Open Journal of Pharmaceutical Science and Research

Review Article Open Access

A Critical Review on Adaptive Sample Size Re-estimation (SSR) Designs for Superiority Trials with Continuous Endpoints

Xiaoyu Cai, Yi Tsong* and Meiyu Shen

Division of Biometrics VI, Office of Biostatistics, Center for Drug Evaluation and Research, FDA, USA

***Correspondig Author:** Yi Tsong, Ph.D, Division of Biometrics VI, Office of Biostatistics, Center for Drug Evaluation and Research, FDA, USA, 10903 New Hampshire Ave, Silver Spring, MD 20903, Tel: 301-796-1013; Email: yi.tsong@fda.hhs.gov

Received Date: *Mar 08, 2019* / ***Accepted Date:*** *Mar 29, 2019*/ ***Published Date:*** *April 01, 2019*

Abstract

Adaptive sample size re-estimation (SSR) methods have been widely used for designing clinical trials, especially during the past two decades. We give a critical review for several commonly used two-stage adaptive SSR designs for superiority trials with continuous endpoints. The objective, design and some of our suggestions and concerns of each design will be discussed in this paper.

Keywords: Adaptive Design; Sample Size Re-estimation; Review

Cite this article as: Xiaoyu Cai, Yi Tsong, Meiyu Shen. 2019. A Critical Review on Adaptive Sample Size Re-estimation (SSR) Designs for Superiority Trials with Continuous Endpoints. Open J Pharm Sci Res. 1: 1-13.

Introduction

Sample size determination is a key part of designing clinical trials. The objective of a good clinical trial design is to achieve the balance between efficiently spending resources and enrolling enough patients to achieve a desired power. At the designing stage of a clinical trial, there usually only have limited information available about the population, so that the sample size calculated at this stage may not be sufficient to address the study objective. Assumed that the data from two parallel treatment groups (e.g. treatment and control groups) are normally distributed with mean treatment effect μ_1 and μ_2, and equal within-group variance σ^2. Let the mean difference (treatment effect) $\delta = \mu_1 - \mu_2$. The efficacy of the treatment will be evaluated by testing the hypothesis.

$$H_0: \delta=0 \text{ against } H_a: \delta>0.$$

Traditional fixed-size designs or group sequential designs calculate the sample size of the trial before it starts based on an assumed treatment effect δ_a and within-group variance $\sigma_a{}^2$, or estimate them from historical data. However, at the planning stage of the study, we may have little information about the parameters or the information we have might be inaccurate, which can lead to grossly overestimation or underestimation of the sample size. The underestimation of the sample size is especially unfavorable, because it will make the trial under-powered and fail to find a significant treatment benefit. It will be helpful if we can re-estimate the sample size after some of the data are observed from the study, so we can re-estimate the unknown parameter accordingly. Because the re-estimated parameter is from the current data, it will represent our current population much better than the parameter estimated from previous information.

There is no doubt that increasing sample size will increase test power, but there are also problems we need to pay attention to. First, how can we control type I error rate? When the re-estimated sample size is depending on observed data, it may bias the final test. Then, how can we control the power at a desired level when the design changes? Moreover, we want the re-estimated sample size to be efficient, because there is no need for the power to be as high as possible, otherwise we can just use the maximum affordable sample size at the beginning of the trial and it may detect a non-clinical meaningful difference.

The purpose of this paper is not to promote or discourage people to use certain SSR design, but to offer some guidance for people who want to use SSR designs (especially for the first time) about the basic ideas, advantages and drawbacks of each design. Many literatures with similar purpose have been published to summarize and to review the existing SSR designs with different focuses. Some old review papers with technical details such as [1,2] were written at least ten years ago. They cannot involve many designs proposed in the literatures published recently, and they focus more on summarizing the authors' own works. The review paper given by [3] five years ago focuses on comparing some commonly used unblinded two-stage SSR designs in terms of their operating characteristics. More recently, the paper published by [4] gives thorough review about the development of SSR designs since 1945. However, they only use a few words to summarize the basic idea of each design, which provide little technical detail or comments about how they perform. In this paper, we focus on reviewing and giving comments on the literatures of two-stage adaptive SSR designs for both blinded and unblinded superiority trials with continuous endpoints, especially those published during the past two decades. Early stop for efficacy or futility will not be discussed here, and here we only consider increasing sample size. To re-estimate sample size based on the information observed from the first stage, a variety of different techniques proposed in past literatures will be summarized in this paper, such as re-estimating within-group variance or treatment effect; adjusting final test statistic, critical value or significance level; giving constraints on adaptive region; and so on. The objective, design details and some of the key suggestions and concerns of each design will be given in this paper. The common adaptive SSR designs can be summarized in the following procedures; it can also be simplified by the flowchart in Figure 1.

a) At the beginning of the trial, calculate the original planned per-group sample size N_0 based on assumed parameters such as the targeted treatment effect δ_a that the experiment wants to detect, the within-group variance $\sigma_a{}^2$ or both.

b) After $n_1 = t * N_0$ $(0 < t < 1)$ patients per-group (for simplicity, here we only consider equal sample size for each group) have been enrolled and have had responses, calculate the test score and exam whether it satisfies certain re-estimation criteria, the criteria can be decided by either practical considerations or theoretical reasons.

c) If the re-estimation criteria are not satisfied, the trial will be finished as its original design. Continue to enroll N_0–n_1 patients per-group, claim superiority if the final test based on the cumulated data of two stages rejects H_0 and stop.

d) If the re-estimation criteria are satisfied, re-estimate the sample size N^*based on the information from the first stage data (such as the re-estimated treatment effect, variance, etc.). Some of the designs may need to adjust test statistic, significance level or critical value to protect the operational characteristics of the final test. Continue to enroll $N^* - n_1$ patients per-group, claim superiority if the final test based on the cumulated data of two stages rejects H_0 and stop.

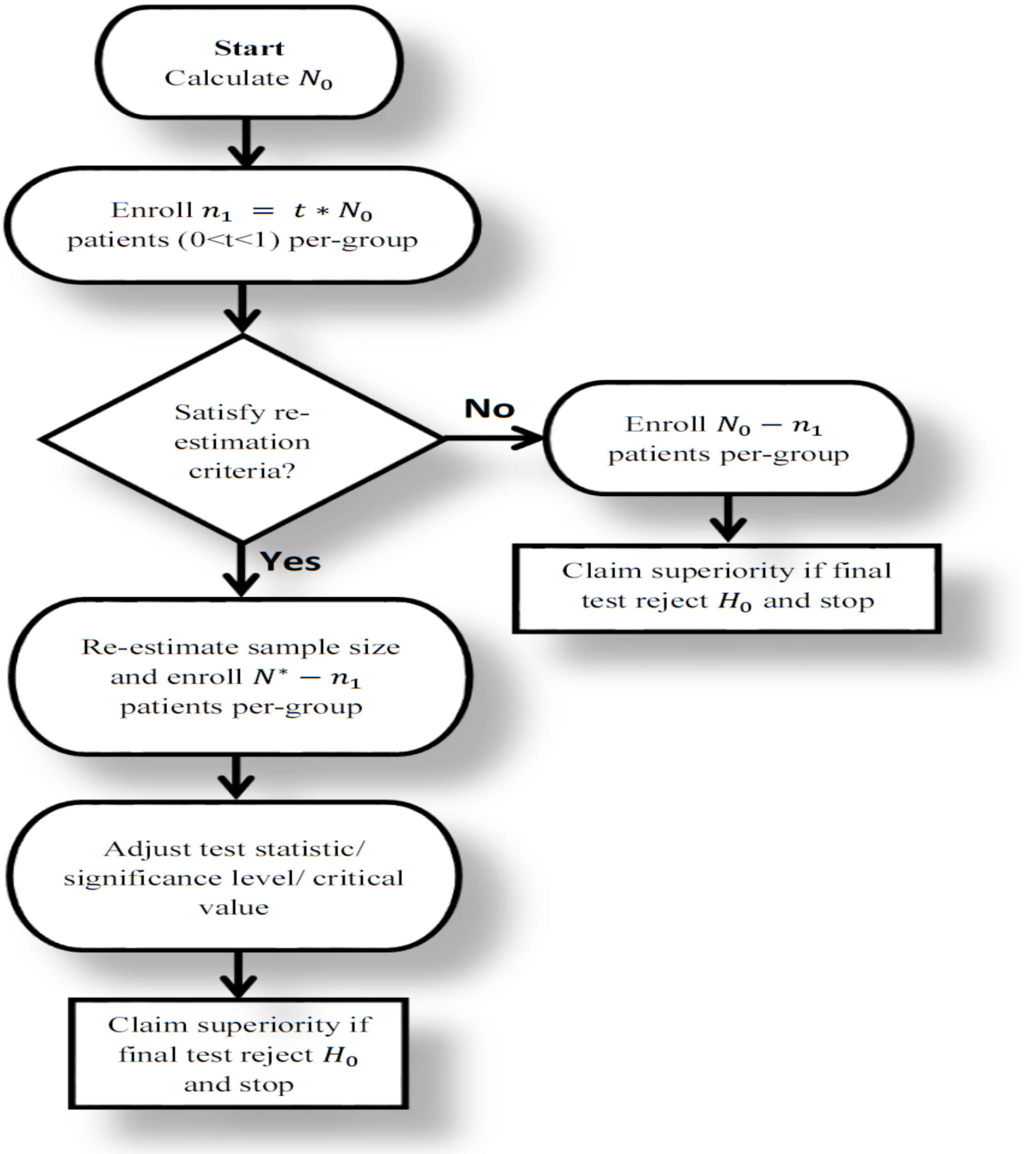
Start
Calculate N_0
Enroll $n_1 = t * N_0$ patients (0<t<1) per-group
Satisfy re-estimation criteria?
No
Enroll $N_0 - n_1$ patients per-group
Claim superiority if final test reject H_0 and stop
Yes
Re-estimate sample size and enroll $N^* - n_1$ patients per-group
Adjust test statistic/ significance level/ critical value
Claim superiority if final test reject H_0 and stop

Figure 1: General Procedure for Adaptive SSR Designs.

Part One: Blinded SSR (BSSR) Designs

In this section, we discuss some well-known blinded SSR designs (BSSR) which do not break treatment code before the trial is finished. The latest FDA draft guidance *Adaptive Designs for Clinical Trials of Drugs and Biologics* [5], refers to these designs as based on non-comparative data. BSSR designs have the advantage of protecting the confidentiality of the treatment effect at the interim study and more acceptable regulatory. The original planned per-group sample size of the designs in this section can be all given by

$$N_0 = \frac{2(z_\alpha + z_\beta)^2}{\delta_a{}^2}\sigma_a{}^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

where z_α and z_β are the (1−α)th and (1−β)th quantiles of the standard normal distribution. Because the treatment result cannot be revealed before the trial complete, only within-group variance $\sigma_\alpha{}^2$ is re-estimated after observing the first stage data to protect trial power against underestimation of the sample size. Two sample t-test will be used at the end of the trial with cumulated data from two stages.

The BSSR with EM-Algorithm Method

Objective and Design

The design proposed by [6,7] conducts interim analysis after part of the original planned data (e.g. N_0 samples for the combined two groups data) are observed. They re-estimate the per-group sample size N^* with the same formula as equation 1, but replace the assumed within-group variance $\sigma_\alpha{}^2$ by the re-estimated value $\hat{\sigma}^2$, where $\hat{\sigma}^2$ is calculated by EM-algorithm method based on the observed first stage data. The sample size increases only if the re-estimated sample size N^* is sufficiently larger than the original planned sample size N_0, say $N^*/N_0 > \lambda > 1$, where λ is a pre-decided value. If the sample size modification is made, the efficacy of the treatment will be claimed with significance level α if the final test with cumulated data from two stages $T^{(N^*)} > t_{2N^*-2,\alpha}$. Where $T^{(N^*)}$ is the two-sample t-test statistic with N^* sample per-group; $t_{2N*-2,\alpha}$ is the (1- α)th quantile of t-distribution with $2N^* - 2$ degrees of freedom. To protect the treatment result, the authors claim, although the EM-algorithm method gives accurate estimation of within-group variance σ^2, it does not estimate standard treatment effect $(\mu_1 - \mu_2)/\sigma$ very well [6]. For instance, we can't get clear evidence from the estimation results of $(\mu_1 - \mu_2)/\sigma$ about how likely the null hypothesis will be rejected, thus protecting the blindness of the trial.

Concerns and Weaknesses of the Design

It was pointed out by [8] that with increased sample sizes, bias and variability of the EM-algorithm estimations of σ^2 and $(\mu_1 - \mu_2)/\sigma$ both decrease. That means although it was claimed by the authors that the estimation of standard treatment effect $(\mu_1 - \mu_2)/\sigma$ is not accurate, it still reveals some information about the test result, especially when the sample size or the mean difference is pretty large. Furthermore, the accuracy of the σ^2 estimation by EM-algorithm greatly depends on the choice of initial values and the procedure sometimes may stop before convergence is reached. It's also shown by [8] that when the true treatment effect is moderate, EM-algorithm dramatically underestimates the within-group variance while the sample variance calculated from the combined two group data (will be introduced in later section of this paper) is much simpler, and the overall variance is only slightly larger than the true within-group variance. Moreover, even though the estimation of σ^2 is accurate, to re-estimate sample size depending on the observed first stage data may bias the final test. Therefore, it might be problematic to still compare the final test with the original planned critical value.

The BSSR with Significance Level Adjustment

Objective and Design

The design given by [9] has flexible choice of first stage sample size $n_1 = t * N_0$ $(0 < t < 1)$ After observing the first stage data, two ways were proposed to re-estimate the within-group variance. The first estimation is denoted by S_{os}^2 (one-sample variance), which is simply the sample variance of the combined data from two groups. The second estimation is

$$s_{adj}^2 = s_{os}^2 - \frac{n_1}{2(2n_1 - 1)} \delta_a{}^2,$$

It adjusts between group variation by a function of the assumed treatment effect. The adjusted variance estimation s_{adj}^2 is unbiased estimation of the true within-group variance if $\delta_a = \delta$ Then re-estimate the total per-group sample size N^* by the same formula as equation 1 but replace $\sigma_a{}^2$ with s_{os}^2 and s_{adj}^2. However, after the sample size re-estimation, the final two sample t-test statistic $T^{(N^*)}$ no longer follows t_{2N^*-2} distribution since N^* is now a random variable. Comparing the final test statistic value with the original planned cut off point $t_{2N^*-2,\alpha}$ may inflate the type I error rate. To evaluate how much re-estimating sample size based on the first stage data affects the control of type I error rate, [10] gives the exact formula for calculating actual type I error rate $\alpha_{act}(\alpha,\ n_1,\ N\ )$ after sample size re-estimation. Their calculation is based on numerical integration. The unknown parameters in the actual type I error rate function $\alpha_{act}(\alpha,\ n_1,\ N\ )$ are significance level α, the first stage sample size n_1 and the unknown actual required sample size N. The authors show when the difference between the true treatment effect δ and the assumed treatment effect δ_a is moderate, there is not practical difference between α_{act} and the nominal level α, thus no adjustment is needed. For other situation, for each fixed n_1 and α, we can find a N to maximize the actual type I error α_{act} say $\alpha_{\max}(\alpha, n_1)$. Then to control the actual type I error rate at α level, for each fixed n_1, we can find an adjusted significance level α^* so that $\alpha_{max}(\alpha^*,\ n_1) = \alpha$. Then, with each pre-determined n_1 and α, using the adjusted significance level α^* for the final test ensures the type I error rate won't exceed α for any arbitrary N. After the sample size modification, efficacy is claimed at significance level α if the final test with cumulated data from two stages $T^{(N^*)} > t_{2N^*-2,\alpha^*}$.

Concerns and Weaknesses of the Design

The advantage of this design is that its type I error rate will not exceed the desired level α. However, the calculation of α_{act} is quite complicated. Furthermore, it might not be comfortable for some people to accept that the final test significance level must be changed to maintain the type I error rate only because the sample size was re-estimated. If the new significance level α^* is smaller than the original one, it feels like that this design leads to a penalty for the final test.

Part Two: Unblinded SSR (UbSSR) Designs Based on Nuisance Parameters

Two commonly used unblinded SSR (UbSSR) designs with re-estimated within-group variance $\sigma_\alpha{}^2$ are introduced in this section. The later one is based on the design of the earlier one with some adjustment on the final test significance level. Two sample t-test is used at the end of the trial with the cumulated data from two stages.

The Naïve UbSSR

Objective and Design

For some non-clinical experiments and clinical designs without blinding requirement, the design proposed by [11] and later further analyzed by [12] was one of the earliest designs that recommended to include the internal pilot study data (i.e. the first stage data) in the final test. The initial planned per-group sample size N_0 can again be given by equation 1.

The authors use two-sided test in the original paper, without loss of generosity, we can make some adjustment to make it a one-sided test. They recommend that after the data of $N_0/2$ patients are observed as per-group, increasing sample size if the pooled sample variance S^2 of the two groups based on the first stage data is larger than $\sigma_\alpha{}^2$. Because the sample sizes in their study were small, to make the calculation precisely, they use t-distribution rather than its normal approximation to compute the re-estimated sample size after the internal pilot study. If $\sigma_\alpha{}^2 < S^2$, the re-estimated per-group sample size N_* can be given by

$$N^* = \min\{n : n \geq \frac{2(t_{2n-2,\alpha} + t_{2n-2,\beta})^2}{\delta_a{}^2} S^2\} \text{ …..(2)}$$

Where $t_{2n-2,\alpha}$ and $t_{2n-2,\beta}$ are the (1−α)th and (1−β)th quantiles of the central t-distribution with degree of freedom $2n-2$. If the sample size is changed, the final test claims efficacy if the test score exceeds $t_{2N^*-2,\alpha}$.

Concerns and Weaknesses of the Design

This design greatly improves test power than the fixed size design if the variance $\sigma_\alpha{}^2$ used for calculating N_0 is less than the true variance σ^2. However, since the final test is biased because of the SSR procedure, simulation results shows there is non-negligible type I error inflation when sample size is relatively small, internal pilot is conducted at around half the required sample size and $\sigma_\alpha{}^2$ is close to the true variance σ^2.

The UbSSR with Significance Level Adjustment

Objective and Design

The design of [13,14] is based on the design of [12], but with moderate to large size trial, they used the normal approximation of t-distribution to calculate the re-estimated sample size, say, N_0 is the same as equation 1, and N^* is given by replacing $\sigma_\alpha{}^2$ in equation 1 with the pooled sample variance S^2 if $\sigma_\alpha{}^2 < S^2$. Moreover, to solve the type I error rate inflation problem, they again derived the exact formula of the actually type I error rate α_{act} after sample size adjustment, which is similar as they did for the BSSR design. The unknown parameters in the actual type I error rate function α_{act} (α, n_1, N) are significance level α, first stage sample size n_1 and the unknown actual required sample size N. For each fixed n_1 and α, we can find a N to maximize the actual type I error, say α_{max} (α, n_1). Then to control the type I error rate at α, for each fixed n_1, we can find an adjusted significance level α^* so that $\alpha_{\max}(\alpha^*, n_1) = \alpha$. Because this method has negative effect on test power if we just use the pooled sample variance of the first stage data to re-estimate the sample size, the authors also proposed to use 100(1−γ)% Upper Confident Limit (UCL) for the variance estimation to have a probability of at least 1−γ to achieve a planned power. If the sample size is changed, the final test claims efficacy if the test score exceeds t_{2N^*-2,α^*}.

Concerns and Weaknesses of the Design

Same as the BSSR with significant level adjustment, after adjusting the final test significance level, the design can control the type I error of the final test exactly at the desired level. However, the calculation of α_{act} and $\alpha_{\max}$ is quite complicated. Some people may not be comfortable with changing the final test significance level. Moreover, the method of improving power by inflating estimated variance may lead the study to lose efficiency depending on the choice of γ.

Part Three: Unblinded SSR (UbSSR) Designs Based on Treatment Effect

Three types of unblinded SSR (UbSSR) designs based on the re-estimated treatment effect after observing the first stage data are introduced in this section. Because the within-group variance σ^2 was not re-esitmated in the related literatures, here we assume σ^2=1 is a known value. Thus, the initial assumed per-group sample size can be simplified by

$$N_0 = \frac{2(z_\alpha + z_\beta)^2}{\delta_a^2} \quad \ldots\ldots\ldots\ldots\ldots \quad (3)$$

The final analysis can simply use z-test since the variance is assumed known.

The interim study is conducted at information time t after the data of $n_1 = t * N_0 (0 < t < 1)$ patients are observed per-group. Besides simply re-estimating sample size N^* by re-estimating δ based on the first stage data and substituting it to equation 3, a new method "conditional power function" is widely used in UbSSR based on the re-estimated treatment effect. The re-estimated total per-group sample size N^* now can be given by one of the following conditional power functions:

$$CP(N^*, z_\alpha | z_1) = Pr\left(Z^{(N^*)} > z_\alpha \middle| Z^{(n_1)} = z_1, \delta = \hat{\delta}\right) = 1 - \beta \quad (4)$$

$$CP(N^*, c^* | z_1) = Pr\left(Z^{(N^*)} > c^* \middle| Z^{(n_1)} = z_1, \delta = \hat{\delta}\right) = 1 - \beta \quad (5)$$

Where $Z^{(N^*)}$ is the two-sample z-statistic with sample size N^* per-group; z_α is the $(1-\alpha)th$ quantile of the standard normal distribution; z_1 is the observed first stage z-score; $\hat{\delta}$ is the re-esimated treatment effect based on the first stage data; c^* is the re-estimated final test critical value, which can be solved by combining equation 5 with the conditional error function in equation 6

$$Pr_0\left(Z^{(N^*)} > c^* \middle| Z^{(n_1)} = z_1\right) = Pr_0\left(Z^{(N_0)} > z_\alpha \middle| Z^{(n_1)} = z_1\right). \quad (6)$$

The UbSSR with Re-Designed Final Test Statistic

Objective and Design

Since to re-estimate the sample size depending on the data observed from the first stage could inflate the type I error rate, some of the designs control the type I error rate by re-designing the final test statistics. The distribution of the re-designed final test statistic will not be affected by the sample size modification under the null hypothesis, so we are able to use the original planned critical values for the final test. The re-estimated per-group sample size N^* of the two methods below can be calculated by any appropriate method (such as equation 3 with $\delta_a{}^2$ replaced by estimated value, equation 4, etc.) after first stage data are observed without inflating the type I error rate.

The design in [15] uses the product of stochastically independent uniform [0,1] distributed p-values from tests before and after the preplanned sample size adjustment to construct a single global test statistic. H_0 can be rejected at significance level α at the end of the trial if

$$p_1 p_2 \leq c_\alpha = \exp[-\tfrac{1}{2}\chi_4^2(1-\alpha)],$$

where c_α is the Fisher's product criterion; p_1, p_2 are the observed error probabilities (p-value) for the tests based on the data observed before and after the interim analysis; $\chi_4^2(1-\alpha)$ is the (1-α)th quantile of the central chi-squared distribution with 4 degrees of freedom.

The design proposed by [16] modifies the traditional z-test of two-sample means by changing the weights of independent z-score from before and after the interim analysis (linear summation of z-score from each stage). If the sample size modification is made, the final test statistic can be given by

$$Z_w = W_1 Z_1 + W_2 Z_2^* = \sqrt{\frac{n_1}{N_o}} Z_1 + \sqrt{1-\frac{n_1}{N_o}} Z_2^*,$$

where Z_1 is the z-score calculated based on the first stage data and Z_2^* is the z-score calculated by the second stage data with re-estimated sample size. Note that this approach is equivalent to a combination test with inverse normal combination function in [17]. The sample size modification will not change the distribution of the test statistic Z_W under the null hypothesis, because Z_1 and Z_2^* are independent and follow standard normal distribution; Therefore, as long as the weights W_1 and W_2 are pre-specified, satisfy W_1+W_2=1, and remain unchanged when the sample size changes, then Z_W is also following standard normal distribution. Thus, the rejection criterion Z_W>z_α results in a level-α test.

Concerns and Weaknesses of the Design

Since the distribution of these re-designed test statistics will not be changed by the sample size modification, the type I error probability will be preserved exactly at desired level, its generality and simplicity greatly facilitate the application of these methods. However, the authors of [15] claim their method has a very small loss of power compared to the optimal test in the whole sample. It is not a surprise as it's generally a nonparametric method, which may lead to power loss compared to parametric methods when the distribution information is known. Moreover, it is well known that the method of [16] unequally weighted the patients enrolled before and after the interim study if a decision of increasing sample size is made, which violates the one patient one vote principle [18]. also mentioned that the modified test statistic will cause efficiency loss since it is not a sufficient statistic for mean difference.

The UbSSR with Adjusted Final Critical Value

Objective and Design

The designs proposed by [18-22] may have different representations, but eventually are based on similar idea. Their methods control the type I error rate by using the re-estimated final test critical value $c*$ (or significance level) calculated based on the conditional error function in equation 6 combined with any of the sample size re-estimation function based on re-estimated treatment effect we have introduced in this paper (e.g. equation 3 with δ_α^2 replaced by estimated value, equation 4 or equation 5). H_0 can be rejected at significance level α at the end of the trial if $Z^{(N^*)} > c^*$.

Concerns and Weaknesses of the Design

With the re-estimated sample size $N*$ and re-estimated critical value c^* calculated by the combined solution of equation 6 and some sample size re-estimation functions, the type I error rate and the conditional power of the final test will be preserved at the desired level. However, the methods of [19-21] provided no constraint or didn't give clear criteria about how to find the constraint on the range of conditional power that allows SSR. The numerical example in [23] suggests that no lower boundary of the adaptive region or no upper boundary of the sample size increase will cause design inefficiencies if a very small value of conditional power is obtained at the interim analysis which is equivalent to having a very large re-estimated critical value $c*$ at the final test. On the other hand, although the designs in [18,22] provide constraint on the range of adaptive region, similar as [19-21], re-estimating sample size with their proposed region may lead the final test to be compared with a critical value larger than the original planned critical value z_{α}. As it was mentioned in previous sections, it might be hard for design users to accept the critical value for the final test to be changed only because the sample size is changed. It is especially difficult when we need a larger critical value c^* than the original critical value z_{α}, it's like giving a penalty for the final test [18]. Moreover, it was proved in [18] that even though $Z^{(N^*)} > c^*$ appear to use the sufficient statistic $Z^{(N*)}$ for the final analysis, it is actually functionally equivalent to the test $Z_{w} > z_{\alpha}$ discussed by Cui, Hung and Wang, and the first stage data are hidden in the re-estimated critical value $c*$. Thus, the problem of violate the one patient one vote principle is equally applicable for UbSSR with adjusted final critical value method. Furthermore, it's also problematic that the critical value for the final test cannot be decided before completing the first stage studies.

The (Constrained) Promising Zone UbSSR

Objective and Design

Because changing the final test critical value due to SSR may not be easily accepted, the designs proposed by [24-26] control the type I error by giving a constraint on the range of conditional power (given by in equation 7) that allows SSR. This constraint is the so called "promising zone".

$$CP(N_0, z_\alpha | z_1) = \\ Pr\left(Z^{(N_0)} > z_\alpha \middle| Z^{(n_1)} = z_1, \delta = \hat{\delta}\right) \text{ .... (7)}$$

They claim that if we only increase sample size when the conditional power with original planned sample size N_0 falls in the promising zone, comparing the final test with the original planned critical value z_{α} will not inflate the type I error rate. Thus, no matter whether the sample size is modified after the first stage, H_0 can be rejected at the end of the trial if the final test score exceeds the original planned critical value z_{α}.

Two procedures are proposed and compared by [24]. The first procedure allows increasing sample size if $CP(N_0, z_\alpha|z_1) \in (0.5,\ \Phi(z_\beta/\sqrt{1-t}))$ and the re-estimated sample size is calculate by replacing $\delta_a{}^2$ with $\hat{\delta}^2$ in equation 3. The second procedure allows increasing sample size if $CP(N_0, z_\alpha|\mathrm{z}_1)$ is between 0.5 and $1-\beta$, its new sample size can be given by equation 4. It was proved that both procedures control the type I error rate, but the simulation results in their paper show that the second procedure is more powerful than the first procedure.

The designs given by [18,25,26] were proved to have a wider promising zone than the design of [24]. More specifically, their "Promising Zone" includes all the value of $CP(N_0, z_\alpha|\mathrm{z}_1)$ that

$c^* < z_\alpha$ Their new sample size N^* and c^* are calculated by solving equation 5 and 6 together. The lower bound of their promising zone is always lower than 0.5 while their upper bound is also at $1-\beta$.

The promising zone UbSSR design was further developed in [27] by setting additional constraints on the range of the "Promising Zone", which considering the balance between increasing conditional power and the cost for increasing sample size. Later, [28] also proposed to constrain the range of the "Promising Zone" with the information of maximum allowed sample size and the range of the conditional power achieved with the maximum allowed sample size evaluated at the smallest clinical meaningful treatment benefit.

Concerns and Weaknesses of the Design

The lower boundary of the promising zone of [24] is set fixed at $CP(N_0, z_\alpha|\text{z}_1)$=0.5 which might be too narrow and do not power the final test to the extent we want. On the other hand, the design of [18,25,26] uses equation 5 to re-estimate the total sample size to guarantee a conditional power of the final test with the re-estimated critical value c^* to achieve the desired level 1-β, but at the end of the trial, they actually compare the final test with the original critical value z_α (always larger than c^* in the promising zone). It is conservative, because the re-estimated total sample size does not actually power the test enough to the critical value it compared with at the final test, even in terms of the conditional power. It may result in loss of power if the difference between z_α and c^* is large.

Further Concerns for Adaptive SRR Based on Conditional Power

Although conditional power-based adaptive SSR can save a trial when the original planned total sample size is underestimated, never think it is without penalty. In fact, it can save the trial if the assumed treatment effect slightly overestimates the true treatment effect, adding more samples can improve the power to certain extent. However, for certain situation, the uncertainty of the conditional power function will actually reduce the efficiency of a well-designed trial. It was shown in the paper of [29] when the expected sample size of a fixed size design is equivalent to that of a [26] design, the power of the [26] design can be lower than that of a fixed size design. It was also pointed out in [30], when the true effect size is small, recalculate sample size in mid-trial based on an interim estimate may lead to an overly large price to be paid in average sample size compared to the gain in overall power. On the other hand, if the assumed treatment effect dramatically over-estimates the true value, the conditional power at the interim study will be too low, no sample size modification will be made, and nothing will be gained from the extra procedure. Moreover, due to the randomness of the conditional power, for small sample size, even when the original trial is well designed, there will still be a high chance for the adaptive SSR to increase the sample size to achieve an undesired higher power. It's better to examine the operating characteristics (power and type I error) of the entire procedure, which can be done, for example, by simulating the adaptive design under different values of δ in the range of interest, through such simulations that one may be able to judge whether the adaptive design is worth adopting.

Conclusions and Remarks

There is no doubt that the designs we review in this paper may help some underpowered trials from failing to find a significant treatment benefit, but as it was summarized in each section, none of them is a perfect design. Before we apply the SSR designs to any real clinical trial, it would be better for us to take into consideration their potential problems such as inflation of type I error rate, inefficiency, computation complication, impractical, etc.

Moreover, a few additional points are also worth to be mentioned. 1) If the population is not following normal distribution, most of the designs discussed in this paper are using asymptotic normal distribution derived from central limit theorem to calculate sample size and conduct hypothesis testing, thus they are only applicable with large sample size at both stages. 2) In this paper, we only discuss the SSR designs for superiority tests. The designs may encounter more problems when they are applied to non-inferiority or equivalence hypothesis. 3) We assume equal variance for both treatment groups at the beginning of this paper, which is also the assumption given by most of the papers we reviewed. Formulation may be more complicated, and efficiency may be compromised when the variances are actually different. 4) The designs based on conditional power function we reviewed in this paper assume known variance for both treatment groups, thus they don't have to re-estimate variance and they can use simple z-statistic for the final test. The formulation will be more complicated but also more accurate if the variance is re-estimated and t-statistic is used for the final test. 5) When we compare different SSR designs, besides the basic statistical operating characteristics (type I error rate, power, etc.), we have plenty of different criteria but hard to identify a most important one. We need to take good consideration about the advantages and disadvantages of each design before using it in clinical trials.

References

1. Friede T, Kieser M. 2006. Sample size recalculation in internal pilot study designs: a review. *Biometrical Journal*. 48: 537-555. Ref.: https://bit.ly/2WvxWpJ
2. Proschan MA. 2009. Sample size re-estimation in clinical trials. *Biometrical Journal*. 51: 348-357. Ref.: https://bit.ly/2OzrTxW
3. Menon S, Massaro J, Pencina MJ, et al. 2013. Comparison of operating characteristics of commonly used sample size re-estimation procedures in a two-stage design. *Communications in Statistics-Simulation and Computation*. 42: 1140-1152. Ref.: https://bit.ly/2CCA3k3
4. Pritchett YL, Menon S, Marchenko O, et al. 2015. Sample size re-estimation designs in confirmatory clinical trials-current state, statistical considerations, and practical guidance. *Statistics in Biopharmaceutical Research*. *7*: 309-321. Ref.: https://bit.ly/2WwJut9
5. Guidance Draft. 2018. Adaptive Designs for Clinical Trials of Drugs and Biologics. Center for Biologics Evaluation and Research (CBER). Ref.: https://bit.ly/2TZaq81
6. Gould LA, Shih WJ. 1992. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics-Theory and Methods*. 21: 2833-2853. Ref.: https://bit.ly/2HLiN0c
7. Gould LA. 1997. Issues in blinded sample size re-estimation. *Communications in Statistics-Simulation and Computation*. 26: 1229-1239. Ref.: https://bit.ly/2Yywa9x
8. Friede T, Kieser M. 2002. On the inappropriateness of an EM algorithm based procedure for blinded sample size re-estimation. *Statistics in Medicine*. 21: 65-176. Ref.: https://bit.ly/2Ou39qI
9. Zucker DM, Wittes JT, Schabenberger O, et al. 1999. Internal pilot studies II: comparison of various procedures. *Statistics in Medicine*. 18: 3493-3509. Ref.: https://bit.ly/2HI4Pwh
10. Kieser M, Friede T. 2003. Simple procedures for blinded sample size adjustment that do not affect the type I error rate. *Statistics in Medicine*. 22: 3571-3581. Ref.: https://bit.ly/2FF45pn

11. Stein C. 1945. A two-sample test for a linear hypothesis whose power is independent of the variance. *The Annals of Mathematical Statistics*. 16: 243-258. Ref.: https://bit.ly/2HKKvKM
12. Wittes J, Brittain E. 1990. The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*. 9: 65-72. Ref.: https://bit.ly/2TFTWwy
13. Kieser M, Friede T. 2000. Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine*. 19: 901-911. Ref.: https://bit.ly/2TEpanC
14. Friede T, Kieser M. 2001. Sample size adjustment in clinical trials for proving equivalence. *Drug information journal*. 35: 1401-1408. Ref.: https://bit.ly/2FCVWQX
15. Bauer P, Kohne K. 1994. Evaluation of experiments with adaptive interim analyses. *Biometrics*. 1029-1041. Ref.: https://bit.ly/2JJWyJT
16. Cui L, Hung HM, Wang SJ. 1999. Modification of sample size in group sequential clinical trials. *Biometrics*. 55: 853-857. Ref.: https://bit.ly/2TBx8y1
17. Lehmacher W, Wassmer G. 1999. Adaptive sample size calculations in group sequential trials. Biometrics. 55: 1286-1290. Ref.: https://bit.ly/2HWT5Fx
18. Mehta C, Liu L. 2016. An objective re-evaluation of adaptive sample size re-estimation: commentary on ‘Twenty-five years of confirmatory adaptive designs’. *Statistics in Medicine*. 35: 350-358. Ref.: https://bit.ly/2JKPm0o
19. Proschan MA, Hunsberge SA. 1995. Designed extension of studies based on conditional power. *Biometrics*. 1315-1324. Ref.: https://bit.ly/2HMbHIR
20. Müller HH, Schäfer H. 2001. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*, 57: 886-891. Ref.: https://bit.ly/2FFF7Gq
21. Denne JS. 2001. Sample size recalculation using conditional power. *Statistics in medicine*. 20: 2645-2660. Ref.: https://bit.ly/2TFYkvw
22. Gaffney M, Ware JH. 2017. An evaluation of increasing sample size based on conditional power. *Journal of Biopharmaceutical Statistics*. 1-11. Ref.: https://bit.ly/2HXw2u4
23. Tsiatis AA, Mehta C. 2003. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*. 90: 367-378. Ref.: https://bit.ly/2WvNIkO
24. Chen YH, DeMets DL, Gordon Lan KK. 2004. Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine*. 23: 1023-1038. Ref.: https://bit.ly/2FxU3VT
25. Gao P, Ware JH, Mehta C. 2008. Sample size re-estimation for adaptive sequential design in clinical trials. *Journal of Biopharmaceutical Statistics*. 18: 1184-1196. Ref.: https://bit.ly/2HXirTA
26. Mehta CR, Pocock SJ. 2011. Adaptive increase in sample size when interim results are promising: a practical guide with examples. *Statistics in Medicine*. 30: 3267-3284. Ref.: https://bit.ly/2THliCq
27. Jennison C, Turnbull BW. 2015. Adaptive sample size modification in clinical trials: start small then ask for more?. Statistics in medicine. 34: 3793-3810. Ref.: https://bit.ly/2I2QQAi
28. Hsiao ST, Liu L, Mehta CR. 2018. Optimal promising zone designs. Biometrical Journal. Ref.: https://bit.ly/2JKTihx

29. Emerson SS, Levin GP, Emerson SC. 2011. Comments on 'Adaptive increase in sample size when interim results are promising: A practical guide with examples'. *Statistics in Medicine*, 30: 3285-3301. Ref.: https://bit.ly/2V111Jz
30. Bauer P, Koenig F. 2006. The reassessment of trial perspectives from interim data—a critical view. Statistics in medicine. 25: 23-36. Ref.: https://bit.ly/2FG7MeA
31. Shih WJ, Li G, Wang Y. 2016. Methods for flexible sample-size design in clinical trials: Likelihood, weighted, dual test, and promising zone approaches. *Contemporary Clinical Trials*. 47: 40-48. Ref.: https://bit.ly/2TG0sDx
32. Levin GP, Emerson SC, Emerson SS. 2013. Adaptive clinical trial designs with pre-specified rules for modifying the sample size: understanding efficient types of adaptation. *Statistics in Medicine*, 32: 1259-1275. Ref.: https://bit.ly/2uwNVIf