Adaptive designs with sample size or event count re-estimation

MedianaDesigner package

1. Introduction

This document provides a description of the statistical methodology used in the adaptive design module that supports sample size or event count re-estimation (ADSSMod function).

For more information on the MedianaDesigner package, visit the following web pages at

http://www.mediana.us/medianadesigner

http://medianasoft.github.io/MedianaDesigner

2. Adaptive designs with sample size or event count re-estimation

2.1. Trial design

Consider a Phase III trial with two arms (experimental treatment versus control). The primary efficacy endpoint in this trial could be a continuous, binary or time-to-event endpoint. Two interim analyses will be conducted to perform early futility and efficacy assessments.

The interim analysis data will be analyzed in an unblinded manner to support the following decision rules:

- Futility stopping rule at the first interim analysis: The trial will be stopped for futility if a significant treatment effect is unlikely to be established at the final analysis.
- Sample size/event count re-estimation rule at the second interim analysis: The total sample size (continuous or binary endpoints) or total number of events (time-to-event endpoint) will be increased if the predicted probability of success is lower than expected at this interim analysis.

Note that the futility stopping and sample size/event count re-estimation rules are non-binding and could be overridden by the trial's sponsor or data monitoring committee. Also, the trial will not be terminated at either interim analysis due to superior efficacy.

The futility stopping rule and sample size/event count re-estimation rule are defined in Sections 2.2 and 2.3, respectively. Adaptive design methodology is presented in Section 2.4. The

proposed approach to constructing adaptive designs with sample size or event count re-estimation is illustrated in Section 3.

2.2. Futility stopping rule

The futility stopping rule to be applied at the first interim analysis could be set up using any relevant definition of the predicted probability of success, see, for example, Wassmer and Brannath (2016, Chapter 7). Most commonly, conditional power is used, which is defined as the probability of establishing a significant treatment effect on the primary efficacy endpoint at the final analysis conditional upon the primary endpoint data available at this interim analysis.

The following notation will be used to compute conditional power at the first interim analysis. Assuming the primary efficacy endpoint is continuous or binary and assuming a balanced trial design, let n_1 denote the total sample size at the first interim analysis. Similarly, let n_2 denote the total sample size at the final analysis. Conditional power is given by

$$\text{CP} = \Phi(aZ - bz_{1-\alpha}),$$

where Z is the interim test statistic,

$$a = \sqrt{\frac{n_2 - n_1}{n_1}} + \sqrt{\frac{n_1}{n_2 - n_1}},\ b = \sqrt{\frac{n_2}{n_2 - n_1}},$$

$z_{1-\alpha}$ is the $100(1-\alpha)$ percentile of the standard normal distribution and $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. Finally, α denotes the one-sided significance level in the trial, i.e., $\alpha = 0.025$. The conditional power formula assumes that the future data are consistent with the interim data. Conditional power is derived in a similar way for trials with time-to-event endpoints.

The trial will be terminated for futility at the first interim analysis if the conditional power does not exceed a pre-defined threshold denoted by c_1, where $0 < c_1 < 1$. This futility threshold is typically set to a fairly low value, e.g., c_1 is commonly less than 0.3. In other words, early stopping is recommended if the predicted probability of success is too low.

It is helpful to note that the futility stopping rule is easy to re-write in terms of the interim effect sizes if the primary efficacy endpoint is continuous or binary, or the interim hazard ratios if the primary efficacy endpoint is a time-to-event endpoint. For example, assuming a continuous or binary endpoint, the effect size at the first interim analysis is given by

$$Z/\sqrt{n_1/4}.$$

The effect size corresponding to conditional power of c_1 is equal to

$$\theta = \frac{\Phi^{-1}(c_1) + bz_{1-\alpha}}{a\sqrt{\frac{n_1}{4}}},$$

which means that the trial will be terminated for futility if the interim effect size is less than θ.

2.2. Sample size/event count re-estimation rule

A sample size/event count re-estimation rule will be applied at the second interim analysis to boost the probability of success if the predicted probability of success is not sufficiently high. This rule will also be set up using conditional power and conditional power will be computed at the second interim analysis as in Section 2.1.

The sample size/event count re-estimation rule is constructed using two thresholds for conditional power, denoted by c_2 and c_3. These thresholds define the "promising interval" where it is most sensible to increase the trial's sample size or target number of events. Assuming that the trial was not terminated due to futility at the first interim analysis, the total number of patients or events may be modified at the second interim analysis as follows

- If $\mathrm{CP} \leq c_2$, retain the original sample size or number of events.
- If $\mathrm{CP} > c_2$ and $\mathrm{CP} \leq c_3$, increase the sample size or number of events to achieve the desirable level of conditional power up to a pre-defined cap.
- If $\mathrm{CP} > c_3$, retain the original sample size or number of events.

The desirable level of conditional power is typically set to the anticipated probability of success in the trial, e.g., if the trial is powered at 90%, the desirable level of conditional power will be set to 0.9. Using the notation introduced in Section 2.1, the updated number of patients or events at the final, denoted by $\tilde{n}_2$, is given by

$$\tilde{n}_2 = n_1 + \frac{n_1}{Z^2}\left(z_{0.9} + z_{1-\alpha}\sqrt{\frac{n_2}{n_2 - n_1}} - Z\sqrt{\frac{n_1}{n_2 - n_1}}\right)^2$$

if the desirable level of conditional power is 0.9. As indicated above, the updated sample size or event count will be capped at a pre-defined value. This cap is often set to 20% or 30%, i.e., the sample size or event count cannot be increased by more than 20% or 30%. However, the cap could sometimes be as high as 50%.

2.3. Adaptive design methodology

The statistical inferences at the final analysis need to be adjusted to account for the data-driven re-estimation of the total sample size or total number of events at the second interim analysis. This adjustment is required since data-driven design changes are known to inflate the Type I

error rate; however, if no changes are made at the second interim analysis, the final analysis will be performed without any adjustments.

To apply the statistical adjustment, the treatment effect needs to be evaluated separately in two trial stages. For example, assuming a continuous or binary endpoint, the two stages are defined as follows:

- Stage 1 includes all patients who complete the treatment period or drop out of the trial before completing the treatment period by the second interim analysis.
- Stage 2 includes all patients who complete the treatment period or drop out of the trial before completing the treatment period after the second interim analysis.

The evidence of treatment effectiveness from the two trial stages will be pooled using the combination function principle (Wassmer and Brannath, 2016, Chapter 6) as explained below. Let p_1 and p_2 denote the one-sided treatment effect p-values computed from the Stage 1 and Stage 2 data, respectively. The significance test at the final analysis will be performed using the combined p-value, which is defined as

$$p = c(p_1, p_2),$$

and a significant treatment effect will be established at the final analysis if $p \leq \alpha$. The stagewise p-values are combined using the weighted inverse-normal combination function, i.e.,

$$c(x, y) = 1 - \Phi\left(\sqrt{w}\Phi^{-1}(1 - x) + \sqrt{1 - w}\Phi^{-1}(1 - y)\right),$$

$\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution, and w and $1 - w$ are the pre-defined weights assigned to Stages 1 and 2. Most commonly, the weight of Stage 1 (w) is equal to the anticipated information fraction at the interim analysis. For example, if the second interim look is expected to be taken after 60% of the patients complete the treatment period or drop out of the trial before completing the treatment period, w will be set to 0.6.

3. Case study

Adaptive designs with sample size re-estimation will be illustrated using a Phase III trial in patients with rheumatoid arthritis. The efficacy and safety profiles of a single dose of an experimental treatment will be compared to those of placebo. The primary efficacy endpoint in this trial is binary; it is based on the American College of Rheumatology definition of improvement. The experimental treatment is expected to improve the response rate.

Patients will be randomized equally to the two trial arms and the total number of enrolled patients in the trial is 240 patients. The sample size calculation assumed that the true response rates in the placebo and treatment arms are equal to 35% and 60%, respectively. Using a one-

sided $\alpha = 0.025$ and assuming a 15% patient dropout rate, the sample size of 240 patients guarantees at least 90% power; however, the sponsor is concerned about the fact that the placebo response rate could be higher than anticipated and a larger sample size would be required to maintain 90% power.

To improve the robustness of this trial, an adaptive design with two interim analyses will be employed. The first interim look, aimed at a futility assessment, will be taken after 40% of the patients complete the treatment period or drop out of the trial prior to completing the treatment period. An option to increase the number of enrolled patients up to 30%, i.e., up to 312 patients, will be enabled at the second interim analysis. This interim analysis will be taken after 60% of the patients complete the treatment period or drop out of the trial.

The futility stopping and sample size re-estimation rules will be set up using conditional power. The futility threshold (c_1) at the first interim analysis was selected using an optimal approach defined in the futility module (FutRule function). Briefly, an optimal threshold was derived by maximizing the rule's sensitivity and specificity rates. The sensitivity rate was computed under the assumption that the true placebo and treatment response rates are equal to 35% and 55%, respectively. The resulting optimal futility threshold is approximately 0.3 and guarantees that the sensitivity and specificity rates of the futility stopping rule are at least 80%.

The sample size re-estimation rule at the second interim analysis was constructed using the following thresholds for conditional power:

- $c_2 = 0.4$.
- $c_3 = 0.9$.

The thresholds define the promising interval at this interim look.

Figure 1 presents a graphical summary of the sample size re-estimation rule. The total sample size at the final analysis is plotted in this figure as a function of the effect size at the interim analysis. It is shown in this figure that the total sample size at the final analysis is set at 240 patients if the interim effect size is greater than 0.358 or less than 0.232, the total sample size grows linearly to 312 patients if the interim effect size is between 0.318 and 0.358 and, finally, the total sample size is 312 patients if the interim effect size is between 0.232 and 0. 318.

The key operating characteristics of the proposed adaptive design under two alternative treatment effect scenarios are summarized in Tables 1 and 2. Under these scenarios, the treatment response rate is consistent with the original assumptions, i.e., it is fixed at 60%, but a higher placebo response rate is assumed, namely,

- Scenario 1: Placebo response rate is 37.5%.
- Scenario 2: Placebo response rate is 40%.

The tables also provide information on a reference design (traditional design) with the same futility stopping rule at the first interim analysis and a fixed sample size of 240 patients.

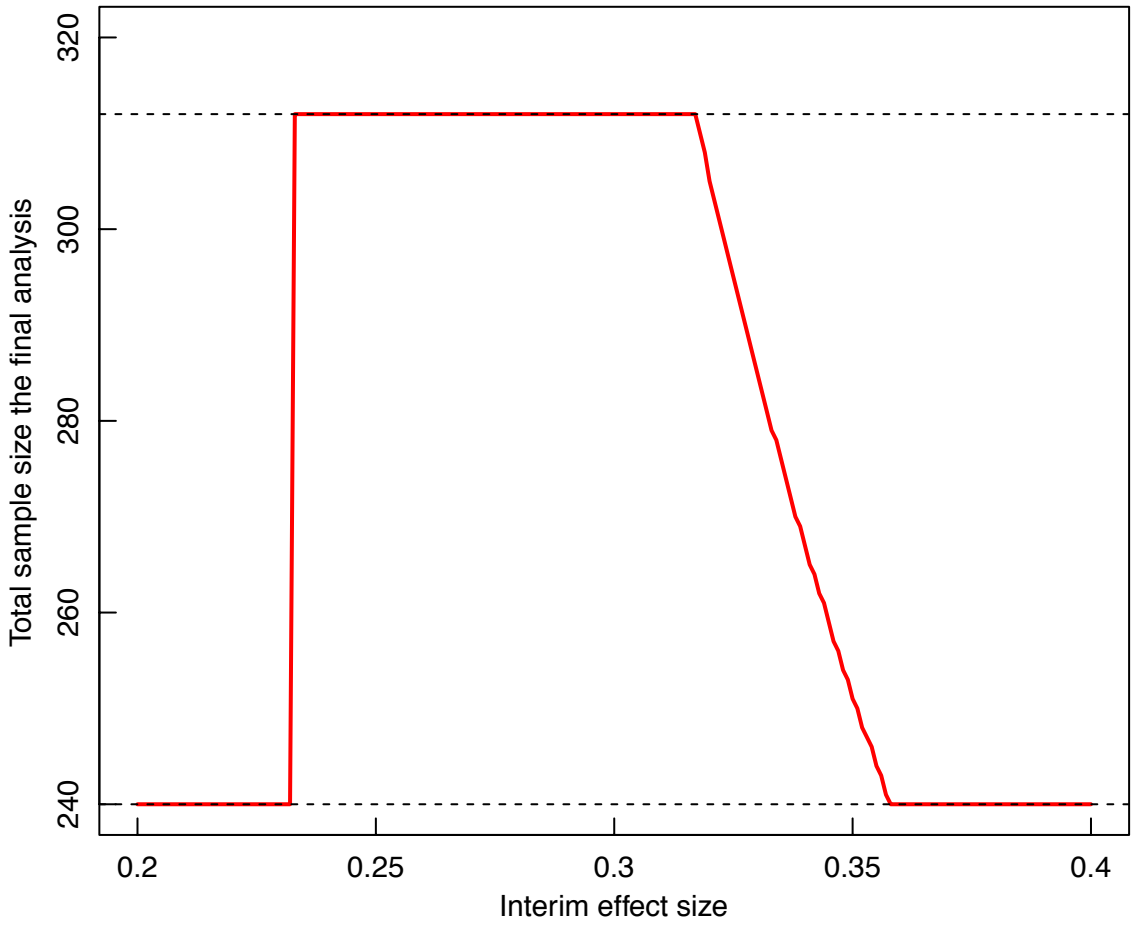
Beginning with Table 1, the futility stopping rule is more likely to be triggered under Scenario 2, namely, the probability of stopping due to futility increases from 11.8% under Scenario 1 to 17.5% under Scenario 2. This is not surprising since the treatment effect is weaker under Scenario 2 and thus there is a higher chance that the effect size will be smaller at the first interim analysis. The probability of increasing the sample size is also higher under Scenario 2, which means that the interim results are more likely to fall within the promising interval. Both trial designs are slightly underpowered under Scenario 1 (power is equal to 83.6% or 84.9%) and considerably underpowered under Scenario 2 (power is equal to 74.0% or 76.0%). These results would suggest that a larger sample size should be considered to ensure adequate power under either scenario.

Table 2 shows that, instead of committing upfront to a larger sample size, the trial's sponsor can rely on the adaptive design to increase the total number of patients in the trial when the results look promising at the second interim analysis. Under Scenario 1, if the interim results end up within the promising window, the traditional design with 240 patients, cannot guarantee power of 90%. With this design, power at the final analysis will be reduced by about two percentage points. By employing a data-driven sample size re-estimation rule, the adaptive design can easily improve power, in fact, power at the final analysis is expected to exceed 95%. The same pattern is observed under Scenario 2 when the treatment effect is even weaker. When the traditional design is applied, power within the promising window is substantially reduced, down to 82.4% whereas the adaptive design can still boost power and guarantee the probability of success over 90%.

References

Wassmer, G., Brannath, W. (2016). Group Sequential and Confirmatory Adaptive Designs in Clinical Trials. New York: Springer.

Figure 1. Sample size re-estimation rule at the second interim analysis



The dashed lines are drawn at the original sample size of 240 patients and maximum sample size of 312 patients.

Table 1. General characteristics of the traditional and adaptive designs

Treatment effect scenario	Parameter	Value
Scenario 1	Probability of stopping for futility at the first interim analysis	11.8%
	Probability of increasing the sample size at the second interim analysis	17.6%
	Traditional design: Power	83.6%
	Adaptive design: Power	84.9%
Scenario 2	Probability of stopping for futility at the first interim analysis	17.5%
	Probability of increasing the sample size at the second interim analysis	20.8%
	Traditional design: Power	74.0%
	Adaptive design: Power	76.0%

Scenario 1: Placebo response rate is 37.5% and treatment response rate is 60%. Scenario 2: Placebo response rate is 40% and treatment response rate is 60%.

Table 2. Power of the traditional and adaptive designs within the promising interval

Treatment effect scenario	Design	Power
Scenario 1	Traditional design	88.2%
	Adaptive design	95.5%
Scenario 2	Traditional design	82.4%
	Adaptive design	91.8%

Scenario 1: Placebo response rate is 37.5% and treatment response rate is 60%. Scenario 2: Placebo response rate is 40% and treatment response rate is 60%.