Published in final edited form as:

J Biopharm Stat. 2017 ; 27(1): 1–24. doi:10.1080/10543406.2016.1148710.

Group-sequential three-arm noninferiority clinical trial designs

Toshimitsu Ochiai^{1,2}, Toshimitsu Hamasaki^{2,3,4,*}, Scott R. Evans⁵, Koko Asakura^{3,4}, and Yuko Ohno²

¹Biostatistics Department, Shionogi & Co., Ltd

²Department of Mathematical Health Science, Osaka University Graduate School of Medicine, Japan

³Office of Biostatistics and Data Management, National Cerebral and Cardiovascular Center, Japan

⁴Department of Innovative Clinical Trials and Data Sciences, Osaka University Graduate School of Medicine, Japan

⁵Department of Biostatistics and the Center for Biostatistics in AIDS Research, Harvard T.H. Chan School of Public Health, USA

Abstract

We discuss group-sequential three-arm noninferiority clinical trial designs that include active and placebo controls for evaluating both assay sensitivity and noninferiority. We extend two existing approaches, the fixed margin and fraction approaches, into a group-sequential setting with two decision-making frameworks. We investigate the operating characteristics including power, Type I error rate, maximum and expected sample sizes, as design factors vary. In addition, we discuss sample size recalculation and its' impact on the power and Type I error rate via a simulation study.

Keywords

Average sample number; Assay sensitivity; Cui-Hung-Wang Statistics; Fixed margin approach; Fraction approach; Maximum sample size; Sample size recalculation; Type I error

1. Introduction

Active-controlled noninferiority trial designs are an alternative to placebo-controlled superiority designs when a use of the placebo control is ethically undesirable due to the availability of a proven effective medical intervention. Active-controlled noninferiority trial designs include an existing effective intervention such as an effective standard of care. In contrast to superiority trials where there is interest in evaluating if an intervention is superior to a control (e.g., placebo), noninferiority trials evaluate if an intervention is noninferior to the control. In a noninferiority trial, the null hypothesis of inferiority is assumed to be true unless there is sufficient data to reject it in favor of the alternative (noninferiority).

*Corresponding author: toshi.hamasaki@ncvc.go.jp, Office of Biostatistics and Data Management, National Cerebral and Cardiovascular Center, 5-7-1 Fujishirodai, Suita, Osaka 565-8565, Japan.

Noninferiority is assessed by evaluating whether inferiority of a pre-specified magnitude (called a noninferiority margin) can be ruled out with reasonable confidence using confidence intervals. The noninferiority margin is carefully selected to ensure that a noninferiority result would: (1) imply retention of the some of the effect that the active control has historically displayed (i.e., when compared to placebo), and (2) rule out clinically important levels of inferiority so that clinical application would be ethical and clinically acceptable.

For example, EMERALD 1 (conducted in the United States) and EMERALD 2 (conducted in Europe) are randomized, controlled, open-label, noninferiority clinical trials to evaluate the efficacy and safety of peginesatide as the maintenance treatment of anemia in patients with chronic renal failure who were receiving hemodialysis and previously treated with epoetin (Fishbane et al., 2013). Both trials included a 6-week screening period, a 28-week initial dose-adjustment period, an 8-week evaluation period, and a longer-term follow-up period (16 additional weeks). Eligible participants were randomly assigned, in a 2:1 ratio, to receive peginesatide once every 4 weeks or to continue to receive epoetin (epoetin alfa in the EMERALD 1, and epoetin beta in the EMERALD 2) one to three times a week. The frequency and route of administration of epoetin was determined based on the treatment regimen during the screening period. The primary efficacy endpoint was the change from the baseline hemoglobin level during the evaluation period. Noninferiority for both trials would be established if the lower limit of the two-sided 95% confidence interval was -1.0 g per deciliter or higher, indicated that inferiority of greater than -1.0 could be ruled out with reasonable confidence, compared to epoetin.

For noninferiority clinical trials to be valid, two assumptions (constancy and assay sensitivity) must be satisfied (D'Agostino, Massaro, Sullivan, 2003; ICH 2000; FDA, 2010). An active intervention which has been shown to be efficacious (e.g., superior to placebo) in a historical trial may be considered as the active control in a noninferiority trial but the most effective should be selected. The constancy assumption states that the demonstrated effect of the active control over placebo in the historical trial has not changed over time, i.e., would be the same as the effect in the current trial if a placebo group was included. This may not be the case if there were differences in trial conduct (e.g., differences in treatment administration, endpoints, or population) between the historical and current trials. This assumption is not testable in a trial without a concurrent placebo group.

Another important design assumption is assay sensitivity, i.e., the ability for the trial to be able to detect differences between strategies if they truly exist. Otherwise noninferiority may be concluded simply due to insensitivity of the trial to detect differences. In noninferiority trials, assay sensitivity (essentially making strategies appear similar) can be reduced (intentionally or unintentionally) by diluting effects through subtle choices about design and conduct. Many factors can affect assay sensitivity including: poor disease diagnosis, endpoint selection and timing, poor adherence, loss-to-follow-up, prior therapy, inclusion of subgroups where treatment effects may be small, and use of concomitant therapies. Furthermore the active-control nature of the most noninferiority trials, can make clinicians and participants more likely to rate positive outcomes, driving the results toward noninferiority.

The methodologies for two-arm (an experimental intervention and an effective active control) noninferiority clinical trials have been well-established. However, two-arm noninferiority trials often lack the necessary support for the assay sensitivity and constancy assumptions. As a result, inclusion of a third arm (placebo) into the trial has been proposed to address these concerns (Pigeot et al., 2003; Koch and Röhmle, 2004; Hauschke and Pigeot, 2005a). Regulatory authorities often recommend a use of such a three-arm (experimental intervention, active control, and placebo) noninferiority trial design (ICH, 2000; CHMP, 2005; FDA, 2010). The three-arm noninferiority trial offers several scientific advantages (ICH, 2000). Specially these designs provide the opportunity of establishing the validity of the assay sensitivity via a comparison of the placebo to the active control intervention within the trial. Although the three-arm noninferiority design provides such scientific advantages, it also provides challenges: (1) there may be ethical constraints to using a placebo, and (2) there is the added complexity of evaluating two distinct objectives: evaluation of (i) the superiority of the active control intervention to placebo (assay sensitivity) and (ii) the noninferiority of the experimental intervention to the active control intervention (noninferiority). This may result in a sample size that is too large and impractical to conduct. One approach to address this concern is the use of group-sequential designs. The group-sequential design offers the possibility to stop a trial early when evidence is overwhelming and thus offers efficiency (i.e., potentially fewer trial participants and minimizing the amount time that participants receive a placebo, compared to fixed-sample designs).

In this paper, we discuss group-sequential designs for three-arm noninferiority clinical trials. We extend two existing approaches for evaluating noninferiority and assay sensitivity into a group-sequential setting. One approach is discussed by Koch and Röhmle (2004), and Hida and Tango (2011a, 2013) (hereafter we call this “fixed margin approach”), and the other is so-called “fraction approach” proposed by Pigeot et al. (2003). We consider a three-arm noninferiority trial that has two co-primary objectives: (i) to evaluate if the control intervention is superior to placebo (assay sensitivity: AS) and (ii) to evaluate if the experimental intervention is not less effective than the control intervention by a prespecified non-inferiority margin (noninferiority: NI). Objective (ii) is relevant when the experimental intervention has advantages over the control (e.g., safer, more convenience, or less costly). On the other hand, in many noninferiority clinical trials, especially in a regulatory setting, demonstrating the superiority of the experimental intervention to placebo is desirable. However, as Gao and Ware (2008) discuss, if the assay sensitivity assumption does not hold, then there will be uncertainty regarding whether a noninferiority result means that they are similarly effective or similarly ineffective. In this paper, when there is a concern about the assay sensitivity, to make the evaluation of objective (ii) more interpretable, we evaluate a direct comparison of the control intervention to the placebo. For related discussions, please see Hauschke and Pigeot (2005a, 2005b) and Stucke and Kieser (2012).

Three-arm noninferiority clinical trials in a group-sequential setting have been discussed (Li and Gao, 2010; Schlömer and Brannath, 2013), but methodologies are still needed. Extensions of the fraction approach are discussed by Li and Gao (2010) and the fixed margin approach by Schlömer and Brannath (2013), in a setting of two-stage group-sequential three-arm noninferiority clinical trials with continuous or binary outcomes. We discuss two

decision-making frameworks for the two approaches when the primary endpoint is continuous. We also discuss a method for sample size recalculation based on the observed effect size at an interim timepoint of the trial.

This paper is structured as follows: in Section 2, we describe the statistical settings and provide methods for the overall power for rejecting the null hypotheses for assay sensitivity and noninferiority when using the two decision-making frameworks for the fixed margin and the fraction approaches in a group-sequential setting. Then, we evaluate the operating characteristics including power, Type I error rate, and sample sizes, as design factors vary in Section 3. In Section 4, we discuss sample size recalculation and consider its' impact on the power and Type I error rate via a simulation study. In Section 5, we discuss a further extension and summarize the findings in Section 6.

2. Assessment of assay sensitivity and noninferiority in group-sequential designs

Consider a three-arm noninferiority group sequential clinical trial with a maximum of K planned analyses ($K \geq 2$). Let n_{Ek} , n_{Rk} and n_{Pk} be the cumulative numbers of participants on the experimental intervention (E), control intervention (R), and placebo (P) groups respectively at the k th analysis ($k = 1, \dots, K$). Let the allocation ratios of the active intervention and placebo groups relative to the experimental intervention group be $n_{Rk} : n_{Pk} = 1 : C_R : C_P$, where $C_R(> 0)$ and $C_P(> 0)$. When the groups are equally sized, then $C_R = C_P = 1$. Hence up to n_{Ek} , $n_{Rk} = C_R n_{Ek}$ and $n_{Pk} = C_P n_{Ek}$ participants are recruited and randomly assigned to either of the intervention groups. So that the sample size required for the final analysis N_K is $N_K = n_{Ek} + n_{Rk} + n_{Pk} = (1 + C_R + C_P)n_{Ek}$.

Assume that the group outcomes X_{Ej_E} , X_{Rj_R} and X_{Pj_P} are independently and normally distributed with common variance σ^2 as $X_{Ej_E} \sim N(\mu_E, \sigma^2)$, $X_{Rj_R} \sim N(\mu_R, \sigma^2)$ and $X_{Pj_P} \sim N(\mu_P, \sigma^2)$, respectively ($j_E = 1, \dots, n_{Ek}$; $j_R = 1, \dots, n_{Rk}$; $j_P = 1, \dots, n_{Pk}$), where μ_E , μ_R and μ_P are the means of the experimental intervention, active control and placebo groups respectively, and that a larger mean represents a more preferable outcome. For simplicity, the variance σ^2 is assumed to be known.

2.1 The fixed margin approach

For the fixed margin approach, the hypotheses for evaluating AS and NI are:

$$H_0^{\text{AS}}: \mu_R - \mu_P \leq \Delta \text{ versus } H_1^{\text{AS}}: \mu_R - \mu_P > \Delta, \quad (1)$$

$$H_0^{\text{NI}}: \mu_E - \mu_R \leq -\Delta \text{ versus } H_1^{\text{NI}}: \mu_E - \mu_R > -\Delta, \quad (2)$$

where $\Delta (> 0)$ is a pre-specified noninferiority margin (Hida and Tango, 2011a). This approach imposes an extra condition on the hypothesis testing for the AS, that is superiority of the control intervention to the placebo is demonstrated with a margin Δ . However, the key

feature of the approach is that the inequalities $\mu_P < \mu_R - \Delta < \mu_E$ hold for any value of Δ if both of the null hypotheses H_0^{AS} and H_0^{NI} are rejected at the significance level of α for a one-sided test. This means that the superiority of the experimental intervention relative to the placebo can be indirectly demonstrated if H_0^{AS} and H_0^{NI} are rejected, without direct comparison of the experimental intervention to the placebo. This avoids introduction of further complexities in adjustment to the Type I or Type II error (Hida and Tango, 2011a).

We are now interested in hypothesis testing for AS and NI based on the fixed margin approach within a group-sequential setting. The corresponding statistics for testing hypotheses (1) and (2) at the k th analysis are given by

$$T_k^{AS} = \frac{\bar{X}_{Rk} - \bar{X}_{Pk} - \Delta}{\sigma \sqrt{1/n_{Rk} + 1/n_{Pk}}} \text{ and } T_k^{NI} = \frac{\bar{X}_{Ek} - \bar{X}_{Rk} + \Delta}{\sigma \sqrt{1/n_{Ek} + 1/n_{Rk}}},$$

where $\bar{X}_{Ek}$, $\bar{X}_{Rk}$ and $\bar{X}_{Pk}$ are the sample means in the experimental intervention, active control intervention and placebo groups at the k th analysis, given by

$\bar{X}_{Ek} = (\sum_{i_E=1}^{n_{Ek}} X_{Ei_E})/n_{Ek}$, $\bar{X}_{Rk} = (\sum_{i_R=1}^{n_{Rk}} X_{Ri_R})/n_{Rk}$ and $\bar{X}_{Pk} = (\sum_{i_P=1}^{n_{Pk}} X_{Pi_P})/n_{Pk}$. Then (T_k^{AS}, T_k^{NI}) is bivariate normally distributed with the correlation

$$\text{corr}[T_k^{AS}, T_k^{NI}] = \rho = -\sqrt{\frac{n_{Ek}n_{Pk}}{(n_{Ek}+n_{Rk})(n_{Rk}+n_{Pk})}} = -\sqrt{\frac{C_P}{(1+C_R)(C_R+C_P)}}.$$

The correlation is determined by the allocation ratios C_P and C_R (Hida and Tango, 2011a).

The test statistics (T_k^{AS}, T_k^{NI}) are negatively correlated and the correlation is $\rho = -0.5$ if the intervention groups are equally sized, i.e., $C_R = C_P = 1$. Furthermore, the joint distribution of $(T_1^{AS}, \dots, T_K^{AS}, T_1^{NI}, \dots, T_K^{NI})$ are $2K$ multivariate normal distributed with correlations

given by $\text{corr}(T_{k'}^{AS}, T_k^{AS}) = \text{corr}(T_{k'}^{NI}, T_k^{NI}) = \sqrt{n_{Ek'}/n_{Ek}}$, and $\text{corr}(T_{k'}^{AS}, T_k^{NI}) = \text{corr}(T_{k'}^{NI}, T_k^{AS}) = \rho \sqrt{n_{Ek'}/n_{Ek}} (1 \leq k' \leq k \leq K)$ since T_k^{AS} and T_k^{NI} can be rewritten as $T_k^{AS} = \sqrt{n_{Ek}}(\bar{X}_{Rk} - \bar{X}_{Pk} - \Delta)/(\sigma \sqrt{1/C_R + 1/C_P})$ and $T_k^{NI} = \sqrt{n_{Ek}}(\bar{X}_{Rk} - \bar{X}_{Pk} - \Delta)/(\sigma \sqrt{1 + 1/C_R})$.

2.2 The fraction approach

For the fraction approach, the hypotheses for evaluating AS and NI are as follows;

$$H_0^{AS}: \mu_R - \mu_P \leq 0 \text{ versus } H_1^{AS}: \mu_R - \mu_P > 0, \quad (3)$$

$$H_0^{NI}: (\mu_E - \mu_P) / (\mu_R - \mu_P) \leq \theta \text{ versus } H_1^{NI}: (\mu_E - \mu_P) / (\mu_R - \mu_P) > \theta, \quad (4)$$

where $\theta (0 < \theta < 1)$ is pre-specified and determined by $\theta = 1 - \alpha / (\mu_R - \mu_P)$ as a fraction of the difference between μ_R and μ_P , using the noninferiority margin (Pigeot et al., 2003). In addition, hypothesis testing is logically ordered, i.e., H_0^{AS} is tested first and then H_0^{NI} is tested if and only if H_0^{AS} is rejected at the prespecified significance level of α . If both null hypotheses H_0^{AS} and H_0^{NI} are rejected, then $\mu_E > \mu_P$ irrespective of θ since $\mu_E - \mu_P > \theta(\mu_R - \mu_P) > 0$. Many authors have discussed the fraction approach in fixed-sample designs; binary outcomes are discussed by Tang and Tang (2004) and Kieser and Friede (2007), time to event outcomes by Mielke et al. (2008) and Kombrink et al. (2013), and continuous outcomes with heterogeneous variances by Hasler et al. (2008).

We focus on hypothesis testing based on the fraction approach within a group-sequential setting. Assuming $\mu_R - \mu_P > 0$, the hypothesis (4) can be rewritten as

$$H_0^{NI}: \mu_E - \theta\mu_R - (1-\theta)\mu_P \leq 0 \text{ versus } H_1^{NI}: \mu_E - \theta\mu_R - (1-\theta)\mu_P > 0.$$

The corresponding statistics for testing hypotheses (3) and (4) at the k th analysis are given by

$$T_k^{AS} = \frac{\bar{X}_{Rk} - \bar{X}_{Pk}}{\sigma \sqrt{1/n_{Rk} + 1/n_{Pk}}} \text{ and } T_k^{NI} = \frac{\bar{X}_{Ek} - \theta\bar{X}_{Rk} - (1-\theta)\bar{X}_{Pk}}{\sigma \sqrt{1/n_{Ek} + \theta^2/n_{Rk} + (1-\theta)^2/n_{Pk}}}.$$

The joint distribution of $(T_1^{AS}, \dots, T_K^{AS}, T_1^{NI}, \dots, T_K^{NI})$ are $2K$ multivariate normal distributed with their correlations given by the same correlation structure as the fixed margin approach, but the correlation of T_k^{AS} and T_k^{NI} is given by

$$\text{corr}[T_k^{AS}, T_k^{NI}] = \rho = \frac{-\theta/C_R + (1-\theta)/C_P}{\sqrt{1 + \theta^2/C_R + (1-\theta)^2/C_P} \sqrt{1/C_R + 1/C_P}}.$$

The correlation is determined by the fraction θ and the allocation ratios C_P and $C_R: \rho = (1-2\theta)/2 \sqrt{1-\theta+\theta^2}$ if the intervention groups are equally sized, i.e., $C_R = C_P = 1$.

There are important differences in the two approaches (Röhmel and Pigeot, 2011; Hida and Tango, 2011a, 2011b; Stucke and Kieser, 2012). Specifically the concept of “assay sensitivity” is different. A different conclusion is driven from the two approaches when $\mu_R - \mu_P < \mu_P < \mu_R$ is true (Hida and Tango, 2011b). The fraction approach can reject H_0^{NI} , but the fixed margin approach cannot. Whether the fraction approach can allow demonstration of

noninferiority of the experimental intervention to the control intervention is questionable under $\mu_R - < \mu_P$. For further discussion, please see Röhmel and Pigeot (2011), Hida and Tango (2011b), and Stucke and Kieser (2012).

2.3 Decision-making frameworks, stopping rules and powers

We consider the two decision-making frameworks associated with hypothesis testing. The first decision-making framework is flexible, where testing hypotheses for AS and NI are logically ordered similarly as in the fraction approach, i.e., NI is evaluated only after the AS is demonstrated and a trial is terminated if H_0^{AS} and H_0^{NI} are rejected at any analysis (i.e., not necessarily simultaneously) (DF-A). The other framework is relatively simple and a special case of DF-A, where a clinical trial is terminated if and only if both H_0^{AS} and H_0^{NI} are rejected simultaneously at the same analysis (DF-B). We separately describe the two decision-making frameworks, corresponding stopping rules and power definitions.

DF-A—Under DF-A, a trial stops if the AS and the NI are achieved at any analysis (i.e., not necessarily simultaneously). NI is evaluated only after the AS is demonstrated. If AS is demonstrated but NI is not, then the trial continues and subsequent hypothesis testing is repeatedly conducted only for NI until the NI is demonstrated. The stopping rule for DF-A is formally given as follows:

At the k th analysis ($k = k', \dots, K - 1$),

if $T_{k'}^{AS} > c_{k'}^{AS}$ for some $k' (1 \leq k' \leq k)$ and $T_k^{NI} > c_k^{NI}$, then reject H_0^{NI} and stop the trial

otherwise, continue the trial,

at the K th analysis,

if $T_{k'}^{AS} > c_{k'}^{AS}$ for some $k' (1 \leq k' \leq K)$ and $T_K^{NI} > c_K^{NI}$, then reject H_0^{NI} ,

otherwise, do not reject H_0^{NI} ,

where c_k^{AS} and c_k^{NI} are the critical boundaries at the k th analysis, which are constant and selected separately for AS and NI to preserve the Type I error of α for each hypothesis, using any group-sequential method such as Lan-DeMets (LD) alpha-spending method (Lan and DeMets, 1983), analogously to a trial with a single primary objective. For example, consider a three-arm noninferiority clinical trial with a maximum number of analyses $K = 4$ and equally spaced increments of information, and the O'Brien-Fleming boundary (O'Brien and Fleming, 1979) is used to reject the null hypothesis for the AS and the NI tests with the same significance level of $\alpha = 2.5\%$ for a one-sided test. The boundaries for each analysis are 4.3326, 2.9631, 2.3590, and 2.0141, respectively. If the AS test is statistically significant at the third analysis, then the NI test is evaluated twice with the boundary of 2.3590 at the third analysis and 2.0141 at the final analysis as if the Type I error for the NI test has been already spent at the first and second analyses despite no test being conducted. Even if the AS test is statistically significant at the third analysis, the remaining Type I error of 1.5% ($=2.5-1.0$) is not reallocated to the hypothesis test for NI. If the remaining Type I error rate

of 1.5% for the AS test is reallocated to the hypothesis test for NI, then the size of the hypothesis tests for AS and NI are at most $\alpha = 4.0\%$ ($=1.5+2.5$) since the test is the intersection-union.

Therefore the overall power for rejecting the both H_0^{AS} and H_0^{NI} under H_1^{AS} and H_1^{NI} in DF-A is

$$1-\beta = \Pr \left[\bigcup_{1 \leq k' \leq K} \left\{ \{T_{k'}^{AS} > c_{k'}^{AS}\} \cap \{T_{k'}^{NI} > c_{k'}^{NI}\} \right\} \mid H_1^{AS} \cap H_1^{NI} \right].$$

This power can be evaluated using the numerical integration method in Genz (1992) or other methods.

When using the fixed margin approach, DF-A allows for dropping of the placebo group if AS is demonstrated at the interim. However, when using the fraction approach, DF-A cannot allow this as the test statistics for the NI includes the amount of $\bar{X}_{PK}$.

DF-B—Under DF-B, a trial is stopped if AS and NI are demonstrated at the same analysis simultaneously. Otherwise the trial will continue and the subsequent hypothesis testing is repeatedly conducted for both AS and NI until simultaneous significance is reached. The stopping rule for DF-B is formally given as follows:

At the k th analysis ($k = 1, \dots, K-1$),

if $T_k^{AS} > c_k^{AS}$ and $T_k^{NI} > c_k^{NI}$ simultaneously, then reject H_0^{AS} and H_0^{NI} , and stop the trial,

otherwise, continue the trial,

at the K th analysis

if $T_K^{AS} > c_K^{AS}$ and $T_K^{NI} > c_K^{NI}$ then reject H_0^{AS} and H_0^{NI} ,

otherwise, do not reject H_0^{AS} and H_0^{NI} .

Similarly as in the DF-A, the critical boundaries at the k th analysis c_k^{AS} and c_k^{NI} are constant and selected separately for the AS and the NI tests to preserve the Type I error of α for each hypothesis, using any group-sequential method, analogously to a trial with a single primary objective. Therefore, the overall power for rejecting both H_0^{AS} and H_0^{NI} under H_1^{AS} and H_1^{NI} in DF-B is

$$1-\beta = \Pr \left[\bigcup_{k=1}^K \left\{ \{T_k^{AS} > c_k^{AS}\} \cap \{T_k^{NI} > c_k^{NI}\} \right\} \mid H_1^{AS} \cap H_1^{NI} \right].$$

Power can also be numerically assessed by using multivariate normal integrals.

Based on the powers for DF-A and DF-B discussed above, in a group-sequential setting, we describe two sample size concepts, i.e., 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 overall power $1 - \beta$ for rejecting both null hypotheses for AS and NI. The MSS is the smallest integer not less than N_K satisfying the desired power for a group-sequential strategy at the prespecified hypothetical values of parameters μ_E , μ_R and μ_P , σ^2 , and with Fisher's information time for the interim analyses. The ASN is the expected sample size under hypothetical reference values and provides information regarding the number of participants anticipated in a group-sequential design in order to reach a decision point. The definitions of ASNs corresponding to the two decision-making frameworks for the fixed margin and fraction approaches are given in the Appendix.

To identify the value of $n_{E,K}$ or N_K , a simple strategy is to implement a grid search to gradually increase (or decrease) $n_{E,K}$ until the power under $n_{E,K}$ exceeds (or falls below) the desired power. The grid search often requires considerable computing time, especially with a larger number of analyses, or a small effect size. To reduce the computing time, the Newton–Raphson algorithm in Sugimoto et al. (2012) or the basic linear interpolation algorithm in Hamasaki et al. (2013) may be utilized. In this paper, we use of the basic linear interpolation algorithm to reduce the computing time.

3. Operating characteristics

In this section, we investigate the operating characteristics of the fixed margin and fraction approaches for group-sequential designs based on the two decision-making frameworks, where the number of planned analyses is $K = 4$. Specifically we evaluate the overall Type I error rate and overall power under a given sample size. Referring to the settings discussed in Hida and Tango (2011a), assume the means (μ_E , μ_R , μ_P) are (10,10,5) with a common standard deviation $\sigma = 6.5$. The pre-specified noninferiority margin for the fixed margin approach is $\delta = 2.5$ and the corresponding fraction for the fraction approach is $\theta = 0.5$. The three allocation ratios $n_{E,K}:n_{R,K}:n_{P,K}$, considered are: (i) 1:1:1 ($C_R = C_P = 1$), (ii) 2:1:1 ($C_R = C_P = 1/2$), and (iii) 5:4:1 ($C_R = 4/5$, $C_P = 1/5$). The critical values are determined based on the O'Brien–Fleming-type boundary (OF), Pocock-type boundary (PC) (Pocock, 1977) or their combinations, using the LD alpha-spending method with equally spaced increment in information. Four stopping boundary combinations are considered: (i) the OF for both AS and NI (OF-OF), (ii) the OF for AS and the PC for NI (OF-PC), (iii) the PC for AS and the OF for NI (PC-OF), and (iv) the PC for AS and NI (PC-PC). Under these parameter configurations and settings, for the overall power evaluation, the probability of rejecting both of the null hypotheses for AS and NI is calculated at the significance level of $\alpha = 2.5\%$ for a one-sided test under H_1^{AS} and H_1^{NI} . For the overall Type I error rate evaluation, the probability is calculated under the three situations: (i) H_0^{AS} and H_0^{NI} , (ii) H_0^{AS} or (iii) H_0^{NI} . We also evaluate the behavior of MSS and ASN to demonstrate both AS and NI with the desirable power of $1 - \beta = 80\%$ at the prespecified significance level of $\alpha = 2.5\%$ for a one-sided test.

3.1 Behavior of the power and Type I error rate

Figures 1 and 2 illustrate the behavior of the power for rejecting both null hypotheses as a function of the experimental intervention group sample size, when using DF-A and DF-B for the fixed margin and fraction approaches.

For the fixed margin approach, there is no practical difference in the overall power between DF-A and DF-B although DF-A provides a slightly higher power than DF-B in all of the stopping boundary combinations. In all three allocation ratios for DF-A and DF-B, the highest power is given by OF-OF and the lowest is by PC-PC. For the fraction approach, there is also no practical difference in the overall power between DF-A and DF-B although DF-A provides a slightly higher power than DF-B in all of the stopping boundary combinations. In all three allocation ratios for DF-A and DF-B, the highest power is given by OF-OF or PC-OF and the lowest by OF-PC or PC-PC. Comparing the powers for the fixed margin and fraction approaches, the fixed margin approach provides consistently lower power than the fraction approach in all of the decision-making frameworks, the stopping boundary combinations, and allocation ratios.

Figures 3 to 6 illustrate the behavior of the Type I error rate for rejecting both null hypotheses as a function of the sample size for an experimental intervention group, when using DF-A and DF-B for the fixed margin and fraction approaches. For the fixed margin approach, the maximum of the Type I error is not inflated over the targeted significance level of $\alpha = 2.5\%$ in any of the decision-making frameworks, the stopping boundary combinations, or allocation ratios, but the Type I error rate is small and conservative, especially when H_0^{AS} and H_0^{NI} are true. There is no significant difference in the Type I error rates between DF-A and DF-B, but DF-B provides a smaller Type I error rate than DF-A. In all three allocation ratios and null hypothesis settings, the largest Type I error rate is given by OF-OF or PC-OF for DF-A, and OF-OF for DF-B. For the fraction approach, the maximum of the Type I error is similarly not inflated over the prespecified significance level of $\alpha = 2.5\%$ in any of the decision-making frameworks, the stopping boundary combinations or allocation ratios, but the Type I error rate is small especially when H_0^{AS} and H_0^{NI} are true. There is no practical difference between the Type I error rates of DF-A and DF-B, but DF-B provides a smaller Type I error rate than DF-A. In all of the three allocation ratios and null hypothesis settings, the largest Type I error rate is given by OF-OF for both DF-A and DF-B.

Comparing the fixed margin and fraction approaches, the fixed margin approach provides consistently smaller power than the fraction approach in all of the decision-making frameworks, the stopping boundary combinations and allocation ratios.

3.2 Behavior of the sample size

Table 1 displays the MSS and ASN required for evaluating AS and NI with the power of $1 - \beta = 80\%$ at the significance level of $\alpha = 2.5\%$ for a one-sided test, when using the fixed margin and fraction approaches based on DF-A and DF-B. For the fixed margin approach based on DF-A, the ASN is calculated under H_1^{AS} and H_1^{NI} in two ways: in one strategy the

placebo group is not discontinued until NI is demonstrated even when AS is demonstrated at an analysis (ASN1); while in the other strategy the placebo group is discontinued when AS is demonstrated at an analysis (ASN2). The definitions of ASN1 and ASN2 are given in Appendix.

For both the fixed margin and fraction approaches, in all of the stopping boundary combinations and allocation ratios, there is a modest difference in the MSS and ASN between the DF-A and DF-B although DF-A provides a slightly smaller sample size than DF-B. For the fixed margin approach, the smallest MSS is given by OF-OF and the largest by PC-PC in all of the allocation ratios. The smallest ASN1 is associated with OF-OF or PC-PC and the largest with PC-OF or OF-PC. The largest ASN2 is provided by PC-OF or OF-PC in all of the allocation ratios. For the fraction approach, the smallest MSS is provided by OF-OF or PC-OF and the largest by OF-PC or PC-PC in all of the allocation ratios. The smallest ASN1 is consistently produced with PC-PC and the largest with OF-PC. Comparing the fixed margin and fraction approaches, the fraction approach provides smaller MSS and ASN than the fixed margin approach in all of the decision-making frameworks, the stopping boundary combinations and allocation ratios.

4. Sample size recalculation

Clinical trials are designed based on assumptions often constructed based on prior data. However, prior data may be limited or an inaccurate indication of future data, resulting in trials that are over- or underpowered. Interim analyses at accumulating data 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 of assay sensitivity and noninferiority. However, more modern adaptive designs may also allow for increases (or decreases) in the sample size if effects are smaller (or larger) than assumed. Such adjustments must be conducted carefully for several reasons, especially to maintain control of statistical error rates. In this section, we discuss sample size recalculation based on the observed intervention's effects at an analysis with a focus on the control of statistical error rates.

4.1 Cui-Hung-Wang test statistics and conditional power

We now consider a scenario where the maximum sample size n_{EK} in the experimental intervention group is recalculated to n'_{EK} at the k th analysis, by allowing both an increase or a decrease in sample size. Suppose n'_{EK} is subject to $n_{EK} < n'_{EK} \leq \lambda n_{EK}$, where λ is a pre-specified constant for the maximum allowable sample size. In addition, let the sample size at the $(k + l)$ th analysis be $n'_{Ek+l} = (n'_{EK} - n_{EK}) / (n_{EK} - n_{Ek}) (n_{Ek+l} - n_{Ek}) + n_{EK}$ ($l = 1, \dots, K - k$).

Consider the Cui-Hung-Wang (CHW) statistics (Cui et al., 1999) for sample size recalculation in group-sequential designs for three-arm clinical trials to preserve the overall Type I error rate at a prespecified significance level of α even when the sample size is increased and conventional test statistics are used. When using the fixed margin approach, the CHW statistics for the AS and the NI are

$$T_{k+l}^{AS} = \sqrt{\frac{n_{Ek}}{n_{Ek+l}}} T_k^{AS} + \sqrt{\frac{n_{Ek+l}-n_{Ek}}{n_{Ek+l}}} \frac{\bar{X}'_{Rk+l} - \bar{X}'_{Pk+l} - \Delta}{\sigma \sqrt{1/(n'_{Ek+l} - n_{Ek})} \sqrt{1/C_R + 1/C_P}}, \text{ and}$$

$$T_{k+l}^{NI} = \sqrt{\frac{n_{Ek}}{n_{Ek+l}}} T_k^{NI} + \sqrt{\frac{n_{Ek+l}-n_{Ek}}{n_{Ek+l}}} \frac{\bar{X}'_{Ek+l} - \bar{X}'_{Rk+l} + \Delta}{\sigma \sqrt{1/(n'_{Ek+l} - n_{Ek})} \sqrt{1+1/C_R}},$$

where $\bar{X}'_{Ek+l} = (\sum_{i=n_{Ek}+1}^{n'_{Ek+l}} X_{Ei}) / (n'_{Ek+l} - n_{Ek})$, $\bar{X}'_{Rk+l} = (\sum_{i=n_{Rk}+1}^{n'_{Rk+l}} X_{Ei}) / (n'_{Rk+l} - n_{Rk})$, and

$\bar{X}'_{Pk+l} = (\sum_{i=n_{Pk}+1}^{n'_{Pk+l}} X_{Pi}) / (n'_{Pk+l} - n_{Pk})$. The sample size is increased or decreased when the conditional power evaluated at the k th analysis is lower or higher than the desired power $1 - \beta$. Under the planned maximum sample size and a given observed value of (T_k^{AS}, T_k^{NI}) , if the decision-making for rejecting the null-hypotheses H_0^{AS} and H_0^{NI} is based on DF-A, then the conditional power at the k th analysis is given by

$$CP(\delta_1, \delta_2) = \begin{cases} \Phi_2 \left(-\frac{c_K^{AS} - \sqrt{mt_k^{AS}}}{\sqrt{1-m}} + \sqrt{n_{Ek} - n_{Ek}} \frac{\delta_1}{\sqrt{1/C_R + 1/C_P}}, -\frac{c_K^{NI} - \sqrt{mt_k^{NI}}}{\sqrt{1-m}} + \sqrt{n_{Ek} - n_{Ek}} \frac{\delta_2}{\sqrt{1+1/C_R}} \right), & \text{if } T_l^{NI} \leq c_l^{NI} \text{ and } l \\ 1 - \Phi_1 \left(\frac{c_K^{NI} - \sqrt{mt_k^{NI}}}{\sqrt{1-m}} - \sqrt{n_{Ek} - n_{Ek}} \frac{\delta_2}{\sqrt{1+1/C_R}} \right), & \text{if } T_l^{NI} \leq c_l^{NI} \text{ for all } l=1, \dots \end{cases}$$

where $m = n_{Ek}/n_{Ek}$, $\delta_1 = (\mu_R - \mu_P -)/\sigma$ and $\delta_2 = (\mu_E - \mu_R +)/\sigma$. In addition, $\Phi_1(\cdot)$ and $\Phi_2(\cdot)$ are the cumulative distribution functions of the standardized univariate and bivariate normal distributions. The critical values c_l^{AS} and c_l^{NI} are the same critical values utilized for the case without sample size recalculation. If the decision making is based on DF-B, then the conditional power is

$$CP(\delta_1, \delta_2) = \Phi_2 \left(-\frac{c_K^{AS} - \sqrt{mt_k^{AS}}}{\sqrt{1-m}} + \sqrt{n_{Ek} - n_{Ek}} \frac{\delta_1}{\sqrt{1/C_R + 1/C_P}}, -\frac{c_K^{NI} - \sqrt{mt_k^{NI}}}{\sqrt{1-m}} + \sqrt{n_{Ek} - n_{Ek}} \frac{\delta_2}{\sqrt{1+1/C_R}} \right).$$

The details of the derivation for the conational powers are given in Appendix. On the other hand, when using the fraction approach, the CHW statistics are given by

$$T_{k+l}^{AS} = \sqrt{\frac{n_{Ek}}{n_{Ek+l}}} T_k^{AS} + \sqrt{\frac{n_{Ek+l}-n_{Ek}}{n_{Ek+l}}} \frac{\bar{X}'_{Rk+l} - \bar{X}'_{Pk+l}}{\sigma \sqrt{1/(n'_{Ek+l} - n_{Ek})} \sqrt{1/C_R + 1/C_P}} \text{ and}$$

$$T_{k+l}^{NI} = \sqrt{\frac{n_{Ek}}{n_{Ek+l}}} T_k^{NI} + \sqrt{\frac{n_{Ek+l}-n_{Ek}}{n_{Ek+l}}} \frac{\bar{X}'_{Ek+l} - \theta \bar{X}'_{Rk+l} - (1-\theta) \bar{X}'_{Pk+l}}{\sigma \sqrt{1/(n'_{Ek+l} - n_{Ek})} \sqrt{1+\theta^2/C_R + (1-\theta)^2/C_P}}.$$

The conditional power can be calculated in the same manner as the fixed margin approach except for $\delta_1 = (\mu_R - \mu_P)/\sigma$ and $\delta_2 = (\mu_E - \theta\mu_R - (1 - \theta)\mu_P)/\sigma$, and the coefficient of δ^2 changes $\sqrt{1+1/C_R}$ to $\sqrt{1+\theta^2/C_R+(1-\theta)^2/C_P}$.

When recalculating the sample size, we consider three possible options, i.e., (i) only allowing an increase in the sample size, (ii) only allowing a decrease in the sample size, and (iii) allowing an increase or a decrease in sample size. The recalculated maximum sample size n'_{EK} required for the experimental intervention group for each respective option is:

$$\begin{aligned} \text{i.} \quad n'_{EK} &= \begin{cases} \min(n_{EK}^*, \lambda n_{EK}), & \text{if } CP(\delta_1, \delta_2) < \gamma(1-\beta), \\ n_{EK}, & \text{otherwise;} \end{cases} \\ \text{ii.} \quad n'_{EK} &= \begin{cases} n_{EK}^*, & \text{if } CP(\delta_1, \delta_2) > \eta(1-\beta), \\ n_{EK}, & \text{otherwise;} \end{cases} \\ \text{iii.} \quad n'_{EK} &= \begin{cases} \min(n_{EK}^*, \lambda n_{EK}), & \text{if } CP(\delta_1, \delta_2) < \gamma(1-\beta), \\ n_{EK}^*, & \text{if } CP(\delta_1, \delta_2) > \eta(1-\beta), \\ n_{EK}, & \text{otherwise;} \end{cases} \end{aligned}$$

where n_{EK}^* is the calculated sample size in which the conditional power based on the observed effect sizes achieves the target power of $1 - \beta$, and $\gamma(0 < \gamma < 1)$ and $\eta(\eta > 1)$ are the pre-specified constant for allowance.

4.2 Simulation study

We investigate the impact of the sample size recalculation on the overall power and the Type I error rate for rejecting both null hypotheses using Monte-Carlo simulation.

Consider group sequential designs with four analyses (i.e., three interim and one final analysis) for the fixed margin and the fraction approaches with the decision-making frameworks DF-A or DF-B, where analyses are conducted with equally spaced increments in information. The original planned total sample size is calculated to test both null hypotheses for the AS and the NI with the power of $1 - \beta = 80\%$ at the prespecified significance level of $\alpha = 2.5\%$ for a one-sided test in the fixed-sample designs, where $(\mu_E, \mu_R, \mu_P) = (10, 10, 5)$, $\sigma = 6.5$, and $\lambda = 2.5$ for the fixed margin approach, and $\theta = 0.5$ for the fraction approach. The total sample sizes are 426 for the fixed margin approach and 240 for the fraction approach.

One sample size recalculation is considered based on the observed effects with three options: (i) only allowing an increase in the sample size, (ii) only allowing a decrease in the sample size, and (iii) allowing an increase or decrease in sample size, evaluated at the first, second or third interim analyses. The sample size calculation is performed when the conditional power is less than 70% (i.e., γ is set as 0.875) and/or exceeds 90% (i.e., η is set as 1.125). The sample size for the experimental intervention group can be increased up to 1.5 times of the original planned sample size (i.e., λ is set as 1.5). Similarly as in Section 3, the critical values are determined based on the three stopping boundary combinations considered using the LD alpha-spending method with equally spaced increments in

information: (i) the OF for both AS and NI (OF-OF), (ii) the PC for AS and the OF for NI (PC-OF), and (iii) the PC for both AS and NI (PC-PC). The empirical overall power is evaluated under the situation where both H_1^{AS} and H_1^{NI} are true with $(\mu_E, \mu_R, \mu_P) = (10, 10, 5)$, and the actual overall Type I rate is evaluated under the three situations, i.e., (i) H_0^{AS} and H_0^{NI} , (ii) H_0^{AS} or (iii) H_0^{NI} . The number of replications for the simulation is set to 1,000,000 for the evaluation of the Type I error rate and 100,000 replications for the power. The number of replications for the simulation was determined based on the precision. 1,000,000 replications provides a two-sided 95% confidence interval with a width equal to 0.001 when the proportion is 2.5%, while 100,000 replications provides a two-sided 95% confidence interval with a width equal to 0.005 when the proportion is 80%. We limit the discussion to the overall empirical power and Type I error rate for DF-A with equally sized groups of $C_R = C_P = 1$ as there is no appreciable differences in the power and Type I error rates between DF-A and DF-B.

Figure 7 summarizes the overall empirical power for the fixed margin and fraction approaches when sample size recalculation is performed based on DF-A under the design setting and parameter configurations described above. For both the fixed margin and fraction approach, regardless of the stopping boundary combinations, the overall powers increase up to 15% compared to a design without sample size recalculation. Larger increases are observed with a later timing of the sample size recalculation when only allowing an increase in the sample size, and allowing an increase or decrease in the sample size. As shown in Table 2, the ASN are approximately equal or smaller than the planned sample size. When only allowing a decrease in the sample size, the overall power cannot maintain the targeted power of 80% although the expected sample size can be reduced more than the other recalculation options. Figures 8 and 9 summarize the actual overall Type I error rates. For both the fixed margin and fraction approaches, regardless of the stopping boundary combinations, in all three recalculation options, the actual Type I error rates do not exceed the prespecified significance level of 2.5%, but are small and conservative, especially in the fixed margin approach.

5 A further extension

When constructing efficient group-sequential designs in three arm noninferiority clinical trials, a major issue is that the Type I error rate is small and conservative as the rejection region of the null hypotheses H_0^{AS} and H_0^{NI} is restricted even when group-sequential designs are used for evaluating the AS and the NI, as with the fixed-sample designs. This is due to the requirement that the allocation of the Type I error to each analysis for the AS and the NI should be prespecified and determined, using an alpha-spending method.

To overcome this issue, the DF-A can be modified to allocate adaptively the Type I error to each analysis for the NI although the Type I error allocation for the AS is prespecified. This idea is first discussed by Tsong et al. (2004) in group-sequential three-arm clinical trials when assessing the equivalence and efficacy of a generic product, where the co-primary objectives of the trial are to assess whether the generic and reference product are effective relative to placebo and whether the generic is equivalent to the reference product with a

prespecified equivalence margin. Their method evaluates equivalence only after both null hypotheses of efficacy are rejected and then specifies the Type I error allocation before the equivalence evaluation is performed. In the three-arm noninferiority clinical trials for the assessment of the AS and the NI, the NI is evaluated only after the AS is demonstrated and the Type I error allocation for the NI is specified just before the NI evaluation is performed. Figure 10 illustrates the behavior of the Type I error rate for rejecting both null hypotheses for the AS and the NI as a function of sample size for an experimental intervention group, when using the fixed margin approach based on DF-A with adaptive Type I error allocation for the NI in a group-sequential setting, where the parameter settings and configuration are same as in Figure 3, but for $n_{E,k}:n_{R,k}:n_{P,k} = 1:1:1$ ($C_R = C_P = 1$). The Type I error rate is improved compared to that seen in Figure 3, but it is not inflated over the targeted significance level. This improvement in the Type I error rate is expected to provide smaller sample sizes but will increase complexity. Further investigation will be required to evaluate how much the method may improve the efficiency and whether this outweighs the complexity. The methodology for adaptive Type I error allocation may be applicable to the fraction approach. However, as seen in Section 2, the correlation between the two test statistics for the AS and the NI are determined by the allocation ratio and the fraction margin and the test statistics may be positively correlated in some situations (e.g., $n_{E,k}:n_{R,k}:n_{P,k} = 1:1:1$ and $\theta < 0.5$). By analogy to work by Hamasaki et al (2015), the Type I error rate is inflated over the targeted significance level when the test statistics are positively correlated. Therefore use of adaptive Type I error allocation for the fraction approach should be carefully considered.

6 Summary

Noninferiority clinical trials recently have received a great deal of attention by regulatory authorities (CHMP, 2005; FDA, 2010) and in the clinical trials literature (e.g., extensive reference found in Rothman et al. (2012)). Noninferiority clinical trials have complexities requiring careful design, monitoring, analyses, and reporting. When designing noninferiority clinical trials, the constancy and assay sensitivity are the important assumptions. The selection of the active control for a noninferiority trial should be done carefully, ensuring that it has demonstrated and precisely measured superiority over placebo and that its effect has not changed compared to the historical trials that demonstrated its efficacy (constancy assumption). To assess these issues in regulatory medical product development, a use of three-arm noninferiority design including an experimental intervention, an active control intervention, and a placebo has been considered as a gold-standard design. But this trial may raise ethical issues and result in a sample size that is too large and impractical to conduct the clinical trial.

In this paper, we discuss three-arm noninferiority clinical trials and extend two existing approaches, i.e., the fixed margin and fraction approaches, for evaluating noninferiority and assay sensitivity to a group-sequential setting with two decision-making frameworks. We evaluate the operating characteristics including power, Type I error rate, maximum and expected sample sizes as design factors vary. We also discuss sample size recalculation and consider its' impact on the power and Type I error rate via a simulation study. Our findings are summarized as follows:

- The decision-making frameworks of DF-A and DF-B for the fixed margin and the fraction approaches provide the possibility of stopping a trial early when evidence is overwhelming, thus offering efficiency (e.g., an ASN potentially 4% to 15% fewer than the fixed-sample designs with equally sized groups and four analyses)
- There are no major differences in both MSS and ASN between DF-A and DF-B for the fixed margin and the fraction approaches, although DF-A is slightly more powerful than DF-B. By using the DF-A for the fixed margin approach, the time that participants are exposed to placebo can be minimized as the DF-A allows dropping of the placebo group if assay sensitivity has been demonstrated at an analysis.
- For the fixed margin approach, selecting the O'Brien-Fleming-type boundary for both AS and NI could lead to fewer participants for the MSS and the ASN compared with other boundary combinations. On the other hand, for the fraction approach, selecting the O'Brien-Fleming-type boundary for both AS and NI, or the Pocock-type boundary for AS and the O'Brien-Fleming-type boundary for NI provides better efficiency with respect to the MSS and the ASN compared with other boundary combinations.
- When considering sample size recalculation during a trial, only allowing a decrease in the sample size may be not a desirable option in both the fixed margin and fraction approaches as the power does not reach desired levels, although the expected sample size can be reduced more than the other recalculation options. In addition, the timing of the sample size recalculation should also be carefully considered. Power is increased if the sample size recalculation is carried out later in the trial, but the expected sample size is larger.

We caution that these findings are based on one set of design parameter configurations except for the allocation ratio. Further investigation will be required to evaluate how the power and Type I error rate behave with other design assumptions.

Acknowledgments

The authors are grateful to the two anonymous referees and the associate editor for their valuable suggestions and helpful comments that improved the content and presentation of the paper. Research reported in this publication was supported by JSPS KAKENHI under Grant Number 26330038 and the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Numbers UM1AI104681. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

- Committee for medical products for human use (CHMP). [Accessed November 12, 2015] Guideline on the choice of the non-inferiority margin. 2005. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003636.pdf
- Cui L, Hung HMJ, Wang SJ. Modification of sample size in group sequential clinical trials. *Biometrics*. 1999; 55:853–857. DOI: 10.1111/j.0006-341X.1999.00853.x [PubMed: 11315017]

- D'Agostino RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues: the encounters of academic consultants in statistics. *Statistics in Medicine*. 2003; 22:169–186. DOI: 10.1002/sim.1425 [PubMed: 12520555]
- Fishbane S, Schiller B, Locatelli F, Covic AC, Provenzano R, Wiecek A, Levin NW, Kaplan M, Macdougall IC, Francisco C, Mayo MR, Polu KR, Duliege AM, Besarab A. for the EMERALD Study Groups. Peginesatide in patients with anemia undergoing hemodialysis. *New England Journal of Medicine*. 2013; 368:307–19. DOI: 10.1056/NEJMoa1203165 [PubMed: 23343061]
- Food and Drug Administration (FDA). Guidance for industry non-inferiority trials. Rockville, MD: Food and Drug Administration; 2010. Available at: <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM202140.pdf> [Accessed July 14, 2015]
- Gao P, Ware JH. Assessing non-inferiority: a combination approach. *Statistics in Medicine*. 2008; 27:392–406. DOI: 10.1002/sim.2938 [PubMed: 17575568]
- Genz A. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*. 1992; 1:141–149. DOI: 10.1080/10618600.1992.10477010
- Hauschke D, Pigeot I. Establishing efficacy of a new experimental treatment in the ‘gold standard’ design. *Biometrical Journal*. 2005a; 47:782–786. DOI: 10.1002/bimj.200510169 [PubMed: 16450851]
- Hauschke D, Pigeot I. Rejoinder to “Establishing efficacy of a new experimental treatment in the ‘gold standard’ design”. *Biometrical Journal*. 2005b; 47:797–798. DOI: 10.1002/bimj.200510179
- Hasler M, Vonk R, Hothorn LA. Assessing non-inferiority of a new treatment in a three-arm trial in the presence of heteroscedasticity. *Statistics in Medicine*. 2008; 27:490–503. DOI: 10.1002/sim.3052 [PubMed: 17853384]
- Hamasaki T, Asakura K, Evans SR, Sugimoto T, Sozu T. Group-sequential strategies in clinical trials with multiple co-primary endpoints. *Statistics in Biopharmaceutical Research*. 2015; 7:36–54. DOI: 10.1080/19466315.2014.1003090 [PubMed: 25844122]
- Hamasaki T, Sugimoto T, Evans SR, Sozu T. Sample size determination for clinical trials with co-primary outcomes: Exponential event times. *Pharmaceutical Statistics*. 2013; 12:28–34. DOI: 10.1002/pst.1545 [PubMed: 23081932]
- Hida E, Tango T. On the three-arm non-inferiority trial including a placebo with a prespecified margin. *Statistics in Medicine*. 2011a; 30:224–231. DOI: 10.1002/sim.4099 [PubMed: 21213340]
- Hida E, Tango T. Response to Joachim Röhm and Iris Pigeot. *Statistics in Medicine*. 2011b; 30:3165.doi: 10.1002/sim.4313
- Hida E, Tango T. Three-arm noninferiority trials with a prespecified margin for inference of the difference in the proportions of binary endpoints. *Journal of Biopharmaceutical Statistics*. 2013; 23:774–789. DOI: 10.1080/10543406.2013.789893 [PubMed: 23799810]
- International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). [Accessed July 14, 2015] ICH Harmonised Tripartite Guideline E10: Choice of control group and related issues in clinical trials. 2000 Jul. Available at: <http://www.ich.org/pdf/ICH/e10step4.pdf>
- Kieser M, Friede T. Planning and analysis of three-arm non-inferiority trials with binary endpoints. *Statistics in Medicine*. 2007; 26:253–273. DOI: 10.1002/sim.2543 [PubMed: 16685703]
- Koch A, Röhm J. Hypothesis testing in the “gold standard” design for proving the efficacy of an experimental treatment. *Journal of Biopharmaceutical Statistics*. 2004; 14:315–325. DOI: 10.1081/BIP-120037182 [PubMed: 15206529]
- Kombrink K, Munk A, Friede T. Design and semiparametric analysis of non-inferiority trials with active and placebo control for censored time-to-event data. *Statistics in Medicine*. 2013; 32:3055–3066. DOI: 10.1002/sim.5769 [PubMed: 23508697]
- Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983; 70:659–663. DOI: 10.1093/biomet/70.3.659
- Li G, Gao S. A group sequential type design for three-arm non-inferiority trials with binary endpoints. *Biometrical Journal*. 2010; 52:504–518. DOI: 10.1002/bimj.200900188 [PubMed: 20645389]
- Mielke M, Munk A, Schacht A. The assessment of non-inferiority in a gold standard design with censored, exponentially distributed endpoints. *Statistics in Medicine*. 2008; 27:5093–5110. DOI: 10.1002/sim.3348 [PubMed: 18570271]

- O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979; 35:549–556. DOI: 10.2307/2530245 [PubMed: 497341]
- Pigeot I, Schäfer J, Röhmel J, Hauschke D. Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. *Statistics in Medicine*. 2003; 22:883–899. DOI: 10.1002/sim.1450 [PubMed: 12627407]
- Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. 1977; 64:191–199. DOI: 10.1093/biomet/64.2.191
- Röhmel J, Pigeot I. Statistical strategies for the analysis of clinical trials with an experimental treatment, an active control and placebo, and a prespecified fixed non-inferiority margin for the difference in means. *Statistics in Medicine*. 2011; 30:3162–3164. DOI: 10.1002/sim.4299 [PubMed: 22020727]
- Rothmann, MD., Wiens, BL., Chan, ISF. *Design and Analysis of Non-Inferiority Trials*. Chapman & Hall/CRC; 2012.
- Schlömer P, Brannath W. Group sequential designs for three-arm 'gold standard' non-inferiority trials with fixed margin. *Statistics in Medicine*. 2013; 32:4875–4899. DOI: 10.1002/sim.5950 [PubMed: 23970488]
- Stucke K, Kieser M. A general approach for sample size calculation for the three-arm 'gold standard' non-inferiority design. *Statistics in Medicine*. 2012; 31:3579–3596. DOI: 10.1002/sim.5461 [PubMed: 22764073]
- Sugimoto T, Sozu T, Hamasaki T. A convenient formula for sample size calculations in clinical trials with multiple co-primary continuous endpoints. *Pharmaceutical Statistics*. 2012; 11:118–128. DOI: 10.1002/pst.505 [PubMed: 22415870]
- Tang ML, Tang NS. Test of noninferiority via rate difference for three-arm clinical trials with placebo. *Journal of Biopharmaceutical Statistics*. 2004; 14:337–347. DOI: 10.1081/BIP-120037184 [PubMed: 15206531]
- Tsong Y, Zhang J, Wang SJ. Group sequential design and analysis of clinical equivalence assessment for generic nonsystematic drug products. *Journal of Biopharmaceutical Statistics*. 2004; 14:359–373. DOI: 10.1081/BIP-120037186 [PubMed: 15206533]

Appendix A: Average sample number

The ASN is the expected sample size under hypothetical reference values and provides information regarding the number of participants anticipated in a group-sequential design in order to reach a decision point. We briefly describe the several definitions of the ASN corresponding to the decision-making frameworks.

When using DF-A or DF-B for the fixed margin and fraction approaches, if the placebo group is not terminated until NI is demonstrated even when the AS is demonstrated at an analysis, then the ASN can be calculated by

$$\begin{aligned} \text{ASN}_1 &= \sum_{k=1}^K N_k P_k(\mu_E, \mu_R, \mu_P, \Delta, \sigma^2) + \left(1 - \sum_{k=1}^K P_k(\mu_E, \mu_R, \mu_P, \Delta, \sigma^2)\right) N_K \\ &= N_K + \sum_{k=1}^{K-1} (N_k - N_K) P_k(\mu_E, \mu_R, \mu_P, \Delta, \sigma^2), \end{aligned}$$

where $N_k (=n_{Ek} + n_{Rk} + n_{Pk})$ is the cumulative number of participants at the k th analysis, and $P_k(\mu_E, \mu_R, \mu_P, \Delta, \sigma^2)$ is the stopping probability at the k th analysis assuming that the true values of the intervention's means are (μ_E, μ_R, μ_P) . If the analysis is conducted with equally-spaced increments in information, then N_k can be rewritten as $(k/K)N_K$.

The stopping probability P_k based on DF-A is given by

$$P_k(\mu_E, \mu_R, \mu_P, \Delta, \sigma^2) = \begin{cases} \Pr[\{T_1^{AS} > c_1^{AS}\} \cap \{T_1^{NI} > c_1^{NI}\}], & k=1, \\ \Pr[\bigcap_{l=1}^{k-1} \{T_l^{AS} \leq c_l^{AS}\} \cap \{T_k^{AS} > c_k^{AS}\} \cap \{T_k^{NI} > c_k^{NI}\}] \\ + \Pr[\{T_1^{AS} > c_1^{AS}\} \cap \bigcap_{l=1}^{k-1} \{T_l^{NI} \leq c_l^{NI}\} \cap \{T_k^{NI} > c_k^{NI}\}] \\ + \sum_{2 \leq l < k} \Pr[\bigcap_{m=1}^{l-1} \{T_m^{AS} \leq c_m^{AS}\} \cap \{T_l^{AS} > c_l^{AS}\} \cap \bigcap_{n=l}^{k-1} \{T_n^{NI} \leq c_n^{NI}\} \cap \{T_k^{NI} > c_k^{NI}\}], & k \geq 2. \end{cases}$$

For instance, at $K = 2$, the stopping probabilities P_1 and P_2 based on DF-A are calculated by multivariate normal integrals as follows:

$$\begin{aligned} P_1(\mu_E, \mu_R, \mu_P, \Delta, \sigma^2) &= \Pr[\{T_1^{AS} > c_1^{AS}\} \cap \{T_1^{NI} > c_1^{NI}\}] = \int_{c_1^{AS}}^{\infty} \int_{c_1^{NI}}^{\infty} f_2(t_1^{AS}, t_1^{NI}) dt_1^{NI} dt_1^{AS} \text{ and} \\ P_2(\mu_E, \mu_R, \mu_P, \Delta, \sigma^2) &= \Pr[\{T_1^{AS} \leq c_1^{AS}\} \cap \{T_2^{AS} > c_2^{AS}\} \cap \{T_2^{NI} > c_2^{NI}\}] + \Pr[\{T_1^{AS} > c_1^{AS}\} \cap \{T_1^{NI} \leq c_1^{NI}\} \cap \{T_2^{NI} > c_2^{NI}\}] \\ &= \Pr[\{T_1^{AS} \leq c_1^{AS}\} \cap \{T_1^{NI} \leq c_1^{NI}\} \cap \{T_2^{AS} > c_2^{AS}\} \cap \{T_2^{NI} > c_2^{NI}\}] \\ &\quad + \Pr[\{T_1^{AS} \leq c_1^{AS}\} \cap \{T_1^{NI} > c_1^{NI}\} \cap \{T_2^{AS} > c_2^{AS}\} \cap \{T_2^{NI} > c_2^{NI}\}] \\ &\quad + \Pr[\{T_1^{AS} > c_1^{AS}\} \cap \{T_1^{NI} \leq c_1^{NI}\} \cap \{T_2^{AS} > c_2^{AS}\} \cap \{T_2^{NI} > c_2^{NI}\}] \\ &\quad + \Pr[\{T_1^{AS} > c_1^{AS}\} \cap \{T_1^{NI} \leq c_1^{NI}\} \cap \{T_2^{AS} \leq c_2^{AS}\} \cap \{T_2^{NI} > c_2^{NI}\}] \\ &= \int_{-\infty}^{c_1^{AS}} \int_{-\infty}^{c_1^{NI}} \int_{c_2^{AS}}^{\infty} \int_{c_2^{NI}}^{\infty} f_4(t_1^{AS}, t_1^{NI}, t_2^{AS}, t_2^{NI}) dt_2^{NI} dt_2^{AS} dt_1^{NI} dt_1^{AS} \\ &\quad + \int_{-\infty}^{c_1^{AS}} \int_{c_1^{NI}}^{\infty} \int_{c_2^{AS}}^{\infty} \int_{c_2^{NI}}^{\infty} f_4(t_1^{AS}, t_1^{NI}, t_2^{AS}, t_2^{NI}) dt_2^{NI} dt_2^{AS} dt_1^{NI} dt_1^{AS} \\ &\quad + \int_{c_1^{AS}}^{\infty} \int_{-\infty}^{c_1^{NI}} \int_{c_2^{AS}}^{\infty} \int_{c_2^{NI}}^{\infty} f_4(t_1^{AS}, t_1^{NI}, t_2^{AS}, t_2^{NI}) dt_2^{NI} dt_2^{AS} dt_1^{NI} dt_1^{AS} \\ &\quad + \int_{c_1^{AS}}^{\infty} \int_{c_1^{NI}}^{\infty} \int_{-\infty}^{c_2^{AS}} \int_{-\infty}^{c_2^{NI}} f_4(t_1^{AS}, t_1^{NI}, t_2^{AS}, t_2^{NI}) dt_2^{NI} dt_2^{AS} dt_1^{NI} dt_1^{AS}, \end{aligned}$$

where $f_k(\cdot)$ is the probability density function of k multivariate normal distribution under the alternative hypotheses H_1^{AS} and H_1^{NI} . On the other hand, the stopping probability P_k based on DF-B is given by

$$P_k(\mu_E, \mu_R, \mu_P, \Delta, \sigma^2) = \begin{cases} \Pr[\{T_1^{AS} > c_1^{AS}\} \cap \{T_1^{NI} > c_1^{NI}\}], & k=1, \\ \Pr\left[\bigcap_{l=1}^{k-1} \{\{T_l^{AS} \leq c_l^{AS}\} \cup \{T_l^{NI} \leq c_l^{NI}\}\} \cap \{T_k^{AS} > c_k^{AS}\} \cap \{T_k^{NI} > c_k^{NI}\}\right], & k \geq 2. \end{cases}$$

When using DF-A for the fixed margin approach, we have an option for discontinuing the placebo group at the interim when the AS is demonstrated. In this situation, the ASN can be calculated by

$$\begin{aligned} \text{ASN}_2 &= \sum_{k=1}^K \sum_{l=1}^k \{N_k - (n_{Pk} - n_{Pl})\} P_{k|l}(\mu_E, \mu_R, \mu_P, \Delta, \sigma^2) + \left(1 - \sum_{k=1}^K P_k(\mu_E, \mu_R, \mu_P, \Delta, \sigma^2)\right) N_K \\ &= N_K + \sum_{k=1}^K \sum_{l=1}^k \{N_k - N_K - (n_{Pk} - n_{Pl})\} P_{k|l}(\mu_E, \mu_R, \mu_P, \Delta, \sigma^2), \end{aligned}$$

where $P_{k|l}(\mu_E, \mu_R, \mu_P, \Delta, \sigma^2)$ is given by

$$P_{k|l}(\mu_E, \mu_R, \mu_P, \Delta, \sigma^2) = \begin{cases} \Pr[\{T_1^{AS} > c_1^{AS}\} \cap \{T_1^{NI} > c_1^{NI}\}], & \text{if } k=l=1, \\ \Pr \left[\bigcap_{l=1}^{k-1} \{T_l^{AS} \leq c_l^{AS}\} \cap \{T_k^{AS} > c_k^{AS}\} \cap \{T_k^{NI} > c_k^{NI}\} \right], & \text{if } k=l \geq 2, \\ \Pr \left[\{T_1^{AS} > c_1^{AS}\} \cap \bigcap_{l=1}^{k-1} \{T_l^{NI} \leq c_l^{NI}\} \cap \{T_k^{NI} > c_k^{NI}\} \right], & \text{if } k>l=1, \\ \Pr \left[\bigcap_{m=1}^{l-1} \{T_m^{AS} \leq c_m^{AS}\} \cap \{T_l^{AS} > c_l^{AS}\} \cap \bigcap_{n=l}^{k-1} \{T_n^{NI} \leq c_n^{NI}\} \cap \{T_k^{NI} > c_k^{NI}\} \right], & \text{if } k>l \geq 2. \end{cases}$$

Appendix B: Conditional power derivation

We briefly describe the derivation of conditional powers discussed in Section 4. As the powers based on the DF-A or DF-B for the fixed margin and the fraction approaches can be derived in the same way, we only focus on the conditional powers based on the DF-A for the fixed margin approach.

Under the planned maximum sample size and a given observed value of (T_k^{AS}, T_k^{NI}) , the conditional power based on the DF-A evaluated at the k th analysis is

$$CP = \begin{cases} \Pr[\{T_K^{AS} > c_K^{AS}\} \cap \{T_K^{NI} > c_K^{NI}\} | T_k^{AS} = t_k^{AS}, T_k^{NI} = t_k^{NI}], & \text{if } T_l^{NI} \leq c_l^{NI} \text{ and } T_l^{AS} \leq c_l^{AS} \text{ for all } l=1, \dots, k, \\ \Pr[T_K^{NI} > c_K^{NI} | T_k^{NI} = t_k^{NI}], & \text{if } T_l^{NI} \leq c_l^{NI} \text{ for all } l=1, \dots, k, \text{ and } T_{l'}^{AS} > c_{l'}^{AS} \text{ for any } l'=1, \dots, k \end{cases}$$

(B1)

For the fixed margin approach, the conditional distribution of

$(T_K^{AS}, T_K^{NI} | T_k^{AS} = t_k^{AS}, T_k^{NI} = t_k^{NI})$ is a bivariate normal distribution with mean vector given as

$$\left(\sqrt{m}t_k^{AS} + \sqrt{1-m} \sqrt{n_{EK} - n_{Ek}} \frac{\delta_1}{\sqrt{1/C_R + 1/C_P}}, \sqrt{m}t_k^{NI} + \sqrt{1-m} \sqrt{n_{EK} - n_{Ek}} \frac{\delta_2}{\sqrt{1+1/C_R}} \right)^T$$

and covariance matrix given as $(1-m) \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, where $m = n_{EK}/n_{Ek}$, $\delta_1 = (\mu_R - \mu_P -)/\sigma$ and $\delta_2 = (\mu_E - \mu_R +)/\sigma$. Therefore, the conditional power (B1) can be rewritten as

$$CP(\delta_1, \delta_2) = \begin{cases} \Phi_2 \left(-\frac{c_K^{AS} - \sqrt{m}t_k^{AS}}{\sqrt{1-m}} + \sqrt{n_{EK} - n_{Ek}} \frac{\delta_1}{\sqrt{1/C_R + 1/C_P}}, -\frac{c_K^{NI} - \sqrt{m}t_k^{NI}}{\sqrt{1-m}} + \sqrt{n_{EK} - n_{Ek}} \frac{\delta_2}{\sqrt{1+1/C_R}} \right), & \text{if } T_l^{NI} \leq c_l^{NI} \text{ and } T_l^{AS} \leq c_l^{AS} \text{ for all } l=1, \dots, k, \\ 1 - \Phi_1 \left(\frac{c_K^{NI} - \sqrt{m}t_k^{NI}}{\sqrt{1-m}} - \sqrt{n_{EK} - n_{Ek}} \frac{\delta_2}{\sqrt{1+1/C_R}} \right), & \text{if } T_l^{NI} \leq c_l^{NI} \text{ for all } l=1, \dots, k, \text{ and } T_{l'}^{AS} > c_{l'}^{AS} \text{ for any } l'=1, \dots, k, \end{cases}$$

where $\Phi_1(\cdot)$ and $\Phi_2(\cdot)$ are the cumulative distribution functions of the standardized univariate and bivariate normal distributions.

Figure 1.

Behavior of the power for rejecting both (i) the null hypotheses for the AS and (ii) NI, as a function of the experimental intervention group sample size, when using DF-A or DF-B for the fixed margin approach, where the number of planned analyses is $K = 4$. Three allocation ratios of $n_{Ek}:n_{Rk}:n_{Pk}$, are considered i.e., (i) 1:1:1 ($C_R = C_P = 1$), (ii) 2:1:1 ($C_R = C_P = 1/2$), and (iii) 5:4:1 ($C_R = 4/5$, $C_P = 1/5$). Four stopping boundary combinations are considered: (i) the OF for both AS and NI (OF-OF), (ii) the OF for AS and the PC for NI (OF-PC), (iii) the PC for AS and the OF for NI (PC-OF), and (iv) the PC for AS and NI (PC-PC), using the LD alpha-spending method, with equally spaced increments in information. The means (μ_E , μ_R , μ_P) are (10, 10, 5) with a common standard deviation $\sigma = 6.5$. The pre-specified noninferiority margin is $\delta = 2.5$. Under these parameter configurations, the power for rejecting both null hypotheses is calculated at the significance level of $\alpha = 2.5\%$ for a one-sided test under H_1^{AS} and H_1^{NI} .

Figure 2.

Behavior of the power for rejecting both (i) the null hypotheses for the AS and (ii) NI, as a function of the experimental intervention group sample size, when using DF-A or DF-B for the fraction approach, where the number of planned analyses is $K = 4$. Three allocation ratios of $n_{Ek}:n_{Rk}:n_{Pk}$, are considered i.e., (i) 1:1:1 ($C_R = C_P = 1$), (ii) 2:1:1 ($C_R = C_P = 1/2$), and (iii) 5:4:1 ($C_R = 4/5$, $C_P = 1/5$). Four stopping boundary combinations are considered: (i) the OF for both AS and NI (OF-OF), (ii) the OF for AS and the PC for NI (OF-PC), (iii) the PC for AS and the OF for NI (PC-OF), and (iv) the PC for AS and NI (PC-PC), using the LD alpha-spending method, with equally spaced increments in information. The means (μ_E , μ_R , μ_P) are (10, 10, 5) with a common standard deviation $\sigma = 6.5$. The pre-specified noninferiority fraction is $\theta = 0.5$. Under these parameter configurations, the power for rejecting both null hypotheses is calculated at the significance level of $\alpha = 2.5\%$ for a one-sided test under H_1^{AS} and H_1^{NI} .

Figure 3.

Behavior of the Type I error rate for rejecting both null hypotheses (i) for the AS and (ii) NI as a function of the experimental intervention group sample size, when using DF-A for the fixed margin approach, where the number of planned analyses is $K = 4$. Three allocation ratios of $n_{E,K}:n_{R,K}:n_{P,K}$ are considered i.e., (i) 1:1:1 ($C_R = C_P = 1$), (ii) 2:1:1 ($C_R = C_P = 1/2$), and (iii) 5:4:1 ($C_R = 4/5$, $C_P = 1/5$). Four stopping boundary combinations are considered: (i) the OF for both AS and NI (OF-OF), (ii) the OF for AS and the PC for NI (OF-PC), (iii) the PC for AS and the OF for NI (PC-OF), and (iv) the PC for AS and NI (PC-PC), using the LD alpha-spending method, with equally spaced increments in information. The means (μ_E , μ_R , μ_P) are (10,10,5), with a common standard deviation $\sigma = 6.5$. The pre-specified noninferiority margin is $\delta = 2.5$. Under these parameter configurations, the Type I error for the power for rejecting both null hypotheses is calculated at the significance level of $\alpha = 2.5\%$ for one-sided test under three situations, i.e., (i) H_0^{AS} and H_0^{NI} , (ii) H_0^{AS} or (iii) H_0^{NI} .

Figure 4.

Behavior of the Type I error rate for rejecting both null hypotheses (i) for the AS and (ii) NI as a function of the experimental intervention group sample size, when using DF-B for the fixed margin approach, where the number of planned analyses is $K = 4$. Three allocation ratios of $n_{E,k}:n_{R,k}:n_{P,k}$ are considered i.e., (i) 1:1:1 ($C_R = C_P = 1$), (ii) 2:1:1 ($C_R = C_P = 1/2$), and (iii) 5:4:1 ($C_R = 4/5$, $C_P = 1/5$). Four stopping boundary combinations are considered: (i) the OF for both AS and NI (OF-OF), (ii) the OF for AS and the PC for NI (OF-PC), (iii) the PC for AS and the OF for NI (PC-OF), and (iv) the PC for AS and NI (PC-PC), using the LD alpha-spending method, with equally spaced increments in information. The means (μ_E , μ_R , μ_P) are (10,10,5), with a common standard deviation $\sigma = 6.5$. The pre-specified noninferiority margin is $\delta = 2.5$. Under these parameter configurations, the Type I error for the power for rejecting both null hypotheses is calculated at the significance level of $\alpha = 2.5\%$ for one-sided test under three situations, i.e., (i) H_0^{AS} and H_0^{NI} , (ii) H_0^{AS} or (iii) H_0^{NI} .

Figure 5.

Behavior of the Type I error rate for rejecting both null hypotheses (i) for the AS and (ii) NI as a function of the experimental intervention group sample size, when using DF-A for the fraction approach, where the number of planned analyses is $K = 4$. Three allocation ratios of $n_{E,k}:n_{R,k}:n_{P,k}$ are considered i.e., (i) 1:1:1 ($C_R = C_P = 1$), (ii) 2:1:1 ($C_R = C_P = 1/2$), and (iii) 5:4:1 ($C_R = 4/5, C_P = 1/5$). Four stopping boundary combinations are considered: (i) the OF for both AS and NI (OF-OF), (ii) the OF for AS and the PC for NI (OF-PC), (iii) the PC for AS and the OF for NI (PC-OF), and (iv) the PC for AS and NI (PC-PC), using the LD alpha-spending method, with equally spaced increments in information. The means (μ_E, μ_R, μ_P) are (10,10,5), with a common standard deviation $\sigma = 6.5$. The pre-specified noninferiority fraction is $\theta = 0.5$. Under these parameter configurations, the Type I error for the power for rejecting both null hypotheses is calculated at the significance level of $\alpha = 2.5\%$ for one-sided test under three situations, i.e., (i) $H_0^{AS} \cap H_0^{NI}$, (ii) H_0^{AS} or (iii) H_0^{NI} .

Figure 6.

Behavior of the Type I error rate for rejecting both null hypotheses (i) for the AS and (ii) NI as a function of the experimental intervention group sample size, when using DF-B for the fraction approach, where the number of planned analyses is $K = 4$. Three allocation ratios of $\eta_{EK}:\eta_{RK}:\eta_{PK}$, are considered i.e., (i) 1:1:1 ($C_R = C_P = 1$), (ii) 2:1:1 ($C_R = C_P = 1/2$), and (iii) 5:4:1 ($C_R = 4/5$, $C_P = 1/5$). Four stopping boundary combinations are considered: (i) the OF for both AS and NI (OF-OF), (ii) the OF for AS and the PC for NI (OF-PC), (iii) the PC for AS and the OF for NI (PC-OF), and (iv) the PC for AS and NI (PC-PC), using the LD alpha-spending method, with equally spaced increments in information. The means (μ_E, μ_R, μ_P) are (10,10,5), with a common standard deviation $\sigma = 6.5$. The pre-specified noninferiority fraction is $\theta = 0.5$. Under these parameter configurations, the Type I error for the power for rejecting both null hypotheses is calculated at the significance level of $\alpha = 2.5\%$ for one-sided test under three situations, i.e., (i) H_0^{AS} and H_0^{NI} , (ii) H_0^{AS} or (iii) H_0^{NI} .

Figure 7.

The impact of the sample size recalculation on the overall empirical power for rejecting both null hypotheses for the AS and the NI, when using DF-A for the fixed margin and fraction approaches, where the number of planned analyses is $K = 4$. One sample size recalculation is conducted based on the observed effect at either of 1/4, 1/2, 3/4 of information time, where $\lambda = 1.5$, $\gamma = 0.875$ and $\eta = 1.125$. The planned maximum samples size $N = 426$ for the fixed margin approach and $N = 240$ for the fraction approach (equally sized groups $C_R = C_P = 1$) are calculated to evaluate the AS and the NI with the power of 80% at the significance level of 2.5% for a one-sided test in the fixed-sample design, where $(\mu_E, \mu_R, \mu_P) = (10, 10, 5)$, $\sigma = 6.5$ and $\theta = 2.5$ for the fixed margin approach and $\theta = 0.5$ for the fraction approach. 100,000 replications were used for the empirical power assessment.

Figure 8.

The impact of sample size recalculation on overall actual Type I error rate for rejecting both null hypotheses for the AS and the NI, when using DF-A for the fixed margin approach, where the number of planned analyses is $K = 4$. One sample size recalculation is conducted based on the observed effect at either of 1/4, 1/2, 3/4 of information time, where $\lambda = 1.5$, $\gamma = 0.875$ and $\eta = 1.125$. The planned maximum sample size $N = 426$ (equally sized groups $C_R = C_P = 1$) is calculated to demonstrate the AS and the NI with the power of 80% at the significance level of 2.5% for a one-sided test in the fixed-sample design, where $(\mu_E, \mu_R, \mu_P) = (10, 10, 5)$, $\sigma = 6.5$ and $\delta = 2.5$. The actual Type I error rate is evaluated at the significance level of $\alpha = 2.5\%$ for one-side test under three situations, i.e., (i) H_0^{AS} and H_0^{NI} , (ii) H_0^{AS} or (iii) H_0^{NI} . 1,000,000 replications were used for the actual Type I error assessment.

Figure 9.

The Impact of sample size recalculation on overall actual Type I error rate for rejecting both null hypotheses for the AS and the NI, when using DF-A for the fraction approach, where the number of planned analyses is $K = 4$. One sample size recalculation is conducted based on the observed effect at either of 1/4, 1/2, 3/4 of information time, where $\lambda = 1.5$, $\gamma = 0.875$ and $\eta = 1.125$. The planned maximum samples size $N = 240$ (equally sized groups $C_R = C_P = 1$) is calculated to demonstrate the AS and the NI with the power of 80% at the significance level of 2.5% for a one-sided test in the fixed-sample design, where $(\mu_E, \mu_R, \mu_P) = (10, 10, 5)$, $\sigma = 6.5$ and $\theta = 0.5$. The actual Type I error rate is evaluated at the significance level of $\alpha = 2.5\%$ for one-side test under three situations, i.e., (i) H_0^{AS} and H_0^{NI} , (ii) H_0^{AS} or (iii) H_0^{NI} . 1,000,000 replications were used for the actual Type I error assessment.

Figure 10.

Behavior of the Type I error for rejecting both null hypotheses (i) for AS and (ii) NI as a function of the experimental intervention group sample size, when using adaptive Type I error allocation for the fixed margin approach, where the planned number of analyses is $K = 4$, and the allocation ratio $n_{E,k}:n_{R,k}:n_{P,k}$ is 1:1:1 ($C_R = C_P = 1$). Four stopping boundary combinations are considered: (i) the OF for both AS and NI (OF-OF), (ii) the OF for AS and the PC for NI (OF-PC), (iii) the PC for AS and the OF for NI (PC-OF), and (iv) the PC for AS and NI (PC-PC), using the LD alpha-spending method, with equally spaced increments in information. The means (μ_E, μ_R, μ_P) are (10, 10, 5), with a common standard deviation of $\sigma = 6.5$. The pre-specified noninferiority margin is $\delta = 2.5$. Under these parameter configurations, the Type I error rate is calculated under the three situations, (i) H_0^{AS} and H_0^{NI} , (ii) H_0^{AS} or (iii) H_0^{NI} .

Table 1

The MSS and ASN for demonstrating the AS and the NI with the power of $1 - \beta = 80\%$ at the significance level of $\alpha = 2.5\%$ for a one-sided test, where the maximum planned number of analyses is $K = 4$ and the means (μ_E, μ_R, μ_P) are $(10, 10, 5)$ with a common standard deviation of $\sigma = 6.5$. The pre-specified noninferiority margin for the fixed margin approach is $= 2.5$ and the corresponding fraction for the fraction approach is $\theta = 0.5$. The three allocation ratios of $\eta_{E|K}:\eta_{R|K}:\eta_{P|K}$ considered are (i) 1:1:1 ($C_R = C_P = 1$), (ii) 2:1:1 ($C_R = C_P = 1/2$), and (iii) 5:4:1 ($C_R = 4/5, C_P = 1/5$). The critical values are based on the O'Brien–Fleming-type boundary (OF), Pocock-type boundary (PC) or their combinations, using the LD alpha-spending method with equally spaced increment in information. Four stopping boundary combinations considered are: (i) the OF for both AS and NI (OF-OF), (ii) the OF for AS and the PC for NI (OF-PC), (iii) the PC for AS and the OF for NI (PC-OF), and (iv) the PC for AS and NI (PC-PC).

Approach	DF	Boundary	$C_R = C_P = 1$			$C_R = C_P = 1/2$			$C_R = 4/5, C_P = 1/5$			
			MSS	ASNI	ASN2	MSS	ASNI	ASN2	MSS	ASNI	ASN2	
Fixed margin	DF-A	OF-OF	435	388	375	510	456	445	682	582	580	
		OF-PC	480	410	399	548	474	466	690	579	578	
		PC-OF	471	402	371	564	477	452	796	610	602	
		PC-PC	507	400	376	594	470	450	798	577	573	
	DF-B	OF-OF	438	391	-	512	457	-	684	583	-	
		OF-PC	480	411	-	548	475	-	690	580	-	
		PC-OF	480	411	-	578	490	-	812	626	-	
		PC-PC	513	406	-	602	478	-	808	586	-	
	Fraction	DF-A	OF-OF	246	206	-	226	193	-	294	248	-
			OF-PC	294	220	-	266	208	-	342	265	-
			PC-OF	246	204	-	232	194	-	304	250	-
			PC-PC	291	204	-	266	194	-	344	244	-
DF-B		OF-OF	246	206	-	226	193	-	294	248	-	
		OF-PC	294	220	-	266	208	-	342	265	-	
		PC-OF	246	204	-	234	196	-	306	252	-	
		PC-PC	291	204	-	266	194	-	344	244	-	

ASN1: expected sample size without dropping of the placebo group

ASN2: expected sample size with dropping of the placebo group

Table 2

The ASN when sample size recalculation is performed, using DF-A for the fixed margin and fraction approaches, where the number of planned analyses is $K = 4$. One sample size recalculation is conducted based on the observed effect at either of $1/4$, $1/2$, $3/4$ of information time, where $\lambda = 1.5$, $\gamma = 0.875$ and $\eta = 1.125$. The planned maximum samples size $N = 426$ for the fixed margin approach and $N = 240$ for the fraction approach (equally sized groups $\bar{C}_R = \bar{C}_P = 1$) are calculated to evaluate the AS and the NI with the power of 80% at the significance level of 2.5% for a one-sided test in the fixed-sample design, where $(\mu_E, \mu_R, \mu_P) = (10, 10, 5)$, $\sigma = 6.5$ and $\theta = 2.5$ for the fixed margin approach and $\theta = 0.5$ for the fraction approach.

Approach	Boundary	Timing for Recalculation	Increase only		Decrease only		Increase or decrease	
			ASN1	ASN2	ASN1	ASN2	ASN1	ASN2
Fixed margin	OF-OF	No-change	383	364	383	364	383	364
		1/4	456	433	363	346	437	415
		1/2	455	428	367	350	439	415
		3/4	459	430	380	362	457	427
	PC-OF	No-change	375	341	375	341	375	341
		1/4	448	404	362	331	436	395
		1/2	453	409	363	333	440	400
		3/4	463	417	374	340	460	415
	PC-PC	No-change	362	332	362	332	362	332
		1/4	429	392	355	326	423	386
		1/2	448	406	361	331	445	404
		3/4	462	416	362	332	463	417
Fraction	OF-OF	No-change	202	-	202	-	202	-
		1/4	235	-	187	-	219	-
		1/2	236	-	193	-	226	-
		3/4	238	-	202	-	237	-
	PC-OF	No-change	200	-	200	-	200	-
		1/4	233	-	186	-	219	-
		1/2	235	-	192	-	226	-
		3/4	237	-	200	-	237	-
	PC-PC	No-change	183	-	183	-	183	-
		1/4	213	-	177	-	207	-
		1/2	221	-	182	-	220	-

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Approach	Boundary	Timing for Recalculation	Increase only		Decrease only		Increase or decrease	
			ASN1	ASN2	ASN1	ASN2	ASN1	ASN2
		3/4	228	-	182	-	228	-

ASN1: expected sample size without dropping of the placebo group

ASN2: expected sample size with dropping of the placebo group