Preliminary Report

Two-Phase, Three-Stage Adaptive Designs in Clinical Trials

Hiroyuki Uesaka^{*1}, Toshihiko Morikawa^{*2} and Akiko Kada^{*3}

^{*1} The Center for Advanced Medical Engineering and Education,
Osaka University (formerly)

^{*2} Biostatistics Center, Kurume University (formerly)

^{*3} Clinical Research Center, National Hospital Organization
Nagoya Medical Center

e-mail:pyracantha@feel.ocn.ne.jp

In clinical investigator initiated clinical trials, we frequently encounter the situation where it is very difficult to estimate the effect size and the clinically meaningful difference between the treatment and control groups. In this paper we explore various two-phase, three-stage adaptive designs which can be applied to this situation. The first phase determines whether the trial should proceed or not. If the decision is to proceed, then the sample size is re-estimated. The second phase consists of two stages, but the sample size is not re-estimated. We introduce hybrid and alpha-split designs, adding to two existing adaptive designs: Bauer-Köhne design and Lehman-Wassmer design. Main findings are: 1) the differences in the overall powers and the average sample number (ASN)s among these designs are small, except for the design which includes O'Brien-Fleming boundaries and the alpha-split design, 2) the two-phase, three-stage design suffers a relative loss of power by 15% but the ASN is less than 50%, as compared to the single stage design under the optimal condition, 3) two-phase, three-stage design compares with the three-stage group sequential design. We conclude that the design can be a candidate when there is no useful information on the effect size.

Key words: adaptive design, sample size re-estimation, group sequential design, conditional error function, Fisher's product criterion, clinical investigator initiated trial.

1. Introduction

Clinical investigators plan clinical trials for various purposes, such as to establish combination therapies for existing treatments, to test the effect of an existing therapy on a restricted population, or to test the effectiveness of existing therapies that are widely used but their effectiveness have not been confirmed through clinical trials. In these trials, useful information for determining sample sizes is often unavailable. In the ordinary clinical development of a new

drug by the pharmaceutical industry, the sample size required for the confirmatory clinical trial is estimated by integrating the information on the effect sizes that were obtained from preceding clinical trials for that drug, the effect sizes of the existing therapies, and a clinically or commercially meaningful difference (CMD) between the test drug and the control or the smallest CMD (SCMD).

If a CMD is available, a sample size may be determined to give a pre-defined power at the effect size of the CMD, even if the information on the effect of the test therapy is not available. This type of a clinical trial can be referred to as a pass/fail trial, because it is possible to judge that if the trial does not give significant result the effect size of the test therapy is probably less than the CMD. When neither CMD nor information on the effect of the test therapy is available, a sample size of the exploratory trial may also be difficult to be determined. Furthermore, clinical investigators who plan clinical trials often confront various difficulties, such as a lack of funds, a restricted study period, and a lack of clinical trial experts. Therefore, even though neither the information on the CMD nor the size of effect of the test therapy is available, still a clinical trial may be planned as a confirmatory trial (for example, see Ogawa et al., 2008, and Minamino et al., 2012). This type of clinical trial can be referred to as a first, definitive trial. A pragmatic trial is mainly conducted in phase IV, and its primary endpoint is often different from that of phase II and III trials. Such a pragmatic trial is also characterized as the first, definitive trial.

When the information on determining a sample size is very scarce or it is very difficult to determine a CMD or SCMD, one approach is to apply adaptive designs that allow mid-trial sample size estimation and determining termination or continuation of the trial, such as those by Bauer and Köhne (1994), Lehmacher and Wassmer (1999), Cui, Hung, and Wang (1999), and Müller and Schäfer (2004). However, The adaptive approach has been strongly criticized by Shih (2001, 2006), Fleming (2006), and Emerson, Levin, and Emerson (2011). Instead, they claim that the SCMD should be determined even when doing so is difficult and advocate the use of a group sequential design with the sample size calculated based on the SCMD. Several authors have discussed that adaptive designs are not as efficient as group sequential designs, or that the reduction in average sample number (ASN) compared to the ordinary group sequential design is small (see, for example, Jennison and Turnbull, 2003, 2006; Emerson, 2006; Jahn-Eimermacher and Hommel, 2007; and Levin, Emerson, and Emerson, 2011). Tsiatis and Mehta (2003) demonstrated that, for any adaptive design, there exists a group sequential design that is better than the adaptive design, although this group sequential design is not realistic in actual situations. Jennison and Turnbull (2006) and Levine, Emerson, and Emerson (2013) proposed modifications to the group sequential design by adjusting the sample size based on interim results and the (S)CMD. Emerson (2006) and Emerson, Levin, and Emerson (2011) discussed various aspects of optimality conditions of a trial design and conclude that adaptive designs cannot prove its superiority to group sequential designs. The reader may also refer to Posch, Bauer, and

Brannath (2003), Bauer and Koenig (2006), Mehta and Pocock (2011).

However, usually it is very difficult to determine a (S)CMD because enough information on the existing therapies needs to be available. Furthermore, the (S)CMD can be affected by various factors, such as the risk-to-benefit ratio of existing therapies, new findings about the efficacy or safety of existing therapies (see, for example, Posch, Bauer, and Brannath, 2003, and Lan, Soo, Siu, and Wang, 2005), the statistical method and patient population to determine the CMD, and the external criterion of the important change in the health status, etc. (see Copay, Subach, Glassman, Polly and Schuler, 2007, Dyck and O'Brien, 1989).

Thus there still exist situations where there are no reasonable information for sample size determination and clinical investigators plan first, definitive clinical trials. Therefore we need a design which can reduce the risk that the trial gives an equivocal result. The adaptive design seems to be useful for this purpose. As many authors of adaptive designs do, one may start with a sample size which is determined by assuming an optimistic effect size, and perform interim analyses to re-estimate a sample size and determine whether the trial is to be continued. However, if the initial sample size is not sufficiently large, the estimated treatment effect may be so unstable that the estimated sample size may be too large or too small to reject the null hypothesis. Thus, it is important to perform interim analysis at least once in the next stage. Concerning the group sequential design, it is known that a design involving four or five interim analyses reduces the ASN compared to a design involving one or two interim analyses. However, it is important to note that the greater the frequency with which the interim results are evaluated, the greater the unreliability of the results of the trial. This is because frequent interim analysis increases the possibility of bias of the results. Furthermore, in general frequent interim analysis prolongs the study period and increases the cost and workloads. Thus, the number of interim analyses should be kept as small as possible. In addition, the number of re-estimations of the sample size should also be minimized, because the more frequently the sample size is re-estimated, the greater the range for interpreting the results of the trial. These considerations and the ethical, scientific and practical considerations lead us to the following conditions about the trial design:

- (1) Since we have little or no information on the effect size and (S)CMD, the sample size is to be re-estimated.
- (2) The number of re-estimations of the sample size should be as small as possible.
- (3) The instability of the re-estimated effect size, i.e., the possibility of great overestimating or underestimating the effect size, is to be taken into account.
- (4) The number of interim analyses should be as small as possible.
- (5) The number of subjects should be as small as possible.
- (6) The power should be as high as possible.
- (7) The allowable maximum sample size should be set.

As it is difficult to satisfy all these conditions simultaneously, we focus on (1), (2), (3), (4)

and (5). This consideration suggests a design which consists of two phases. At the end of the first phase we perform an analysis in order to judge whether it would be reasonable to proceed to the second phase. In case of proceeding to the second phase, we estimate the effect size and revise the sample size. Based on the condition (2), we perform sample size re-estimation only once, that is, it is performed at the end of the first phase. Considering (3), (4) and (5), in the second phase a group sequential design of two stages is adopted so that the trial allows early termination due to futility or efficacy. We refer to this design as a two-phase, three-stage design. The conditions (5) and (6) are measures of the performance of the designs. The condition (7) is imposed for preventing impractically large sample size. Existing adaptive designs, such as the Bauer-Köhne product probability design with three stages, and the Lehmache-Wassmer adaptive group sequential design can be used as the two-phase, three-stage design.

So far we have discussed late stage clinical trials. However, a clinical investigator may also try to develop a new therapy for a rare, intractable, or incurable disease for which there are no effective therapies, or try to explore a biological variable that heralds a clinical response. In these situations neither the information on the effect size of the test treatment nor (S)CMD exists. Even though, the clinical investigator plans a clinical trial to determine whether the test treatment generates some response. It is important that the design of such a clinical trial should allow for the determination of whether conducting further studies will be worthwhile. For this purpose, one approach is to design a clinical trial to attain statistically significant results with a higher type-I error rate than the confirmatory trial, for example, a one-sided significance level of 5 to 20% and a low type-II error rate such as 10%. Due to a lack of information regarding the effect size, one possibility is to use a two-phase, three-stage design.

As we demonstrate in later sections, a two-phase, three-stage design is defined by numerous design parameters, such as the overall type-I error rate of the trial, type-II error rate of each phase, the futility stopping criterion at each phase, the critical points of the group sequential design of the second phase, the method of sample size re-estimation, the sample size of the first phase, and so on. We need to determine appropriate design parameters for various situations in order to provide some guidance on how to determine design parameters. Therefore, we consider various scenarios for clinical trials and evaluate the performance of various designs.

In Section 2, we explain the basic concept of the two-phase, three-stage designs and define four types of designs. In Section 3, we discuss methods of sample-size determination, and in Section 4, we explain the criteria used to determine the performance of these designs. Evaluation results are given in Section 5, and we determine the nearly optimal design parameters for each design type. In Section 6, the selected designs are compared. In Section 7 our design is compared to a few other designs. Finally, in Section 8, we discuss miscellaneous problems for the two-phase, three-stage designs and summarize the results.

2. Three-stage designs consisting of two phases

2.1 Design considerations and notation

Let us consider a randomized, parallel-group, controlled clinical trial of a test treatment T, in which C is the control. Let X be the observed random variable, and we will assume that it follows a normal distribution with mean μ_T for the test treatment and μ_C for the control, and that it has a known variance σ^2 . Let $\delta = \mu_T - \mu_C$ be the unknown difference between the two means. We will test a null hypothesis $H_0: \delta = 0$ against an alternative $H_1: \delta > 0$. Usually, in a confirmatory trial, there is an estimated value, $\delta = \delta_1$, for which the sample size of the trial is determined so that a prespecified power, say $1 - \beta$, is attained. However, as explained in the introduction, we often encounter a situation where it is extremely difficult to determine either a reasonable estimate for δ_1 or the CMD. For this situation, we propose two-phase, three-stage design. In the first phase the sample size is determined to be a value that will attain a high power of $1 - \beta_1$ at an initial estimate of δ_1 . When the test is not significant, we judge that the true effect size is likely to be less than δ_1 and then we judge whether it would be reasonable to proceed to the second phase. In the second phase, we calculate a sample size by re-estimating the effect size from the first phase and we adopt a two stage group sequential design for the one-sided hypothesis.

Adaptive designs for testing H_0 against H_1 have already been proposed. Bauer and Köhne (1994) proposed a two-stage adaptive design, and Wassmer (1999) extended the Bauer-Köhne design to include three or more stages. We refer to the extended design as the Bauer-Köhne-Wassmer design. Lehmacher and Wassmer (1999) proposed an adaptive group sequential design that was based on the unweighted cumulative test statistics. We refer to this design as Lehmacher-Wassmer design. Each of these designs can be used as a two-phase, three-stage design. Therefore, after describing the notation used herein, we first describe these two designs and then define two new designs: hybrid design and alpha-split design.

For the two-phase, three-stage design, we have three stages, and we assume that the sample sizes n_k , $k = 1, 2, 3$, are equal for the two treatment groups at each stage. We denote the cumulative sample size by $N_k = \sum_{j=1}^k n_j$, $k = 1, 2, 3$. Let the sample mean at the k -th stage of the i -th group be $\bar{X}_{i.}^{(k)}$, $i = T, C$, and let the stage-wise test statistic of H_0 against H_1 be $Z_k = \frac{\sqrt{n_k/2}(\bar{X}_{T.}^{(k)} - \bar{X}_{C.}^{(k)})}{\sigma}$. Let us write $p_k(z_k) = 1 - \Phi(Z_k)$, $k = 1, 2, 3$, where Φ represents the cumulative distribution function of the standard normal distribution. Also, we define a cumulative test statistic up to the k -th stage by $Y_k = \frac{1}{\sqrt{k}} \sum_{l=1}^k Z_l$, and its p-value is $p_k(Y_k) = 1 - \Phi(Y_k)$, $k = 1, 2, 3$.

Below, we discuss adaptive two-phase, three-stage designs, each of which have an overall type-I error rate α . We denote the upper α points of the standard normal distribution and the chi-square distribution with f degrees of freedom by Z_α and $\chi_f^2(\alpha)$, respectively. The probability density function of the standard normal distribution is denoted by $\phi(\cdot)$.

2.2 The Bauer-Köhne-Wassmer design and an adaptive group sequential design

For the three-stage Bauer-Köhne-Wassmer design, let us write $C_k(\alpha) = \exp(-\chi_{2k}^2(\alpha)/2)$ for any α such that $0 < \alpha < 1$, $k = 1, 2, 3$. Let α_k , $k = 1, 2, 3$ be real numbers between 0 and 1. We will refer to α_k as the local significance level and take a value α_0 , such that $0 < \alpha_0 < 1 - \alpha_1$, as the stopping criterion for futility. Let the type-I error probability of the k -th stage be denoted by EP_k , $k = 1, 2, 3$, where $EP_1 = \alpha_1$ and $EP_1 + EP_2 + EP_3 = \alpha$. Then, we have the following equations in terms of $C_k(\alpha)$:

$$\begin{aligned} EP_1 &= C_1(\alpha_1), \\ EP_2 &= C_2(\alpha_2)(\log \alpha_0 - \log \alpha_1), \\ EP_3 &= C_3(\alpha_3)(\log \alpha_0 - \log \alpha_1)\{\log \alpha_0 - \log C_2(\alpha_2) + (\log \alpha_0 + \log \alpha_1)/2\}, \end{aligned}$$

for which the solutions $C_1(\alpha_1)$, $C_2(\alpha_2)$, and $C_3(\alpha_3)$ are denoted by $C_{3,1}(\alpha, \alpha_0)$, $C_{3,2}(\alpha, \alpha_0)$, and $C_{3,3}(\alpha, \alpha_0)$, respectively.

Now, let us define the testing procedure. In the first stage, we stop the trial if we reject the null hypothesis because $p_1(z_1) \leq \alpha_1$, or if we judge the trial to be futile because $p_1(z_1) \geq \alpha_0$. Otherwise, we re-estimate the sample size of the second and third stages based on the result of the first phase and proceed to the second stage. How to determine the sample sizes will be explained in Section 3. In the second stage, we stop the trial if we reject the null hypothesis because $p_2(z_2) \leq \frac{C_{3,2}(\alpha, \alpha_0)}{p_1(z_1)}$, or if we judge the trial to be futile because $p_2(z_2) \geq \alpha_0$. Otherwise, we proceed to the third stage, in which the null hypothesis is rejected if $p_3(z_3) \leq \frac{C_{3,3}(\alpha, \alpha_0)}{p_1(z_1)p_2(z_2)}$.

Wassmer (1999) proposed three criteria for determining the local significance levels:

CR1: criterion of the constant local significance level, where $\alpha_1 = \alpha_2 = \alpha_3$;

CR2: criterion of the full level α in the final stage, where $\alpha_1 = \alpha_2$, $\alpha_3 = \alpha$;

CR3: criterion of without-treatment-stage interaction, where

$$C_3(\alpha, \alpha_0) = \exp(-\chi_6^2(\alpha)/2), \quad C_k(\alpha, \alpha_0) = C_{k+1}(\alpha, \alpha_0)/\alpha_0, \quad k = 1, 2.$$

We define the conditional significance level of the second stage under the condition that $Z_1 = z_1$ as $\alpha_2(z_1, \alpha_0) = \frac{C_{3,2}(\alpha, \alpha_0)}{p_1(z_1)}$, and the conditional significance level of the third stage, given that $Z_1 = z_1$ and $Z_2 = z_2$, as $\alpha_3(z_1, z_2, \alpha_0) = \frac{C_{3,3}(\alpha, \alpha_0)}{p_1(z_1)p_2(z_2)}$. In order to adjust this design to the two-phase, three-stage design, the first phase consists of the first stage, and the second and third stages make up the second phase.

Lehmacher and Wassmer (1999) proposed an adaptive group sequential design based on unweighted cumulative test statistics Y_k , $k = 1, 2, 3$. Since our hypothesis is one-sided, the trial should allow stopping due to futility after either the first or second stage. The critical points for rejection and acceptance for a futility stop for a one-sided hypothesis have been proposed by DeMets and Ware (1980, 1982), Emerson and Fleming (1989), and Pampallona and Tsiatis

(1994), for example. Among these criteria, the asymmetric boundary of DeMets and Ware (1980) is very simple and accepts the null hypothesis when the p-value $p_k(Y_k)$, $k = 1, 2$ is greater than or equal to the constant value of α_0 , i.e., the trial is terminated when $Y_k \leq Z_{\alpha_0}$, $k = 1, 2$. Boundaries for rejecting the null hypothesis have also been defined in various manners. One useful boundary is the power family proposed by Wang and Tsiatis (1987). The boundary of the k -th stage is defined by $C_{K,k}(\alpha, \alpha_0) = C_{K,K}(\alpha, \alpha_0)(k/K)^{\Delta-0.5}$, where K is the total number of stages of the group sequential design and Δ is a real value. We hereinafter refer to Δ as the boundary parameter. The Pocock boundary and the O'Brien-Fleming boundary are given as $\Delta = 0.5$ and 0 , respectively. Although Wang and Tsiatis (1987) originally defined this boundary for testing a two-sided hypothesis, we apply the same idea in order to determine the boundary for testing a one-sided hypothesis in conjunction with the stopping boundary for futility. The critical points are determined as follows: under the null hypothesis,

$$\begin{aligned} \alpha = & \Pr.(C_{3,1}(\alpha, \alpha_0) \leq Y_1) + \Pr.(Z_{\alpha_0} < Y_1 < C_{3,1}(\alpha, \alpha_0) \text{ and } C_{3,2}(\alpha, \alpha_0) \leq Y_2) \\ & + \Pr.(Z_{\alpha_0} < Y_1 < C_{3,1}(\alpha, \alpha_0) \text{ and } Z_{\alpha_0} < Y_2 < C_{3,2}(\alpha, \alpha_0) \text{ and } C_{3,3}(\alpha, \alpha_0) \leq Y_3) \end{aligned}$$

This testing procedure can be equivalently defined using Z_k , $k = 1, 2, 3$. The critical point for Z_k is a function of Z_j , $j < k$ for $k = 2, 3$, as follows:

$$\text{for } Z_1, C_{3,1}(\alpha, \alpha_0); \text{ for } Z_2, \sqrt{2}C_{3,2}(\alpha, \alpha_0) - Z_1; \text{ for } Z_3, \sqrt{3}C_{3,3}(\alpha, \alpha_0) - Z_1 - Z_2,$$

and we have

$$\begin{aligned} \alpha = & 1 - \Phi(C_{3,1}(\alpha, \alpha_0)) \\ & + \int_{Z_{\alpha_0}}^{C_{3,1}(\alpha, \alpha_0)} [1 - \Phi(\sqrt{2}C_{3,2}(\alpha, \alpha_0) - z_1)] \\ & + \int_{\sqrt{2}Z_{\alpha_0} - Z_1}^{\sqrt{2}C_{3,2}(\alpha, \alpha_0) - Z_1} \{1 - \Phi(\sqrt{3}C_{3,3}(\alpha, \alpha_0) - z_1 - z_2)\} \phi(z_2) dz_2 \phi(z_1) dz_1. \end{aligned}$$

Thus, the conditional significance levels for the second phase are

$$\alpha_2(z_1, \alpha_0) = 1 - \Phi^{-1}(\sqrt{2}C_{3,2}(\alpha, \alpha_0) - Z_1) \text{ and } \alpha_3(z_1, z_2, \alpha_0) = 1 - \Phi^{-1}(\sqrt{3}C_{3,3}(\alpha, \alpha_0) - Z_1 - Z_2).$$

We define the stopping boundary for futility for Z_1 as $Z_1 \leq Z_{\alpha_0}$ and that for Z_2 as $Z_2 \leq \sqrt{2}Z_{\alpha_0} - Z_1$, and we write $\alpha_{02}(Z_1) = 1 - \Phi(\sqrt{2}Z_{\alpha_0} - Z_1)$.

2.3 A hybrid design

Bauer and Köhne (1994) proposed the use of Fisher's product probability criterion for a two-stage design. The second stage is carried out only when $Z_{\alpha_0} < Z_1 < Z_{\alpha_1}$, and at the second phase the null hypothesis is rejected when the p-value of the test statistic of the second stage is less than or equal to $C_2(\alpha)/p_1(z_1)$, where $C_2(\alpha) = \exp(-\chi_4^2(\alpha)/2)$. For given α and α_0 , the probability of defining the critical point of the first stage, i.e., α_1 , is determined as the solution to the equation $\alpha_1 + C_2(\alpha)(\log \alpha_0 - \log \alpha_1) = \alpha$. This criterion is given by CR2 for $K = 2$ in the Bauer-Köhne-Wassmer design. Another criterion for determining α_1 is given by the equation

Jpn J Biomet Vol.35, No.2, 2014

$\alpha_1 + C_2(\alpha_1)(\log \alpha_0 - \log \alpha_1) = \alpha$, which is derived by CR1 in the Bauer-Köhne-Wassmer design. Let us write $\alpha(z_1, \alpha, \alpha_0) = \frac{C_2(\alpha, \alpha_0)}{p_1(z_1)}$, where $C_2(\alpha, \alpha_0) = C_2(\alpha_1)$ or $C_2(\alpha, \alpha_0) = C_2(\alpha)$.

Now, we apply the two-stage Bauer-Köhne-Wassmer design to the two-phase, three-stage design. The basic idea is to use the conditional error function defined by Proschan and Hunsberger (1995) for the conditional significance level of the second phase. As Posch and Bauer (1999) explained, the product probability criterion defined by Bauer and Köhne (1994) defines a conditional error function. We begin the first phase with a significance level of α_1 , which is derived by CR1 or CR2. If $Z_1 \geq Z_{\alpha_1}$, we stop the trial by rejecting the null hypothesis, and if $Z_1 \leq Z_{\alpha_0}$, we stop the trial due to futility. When $Z_{\alpha_0} < Z_1 < Z_{\alpha_1}$, we proceed to the second phase, which is a trial of the significance level of $\alpha(z_1, \alpha, \alpha_0)$, given that the p-value of the first phase is $\alpha_1 < p_1 < \alpha_0$. Considering the fourth requirement, we define the second phase as a trial of the two-stage group sequential design. We define the critical points for the one-sided hypothesis by the power family proposed by Wang and Tsiatis (1987), and the futility criterion is defined by the DeMets and Ware's asymmetric boundary.

Let us denote the critical points for the cumulative statistics $Y_{2,1} = Z_2$ and $Y_{2,2} = (Z_2 + Z_3)/\sqrt{2}$ by $C_{2,1}(\alpha(z_1, \alpha, \alpha_0), \alpha_0)$ and $C_{2,2}(\alpha(z_1, \alpha, \alpha_0), \alpha_0)$, respectively. The null hypothesis is rejected at the second phase if $Z_2 \geq C_{2,1}(\alpha(z_1, \alpha, \alpha_0), \alpha_0)$, or $Z_{\alpha_0} < Z_2 < C_{2,1}(\alpha(z_1, \alpha, \alpha_0), \alpha_0)$ and $Z_3 \geq \sqrt{2}C_{2,2}(\alpha(z_1, \alpha, \alpha_0), \alpha_0) - Z_2$. These critical points are determined as the solution to the equations

$$\begin{aligned} \alpha(z_1, \alpha, \alpha_0) &= \Pr.(C_{2,1}(\alpha(z_1, \alpha, \alpha_0), \alpha_0) \leq Y_{2,1}) \\ &\quad + \Pr.(Z_{\alpha_0} < Y_{2,1} < C_{2,1}(\alpha(z_1, \alpha, \alpha_0), \alpha_0) \text{ and } C_{2,2}(\alpha(z_1, \alpha, \alpha_0), \alpha_0) \leq Y_{2,2}) \\ &= 1 - \Phi(C_{2,1}(\alpha(z_1, \alpha, \alpha_0), \alpha_0)) \\ &\quad + \int_{Z_{\alpha_0}}^{C_{2,1}(\alpha(z_1, \alpha, \alpha_0), \alpha_0)} \{1 - \Phi(\sqrt{2}C_{2,2}(\alpha(z_1, \alpha, \alpha_0), \alpha_0) - z_2)\} \phi(z_2) dz_2. \end{aligned}$$

Since $\int_{Z_{\alpha_0}}^{Z_{\alpha_1}} \alpha(z_1, \alpha, \alpha_0) \phi(z_1) dz_1 = \alpha - \alpha_1$, the overall type-I error rate is equal to α . These critical points are calculated by numerical integration.

2.4 Alpha-split design

When testing a null hypothesis by combining two trials, the type-I error rate can be split into two components and a different component be allocated to each of the two trials. The null hypothesis can be rejected if at least one trial rejects the null hypothesis based on the component significance level. Like the Bonferroni inequality procedure for multiple comparisons, this procedure is not as powerful as a single trial. However, when the two trials are conducted sequentially, and the second trial is conducted only if the first trial cannot reject the null hypothesis and its p-value is less than α_0 , then in order to maintain the overall type-I error rate under the null hypothesis at α , the significance level of the second trial can be $\alpha_2^* = \alpha_2^*(\alpha, \alpha_1, \alpha_0) = (\alpha - \alpha_1)/(\alpha_0 - \alpha_1)$

in the interval $(Z_{\alpha_0}, Z_{\alpha_1})$. The first and second trials are referred to as the first and second phases, respectively. The second phase is defined as a two-stage group sequential design for a one-sided hypothesis. Thus, the futility stopping criterion at the first stage of the second phase is defined as $Z_2 \leq Z_{\alpha_0}$, and the conditional critical points for the stage-wise test statistics of the second phase are given by the solution to the following equation:

$$\begin{aligned}\alpha_2^* &= \int_{C_{2,2}(\alpha_2^*, \alpha_0)2^{0.5-\Delta}}^{\infty} \phi(z_2) dz_2 + \int_{Z_{\alpha_0}}^{C_{2,2}(\alpha_2^*, \alpha_0)2^{0.5-\Delta}} \int_{\sqrt{2}C_{2,2}(\alpha_2^*, \alpha_0) - z_2}^{\infty} \phi(z_3) \phi(z_2) dz_3 dz_2 \\ &= 1 - \Phi(C_{2,2}(\alpha_2^*, \alpha_0)2^{0.5-\Delta}) + \int_{Z_{\alpha_0}}^{C_{2,2}(\alpha_2^*, \alpha_0)2^{0.5-\Delta}} \{1 - \Phi(\sqrt{2}C_{2,2}(\alpha_2^*, \alpha_0) - z_2)\} \phi(z_2) dz_2.\end{aligned}$$

Let us denote the critical points for the stage-wise test statistics by

$$Z_{\alpha_2(Z_1, \alpha_0)} = C_{2,1}(\alpha_2^*, \alpha_0) = C_{2,2}(\alpha_2^*, \alpha_0)2^{0.5-\Delta}, \quad Z_{\alpha_3(Z_1, Z_2, \alpha_0)} = \sqrt{2}C_{2,2}(\alpha_2^*, \alpha_0) - Z_2,$$

and the conditional significance levels by

$$\alpha_2(z_1, \alpha_0) = 1 - \Phi^{-1}(C_{2,1}(\alpha_2^*, \alpha_0)) \text{ and } \alpha_3(z_1, z_2, \alpha_0) = 1 - \Phi^{-1}(\sqrt{2}C_{2,2}(\alpha_2^*, \alpha_0) - Z_2).$$

The critical points of the second phase are calculated by numerical integration.

In our adaptive designs, the power of the first phase is set to be rather high, say 80 or 90% for the effect size that is initially assumed. Therefore, proceeding to the second phase implies that the true effect size is probably less than the set value. In the second phase, we need to revise the assumed effect size and the sample size. The significance level of the second phase is constant over the range $(Z_{\alpha_0}, Z_{\alpha_1})$. However, the sample size of the second phase depends on the result of the first phase. For the clinical trial to be interpreted as confirmative whenever possible, the significance level should be as low as 2.5% and should not depend on the results of the first phase. In this sense, the alpha-split design more rigorously resembles the stepwise process of exploration and confirmation than do the other adaptive designs discussed herein.

It may be true that an exploratory trial at any level of significance should be conducted and the effect size should be estimated. The second trial can then be conducted based on the sample size that was calculated based on the estimated effect size. The second trial can adopt the significance level of α and be designed as a two-stage group sequential trial. Let us discuss this point from a few aspects. The overall type-I error rate of the two-phase, three-stage design is maintained at α . For the stepwise procedure, assume that we conduct the confirmatory trial after the exploratory trial only if the p-value of the exploratory trial is less than α_0 . Then, the overall type-I error rate of this step-wise procedure is $(1 - \alpha_0)\alpha$. Thus, the overall type-I error rate of the stepwise procedure is less than that of the alpha-split design. On the other hand, the type-I error rate of the second phase of the alpha-split design is $\alpha_2^* = (\alpha - \alpha_1)/(\alpha_0 - \alpha_1)$. Here, α_2^* is greater than α if $\alpha_1 < \alpha(1 - \alpha_0)/(1 - \alpha)$. For example, when $\alpha_0 \leq 0.5$ and $\alpha_1 \leq \alpha/2$, α_2^* is greater than α . When the stepwise procedure is used, the go/no go criterion for the confirmatory trial, the method of sample size determination for each trial, and other design aspects may not be

fully formalized at the beginning of the exploratory trial, whereas when we adopt the alpha-split design, everything should be determined during the design stage.

3. Methods of determining sample size

3.1 Critical and acceptance regions

Now let us denote the rejection region of the k -th stage by R_k , $k = 1, 2, 3$, and the futility-stopping region by A_1 and A_2 . The criterion for futility stopping for the second stage is defined by the p-value and is denoted by $\alpha_{02}(z_1)$. This is because in the case of the Lehman-Wassmer design, the p-value for futility stopping depends on the results of the first phase. For other designs, $\alpha_{02}(z_1) = \alpha_0$. The critical and acceptance regions are

$$\begin{aligned} R_1 &= \{Z_1 \geq Z_{\alpha_1}\}, A_1 = \{Z_1 \leq Z_{\alpha_0}\}; \\ R_2 &= \{Z_{\alpha_0} < Z_1 < Z_{\alpha_1}, Z_{\alpha_2(z_1, \alpha_0)} \leq Z_2\}, A_2 = \{Z_{\alpha_0} < Z_1 < Z_{\alpha_1}, Z_2 < Z_{\alpha_{02}(z_1)}\}; \\ R_3 &= \{Z_{\alpha_0} < Z_1 < Z_{\alpha_1}, Z_{\alpha_{02}(z_1)} < Z_2 < Z_{\alpha_2(z_1, \alpha_0)}, Z_{\alpha_3(z_1, z_2, \alpha_0)} \leq Z_3\}. \end{aligned}$$

The conditional rejection and acceptance regions of the stages in the second phase, given the observed test statistic in the first phase, are denoted by

$$\begin{aligned} R_2(z_1) &= \{Z_1 = z_1, Z_{\alpha_2(z_1, \alpha_0)} \leq Z_2\}, A_2(z_1) = \{Z_1 = z_1, Z_2 \leq Z_{\alpha_{02}(z_1)}\}; \\ R_3(z_1) &= \{Z_1 = z_1, Z_{\alpha_{02}(z_1)} < Z_2 < Z_{\alpha_2(z_1, \alpha_0)}, Z_{\alpha_3(z_1, z_2, \alpha_0)} \leq Z_3\}. \end{aligned}$$

3.2 Estimation of the effect size from the results of the first phase

Many authors use a simple point estimate of the effect size which is defined as the ratio of the observed mean difference to the standard deviation. We denote this estimate by $\hat{\lambda}_{raw}$. However, since the probability is 50% that the observed mean difference is greater than the true value, and the power curve is asymmetric about the true effect size, the expected true conditional power at the second phase is less than the assumed value. To prevent this loss of power, a conservative estimate may be used for the effect size. To keep the expected conditional power at least $1 - \alpha_{CL}$, it is appropriate to use the lower limit of the one-sided confidence interval of the effect size, with the confidence coefficient $1 - \alpha_{CL}$. Uesaka (2003) mentions that the level of α_{CL} of 20% to 30% is generally appropriate. We denote the lower confidence limit of the one-sided confidence interval with confidence coefficient $1 - \alpha_{CL}$ by $\hat{\lambda}_{\alpha_{CL}}$. We will briefly summarize some results of using the conservative estimate $\hat{\lambda}_{\alpha_{CL}}$ in Section 8.

3.3 Sample size of the first phase

The sample size of the first phase is determined by a somewhat optimistic assumption on the effect size, such as a point estimate from the results of a small trial or a simple guessed value. Once we have determined the significance level α_1 and the type-I error rate β_1 for the first phase, and have tentatively determined the effect size λ_0 , then the sample size is given by $n_1 = 2(Z_{\alpha_1} + Z_{\beta_1})^2 / \lambda_0^2$.

3.4 Determining the upper limit of the sample size

The minimum z_1 value for which the second phase is performed is Z_{α_0} . Thus, when $\alpha_0 \geq 0.5$, the sample size tends to infinity as z_1 approaches zero. When $\alpha_0 > 0.5$, the estimated effect size can be negative, and in that case, determining the sample size of the second phase using $\hat{\lambda}_{raw}$ or $\hat{\lambda}_{\alpha_{CL}}$ does not make sense. To avoid such awkward situations, it is appropriate to set an upper limit on the total sample size or on that for the second phase. In what follows we consider the former case. We refer to the upper limit of the sample size as allowable maximum sample size and denote it by M .

3.5 Sample-size re-estimation based on the conditional power

If the results of the first phase are such that the decision is to proceed with the trial, then the sample size is re-estimated at the end of the first phase. However, the sample size is not re-estimated following the interim analysis during the second phase. Thus, the sample size of the second phase is determined such that the conditional power is $1 - \beta_2$, given the first phase test statistic of z_1 . For the second phase, let us denote the conditional rejection probability, that is, the conditional power at a given z_1 , by $CRP(z_1)$, and the total sample size by $N(z_1)$. We will denote the sample sizes of the second and third stages by $n_2(z_1)$ and $n_3(z_1)$, respectively. Let r be a predefined proportion of the sample size of the second stage. We define $n_2(z_1) = \lceil rN(z_1) \rceil$ and $n_3(z_1) = N(z_1) - n_2(z_1)$, where $\lceil x \rceil$ stands for the smallest integer which is not less than x . Then, $N(z_1)$ is defined as the minimum sample size which satisfies $CRP(z_1) \geq 1 - \beta_2$. The $CRPs$ at the second and third stages are

$$CRP_2(z_1) = \Pr(R_2(z_1)) = \Phi(\sqrt{n_2(z_1)}/2\lambda - Z_{\alpha_2(Z_1, \alpha_0)}),$$

and

$$CRP_3(z_1) = \int_{Z_{\alpha_{02}(z_1)}}^{Z_{\alpha_2(z_1, \alpha_0)}} \phi(z_2 - \sqrt{n_2(z_1)}/2\lambda) \Phi(\sqrt{n_3(z_1)}/2\lambda - Z_{\alpha_3(Z_1, Z_2, \alpha_0)}) dz_2,$$

respectively. Thus, the conditional power at the second phase is

$$\begin{aligned} CRP(z_1) &= CRP_2(z_1) + CRP_3(z_1) \\ &= \Phi(\sqrt{n_2(z_1)}/2\lambda - Z_{\alpha_2(Z_1, \alpha_0)}) \\ &\quad + \int_{Z_{\alpha_{02}(z_1)}}^{Z_{\alpha_2(z_1, \alpha_0)}} \phi(z_2 - \sqrt{n_2(z_1)}/2\lambda) \Phi(\sqrt{n_3(z_1)}/2\lambda - Z_{\alpha_3(Z_1, Z_2, \alpha_0)}) dz_2. \end{aligned}$$

The conditional acceptance probability, denoted by $CAP_2(z_1)$, is

$$CAP_2(z_1) = \Pr(A_2(z_1)) = \Phi(Z_{\alpha_{02}(z_1)} - \sqrt{n_2(z_1)}/2\lambda).$$

When the calculated sample size $N(z_1)$ exceeds $M - n_1$, we set $N(z_1) = M - n_1$. Now, let $Z_1(N) = \{z_1; N(z_1) = N\}$, and denote the minimum and maximum sample sizes of the second phase by $N_{\min}$ and $N_{\max}$, respectively. The sample sizes of the second phase are any integer in the interval $[N_{\min}, N_{\max}]$.

3.6 Relationship between the sample sizes of the first and second phases

The minimum sample size $N_{\min}$ is given at $Z_1 = Z_{\alpha_1}$, and the maximum value $N_{\max}$ at $Z_1 = Z_{\alpha_0}$. $N(z_1)$ reaches the maximum in the interval $(Z_{\alpha_0}, Z_{\alpha_1})$. Let us define $Z_{1, N_{\max}}(N_{\max}) = \sup_{z_1} Z_1(N_{\max})$. The sample size decreases as a step function in the interval $(Z_{1, N_{\max}}(N_{\max}), Z_{\alpha_1})$. When M is not very large and α_1 is small, it may happen that $N_{\min} = N_{\max}$, because the first-phase sample size becomes large thus $M - n_1$ becomes small. In this case, the sample size of the second phase is constant, and $n_2(z_1) + n_3(z_1) = M - n_1$ for $Z_{\alpha_0} < z_1 < Z_{\alpha_1}$. In an extreme case, it may happen that $n_1 \geq M$. In this case, the design cannot be accepted. We will show these extreme cases in Section 5.

4. Evaluation of the designs

Characteristics of the designs can be described by various measures. For the first phase, the rejection probability and the probability of futility stopping are $1 - \Phi(Z_{\alpha_1} - \sqrt{n_1/2}\lambda)$ and $\Phi(Z_{\alpha_0} - \sqrt{n_1/2}\lambda)$, respectively. Thus, the probability of continuing to the second phase, denoted by $P(\text{stage } 2)$, is given by

$$P(\text{stage } 2) = \Phi(Z_{\alpha_1} - \sqrt{n_1/2}\lambda) - \Phi(Z_{\alpha_0} - \sqrt{n_1/2}\lambda).$$

The conditional error function for the second phase was described in Section 3. This function gives the conditional significance level for the second phase under the null hypothesis, and the conditional power under an alternative hypothesis. The expected rejection probabilities for the second and third stages and the probability of futility stopping can be calculated exactly by

$$\begin{aligned} RP_2 &= \sum_{n=N_{\min}}^{n=N_{\max}} \int_{Z_1(n)} \phi(z_1 - \sqrt{n_1/2}\lambda) \Phi(\sqrt{[n/2]}/2\lambda - Z_{\alpha_2(z_1, \alpha_0)}) dz_1. \\ RP_3 &= \sum_{n=N_{\min}}^{n=N_{\max}} \int_{Z_1(n)} \phi(z_1 - \sqrt{n_1/2}\lambda) \int_{Z_1(n)} \phi(z_2 - \sqrt{[n/2]}/2\lambda) \\ &\quad \Phi(\sqrt{(n - [n/2])}/2\lambda - Z_{\alpha_3(z_1, z_2, \alpha_0)}) dz_2 dz_1. \\ AP_2 &= \sum_{n=N_{\min}}^{n=N_{\max}} \int_{Z_1(n)} \phi(z_1 - \sqrt{n_1/2}\lambda) \Phi(Z_{\alpha_{02}(z_1)} - \sqrt{[n/2]}/2\lambda) dz_1. \end{aligned}$$

For the second phase, the overall rejection probability is $RP = RP_2 + RP_3$. The conditional average sample size of the second phase at z_1 is given by

$$CASN(z_1) = n_2(z_1) + n_3(z_1) \{ (1 - CRP_2(z_1) - CAP_2(z_1)) \},$$

and the overall average sample number (ASN) is given by

$$ASN = n_1 + \int_{Z_{\alpha_0}}^{Z_{\alpha_1}} CASN(z_1) \phi(z_1 - \sqrt{n_1/2}\lambda) dz_1.$$

In general, the sample size and the power are positively correlated. To balance these two measures, it may be useful to consider the ratio of the overall power to the overall ASN. We refer to

this as the power-to-ASN ratio.

5. Selection of the design parameters for each of the four types of designs

We will investigate the four types of designs in detail and determine the nearly optimal design parameters for each of them. We will use the following notation for the new designs: AD1, the hybrid design of the Bauer-Köhne and group sequential designs; and AD2, the alpha-split design. The two existing designs will be denoted as AD3 for the Bauer-Köhne-Wassmer design and AD4 for the Lehmacher-Wassmer design. The design parameters are specified as follows:

DP1: The assumed effect size for the first phase: 1.0.

DP2: The true effect sizes: 0.7, 0.4.

DP3: The overall significance levels: 10%, 5%, and 2.5%. The significance level of 10% is selected for the case of exploratory trial.

DP4: The first-stage power and the conditional second-phase power: we consider both as either 80% or 90%. This means that $\beta_1 = \beta_2 = \beta = 10\%$ or 20% .

DP5: The futility stopping criteria, p-value: 30%, 40%, 50%, 60%, and 70%.

DP6: The allowable maximum sample size is that for which the single-stage design has a power of $1 - \beta$ at the true effect size.

Important measures for evaluating an optimal design are the overall power, the ASN, and the power-to-ASN ratio. Since the ASN increases with increasing power, it is important to balance the power and the ASN. The maximum allowable sample size makes it possible to compare the performances of the adaptive designs and the single-stage design by using the differences between their powers. The futility stopping criterion affects the power and the ASN. However, in this section, we will only show results for when the criterion is 50%, and the type-II error rate of the first phase and the conditional type II error rate of the second phase are 10%. The maximum allowable sample sizes at the effect size 0.7 are 27, 35 and 43 and at the effect size 0.4, they are 83, 108 and 132, for the significance level 10, 5 and 2.5%, respectively. The relationship between the criterion and power is given in Section 7.3.

5.1 A hybrid of the Bauer-Köhne and group sequential designs (hybrid design: AD1)

The results are shown in Table 1. During the second phase, regardless of the values set for DP1 through DP6, as the boundary parameter Δ of the group sequential design increases, the power and the ASN decrease. However, the effect of Δ is small, especially when the difference between the assumed and true effect sizes is small. The CR2 criterion for the local significance levels is better than the CR1. In general, we conclude that it is best to use CR2 and $\Delta = 0.2$ or 0.3 .

Table 1. Powers, ASNs, and power-to-ASN ratios of the hybrid of the Bauer-Köhne and adaptive group sequential designs (AD1).

The true effect size: 0.7 and 0.4; significance level: 10%, 5%, and 2.5%; the powers of the first phase and the conditional powers are 90%; the Wang-Tsiatis boundary parameter is set to $\Delta = 0, 0.1, \dots, 0.5$. The allowable maximum overall sample size is that for which the single-stage design has an overall power of 90% at the true effect size.

λ	α	Criterion		Δ					
				0	0.1	0.2	0.3	0.4	0.5
0.7	0.1	CR1	Power	85.8	85.7	85.5	85.4	85.1	84.9
			ASN	17.9	17.9	17.8	17.8	17.7	17.7
			Power/ASN	4.79	4.80	4.80	4.80	4.80	4.79
		CR2	Power	87.6	87.5	87.4	87.2	87.1	86.9
			ASN	19.3	19.3	19.3	19.2	19.2	19.2
			Power/ASN	4.53	4.53	4.53	4.53	4.53	4.53
		CR1	Power	85.9	85.7	85.6	85.3	85.1	84.7
			ASN	24.3	24.2	24.1	24.1	24.0	24.0
			Power/ASN	3.54	3.54	3.55	3.55	3.54	3.54
	0.05	CR2	Power	87.7	87.6	87.4	87.3	87.0	86.8
			ASN	25.7	25.6	25.6	25.5	25.5	25.5
			Power/ASN	3.41	3.42	3.42	3.42	3.41	3.41
		CR1	Power	85.7	85.5	85.4	85.1	84.7	84.3
			ASN	30.2	30.1	30.0	29.9	29.8	29.7
			Power/ASN	2.84	2.84	2.85	2.85	2.84	2.84
		CR2	Power	87.7	87.6	87.4	87.2	86.9	86.6
			ASN	31.6	31.5	31.4	31.4	31.3	31.2
			Power/ASN	2.77	2.78	2.78	2.78	2.78	2.77
	0.025	CR1	Power	74.8	74.7	74.4	74.1	73.8	73.3
			ASN	41.8	41.2	40.7	40.2	39.8	39.4
			Power/ASN	1.79	1.81	1.83	1.85	1.86	1.86
		CR2	Power	77.7	77.5	77.3	77.1	76.8	76.5
			ASN	41.8	41.3	40.8	40.5	40.1	39.9
			Power/ASN	1.86	1.88	1.89	1.91	1.91	1.92
		CR1	Power	75.5	75.4	75.1	74.8	74.3	73.6
			ASN	61.6	60.6	59.7	58.9	58.2	57.6
			Power/ASN	1.23	1.24	1.26	1.27	1.28	1.28
		CR2	Power	78.1	78.0	77.8	77.5	77.1	76.6
			ASN	60.4	59.5	58.7	58.0	57.4	56.9
			Power/ASN	1.29	1.31	1.33	1.34	1.34	1.35
		CR1	Power	74.9	74.8	74.6	74.2	73.6	72.8
			ASN	81.4	80.1	78.7	77.5	76.5	75.6
			Power/ASN	0.92	0.93	0.95	0.96	0.96	0.96
		CR2	Power	77.5	77.4	77.2	76.8	76.3	75.7
			ASN	79.2	77.8	76.6	75.5	74.6	73.8
			Power/ASN	0.98	0.99	1.01	1.02	1.02	1.03

5.2 Alpha-split design (AD2)

Although in the second phase, the conditional error function is constant in the interval $(Z_{\alpha_0}, Z_{\alpha_1})$, the estimated effect size depends on Z_1 , and thus so does the sample size for the

second phase. In the interval $(Z_{\alpha_0}, Z_{1, N_{\max}}(N_{\max}))$, the sample size for the second phase is $N = M - n_1$, and so the power is constant. On the other hand, in the interval $(Z_{1, N_{\max}}(N_{\max}), Z_{\alpha_1})$, the estimated effect size increases and the sample size decreases; as a result, the conditional power monotonically decreases.

We investigated the effect of α_1 and the boundary parameter Δ on the power, the ASN, and the power-to-ASN ratio. In the second phase, we applied a two-stage group sequential design. When the boundary parameter increased from 0 to 0.5, both the power and the ASN gradually decreased. The power-to-ASN ratio decreased for $\lambda = 0.4$, but increased for $\lambda = 0.7$. Since the changes in power and power-to-ASN ratio when Δ is in the interval $[0, 0.5]$ are small, we only show the results for two typical cases: $\Delta = 0$ (the O'Brien-Fleming boundary) and $\Delta = 0.5$ (the Pocock boundary). Table 2 and Table 3 show the results for Δ set to 0 and 0.5, respectively. The effect of the first phase significance level α_1 on the power and the ASN is small when Δ is in

Table 2. Powers, ASNs, and power-to-ASN ratios for the alpha-split design (AD2) with the O'Brien-Fleming boundary for the second phase.

The true effect size: 0.7 and 0.4; significance level: 10%, 5%, and 2.5%; the powers of the first phase and the conditional powers are 90%. The allowable maximum overall sample size is that for which the single-stage design has an overall power of 90% at the true effect size.

λ	α	Significance level of the first phase							
0.7	0.1		0.02	0.03	0.04	0.05	0.06	0.07	0.08
		Power	80.4	83.5	85.0	85.6	85.5	85.0	83.7
		ASN	24.1	22.6	21.1	20.3	19.4	18.7	17.9
		Power/ASN	3.33	3.70	4.03	4.22	4.40	4.55	4.67
	0.05		0.01	0.015	0.02	0.025	0.03	0.035	0.04
		Power	79.2	82.1	83.0	83.6	83.4	82.9	81.5
		ASN	29.7	27.7	26.8	26.0	25.2	24.5	23.8
		Power/ASN	2.67	2.96	3.10	3.21	3.31	3.38	3.42
	0.025		0.007	0.008	0.009	0.01	0.011	0.012	0.013
		Power	79.9	80.3	80.8	81.1	81.4	81.5	81.5
		ASN	33.9	33.6	33.0	32.8	32.3	32.1	31.5
		Power/ASN	2.36	2.39	2.45	2.47	2.52	2.54	2.58
0.4	0.1		0.02	0.03	0.04	0.05	0.06	0.07	0.08
		Power	80.5	80.3	79.5	78.9	77.8	76.2	73.6
		ASN	49.0	47.1	45.4	44.7	44.1	43.7	43.4
		Power/ASN	1.64	1.71	1.75	1.77	1.77	1.75	1.70
	0.05		0.01	0.015	0.02	0.025	0.03	0.035	0.04
		Power	79.8	79.1	78.9	78.4	77.4	75.8	73.2
		ASN	68.7	66.2	65.6	65.2	65.1	65.2	65.7
		Power/ASN	1.16	1.20	1.20	1.20	1.19	1.16	1.12
	0.025		0.007	0.008	0.009	0.01	0.011	0.012	0.013
		Power	77.8	78.1	77.7	77.8	77.3	77.3	76.6
		ASN	87.4	87.5	86.9	87.1	86.5	86.8	86.3
		Power/ASN	0.89	0.89	0.90	0.89	0.89	0.89	0.89

Table 3. Powers, ASNs, and power-to-ASN ratios of the alpha-split design (AD2) with the Pocock boundary.

The true effect size: 0.7 and 0.4; significance level: 10%, 5%, and 2.5%; the powers of the first phase and the conditional powers are 90%. The allowable maximum overall sample size is that for which the single-stage design has an overall power of 90% at the true effect size.

λ	α	Significance level of the first phase								
0.7	0.1		0.02	0.03	0.04	0.05	0.06	0.07	0.08	
		Power	79.7	82.7	84.1	84.8	84.5	84.1	82.6	
		ASN	24.1	22.5	20.9	20.1	19.3	18.5	17.7	
		Power/ASN	3.31	3.68	4.02	4.21	4.39	4.54	4.66	
	0.05		0.01	0.015	0.02	0.025	0.03	0.035	0.04	
		Power	77.9	80.8	81.6	82.3	82.0	81.5	80.0	
		ASN	29.5	27.5	26.5	25.7	24.9	24.2	23.5	
		Power/ASN	2.64	2.94	3.08	3.20	3.30	3.37	3.40	
	0.025		0.007	0.008	0.009	0.01	0.011	0.012	0.013	
		Power	78.1	78.6	78.9	79.2	79.6	79.8	79.6	
		ASN	33.4	33.2	32.5	32.3	31.8	31.6	31.0	
		Power/ASN	2.34	2.37	2.43	2.45	2.51	2.52	2.57	
	0.4	0.1		0.02	0.03	0.04	0.05	0.06	0.07	0.08
			Power	79.7	79.4	78.5	77.8	76.5	74.7	71.7
			ASN	47.6	45.5	43.6	42.7	41.7	41.1	40.5
Power/ASN			1.68	1.75	1.80	1.82	1.83	1.82	1.77	
0.05			0.01	0.015	0.02	0.025	0.03	0.035	0.04	
		Power	78.5	77.7	77.5	76.7	75.6	73.7	70.8	
		ASN	65.4	62.6	61.8	61.1	60.7	60.5	60.9	
		Power/ASN	1.20	1.24	1.25	1.27	1.25	1.22	1.16	
0.025			0.007	0.008	0.009	0.010	0.011	0.012	0.013	
		Power	76.1	76.3	75.9	76.0	75.4	75.3	74.6	
		ASN	81.5	81.5	80.8	80.9	80.2	80.4	79.9	
		Power/ASN	0.93	0.94	0.94	0.94	0.94	0.94	0.93	

the interval $[0, 0.5]$. When the true effect size is 0.7, the largest power is attained when the first phase significance level is about 50% of the overall significance level. On the other hand, when the true effect size is 0.4, the largest power is attained when the first-phase significance level is about 20% to 25% of the overall significance level. However, when α_1 is about 50% of the overall significance level the loss of power and the increase in the ASN are small. Therefore, the first phase significance level can be set to about 50% of the overall significance level.

5.3 Bauer-Köhne-Wassmer three-stage design (AD3)

The results are given in Table 4. For the three criteria for the local significance level (CR1, CR2, and CR3), regardless of the above design parameters, CR2 and CR3 give slightly higher powers and also have a larger ASN than does CR1. The maximum difference in the power is about three percent, and the difference between the ASNs is one to two. CR2 and CR3 result in nearly the same powers and ASNs. Looking at the power-to-ASN ratio, when the true effect

Table 4. Powers, ASNs, and power-to-ASN ratios for the Bauer-Köhne-Wassmer design (AD3).

The true effect size: 0.7 and 0.4; significance level: 10%, 5%, and 2.5%; the powers of the first phase and the conditional powers are 90%; CR1: constant significance levels; CR2: full level α in the final stage; CR3: no treatment-stage interaction. The allowable maximum overall sample size is that for which the single-stage design has an overall power of 90% at the true effect size.

λ	α		CR1	CR2	CR3
0.7	0.1	Power	85.8	87.0	87.0
		ASN	18.5	19.3	19.3
		Power/ASN	4.65	4.51	4.51
	0.05	Power	85.0	86.9	86.9
		ASN	24.1	25.6	25.7
		Power/ASN	3.53	3.39	3.39
	0.025	Power	84.5	86.7	86.7
		ASN	30.3	31.5	31.5
		Power/ASN	2.79	2.75	2.75
0.4	0.1	Power	75.1	77.6	77.6
		ASN	39.7	41.1	41.1
		Power/ASN	1.89	1.89	1.89
	0.05	Power	74.6	78.0	77.9
		ASN	56.8	58.9	58.8
		Power/ASN	1.31	1.32	1.32
	0.025	Power	74.4	77.2	77.2
		ASN	75.1	76.6	76.3
		Power/ASN	0.99	1.01	1.01

size is 0.7, CR1 is better than the other criteria, but when it is 0.4, CR1 results in a ratio that is almost the same or is smaller than those of the other two. In summary, CR2 and CR3 seem to be appropriate.

5.4 Lehman-Wassmer three-stage adaptive group sequential design (AD4)

A summary of the results is given in Table 5. When using a group sequential design, a Pocock or O'Brien-Fleming boundary is often used. However, Wang and Tsiatis (1987) found that for two-sided hypotheses, the optimal parameter from the power family of boundaries was located near the Pocock boundary ($0.3 < \Delta \cong 0.5$), and as the true effect size decreased, the nearly optimal Δ also decreased. Pampallona and Tsiatis (1994) also found that in the power family of boundaries for a one-sided hypothesis, the optimal Δ was located between 0.3 and 0.5. We searched for a nearly optimal boundary in the power family.

It is sometimes impossible to apply the O'Brien-Fleming design, because, when Δ is near 0, the first-phase sample size may exceed the maximum allowable size. For example, let $\alpha = 2.5\%$, $1 - \beta_1 = 90\%$, $\alpha_0 = 50\%$, and let the assumed and true effect sizes be 1.0 and 0.7, respectively. The maximum allowable sample size M is 43 per group. In the O'Brien-Fleming design, the critical point of the first stage is 3.437. Thus the first-stage sample size is 45 and exceeds M .

Table 5. Powers, ASNs, and power-to-ASN ratios of the Lehman-Wassmer adaptive group sequential designs (AD4).

The true effect size: 0.7 and 0.4; significance level: 10%, 5%, and 2.5%; the powers of the first phase and the conditional powers are 90%; the Wang-Tsiatis boundary parameter is set to $\Delta = 0, 0.1, \dots, 0.5$. The allowable maximum overall sample size is that for which the single-stage design has an overall power of 90% at the true effect size.

λ	α		Δ					
			0	0.1	0.2	0.3	0.4	0.5
0.7	0.1	Power	81.5	86.1	87.7	88.3	88.2	87.5
		ASN	26.4	24.9	23.4	21.8	20.3	19.4
		Power/ASN	3.09	3.45	3.75	4.04	4.34	4.50
	0.05	Power	—	83.4	86.5	87.5	87.5	87.1
		ASN	—	32.4	30.3	28.1	26.5	25.6
		Power/ASN	—	2.57	2.85	3.12	3.30	3.40
	0.025	Power	—	80.2	84.9	86.9	87.4	86.8
		ASN	—	40.6	37.9	35.1	32.9	31.3
		Power/ASN	—	1.97	2.24	2.48	2.66	2.77
0.4	0.1	Power	82.0	81.6	80.9	79.8	78.3	77.1
		ASN	47.5	45.9	44.3	42.8	41.3	40.7
		Power/ASN	1.73	1.78	1.83	1.87	1.89	1.90
	0.05	Power	82.3	81.2	80.5	79.3	78.1	77.1
		ASN	67.7	64.2	61.9	59.8	58.5	58.1
		Power/ASN	1.22	1.27	1.30	1.33	1.34	1.33
	0.025	Power	81.0	80.4	79.2	77.7	76.0	73.1
		ASN	83.5	80.5	77.8	76.0	75.4	75.3
		Power/ASN	0.97	1.00	1.02	1.02	1.01	0.97

On the other hand, for the Pocock design, the critical point of the first phase is 2.2826, and the sample size of the first stage is 25.

Let us consider the power and the ASN as a function of Δ . As Δ decreases, the critical point increases, thus the power at the first phase decreases and the probability of proceeding to the second phase increases; as a result, the ASN increases. For a true effect size of 0.7, the power curve has a single peak, but the ASN monotonically decreases. The power-to-ASN ratio increases until $\Delta = 0.5$. For a true effect size of 0.4, as Δ increases, both the power and the ASN monotonically decrease. The power-to-ASN ratio attains its maximum at around $\Delta = 0.5$ when $\alpha = 10\%$; the best Δ decreases to $\Delta = 0.4$ and $\Delta = 0.3$, respectively, when $\alpha = 5\%$ and $\alpha = 2.5\%$. To balance the power and the ASN regardless of the significance level, a Δ of around 0.3 gives an acceptable power and ASN.

The O'Brien-Fleming and Pocock boundaries result in very different powers and ASNs, depending on the conditions. When there is only a small difference between the assumed and true effect sizes, the Pocock design has much higher power and a smaller ASN. On the other hand, when the difference is large, the O'Brien-Fleming design has higher power and a larger ASN than that of the Pocock design. A boundary defined by $\Delta = 0.3$ seems to balance the power

and the ASN.

6. Comparison of four types of designs

For comparing the four types of designs, we select a few nearly optimal designs from each of the four types of designs. We show the results for $\alpha_0 = 50\%$. The selected designs are presented in Table 6.

Table 6. Designs from each type of adaptive design selected for comparison. For AD3, CR2 and CR3 give almost the same results.

Design type	Conditional error function	Second phase group sequential design
AD1 (hybrid design)	CR2	O'Brien-Fleming boundary
AD2 (alpha-split design)	$\alpha_1 = \alpha/2$	O'Brien-Fleming boundary
AD3 (Bauer-Köhne-Wassmer design)	CR2/CR3	
AD4 (adaptive group sequential design)	1) A nearly optimal Δ 2) Pocock boundary 3) O'Brien-Fleming boundary	

Table 6 shows the designs selected from AD1, AD2, AD3, and AD4 which show the highest powers in the respective design types, and also lists the Pocock and O'Brien-Fleming designs from AD4. The results are compared in Table 7. We only show the results for $\alpha = 10\%$, 5% , and 2.5% , $\alpha_0 = 50\%$, $\beta = 10\%$, the conditional power equal to 90% , and naïve estimates for the effect size, because the other conditions have very similar results. When the true effect size is 0.7 and the overall significance level is 10% , AD4 with $\Delta = 0.3$ results in the largest power and also the largest ASN. AD1, AD3 and AD4 with the Pocock boundary result in a slightly smaller power than the optimal design in AD4, and the ASN is smaller than that by about 2 subjects. The AD2 results in a somewhat lower power and a larger ASN than those of the above three designs. The AD4 with the O'Brien-Fleming boundary results in much lower power and a larger ASN than those of the other designs. For alpha of 5% and 2.5% , the hybrid design results in the highest power and has almost the same ASN as do other designs. AD2 results in much less power than do the others, but it has nearly the same ASN as them.

When the true effect size is 0.4 , AD4 with the O'Brien-Fleming boundary results in the maximum power, but it has a much larger ASN than that of the others. The AD2 results in the second highest power and has a much larger ASN than that of the others. The AD3 and AD4 designs with the Pocock boundary and the AD1 with the O'Brien-Fleming boundary result in nearly the same power and ASN.

When the true effect size is 0.7 , the maximum allowable sample size is twice as large as the sample size for a single-stage design at the assumed effect size. But for a true effect size of 0.4 , the maximum allowable sample size is about 6 times larger than the sample size of a single-stage design at the assumed effect size. This difference in the ratio of the maximum allowable sample size to the sample size for the assumed effect size seems to affect the relative performance of the

Jpn J Biomet Vol.35, No.2, 2014

Table 7. Comparisons between the four types of designs.

The values listed under the significance levels are the power, and the ASN is given in the parentheses. For AD4, the first line is with the O'Brien-Fleming boundary, the second is with the nearly optimal Δ , and the third is with the Pocock boundary. The fourth, sixth, and eighth columns contain the design parameters that determined the conditional error functions for the second phase. The numerical values are the α_1 s for AD2, and Δ for AD4. Hyphens (-) indicate that it is impossible to conduct the design since $n_1 \geq M$. The allowable maximum overall sample size is that for which the single-stage design has an overall power of 90% at the true effect size.

λ_0/λ	Design	Significance level					
		10%		5%		2.5%	
1.0/0.7	AD1	87.6 (19.3)	CR2/OF	87.7 (25.7)	CR2/OF	87.7 (31.6)	CR2/OF
	AD2	85.6 (20.3)	5%/OF	83.6 (26.8)	2.5%/OF	81.5 (31.2)	1.2%/OF
	AD3	87.6 (19.3)	CR2	86.9 (25.6)	CR2/CR3	86.7 (31.5)	CR2/CR3
	AD4	81.5 (26.4)	OF	-	-	-	-
		88.3 (21.8)	$\Delta = 0.3$	87.5 (26.5)	$\Delta = 0.4$	87.4 (32.9)	$\Delta = 0.4$
		87.5 (19.4)	Po	87.1 (25.6)	Po	86.8 (31.3)	Po
1.0/0.4	AD1	77.7 (41.8)	CR2/OF	78.1 (60.4)	CR2/OF	77.5 (79.2)	CR2/OF
	AD2	80.5 (49.0)	2%/OF	79.8 (68.7)	1%/OF	78.1 (87.5)	0.8%/OF
	AD3	77.6 (41.1)	CR2	78.0 (58.9)	CR2/CR3	77.2 (76.3)	CR3
	AD4	82.0 (47.5)	OF	82.3 (67.7)	OF	81.0 (83.5)	OF
		82.0 (47.5)	$\Delta = 0(\text{OF})$	82.3 (67.7)	$\Delta = 0(\text{OF})$	81.0 (83.5)	$\Delta = 0(\text{OF})$
		77.1 (40.7)	Po	77.1 (58.1)	Po	76.0 (75.4)	Po
0.517/0.4	AD1	87.6 (58.6)	CR2/OF	87.7 (78.0)	CR2/OF	87.6 (96.9)	CR2/OF
	AD2	85.5 (58.4)	6%/OF	83.7 (78.7)	2.5%/OF	81.6 (96.5)	1.3%/OF
	AD3	87.0 (58.5)	CR2/CR3	86.9 (77.8)	CR2/CR3	86.7 (96.3)	CR2
	AD4	82.7 (80.3)	OF	-	OF	-	OF
		88.3 (66.0)	$\Delta = 0.3$	87.8 (81.8)	$\Delta = 0.4$	87.4 (100.3)	$\Delta = 0.4$
		87.6 (59.6)	Po	87.2 (77.8)	Po	86.7 (95.1)	Po

When $\lambda_0/\lambda = 1.0/0.4$, the nearly optimal boundary parameter is zero (O'Brien-Fleming boundary). Therefore, the fourth and fifth rows coincide.

designs. To clarify this, we investigated a case where the assumed effect size was 0.517 and the true effect size was 0.4. The results are given in Table 7. The relative performance is similar to that when the true effect size is 0.7. Therefore, when the ratio of the true effect size to the assumed effect size is 0.7 or more, the AD1 with CR2 seems appropriate. However, the AD3 and AD4 with the Pocock boundary performed almost as well.

7. Comparison with other designs

In the previous section, we examined in detail the performance of two-phase, three-stage designs. Now we briefly discuss some comparisons between our designs and (1) single stage design, (2) ordinary three-stage group sequential design for a one-sided hypothesis.

7.1 Comparison with the single stage design

First we compare the two-phase, three-stage design with the single-stage design. Although the single stage design is difficult to use in our situation, this comparison tells us how our adaptive design suffers the loss of power as compared to the case that a good effect size of the test therapy is known. The sample size of a single stage design is determined so that the trial achieves a power of 90% at the true effect size, thus it is the maximum allowable sample size of the two-phase, three-stage design. Thus we can evaluate how much an adaptive design loses the power. The powers of the two-phase, three-stage designs are given in Table 1, 2, 3, 4, 5 and 7, and are compared to the power of 90%. We see that the relative loss of the power of the two-phase, three-stage design as compared to the single stage design is at most 15% and in some cases about only 5% with the ASN less than 50% of the sample size needed for a single-stage trial.

7.2 Comparison with the three-stage group sequential design

In order to compare the two-phase, three-stage design to the ordinary group sequential design, we show in Table 8 the power and ASN of the three-stage group sequential design. The sample size is that of the single stage design with a power of 90% at the true effect size, and it is the maximum allowable sample size of the two-phase, three-stage design. The true effect size is 0.4 and 0.7. These results are compared with those of the two-phase, three-stage designs given in Table 7. For a true effect size of 0.4, a type-II error rate of 10%, and a futility stopping criterion of $\alpha_0 = 0.5$, the group sequential design gives higher power and larger ASN than the two-phase, three-stage design. For a true effect size of 0.7, AD1 and O'Brien-Fleming design give similar powers and ASNs, and the Pocock design is inferior in the power to AD1 design and the O'Brien-Fleming design. In general the higher power is associated with the larger ASN.

Although the above scenario for comparisons are very restricted, the authors conclude that a higher power is associated with a higher ASN, and the two-phase, three-stage design gives comparable performance with the group sequential design.

Table 8. Powers and ASNs (in the parentheses) of three stage group sequential designs. The sample size is that of the single stage design with a power of 90% at the true effect size. It is the maximum allowable sample size of the two-phase, three-stage design. The numbers in the parentheses of the rows of the sample size stand for stage-wise sample sizes. The critical boundaries are O'Brien-Fleming (OF) and Pocock (Po) boundaries. The futility stopping criterion for the group sequential design is the DeMets-Ware asymmetric boundary with $\alpha_0 = 0.5$.

True effect size	Significance level	10%	5%	2.5%
0.7	Sample size	27(9,9,9)	35(12,12,11)	43(15,14,14)
	OF	86.3 (17.7)	87.8 (25.5)	87.8 (32.2)
	Po	83.4 (15.3)	84.2 (21.5)	83.7 (27.2)
0.4	Sample size	83(28,28,27)	108(36,36,36)	132(44,44,44)
	OF	86.7 (54.9)	87.8 (77.7)	88.6 (100.8)
	Po	83.9 (47.4)	84.3 (65.8)	84.6 (84.8)

8. Discussion and summary

A two-phase, three-stage design can be adopted for the first, definitive trial. The first, definitive trial implies a confirmatory trial. However, for establishing firm evidence of the efficacy or effectiveness of a therapy, it is important that the trial results are reproducible. Thus, the question remains as to whether affirmative results of the first, definitive trial can provide sufficient evidence for choosing a therapy. The validity of the conclusion derived from the trial and applicability of the results to daily practice must be discussed. Moreover, the appropriateness of the actual current state, in which the clinical investigator initiates a confirmatory clinical trial without any preceding exploratory clinical trials or information on the CMD, must be discussed. These important issues are related to the regulation and policy of funding agents and so are not discussed herein.

In Section 5, we examined how the power and ASN change by the design parameters, except for the futility stopping criterion and the type of the estimate of the effect size, that is, naïve or conservative estimate. We examined how the futility stopping criterion affect the overall power, the ASN, and power-to-ASN ratio for $\alpha_{01} = \alpha_{02} = \alpha_0 = 30\%$ to 70% , under the conditions described in Section 6. Here we briefly summarize the results. In all combinations of the conditions DP1 to DP4, as α_0 increased from 30% to 70% , the power-to-ASN ratio monotonically decreased. Although the degree of influence of α_0 to the power and ASN depends on the assumed and true effect sizes and the maximum allowable sample size, the power and ASN change in parallel and the amount of the change is small. Therefore we conclude that $\alpha_{01} = \alpha_{02} = \alpha_0 = 50\%$ is a good choice. We also investigated how a conservative estimate increases the power as compared to the naïve estimate. The increase in power is a few percent, and thus the use of a naïve estimate may be acceptable.

The relative loss of power of the two-phase, three-stage design as compared to the single stage design was at most 15% with ASN less than 50% under a condition that the single stage design is most powerful. The two-phase, three-stage design gives the power nearly the same or slightly lower than the three-stage group sequential design for the same maximum sample size. However, the group sequential design requires a reasonably estimated effect size or the (S)CMD. If the assumed effect size is larger than the true effect size, group sequential design suffers from a large loss of power. Furthermore, the group sequential design requires an upfront commitment of the maximum sample size by a sponsor.

One may be interested in comparison between the two-phase, three-stage design and the two-phase, K-stage design. The two-phase, K-stage design is defined here as a design that the sample size re-estimation is carried out based on the result of the first phase, and the second phase is designed as a K-1 stage group sequential design. In the group sequential design as the number of stages increases, the maximum sample size increases and the ASN decreases. Therefore the range of the test statistic of the first phase where the re-estimated sample size

exceeds the allowable maximum sample size gets wider. Thus the two-phase, two-stage design seems more powerful than the two-phase, three-stage design. However, this two-stage design does not reduce the conditional ASN in the second phase, because no interim analysis is performed in the second phase. The two-phase, four-stage design may give a slightly smaller ASN and slightly lower power than the three-stage design. Based on the considerations given in the introduction section these designs do not seem to be superior to the two-phase, three-stage design. However, since these discussions are not based on the actual numerical comparisons, further investigation to compare these designs would be desirable.

The four types of the two-phase, three-stage designs show similar powers and ASNs except for the adaptive group sequential design with the O'Brien-Fleming boundary and the alpha-split design. The maximum allowable sample size depends on budget, study period, rate of patient recruitment and other conditions. These factors may vary as the trial proceeds. The adaptive design can modify the sample size based on the interim result and other factors. Thus two-phase, three-stage design can be a good candidate when neither CMD nor reliable information on the effect size of the test treatment is available.

Acknowledgements

The authors thank the referee and the editor for clarifying our motivation of the two-phase, three-stage design. This research was supported by Grants-in-Aids for Scientific Research (C) from the Ministry of Education, Culture, Sports, Science and Technology of Japan (No. 23500349).

REFERENCES

- Bauer, P. and Koenig, F. (2006). The reassessment of trial perspectives from interim data—a critical review. *Statistics in Medicine*, **25**: 23–36.
- Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, **50**: 1029–1041.
- Copay, A. G., Subach, B. R., Glassman, S. D., Polly, D. W. Jr. and Schuler, T. C. (2007). Understanding the minimum clinically important difference: a review of concepts and methods. *Spine Journal*, **7**: 541–546.
- Cui, L., Hung, H. M. J. and Wang, S. J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics*, **55**: 853–857.
- DeMets, D. L. and Ware, J. H. (1980). Group sequential methods for clinical trials with a one-sided hypotheses. *Biometrika*, **67**: 651–660.
- DeMets, D. L. and Ware, J. H. (1982). Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika*, **69**: 661–663.

- Dyck, P. J. and O'Brien, P. C. (1989). Meaningful degree of prevention of improvement of nerve conduction in controlled clinical trials of diabetic neuropathy. *Diabetes Care*, **12**: 649–652.
- Emerson, S. S. (2006). Issues in the use of adaptive clinical trial designs. *Statistics in Medicine*, **25**: 3279–3296.
- Emerson, S. S. and Fleming, T. R. (1989). Symmetric group sequential trial designs. *Biometrics*, **45**: 905–923.
- Emerson, S. S., Levine, G. P. and Emerson, S. C. (2011). Comments on 'Adaptive increase in sample size when interim results are promising: A practical guide with examples' by Mehta and Pocock. *Statistics in Medicine*, **30**: 3285–3301.
- Fleming, T. R. (2006). Standard versus adaptive monitoring procedures: A commentary. *Statistics in Medicine*, **25**: 3305–3312.
- Jahn-Eimermecher, A. and Hommel, G. (2007). Performance of adaptive sample size adjustment with respect to stopping criteria and time of interim analysis. *Statistics in Medicine*, **26**: 1450–1461.
- Jennison, C. and Turnbull, B. W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine*, **22**: 971–993.
- Jennison, C. and Turnbull, B. W. (2006). Efficient group sequential designs when there are several effect sizes under consideration. *Statistics in Medicine*, **25**: 917–923.
- Lan, K. K. G., Soo, Y., Siu, C. and Wang, M. (2005). The use of weighted Z-tests in medical research. *Journal of Biopharmaceutical Statistics*, **15**: 625–639.
- Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, **55**: 1286–1290.
- Levin, G. P., Emerson, S. C., and Emerson, S. S. (2013). Adaptive clinical trial designs with pre-specified rules for modifying the sample size: understanding efficient type of adaptation. *Statistics in Medicine*, **32**: 1259–1275.
- Mehta C. R. and Pocock, S. J. (2011). Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statistics in Medicine*, **30**: 3267–3284.
- Minamino, T., Toba, K., Higo, S., Nakatani, D., Hikoso, S., Umegaki, M., Yamamoto, K., Sawa, Y., Aizawa, Y., Komuro, I., and EPO-AMI-II study investigators. (2012). Design and rationale of low-dose erythropoietin in patients with ST-segment elevation myocardial infarction (EPO-AMI-II Study): A randomized controlled clinical trial. *Cardiovascular Drug Therapy*, **26**: 409–416.
- Müller, H. H. and Schäfer, H. (2004). A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine*, **23**: 2497–2508.

- Ogawa, H., Nakayama, M., Morimoto, T., Uemura, S., Kanauchi, M., Doi, N., Jinnouchi, H., Sugiyama, S., Saito, Y. and Japanese Primary Prevention of Atherosclerosis with Aspirin for Diabetes (JPAD) Trial Investigators. (2008). Low-dose aspirin for primary prevention of atherosclerotic events in patients with type 2 diabetes: A randomized controlled trial. *JAMA*, **300**: 2134–2141.
- Posch, M. and Bauer, P. (1999). Adaptive two stage designs and the conditional error function. *Biometrical Journal*, **41**: 689–696.
- Posch, M., Bauer, P. and Brannath, W. (2003). Issues in designing flexible trials. *Statistics in Medicine*, **22**: 953–969.
- Pampallona, S. and Tsiatis, A. A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference*, **42**: 19–35.
- Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics*, **51**: 1315–1324.
- Shih, W. J. (2001). Sample size re-estimation-journey for a decade. *Statistics in Medicine*, **20**: 515–518.
- Shih, W. J. (2006). Commentary on: Group sequential, sample size re-estimation and two-stage adaptive designs in clinical trials: A comparison. *Statistics in Medicine*, **25**: 933–941.
- Tsiatis, A. A. and Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, **90**: 367–378.
- Uesaka, H. (2003). A personal view of method of sample size estimation for a clinical trial. *Japanese Journal of Biometrics*, **24**: 17–41.
- Wang, S. K. and Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, **43**: 193–199.
- Wassmer, G. A. (1999). Multistage adaptive test procedures based on Fisher's product criterion. *Biometrical Journal*, **41**: 279–293.