Lan and DeMets, 1983), analogously to the single endpoint case, as if they were a single primary endpoint, ignoring the other co-primary endpoint. The power corresponding to this decision-making framework at $\boldsymbol{\delta} = \boldsymbol{\delta}^*$ is

$$1 - \beta = \Pr_{\delta = \delta^*}\left[\left\{\bigcup_{l=1}^{L} Z_{1l} > c^{\mathrm{E}}_{1l}(\alpha)\right\} \cap \ldots \cap \left\{\bigcup_{l=1}^{L} Z_{Kl} > c^{\mathrm{E}}_{Kl}(\alpha)\right\} \middle| \mathrm{H}_1\right].$$

This power can be numerically assessed by using multivariate normal integrals.

The decision-making framework described above is flexible but stopping measurement may also introduce operational challenges into the trial. To avoid the operational difficulties, one may opt for a restriction regarding when H_0 is rejected and the trial is stopped. The simplified version of the former decision-making framework is to reject H_0 if statistical significance is achieved on all of the endpoints at an interim simultaneously (Asakura et al., 2014; Cheng et al., 2014; Hamasaki et al., 2015). If any test of the endpoints is not significant, then the trial continues until the joint significance for all endpoints is established simultaneously. The stopping rule is formally described as follows;

Until the lth analysis ($l = 1, \ldots, L$ ′ 1),

If $Z_{kl} > c_{kl}^{\mathrm{E}}(\alpha)$ for all endpoints, at the same lth interim analysis, then reject H_0 and stop the trial,

otherwise, continue to the ($l + 1$)th analysis

at the Lth analysis,

If $Z_{kL} > c_{kL}^{\mathrm{E}}(\alpha)$ for all endpoints, then reject H_0, otherwise, do not reject H_0.

The efficacy boundaries for each endpoint can be simply prespecified using any group-sequential method such as the Lan–DeMets error-spending method. The power corresponding to this decision-making framework at $\boldsymbol{\delta} = \boldsymbol{\delta}^*$ is

$$1 - \beta = \Pr_{\delta = \delta^*}\left[\bigcup_{l=1}^{L}\left\{\left\{Z_{1l} > c_{1l}^{\mathrm{E}}(\alpha)\right\} \cap \ldots \cap \left\{Z_{Kl} > c_{Kl}^{\mathrm{E}}(\alpha)\right\}\right\} \middle| \mathrm{H}_1\right]$$

This power can be numerically assessed by using multivariate normal integrals. Hamasaki et al. (2015, 2017) summarized advantages and disadvantages of these two decision-making frameworks. Hamasaki et al. (2015) discussed more flexible decision-making frameworks allowing the different time points of analyses among the endpoints and considered the use of hierarchical hypothesis testing methodology for CPE clinical trials.

In a group-sequential setting, we discuss two sample size concepts: the maximum sample size (MSS) and the average sample number (ASN). The MSS is the sample size required for the final analysis to achieve the desired power. The MSS is given by the smallest integer no less than n_L satisfying the above power for prespecified design parameters including the differences in means, correlations among the endpoints, and the Fisher's information time for the interim analyses. Similarly as in fixed-sample designs, an iterative procedure is required to find the maximum sample size.

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

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

where $P_l = P_l\,(\Delta_1,\ldots, \Delta_K|\boldsymbol{\rho}_Z)$ is stopping probability or exit probability as defined the likelihood of crossing the critical boundaries at the lth interim analysis assuming the true values of intervention's standardized mean differences. Calculating power, MSS and ASN using the multivariate normal integrals does not requires expensive computing resources in most practical situations, and can be done within seconds using standard statistical software such as R or SAS® However, when considering more than two endpoints in a group-sequential trial with more than five analyses, the computational time will increase considerably (Hamasaki et al., 2015). In such situations, Monte Carlo simulation-based methods provide an alternative. However the number of replications for the simulations should be carefully chosen to control simulation error in evaluating the empirical power. For more details, please see Asakura et al. (2014) and Hamasaki et al. (2015, 2016).

Table 3 provides MSS and ASN per intervention group (equally-sized groups: $r = 1$) in clinical trials with two co-primary endpoints (K =2), based on the two decision-making frameworks described above, with varying clinical meaningful standardized mean differences (Δ_1^*, Δ_2^*), assumed correlation ρ_{12}^*, and the number of planned analyses L. The MSS is calculated to detect a joint effect on two endpoints with the power of $1 - \beta$ =80% or 90% at the significance level of α =2.5% by a one-sided-test. The O'Brien-Fleming critical boundaries are selected for both endpoints and determined using the Lan-DeMets error spending method with equally-spaced increments of information. For both decision-making frameworks, when effect sizes are equal, i.e., (Δ_1^*, Δ_2^*) = (0.2, 0.2), the MSS increases as the number of the number of analyses increases and with lower correlation. On the other hand, the ASN decreases as the number of analyses increases and larger correlation. When effect sizes are unequal, i.e., $\Delta_1^* > \Delta_2^*$, the MSS increases as the number of planned analyses increases, but it does not change as the correlation varies. On the other hand, the ASN decreases as the number of planned analyses increases independently of the correlation.

4.3 Interim evaluation of efficacy or futility

In many trials, in addition to efficacy assessments, it is often desirable to conduct interim assessments for futility (Gould and Pecore, 1982; Snapinn et al., 2006; Ware et al., 1985). There are two fundamental approaches for the interim futility assessment, based on: (i) the conditional power (Lachin, 2005; Lan et al., 1982;), and (ii) futility boundaries using group-sequential methodology (DeMets and Ware, 1980, 1982; Whitehead and Matsushita, 2003). For CPE clinical trials, methods are limited with group-sequential based methods having been discussed by a few authors (Asakura et al. 2017; Cook and Farewell, 1994; Jennison and Turnbull, 1993; Schüler et al., 2017).

When considering CPE group-sequential trials with the decision-making frameworks evaluating efficacy (rejecting the null hypothesis) or futility (accepting the null hypothesis), efficacy and futility boundaries are prespecified and determined using any group-sequential method. Jennison and Turnbull (1993) provide the fundamentals for this design in clinical trials with two endpoints. When planning interim efficacy and futility assessments in CPE clinical trials, the approach determines efficacy and futility boundaries to preserve the desired Type I and II errors, analogously to the single endpoint case. Jennison and Turnbull (1993) described a simple decision-making framework for rejecting or accepting the null hypothesis associated with CPE. They assumed that both of efficacy and futility assessments are performed at the same interims and determined the efficacy and non-binding futility boundaries based on methods in Emerson and Fleming (1989). Both of the efficacy and futility boundaries are fixed for any values of correlation among the endpoints but the method incorporates the correlations into the power assessment.

As an extension of Asakura (2014), Asakura et al. (2017) discussed more flexible division-making frameworks that allow for different timings for the efficacy and futility assessments with two or more endpoints. The framework provides savings for error spending (Type I and II errors), thus improving the efficiency (increasing power and reducing required sample sizes). The efficacy and non-binding futility boundaries are determined using any error-spending function to spend both Type I and Type II errors. The efficacy boundary for each endpoint is determined independent of the futility boundary. The futility boundary is determined by incorporating the correlations among the endpoints. They showed how the correlations may affect the decision-making for accepting the null hypothesis. Asakura et al. (2017) provide R codes for implementing the methods including efficacy and futility boundary calculation, and MSS and ASN.

Schüler et al. (2017) discussed methods to determine the binding futility boundary based on several optimal criteria in two-stage group-sequential designs with two co-primary endpoints. However, analogously to trials with a single primary endpoint, in general use of binding futility boundaries should be selected carefully in practice since the Type I error will be inflated if the trial is not stopped when at least one test statistic has crossed the futility boundary.

Group-sequential designs and other related methods do not provide formal evaluation regarding potential effect size estimates and associated precision with continuation of the trial to aid in go/no-go decision-making. Using prediction (Evans et al., 2007; Li et al., 2009) could be a flexible and practical approach for monitoring interim data of CPE clinical trials. This approach is appealing in that it provides quantitative evaluation of potential effect sizes and associated precision, with endpoint measurement continuation, thus providing investigators with a better understanding of the pros and cons associated with continuation of endpoint measurement. Asakura et al (2017) discussed an extension to the two endpoint situation, and evaluated the relationship between the prediction with other methods including conditional power, predictive power, and group-sequential designs.

4.4 Sample size recalculation based on the observed effect at an interim look

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/underpowered. Interim analyses provide 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. 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 (Evans and Ting, 2015). Challenges include the following: (i) maintaining control of statistical error rates, (ii) developing a plan to make sure that treatment effects cannot be inferred via back-calculation of a resulting change in the sample size, (iii) consideration of the clinical relevance of the treatment effects, and (iv) practical concerns such as an increased 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.

Asakura et al. (2014) discussed sample size recalculation based on the observed intervention's mean differences at an interim analysis with a focus on the control of statistical error rates, within the two decision-making frameworks discussed in Section 4.1, where all endpoints are continuous. Their method is based on Cui–Hung–Wang (CHW) statistics (Cui et al., 1999) to control the Type I error rate. Incorporating the uncertainty of the estimates at the interim into the sample size recalculation is important. When planning the sample size recalculation in CPE clinical trials, one practical question is whether the sample size can be increased or decreased in sample size recalculation. Referring to Asakura et al. (2014), the option of decreasing the sample size is a suboptimal choice as the power cannot maintain the targeted power although the expected sample size can be reduced more than other recalculation options. For other options, i.e., only allowing an increase in the sample size or allowing an increase or decrease in the sample size, the targeted power is maintained. An important decision regards the optimal timing of the sample size recalculation. The timing should also be carefully considered as the power does not reach desired levels if the sample size recalculation is done too early in the trial, especially when considering a decrease in the sample size. For more details, please see Asakura et al. (2014).

For other endpoint scale such as binary outcomes, Ando et al. (2015) and Asakura et al. (2015) described the methods when the risk difference or relative risks are being evaluated.

5 Considerations when designing co-primary clinical trials

When designing CPE trials in a fixed-sample or group-sequential setting, one important decision is the selection of the correlations among the endpoints in the power evaluation and sample size calculation, i.e., whether the observed correlations from external or pilot data should be utilized. As shown in Section 3, when the standardized mean differences for the endpoints are unequal, the advantage of incorporating the correlation into sample size calculation is less dramatic as the required sample size is primarily determined by the smaller standardized mean difference and does not greatly depend on the correlation. In this situation, the sample size equation for CPE can be simplified using the equation for a single endpoint. When the standardized mean differences among endpoints are approximately equal, one conservative approach is to assume that the correlations are zero even if nonzero correlations are expected. However, this may result in a sample size that is too large and impractical to conduct the clinical trial.

As discussed in Section 4, group-sequential designs provide an alternative solution to overcome this issue although may create implementation and operational issues associated with maintaining confidentiality of interim data (Evans and Ting, 2015). For example, when planning two analyses (one interim and final analyses), group-sequential designs provides very minimal increases in the required sample size compared to fixed-sample designs (see

Tables 2 **and** 3). Table 4 summarizes ASN under a given MSS in clinical trials with two co-primary endpoints with varying clinically meaningful standardized mean differences (Δ_1^*, Δ_2^*), the number of planned analyses L and true correlation $\rho_{12}^{\#}$, where $K = 2$. The given MSS per intervention group (equally sized groups: $r = 1$) is calculated to detect a joint effect on the two endpoints with a power of 80% or 90% using a one-sided test at the significance level of α = 2.5%, based on the clinically meaningful standardized mean difference (Δ_1^*, Δ_2^*) and zero correlation, where the decision-making framework is to reject H_0 at the same analysis for both endpoints. The critical boundaries for both endpoints are determined, using the O'Brien-Fleming function based on the Lan-DeMets error spending method with equal information space. The ASN is calculated under H_1. For example, in a fixed-sample design, 516 participates per intervention group are required to detect a joint effect on both endpoints assuming (Δ_1^*, Δ_2^*) =(0.2, 0.2) with correlation ρ_{12}^* =0.0, at the power of 80% and a significance level of 2.5% by one-sided test. When planning two analyses in a group-sequential design with the same designs parameter configuration, the MSS is 518. Under this MSS, the ASN are 502, 494, 488, 475 and 459 corresponding to true correlations $\rho_{12}^{\#}$ =0.0, 0.3, 0.5, 0.8 and 0.99 respectively and smaller than the fixed sample size. The relative ratio of ASN to fixed sample size are from 3% to 11%. When planning more than two analyses, the ASN is much smaller than the fixed sample size. If four analyses are planned, the ASNs are 459, 449, 442, 428 and 410 corresponding to true correlations $\rho_{12}^{\#}$ =0.0, 0.3, 0.5, 0.8 and 0.99 respectively, and the relative ratio of ASN to fixed sample size are from 11% to 21%. Similar behavior is also observed with unequal standardized mean differences (Δ_1^*, Δ_2^*) =(0.3, 0.2) and (0.4, 0.2). There, if assuming zero correlations among the endpoints in group-sequential designs, careful consideration is required regarding how to deal with the number of planned analyses to reduce the ANS.

However, assuming zero correlation is conservative when there is concrete evidence of higher correlations. In this situation, one approach is to use the confidence limit method discussed in Tamhane et al. (2012), which takes sampling error associated with the correlations into account by using of the upper confidence limit of the correlation. Recently Kunz et al. (2017) considered methods for incorporating the observed correlations among the endpoints into interim decision-making in clinical trials with multiple endpoints in blinded and unblinded settings and evaluated several types of correlation estimators in terms of expected value and mean square error. When standardized mean differences are unequal among the endpoints, the power is not improved with larger correlation. As discussed in Asakura et al. (2014), with unequal standardized mean differences, incorporating the correlation into the sample size calculation at the planning or interim stages may offer no advantage. Careful consideration is required regarding how to deal with correlations among the endpoints in designing clinical trials with multiple endpoints.

When constructing efficient group-sequential designs in CPE clinical trials, another important decision is the choice of the critical boundary based on an error-spending method for each endpoint. Although for illustrative objectives, the same critical boundaries were selected for both endpoints in Section 4. Different critical boundaries can be considered. If the trial was designed to detect effects on at least one endpoint with a prespecified ordering of endpoints, then the selection of different boundaries for each endpoint (i.e., the O'Brien-Fleming-type boundary for the primary endpoint and the Pocock type boundary for the secondary endpoint) can provide higher power than using the same critical boundary for both endpoints (Glimm et al. 2010; Tamhane et al., 2010). However, as Hung et al. (2016) noted, in cardiovascular clinical trials where the primary endpoint is a composite of major adverse cardiac events (MACE) including all-cause death, myocardial infraction, and stroke, and the secondary endpoint is all-cause death, the choice of the Pocock type boundary for the secondary endpoint is impractical as a larger sample sizes (or greater number of events) may be required to detect an effect on all-cause death.

On the other hand, as shown in Asakura et al. (2014) and Hamasaki et al. (2015, 2017), the selection of a different critical boundary has a minimal effect on the power, MSS and ASN. In the decision-making frameworks described in Section 4.2, regardless of equal or unequal standardized mean difference among the endpoints, the largest power is obtained from the O'Brien-Fleming boundary for all of the endpoints, and the lowest from the Pocock-type boundary for all of the endpoints. Regarding the ASN, the smallest is provided by the Pocock-type boundary for all of the endpoints while the largest is provided by the O'Brien-Fleming boundary. One possible scenario for selecting different boundaries is when one endpoint is invasive or costly, and stopping measurement of that endpoint is desirable as soon as possible, e.g., once the superiority for the endpoint has been demonstrated.

6 Additional issues

We have reviewed methods recently developed for the design, data monitoring, and analyses of clinical trials with CPE in fixed-sample and group-sequential settings. In this final section, we briefly discuss further developments with other design characteristics, including: (i) more than two intervention groups; (ii) group-sequential designs with multiple time-to-event endpoints; (iii) enrichment designs and subgroup analysis, and (iv) multi-regional clinical trials.

More than two intervention groups:

In clinical trials with multiple intervention arms, clarification of the trial objective is paramount. Objectives may include evaluating if all of the interventions are superior (or non-inferior) to a control or if at least one intervention is superior (or non-inferior) to a control. For the latter objective, methods for group-sequential and modern adaptive designs for multiple intervention arms have been discussed (e.g., Thall et al. 1989; Follmann et al. 1994; Stallard and Todd 2003, 2008; König et al. 2008; Magirr et al. 2012). Further investigation is needed for group-sequential and adaptive designs in more complex clinical trial settings, e.g., multiple intervention arms with CPE and targeted subpopulations.

Group sequential designs with multiple time-to-event endpoints:

When extending the methods for fixed-sample designs in Sugimoto et al. (2013, 2017) to a group-sequential setting, a complex issue is how to allocate the significance level to each interim analysis between the two endpoints as the amount of information for the endpoints may vary at a particular interim time-point of the trial. One strategy is to allocate the significance level, assuming the same information between the two endpoints based on one of the endpoints, even though they may never be at the same at the interim time-point. Another strategy is to calculate the required maximum number of events, or sample size and determine the timing of interim analyses based on one of the endpoints, and then calculate the maximum number of events required for the other endpoint under this sample size, and the information corresponding the timing of interim analyses for one endpoint, ignoring the relationship among the endpoints. However, when one event is fatal and the other is non-fatal, then the fatal event may impact the information for the nonfatal event as the fatal event censors the non-fatal event. In addition, the required maximum number of events for both endpoints may not be achieved simultaneously, potentially resulting in a situation where the maximum required number of events for one endpoint is observed at a particular time point, but not for the other endpoint. Here, there may be the two options: (i) stop the trial even though the required maximum number of events for one of the endpoints has not yet been observed, or (ii) continue the trial until the required maximum number of events for both of the endpoints are observed. For (ii), the observed number of events may be larger than that required. Here the final critical boundary for the endpoint with the larger number of observed events must be recalculated to control the Type I error rate based on the observed events.

Enrichment designs and subgroup analysis:

When a disease is heterogeneous or the intervention can target a specific mechanism of action related to disease subtypes, use of conventional clinical trial design may not suffice. Conventional trials generally assume homogeneous treatment effect for all participants in the trial. When markers can precisely identify individuals with a high probability of response to an intervention, clinical trials could focus on such individuals. Conducting a trial in subgroup patients with a potentially high response is termed “enrichment.” Advantages of enrichment designs include: increasing the chance of success often with a smaller sample size, directing treatment where it is likely to work best, and avoiding unnecessary harm. In enrichment designs, the statistical challenging task is identifying and confirming that a subgroup of patients with a positive benefit: risk balance when treated with an intervention (Ondra et al. 2015).

Subgroup analyses are common in clinical trials. However, the quality and level of evidence as well as the strength of conclusions regarding a subgroup-specific intervention depends on many factors including the trial design and conduct, and the reliability and predictive ability of the biomarker that defines the subgroup. Wang and Hung (2014) provide a list of criteria for consideration that may affect interpretability of subgroup-specific findings. If participants with and without an enrichment characteristic are studied, then the primary result may be driven by the result in the enriched subgroup. In some enrichment designs that recruit participants with and without the enrichment characteristic, the trial-wise Type I error rate can be shared between a test conducted using only the enriched subgroup and a test conducted using the entire population. The Type I error allocation scheme allows for the assessment of the intervention effect in the entire entered population when there may be some effect in the non-enriched subgroup while also allowing assessment in the enriched subgroup. Determining the required sample size that will provide reasonable power to test these hypotheses while controlling the Type I error including a prespecified order of testing or a multiple testing procedure is challenging. Statistical methods have been discussed and recent developments in the statistical literature regarding identification and confirmation of targeted subgroups can be found in the Journal of Biopharmaceutical Statistics Special Issue, “Subgroup Analysis in Clinical Trials” (2014). In addition, Ondra et al. (2015) provide a systematic review.

Multi-regional clinical trials:

Recently, clinical trials across multiple regions of the world, socalled ‘multiregional clinical trials’ (MRCTs), have become a common practice in the shift to the simultaneous development of drugs on a global scale (Ando and Hamasaki, 2010). The use of MRCTs in drug development will be of great benefit to the creation of solid evidence regarding the safety and efficacy of drugs, to more efficient and cost-effective drug development, and to a resolution of the drug lag with simultaneous worldwide registration, where drug lag means, for example, circumstances in which drugs already approved in the European Union (EU), United States (US) or other regions have not yet been approved and have not been made available to patients in e.g., Japan over a long period of time. However, such trials present considerable challenges as far as quality, design, implementation, analysis, and interpretation are concerned (please see ICH E17 Guideline (2016)). A key issue is sample size determination and evaluation of the consistency of the interventions’ effect among the regions participated in the MRCT. For example, Doody et al (2013) reported a randomized controlled clinical trial to evaluate an effect of semagacestat 100 mg and 140 mg compared to placebo in in patients with probable Alzheimer’s disease, and 19 countries participated the trial. The primary endpoints were (1) Changes in cognition from baseline to week 76 assessed by the cognitive subscale of ADAS-cog, (2) changes in functioning assessed by ADCS-ADL. A sample size of 500 participants per intervention group (1500 participants in total) will have 89% power to detect a difference of 1.8 points change in ADAS-Cog and 98% power to detect a difference of 3.1 points change in ADCS-ADL between the treatment groups after 18 months of treatment. This trial also planned to have an interim analysis of the primary endpoints after 50% of patients had completed 12 months of treatment or had dropped out of the study. Sample size determination and evaluation of consistency in intervention’s effects based on multiple endpoint are more complex in a fixed-sample or group-sequential setting. Huang et al. (2017) discussed sample size determination for a specific region in MRCTs with CPE and evaluated three consistency criteria for evaluating the intervention’s effects on two endpoints in fixed-sample designs. Further investigation is needed for in more complex MRCTs with multiple endpoints in a fixed-sample and group-sequential/adaptive designs.

Acknowledgments:

The authors are indebted to Dr. Alex Dmitrienko for inviting us to contribute this paper to the Journal of Biopharmaceutical Statistics. Research reported in this publication was partially supported by the Japan Agency for Medical Research and Development under the Project Promoting Clinical Trials for Development of New Drugs 17lk0201061h0002 and 17lk0201061h0202, and the National Institute of Allergy and Infectious Diseases under

Award Numbers UM1AI104681 and UM1AI068634. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Reference

1. Alosh M, Bretz F, Huque M (2014). Advanced multiplicity adjustment methods in clinical trials. Statistics in Medicine 33:693–713. DOI: 10.1002/sim.5974 [PubMed: 24105821]
2. American College of Gastroenterology. (2013). Understanding Irritable Bowel Syndrome. available at www.patients.gi.org/gi-health-and-disease/understanding-irritable-bowel-syndromeleaving site icon
3. Ando Y, Hamasaki T. (2010). Practical issues and lessons learned from multi-regional clinical trials via case examples: a Japanese perspective. Pharmaceutical Statistics 9: 190–200. DOI: 10.1002/pst.448 [PubMed: 20737442]
4. Ando Y, Hamasaki T, Asakura K, Evans SR, Sugimoto T, Sozu T, Ohlabel Y (2015). Sample size considerations in clinical trials when comparing two interventions using multiple co-primary binary relative risk contrasts. Statistics in Biopharmaceutical Research 7:81–94. DOI: 10.1080/19466315.2015.1006373 [PubMed: 26167243]
5. Asakura K. Evans SR. Hamasaki T. Interim evaluation of futility in clinical trials with co-primary endpoints; The Joint Conference on Biometrics & Biopharmaceutical Statistics; Vienna, Austria. August 28-September 1, 2017; 2017.
6. Asakura K, Hamasaki T, Evans SR (2017). Interim evaluation of efficacy or futility in group-sequential trials with multiple co-primary endpoints. Biometrical Journal 59, 703–731. DOI: 10.1002/bimj.201600026 [PubMed: 27757980]
7. Asakura K, Hamasaki T, Evans SR, Sugimoto T, Sozu T (2015). Group-sequential designs when considering two binary outcomes as co-primary endpoints. In Applied Statistics in Biomedicine and Clinical Trials Design, Chen Z, Liu A, Qu Y, Tang L, Ting N, Tsong Y (eds.), Chap. 14, 235–262, Springer DOI: 10.1007/978-3-319-12694-4_14
8. Asakura K, Hamasaki T, Sugimoto T, Hayashi K, Evans SR, Sozu T (2014). Sample size determination in group-sequential clinical trials with two co-primary endpoints. Statistics in Medicine 33:2897–2913. DOI: 10.1002/sim.6154 [PubMed: 24676799]
9. Berger RL (1982). Multiparameter hypothesis testing and acceptance sampling. Techlabelmetrics 24:295–300.DOI: 10.2307/1267823
10. Cheng Y, Ray S, Chang M, Melabeln S (2014). Statistical monitoring of clinical trials with multiple co-primary endpoints using multivariate B-value. Statistics in Biopharmaceutical Research 6:241–250. DOI: 10.1080/19466315.2014.923324
11. Cohen A, Gatsonis C, Marden JI (1983). Hypothesis testing marginal probabilities in a 2×2×2 contingency tables with conditional independence. Journal of the American Statistical Association 78:920–929. DOI: 10.1080/01621459.1983.10477041
12. Committee for Medicinal Products for Human Use s (CHMP) (2008) Guideline on medicinal products for the treatment Alzheimer’s disease and other dementias (CPMP/EWP/553/95 Rev.1). European Medicines Agency, London, UK.
13. Committee for Human Medicinal Products (CHMP) (2016). Guideline on multiplicity issues in clinical trials (Draft) (EMA/CHMP/44762/2017). European Medicines Agency, London, UK.
14. Cook RJ, Farewell VT (1994). Guideline for monitoring efficacy and toxicity responses in clinical trials. Biometrics 50:1146–1162. DOI: 10.2307/2533451 [PubMed: 7786995]
15. Chuang-Stein C, Li J (2017). Changes are still needed on multiple co-primary endpoints. Statistics in Medicine (First published online on 19 July 2017 as DOI: 10.1002/sim.7383).
16. Chuang-Stein C, Stryszak P, Dmitrienko A, Offen W (2007). Challenge of multiple co-primary endpoints: a new approach. Statistics in Medicine 26:1181–1192. 10.1002/sim.2604 [PubMed: 16927251]
17. Delorme P, Lafaye MP, Liquet B, Riou J (2016). Type-II generalized family-wise error rate formulas with application to sample size determination. Statistics in Medicine 35, 2687–2714. DOI: 10.1002/sim.6909 [PubMed: 26914402]

18. DeMets DL, Ware JH (1980). Group sequential methods for clinical trials with one-sided hypothesis. Biometrika 67:651–660. DOI: 10.1093/biomet/67.3.651

19. DeMets DL, Ware JH (1982). Asymmetric group sequential boundaries for monitoring clinical trials. Biometrika 69:661–663. DOI: 10.1093/biomet/69.3.661

20. Dmitrienko A, D'Agostilabel RB (2013). Traditional multiplicity adjustment methods in clinical trials. Statistics in Medicine 32: 5172–5218. DOI: 10.1002/sim.5990 [PubMed: 24114861]

21. Dmitrienko A, D'Agostilabel RB, Huque MF (2013). Key multiplicity issues in clinical drug development. Statistics in Medicine 32:1079–1111. DOI: 10.1002/sim.5642 [PubMed: 23044723]

22. Dmitrienko A, Tamhane AC, Bretz F (2010). Multiple Testing Problems in Pharmaceutical Statistics. Chapman & Hall/CRC, Boca Raton

23. Eaton ML, Muirhead RJ (2007). On a multiple endpoints testing problem. Journal of Statistical Planning and Inference 137:3416–3429. DOI: 10.1016/j.jspi.2007.03.021

24. Evans SR, Li L, Wei LJ (2007). Data monitoring in clinical trials using prediction. Drug Information Journal 41: 733–742. DOI: 10.1177/009286150704100606

25. Evans SR, Ting N (2015). Fundamental Concepts for New Clinical Trialists. Chapman & Hall.

26. Follmann DA, Proschan MA, Geller NL (1994) Monitoring pairwise comparisons in multi-armed clinical trials. Biometrics 50:226–325. DOI: 10.2307/2533376

27. Food and Drug Administration (2013). Guidance for industry: Alzheimer's disease: developing drugs for the treatment of early stage disease. U.S. Department of Health and Human Services Food and Drug Administration, Rockville, MD, USA.

28. Food and Drug Administration (2017). Guidance for Industry: Multiple Endpoints in Clinical Trials. U.S. Department of Health and Human Services Food and Drug Administration, Rockville, MD, USA.

29. Gleser LJ (1973). On a theory of intersection-union tests. Institute of Mathematical Statistics Bulletin 2:233.

30. Glimm E, Maurer W, Bretz F (2010). Hierarchical testing of multiple endpoints in group-sequential trials. Statistics in Medicine 29:219–228. DOI: 10.1002/sim.3748 [PubMed: 19827011]

31. Grundmann O, and Yoon SL (2010). Irritable bowel syndrome: epidemiology, diaglabelsis, and treatment: an update for healthcare practitioners. Journal of Gastroenterology and Hepatology 25: 691–699 [PubMed: 20074154]

32. Gould AL, Pecore VJ (1982). Group sequential methods for clinical trials allowing early acceptance of H0 and incorporating costs. Biometrika 69:75–80. DOI: 10.1093/biomet/69.1.75

33. Hamasaki T, Asakura K, Ochiai T, Evans SR (2016). Group-Sequential Clinical Trials with Multiple Co-Objectives. Cham/Heidelberg/ New York: Springer DOI: 10.1007/978-4-431-55900-9

34. Hamasaki T, Asakura K, Evans SR, Sugimoto T, Sozu T (2015). Group sequential strategies for clinical trials with multiple co-primary endpoints. Statistics in Biopharmaceutical Research 7:36–54, 2015. DOI: 10.1080/19466315.2014.1003090 [PubMed: 25844122]

35. Hamasaki T, Sugimoto T, Evans SR, Sozu T (2013). Sample size determination for clinical trials with co-primary outcomes. Exponential event-times. Pharmaceutical Statistics 12:28–34. DOI: 10.1002/pst.1545 [PubMed: 23081932]

36. Huang WS, Hung HN, Hamasaki T, Hsiao CF (2017). Sample size determination for a specific region in multiregional clinical trials with multiple co-primary endpoints. PLoS ONE 12(6):e0180405 DOI: 10.1371/journal.pone.0180405 [PubMed: 28665972]

37. Hung HMJ, Wang SJ (2009). Some controversial multiple testing problems in regulatory applications. Journal of Biopharmaceutical Statistics 19:1–11.DOI: 10.1080/10543400802541693 [PubMed: 19127460]

38. Hung HMJ, Wang SJ (2010). Challenges to multiple testing in clinical trials. Biometrical Journal 52:747–756. DOI: 10.1002/bimj.200900206 [PubMed: 20589856]

39. Hung HMJ, Wang SJ, Yang P, Jin K, Lawrence J, Kordzakhia G, Massie T (2016). Statistical challenges in regulatory review of cardiovascular and CNS clinical trials. Journal of Biopharmaceutical Statistics 26:37–43. DOI:10.1080/10543406.2015.1092025 [PubMed: 26366624]

40. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) (1998). ICH harmonised tripartite guideline E9: statistical principles for clinical trials. February 1998.

41. International Council for Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) (2016). ICH E17: General Principles for Planning and Design of Multi-Regional Clinical Trials, step 2, May 2016.

42. Jennison C, Turnbull BW (1993). Group sequential tests for bivariate response: interim analyses of clinical trials with both efficacy and safety. Biometrics 49:741–752. DOI: 10.2307/2532195 [PubMed: 8241370]

43. Julious SA, McIntyeare NE (2012). Sample sizes for trials involumeving multiple correlated must-win comparisons. Pharmaceutical Statistics 11:177–185. DOI: 10.1002/pst.515 [PubMed: 22383136]

44. König F, Brannath W, Bretz F, Posch M (2008). Adaptive Dunnett tests for treatment selection. Statistics in Medicine 27:1612–1625. DOI: 10.1002/sim.3048 [PubMed: 17876763]

45. Kordzakhia G, Siddiqui O, Huque MF (2010). Method of balanced adjustment in testing co-primary endpoints. Statistics in Medicine 29:2055–2066. DOI: 10.1002/sim.3950 [PubMed: 20683896]

46. Kunz CU, Stallard N, Parsons N, Todd S Friede T (2017). Blinded versus unblinded estimation of a correlation coefficient to inform interim design adaptations. Biometrical Journal 59: 344–357. DOI: 10.1002/bimj.201500233 [PubMed: 27886393]

47. Lachin JM (2005). A review of methods for futility stopping based on conditional power. Statistics in Medicine 24:2747–2764. DOI: 10.1002/sim.2151 [PubMed: 16134130]

48. Lan KKG, DeMets DL (1983). Discrete sequential boundaries for clinical trials. Biometrika 70:659–663. DOI:10.1093/biomet/70.3.659

49. Lan KKG, Simon R, Halperin M (1982). Stochastically curtailed tests in long-term clinical trials. Communications in Statistics: Theory and Methods 1:207–219. DOI: 10.1080/07474948208836014

50. Lafaye MP, Liquet B, Marque S, Riou J (2014). Power and sample size determination in clinical trials with multiple primary continuous correlated endpoints. Journal of Biopharmaceutical Statistics 24:378–97. DOI: 10.1080/10543406.2013.860156 [PubMed: 24605975]

51. Laska EM, Meisner M (1989). Testing whether an identified treatment is best. Biometrics 45, 1139–1151. DOI: 10.2307/2531766 [PubMed: 2611321]

52. Lehmann EL (1952). Testing multiparameter hypotheses. Annals of Mathematical Statistics 23:541–552.

53. Li QH (2009). Evaluating co-primary endpoints collectively in clinical trials. Biometrical Journal 51:137–45.DOI: 10.1002/bimj.200710497 [PubMed: 19219905]

54. Li L, Evans SR, Ulabel H, Wei LJ (2009). Predicted interval plots (PIPS): A graphical tool for data monitoring of clinical trials. Statistics in Biopharmaceutical Research 1: 348–355. DOI: 10.1198/sbr.2009.0041 [PubMed: 21423789]

55. Magirr D, Jaki T, Whitehead J (2012). A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. Biometrika 99:494–501. DOI: 10.1093/biomet/ass002

56. 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 (2007). Multiple co-primary endpoints: medical and statistical solutions. Drug Information Journal 41:31–46. DOI: 10.1177/009286150704100105

57. Ondra T, Dmitrienko A, Friede T, Graf A, Miller F, Stallard N, Posch M (2016). Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. Journal of Biopharmaceutical Statistics 26:99–119. DOI: 10.1080/10543406.2015.1092034 [PubMed: 26378339]

58. Patel HI (1991). Comparison of treatments in a combination therapy trial. Journal of Biopharmaceutical Statistics 1:171–183. 10.1080/10543409108835016 [PubMed: 1844694]

59. Ristl R, Frommlet F, Koch A, Posch M (2016). Fallback tests for co-primary endpoints. Statistics in Medicine 35, 2669–2686. DOI: 10.1002/sim.6911 [PubMed: 26919166]

60. Roy SN (1953). On a heuristic method of test construction and its use in multivariate analysis. Annals of Mathematical Statistics 24:220–238

61. Sarkar SK, Snapinn S, Wang W (1995). On improving the min test for the analysis of combination drug trials. Journal of Statistical Computation and Simulation 51, 197–213. DOI: 10.1080/00949659508811632

62. Schüer S, Kiese M, Rauch G (2017). Choice of futility boundaries for group sequential designs with two endpoints. BMC Medical Research Methodology 17:119 DOI: 10.1186/s12874-017-0387-4 [PubMed: 28789615]

63. Snapinn S, Chen MG, Jiang Q and Koutsoukos T (2006). Assessment of futility in clinical trials. Pharmaceutical Statistics 5:273–281. DOI: 10.1002/pst.216 [PubMed: 17128426]

64. Senn S, Bretz F (2007). Power and sample size when multiple endpoints are considered. Pharmaceutical Statistics 6:161–170. 10.1002/pst.301 [PubMed: 17674404]

65. Song JX (2009). Sample size for simultaneous testing of rate differences in labelninferiority trials with multiple endpoints. Computational Statistics and Data Analysis 53:1201–1207. DOI: 10.1016/j.csda.2008.10.028

66. Song JX (2015). A two-stage design with two co-primary endpoints. Contemporary Clinical Trials Communications 1:2–4. DOI: 10.1016/j.conctc.2015.08.002 [PubMed: 29736433]

67. Sozu T, Kalabelu T, Hamada C, Yoshimura I (2006). Power and sample size calculations in clinical trials with multiple primary variables. Japanese Journal of Biometrics 27:83–96. DOI: 10.5691/jjb.27.83

68. Sozu T, Sugimoto T, Hamasaki T (2010). Sample size determination in clinical trials with multiple co-primary binary endpoints. Statistics in Medicine 29:2169–2179. DOI: 10.1002/sim.3972. [PubMed: 20687162]

69. Sozu T, Sugimoto T, Hamasaki T (2011). Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints. Journal of Biopharmaceutical Statistics 21:650–668.DOI: 10.1080/10543406.2011.551329. [PubMed: 21516562]

70. Sozu T, Sugimoto T, Hamasaki T (2012). Sample size determination in clinical trials with multiple co-primary endpoints including mixed continuous and binary variables. Biometrical Journal 54:716–729. DOI: 10.1002/bimj.201100221 [PubMed: 22829198]

71. Sozu T, Sugimoto T, Hamasaki T (2016). Reducing unnecessary measurements in clinical trials with multiple primary endpoints. Journal of Biopharmaceutical Statistics 26:631–643. DOI: 10.1080/10543406.2015.1052497 [PubMed: 26098617]

72. Sozu T, Sugimoto T, Hamasaki T, Evans SR (2015). Sample Size Determination in Clinical Trials with Multiple Endpoints. Cham/Heidelberg/ New York: Springer DOI: 10.1007/978-3-319-22005-5

73. Sugimoto T, Hamasaki T, Evans SR, Sozu T (2017). Sizing clinical trials when comparing bivariate time-to-event outcomes. Statistics in Medicine 36:1363–1382. DOI: 10.1002/sim.7225 [PubMed: 28120524]

74. Sugimoto T, Sozu T, Hamasaki T (2012). A convenient formula for sample size calculations in clinical trials with multiple co-primary continuous endpoints. Pharmaceutical Statistics 11:118–128. DOI: 10.1002/pst.505 [PubMed: 22415870]

75. Sugimoto T, Sozu T, Hamasaki T, Evans SR (2013). A logrank test-based method for sizing clinical trials with two co-primary time-to-events endpoints. Biostatistics 14:409–421. DOI: 10.1093/biostatistics/kxs057 [PubMed: 23307913]

76. Stallard N, Todd S (2003). Sequential designs for phase III clinical trials incorporating treatment selection. Statistics in Medicine 22:689–703. DOI: 10.1002/sim.1362 [PubMed: 12587100]

77. Stallard N, Todd S (2008). A group-sequential design for clinical trials with treatment selection. Statistics in Medicine 27:6209–6227. DOI: 10.1002/sim.3436 [PubMed: 18792085]

78. Stallard N, Hamborg N, Parsons N, Friede T (2014) Adaptive designs for confirmatory clinical trials with subgroup selection. Journal of Biopharmaceutical Statistics 24:168–187. DOI: 10.1080/10543406.2013.857238 [PubMed: 24392984]

79. Tamhane AC, Mehta CR, Liu L (2010). Testing a primary and secondary endpoint in a group sequential design. Biometrics 66:1174–1184. DOI: 10.1111/j.1541-0420.2010.01402.x [PubMed: 20337631]

80. Thall PF, Simon R, Ellenberg SS (1989) A two-stage design for choosing among several experimental treatments and a control in clinical trial. Biometrics 45:537–547. DOI: 10.2307/2531495 [PubMed: 2765637]

81. Varga S, Tsang YC, Singer J (2017). A simple procedure to estimate the optimal sample size in case of conjunctive coprimary endpoints. Biometrical Journal 59:626–635. DOI: 10.1002/bimj.201500231 [PubMed: 27346828]

82. Ware JH, Muller JE, Braunwald E (1985).The futility index: an approach to the cost-effective termination of randomized clinical trials. American Journal of Medicine 78:635–643. DOI: 10.1016/0002-9343(85)90407-3 [PubMed: 3920906]

83. Wang SJ, Hung HMJ (2014) A regulatory perspective on essential considerations in design and analysis of subgroups when correctly classified. Journal of Biopharmaceutical Statistics 24:19–41.DOI: 10.1080/10543406.2013.856022 [PubMed: 24392976]

84. Whitehead J, Matsushita T (2003). Stopping clinical trials because of treatment ineffectiveness: a comparison of a futility design with a method of stochastic curtailment. Statistics in Medicine 22:677–687. DOI: 10.1002/sim.1429 [PubMed: 12587099]

85. Wiens B (2003). A fixed sequence Bonferroni procedure for testing multiple endpoints. Pharmaceutical Statistics 2:211–215. DOI: 10.1002/pst.64

86. Wiens BL, Dmitrienko A (2005). The fallback procedure for evaluating a single family of hypotheses. Journal of Biopharmaceutical Statistics 15:929–942. DOI: 10.1080/10543400500265660 [PubMed: 16279352]

87. 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. Clinical Trials 2:387–393. DOI: 10.1191/1740774505cn112oa [PubMed: 16317808]

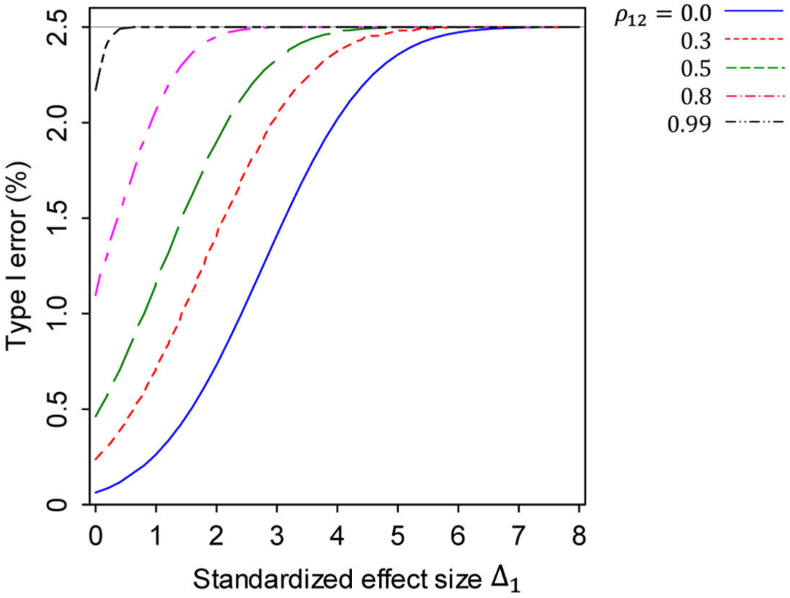

Fig. 1.

Behavior of the Type I error for α =2.5% as a function of the standardized mean effect size Δ_1 and the correlation ρ_{12} for H$_0$ when Δ_2 = 0, where K = 2.

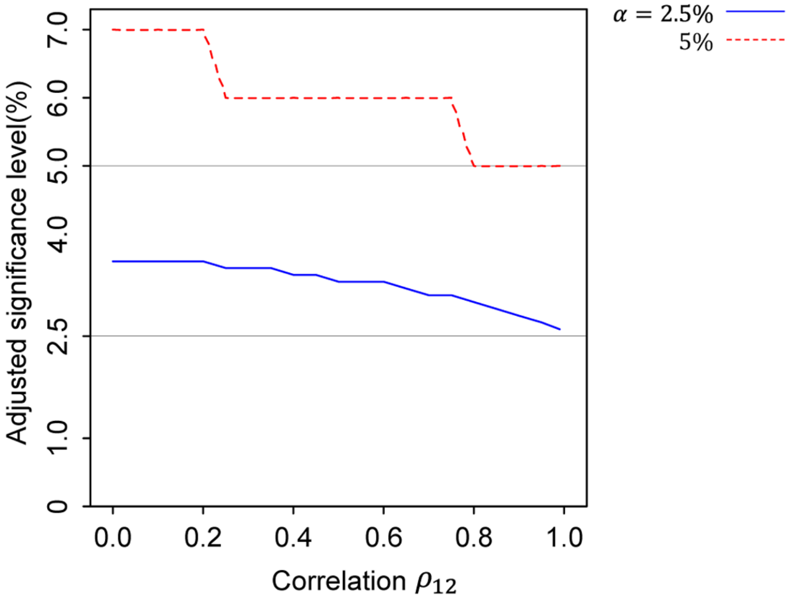

Fig. 2.

Behavior of the adjusted significance level α^* for α =2.5% and α =5.0% as a function of the correlation ρ_{12}, where $K = 2$.

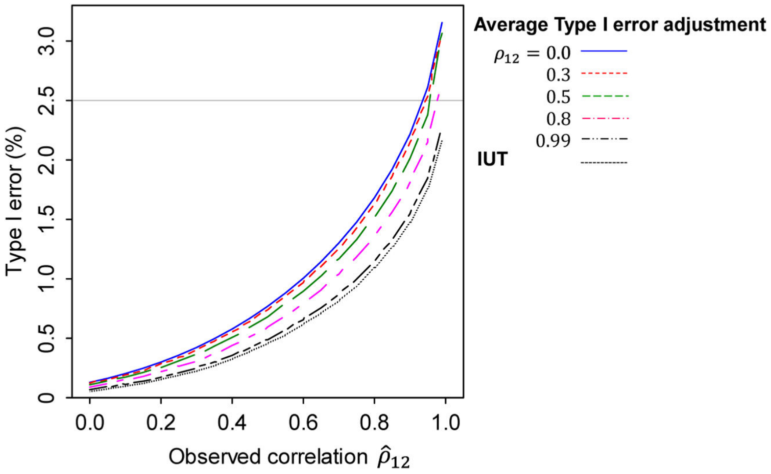

Fig. 3.

Behavior of the Type I error for α^* as a function of the observed correlation $\hat{\rho}_{12}$ for H_0 when $\Delta_1 = \Delta_2 = 0$, where $K = 2$ and $\alpha = 2.5\%$. The adjusted significance levels using the Average Type I error method are α^* =3.6%, 3.5%, 3.3%, 3.0% and 2.6%, corresponding to ρ_{12} =0.0, 0.3, 0.5, 0.8 and 0.99.

Table 1.

Comparison of multiple primary endpoints and co-primary endpoints in clinical trials

	Multiple primary endpoints (MPC)	**Multiple co-primary endpoints (CPE)**
Study objective	To evaluate whether a test intervention has an effect on *at least one* of the primary endpoints.	To evaluate whether a test intervention has an effect on *all* of the primary endpoints
Decision-making criterion	Failure to demonstrate an effect on *all* endpoints implies that effects cannot be concluded.	Failure to demonstrate an effect on *any single* endpoint implies that effects cannot be concluded.
Hypothesis testing principle	Union-Intersection Principle	Intersection–Union Principle
Null hypothesis H_0 (Number of endpoints: k =1, …, K)	An intersection of a family of hypotheses H_{01}, …, H_{0k}, $H_0 = \cap_{k=1}^{K} H_{0k}$	A union of a family of hypotheses H_{01}, …, H_{0k}, $H_0 = \cup_{k=1}^{K} H_{0k}$
Test statistics for rejecting H_0	$Z = \max_{k=1,\ldots,K} Z_k$	$Z = \min_{k=1,\ldots,K} Z_k$
Type I error control	Required: the Type I error increases as the number of endpoints to be tested is increased	Not required
Type II error control	Not required	Required: the Type II error increases as the number of endpoints to be tested is increased
Rejection region for the testing of H_0 (Rejection region for each endpoint R_k)	$\cup_{k=1}^{K} \left\{Z_k \in R_k\right\}$ H_0 is rejected if and only if *at least one* of the hypotheses H_{0k} is rejected.	$\cap_{k=1}^{K} \left\{Z_k \in R_k\right\}$ H_0 is rejected if and only if *all* hypotheses H_{0k} are rejected.

Table 2.

Sample size per intervention group (equally-sized group: $r = 1$) for two co-primary endpoints, using the conventional method (Conventional), average Type I error method (Average) and balanced adjustment method (Balanced) with varying clinically meaningful standardized mean differences (Δ_1^*, Δ_2^*) and correlation ρ_{12}, where the power $1 - \beta = 80\%$ and 90%, and the significance level $\alpha = 2.5\%$

			ρ_{12}				
$1 - \beta$	**(Δ_1^*, Δ_2^*)**	**Methods**	**0.0**	**0.3**	**0.5**	**0.8**	**0.99**
80%	(0.2, 0.2)	Conventional	516	503	490	458	409
		Average	465	457	453	435	404
		Balanced	470	463	458	439	405
	(0.3, 0.2)	Conventional	402	399	397	393	393
		Average	360	360	364	371	388
		Balanced	361	361	365	372	388
	(0.4, 0.2)	Conventional	393	393	393	393	393
		Average	349	353	360	371	388
		Balanced	349	353	360	371	388
90%	(0.2, 0.2)	Conventional	646	637	626	597	544
		Average	589	585	584	570	538
		Balanced	593	590	589	574	539
	(0.3, 0.2)	Conventional	529	528	527	526	526
		Average	479	482	489	501	520
		Balanced	479	482	489	501	520
	(0.4, 0.2)	Conventional	526	526	526	526	526
		Average	475	479	487	501	520
		Balanced	475	479	487	501	520

Table 3.

The MSS and ASN per intervention group (equally-sized group: $r = 1$) in clinical trials with two co-primary endpoints ($K = 2$), with varying clinically meaningful standardized mean differences (Δ_1^*, Δ_2^*), correlation ρ_{12}, and number of analyses L, where the power $1 - \beta$ =80% and 90%, and the significance level α =2.5%. Two decision-making frameworks are considered: (1) reject H_0 at any analysis, and (2) reject H_0 at a same interim simultaneously, The critical boundaries for both endpoints are determine, using the O'Brien-Fleming-type function based on the Lan-DeMets error spending method with equal information space. The ASN is calculated under H_1.

Decision-making framework				**MSS/ASN per intervention group**				
	$1 - \beta$	**(Δ_1^*, Δ_2^*)**	***L***	**ρ_{12} =0.0**	**0.3**	**0.5**	**0.8**	**0.99**
(1) Any analysis	80%	(0.2, 0.2)	2	518/502	505/483	492/466	460/429	410/377
			3	522/469	509/451	496/436	464/402	414/355
			4	525/457	512/439	499/423	467/390	417/344
			5	528/449	515/432	502/417	470/384	419/337
		(0.3, 0.2)	2	403/385	401/378	398/371	395/364	394/362
			3	407/358	404/352	402/347	398/341	398/340
			4	410/349	407/342	404/337	401/331	401/330
			5	412/343	409/336	406/331	403/325	403/324
		(0.4, 0.2)	2	395/368	394/364	394/362	394/362	394/362
			3	398/343	398/341	398/341	398/340	398/340
			4	401/333	401/331	401/330	401/330	401/330
			5	403/327	403/325	403/324	403/324	403/324
	90%	(0.2, 0.2)	2	648/611	639/591	628/573	599/535	546/479
			3	653/555	644/539	633/525	604/493	550/442
			4	657/536	648/520	637/505	608/473	554/424
			5	660/524	651/508	640/494	610/462	556/414
		(0.3, 0.2)	2	531/484	530/475	529/469	528/462	528/461
			3	535/440	534/435	533/432	532/427	532/427
			4	539/424	538/418	537/414	536/410	536/409
			5	541/414	540/408	539/404	538/400	538/399
		(0.4, 0.2)	2	528/465	528/462	528/462	528/461	528/461
			3	532/430	532/428	532/427	532/427	532/427
			4	536/411	536/410	536/409	536/409	536/409
			5	538/402	538/400	538/399	538/399	538/399
(2) Same analysis	80%	(0.2, 0.2)	2	518/502	505/483	492/466	460/429	410/377
			3	524/471	510/452	497/437	465/403	414/355
			4	528/459	514/440	501/425	468/391	417/344
			5	530/451	517/433	503/417	470/384	419/337
		(0.3, 0.2)	2	404/386	401/378	398/371	398/371	395/364
			3	408/359	405/352	402/348	402/348	398/341
			4	411/350	407/342	405/337	405/337	401/331

Decision-making framework	**$1-\beta$**	**(Δ_1^*, Δ_2^*)**	**L**	**MSS/ASN per intervention group**				
				ρ_{12} =0.0	**0.3**	**0.5**	**0.8**	**0.99**
			5	413/344	410/337	407/331	407/331	403/325
		(0.4, 0.2)	2	394/362	395/368	395/365	394/362	394/362
			3	398/340	398/34	398/341	398/341	398/340
			4	401/330	401/333	401/331	401/330	401/330
			5	403/324	403/327	403/325	403/324	403/324
	90%	(0.2, 0.2)	2	648/611	639/591	628/573	599/535	546/479
			3	654/555	645/540	634/526	604/493	550/442
			4	659/538	649/521	638/506	608/473	554/424
			5	662/526	653/510	642/495	611/462	556/414
		(0.3, 0.2)	2	531/484	530/475	529/469	528/462	528/461
			3	535/440	534/435	533/432	532/427	532/427
			4	539/424	538/418	537/414	536/410	535/408
			5	541/415	540/409	539/404	538/400	538/399
		(0.4, 0.2)	2	528/465	528/462	528/462	528/461	528/461
			3	532/430	532/428	532/427	532/427	532/427
			4	536/411	536/410	535/409	535/408	535/408
			5	538/402	538/400	538/399	538/399	538/399

Table 4.

The ASN under a given MSS in clinical trials with two co-primary endpoints ($K = 2$), with varying clinically meaningful standardized mean differences (Δ_1^*, Δ_2^*) the number of planned analyses L and true correlation ρ_{12}^*, where $K = 2$. The given MSS per intervention group (equally sized groups: $r = 1$) is calculated to detect a joint effect on the two endpoints with a power of 80% or 90% using a one-sided test at the significance level of α = 2.5%, assuming standardized mean differences (Δ_1^*, Δ_2^*) and zero correlation where the decision-making framework is to reject H_0 at a same analysis for both endpoints simultaneously. The critical boundaries for both endpoints are determined using the O'Brien-Fleming-type function based on the Lan-DeMets error spending method with equal information space. The ASN is calculated under H_1.

				ASN per intervention group				
$1-\beta$	**(Δ_1^*, Δ_2^*)**	***L***	**MSS**	$\rho_{12}^* = 0.0$	**0.3**	**0.5**	**0.8**	**0.99**
80%	(0.2, 0.2)	1	516					
		2	518	502	494	488	475	459
		3	524	471	462	455	443	426
		4	528	459	449	442	428	410
		5	530	451	441	433	419	400
	(0.3, 0.2)	1	402					
		2	404	386	380	376	371	370
		3	408	359	354	352	348	347
		4	411	349	344	341	337	336
		5	413	343	338	335	331	330
	(0.4, 0.2)	1	393					
		2	394	367	364	362	362	362
		3	398	343	341	341	340	340
		4	401	333	331	330	330	330
		5	403	327	325	324	324	324
90%	(0.2, 0.2)	1	646					
		2	648	611	597	587	569	545
		3	654	555	545	538	522	502
		4	659	538	526	517	500	478
		5	662	526	514	505	487	465
	(0.3, 0.2)	1	529					
		2	531	484	476	470	464	463
		3	535	440	436	433	429	428
		4	539	424	419	415	411	411
		5	541	415	409	405	401	400
	(0.4, 0.2)	1	526					
		2	528	465	462	462	461	461
		3	532	430	428	427	427	427
		4	536	411	410	409	409	409

				ASN per intervention group				
$1 - \beta$	**(Δ_1^*, Δ_2^*)**	***L***	**MSS**	$\rho_{12}^* = 0.0$	**0.3**	**0.5**	**0.8**	**0.99**
		5	538	402	400	399	399	399