Adaptive designs with data-driven treatment selection

MedianaDesigner package

1. Introduction

This document provides a description of the statistical methodology used in the adaptive design module that supports data-driven treatment selection (ADTreatSel 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 data-driven treatment selection

2.1. Trial design

Consider a multi-arm Phase III trial that will be conducted to investigate the efficacy of several doses or regimens of an experimental therapy (the doses or regimens will be referred to as treatments) compared to a control, e.g., placebo. 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 an early futility assessment and identify the most promising treatment, respectively.

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: A futility assessment will be carried out for each treatment and a treatment will be dropped at this interim look if it is unlikely to be effective.
- Treatment selection rule at the second interim analysis: One or more treatments will be selected and the remaining treatments will be dropped at this interim analysis.

The decision rules (futility stopping and treatment selection rules) are non-binding and could be overridden by the trial's sponsor or data monitoring committee. It is assumed in this module that the trial will not be stopped at either interim analysis due to superior efficacy; however, the proposed adaptive design is easily extended to support an option to evaluate the efficacy profile of each treatment and terminate the trial due to superior efficacy if a very strong efficacy signal is detected in at least one treatment arm. Another potential extension relies on a rule for sample size or event count re-estimation introduced in the adaptive design module that supports sample

size or event count re-estimation (ADSSMod function). After the best treatment has been chosen at the second interim analysis, the target number of patients or events could be increased to improve the probability of success based on a pre-defined re-estimation rule.

The futility stopping and treatment selection rules are defined in Sections 2.2 and 2.3, respectively, and the adaptive design methodology is presented in Section 2.4. The proposed approach to defining adaptive designs with data-driven treatment selection is illustrated in Section 3.

2.2. Futility stopping rule

The futility stopping rule to be applied at the first interim analysis will be set up using conditional power, see, for example, Wassmer and Brannath (2016, Chapter 7). It is defined for each individual treatment as the probability of a significant improvement over control with respect to the primary efficacy endpoint at the final analysis conditional upon the interim data in the selected treatment arm and control arm.

Let m denote the total number of treatments in the trial and consider the comparison of the kth treatment versus control, $k = 1, \ldots, m$. The conditional power for this comparison is denoted by CP_k. The derivation of conditional power is provided in the documentation for the ADSSMod function and will be omitted. The kth treatment arm will be dropped at the first interim analysis due to futility if the treatment-specific conditional power does not exceed a pre-defined threshold denoted by c, where $0 < c < 1$, i.e.,

$$CP_k \leq c.$$

The futility threshold is typically set to a fairly low value, e.g., it rarely exceeds 0.3. The trial will be terminated due to futility at this interim look if all treatments are dropped.

2.2. Treatment selection rule

A treatment selection rule will be applied at the second interim analysis to identify the pre-defined number of best performing treatments. The best treatments correspond to the largest effect sizes based on the primary efficacy endpoint (the effect size is defined as the negative log-hazard ratio if the primary efficacy endpoint is a time-to-event endpoint).

2.3. Adaptive design methodology

As in the other adaptive design modules (ADSSMod and ADPopSel functions), a data-driven decision rule will be applied at the second interim analysis, which will inflate the Type I error rate. To address this problem, the comparison of the selected treatment arm to control at the final analysis needs to be adjusted as described below.

To define the adjustment, consider first the case where no treatment selection rule is applied at the second interim analysis and all treatments that were not dropped at the first interim analysis

will be retained. Let p_k denote the one-sided treatment effect p-value corresponding to the comparison of the kth treatment versus control at the final analysis, $k = 1, \ldots, m$. Since there are several opportunities to claim a positive outcome of this trial, a multiplicity adjustment (multiple testing procedure) needs to be applied. The available options include the Bonferroni, Holm and Hochberg procedures. For more information on commonly used multiplicity adjustments, see for example Dmitrienko and D'Agostino (2018).

Returning to the adaptive design with a data-driven treatment selection rule, let $1 \leq s \leq m$ denote the pre-specified number of treatments chosen at the second interim analysis. It is important to note that, even if only one treatment is chosen to be compared to control at the final analysis, there is still a need for a multiplicity adjustment. This adjustment represents a penalty for selection bias due to the option to identify the best treatment at the second interim analysis.

The treatment effect p-values for the selected treatments will be computed in a standard way at the final analysis and the p-values for the remaining treatments will be set to 1. To protect the overall Type I error rate in the trial, a prospectively defined multiple testing procedure will be applied to the resulting p-values. For example, if a single treatment is chosen for the final assessment, all but one p-value are equal to 1 and the resulting multiplicity adjustment typically takes a very simple form. With the Bonferroni, Holm or Hochberg procedures, a statistically significant treatment effect will be established at the final analysis if the p-value for the selected treatment does not exceed α/m, where α denotes the one-sided significance level in the trial ($\alpha = 0.025$).

3. Case study

The use of data-driven treatment selection rules in confirmatory trials will be illustrated using a Phase III trial for the treatment of schizophrenia. Three doses of an investigational treatment will be compared to placebo in this trial. The doses are labeled Dose L (low dose), Dose M (medium dose) and Dose H (high dose). The length of the treatment period is 6 weeks and the primary efficacy analysis is based on the change from baseline to Week 6 in the PANSS (Positive and Negative Syndrome Scale) total score. A larger reduction in the PANSS total score indicates a beneficial effect.

The data collected in the Phase II trial are not very reliable due to high variability and the trial's sponsor would like to take an advantage of an adaptive design to perform an early futility assessment and drop underperforming doses. Another interim look will be taken to select the most promising dose. This dose will be included in the final analysis whereas the other doses will be dropped.

The number of enrolled patients per arm was set to 180. An equal randomization scheme was assumed and a high patient dropout rate of 25% was assumed (patients are likely to be lost to follow up in schizophrenia trials).

The adaptive design will be set up as follows:

- A futility stopping rule based on conditional power will be applied at the first interim analysis, which corresponds to a 30% information fraction. This means that this interim analysis will be conducted after 216 patients complete the treatment period or drop out of the trial prior to completing the treatment period. The futility threshold (c) was set to 20%. This value is lower than an optimal threshold derived using the approach implemented in the futility module (FutRule function). The lower value was chosen because it resulted in a very high sensitivity rate (probability of correctly retaining at least one treatment arm at this interim analysis) assuming a common effect size of 0.4 across the three treatment arms. Under this alternative hypothesis of beneficial effect, the sensitivity rate exceeds 90%.
- The second interim analysis will be performed at 50% of the total sample size, i.e., after 360 patients complete the treatment period or drop out of the trial before completing the treatment period. The most promising dose with the largest effect size will be chosen at this interim analysis.
- The final analysis will be conducted using a multiplicity adjustment based on the Hochberg test, which means that a significant treatment effect will be established if the p-value for the chosen dose is no greater than $\alpha/3$.

It is helpful to note that the overall sample size in the adaptive design is much less than the number of enrolled patients (180 patients) times the number of trial arms (4 arms). An important feature of the adaptive design is that only two arms are retained in the trial after the second interim analysis. If the patient enrollment is sufficiently slow, the total sample size could be reduced by 25%, i.e., as few as 540 patients will be enrolled in the trial.

The following treatment effect scenarios were considered to evaluate operating characteristics of the adaptive design:

- Scenario 1: The common effect size is 0.4.
- Scenario 2: The effect size at Doses L and M is 0.3 and the effect size at Dose H is 0.4.

Summaries of the most important operating characteristics of the adaptive design are presented in Tables 1, 2 and 3. The adaptive design's performance is compared to that of three reference designs. These designs are set up as traditional two-arm designs with 180 enrolled patients per arm. Each design compares a single dose of the investigational treatment (Dose L, Dose M or Dose H) to placebo and applies the same futility stopping rule as in the adaptive design but does not employ treatment selection.

Table 1 presents the probabilities of dropping each treatment arm due to futility at the first interim analysis in the adaptive or any of the three traditional two-arm designs. The probability of dropping a treatment arm is quite low (about 13%) under Scenario 1, which is to be expected

since a strong treatment effect with the true effect size of 0.4 is assumed. Under Scenario 2, the treatment effect is weaker at Doses L and M and, as a direct consequence of that, these two doses are more likely to be removed at the interim look. The probability of dropping each of these doses is about 26%.

Table 2 focuses on the adaptive design and summarizes the characteristics of the treatment selection rule at the second interim analysis. The probabilities of selecting each dose as the best performing treatment are presented. Since the three doses are assumed to be equally effective under Scenario 1, each dose is equally likely to be selected and the three probabilities shown in Table 2 are essentially equal to 33.3%. When unequal effect sizes are considered (Scenario 2), Doses L and M are clearly less likely to chosen at the second interim analysis. The most promising dose (Dose H) will be identified as the best dose at this look over 50% of the time.

Finally, the results of power calculations are reported in Table 3. With the two-arm designs, the power values correspond to the probabilities of establishing a significant effect at each dose compared to placebo without a multiplicity adjustment. Table 3 shows that, under Scenario 1, each traditional design is adequately powered with the power values exceeding 80%. The adaptive design provides a considerable improvement over each two-arm design and power for the comparison of the best dose versus placebo is 90%. Under Scenario 2, the adaptive design performs much better than the two-arm designs corresponding to Doses L and M. Unlike these two designs, the adaptive design has an option to switch to Dose H with a larger effect size, which boosts the probability of success. When the two-arm design corresponding to Dose H is compared to the adaptive design, the former provides a small power advantage over the latter (81.2% versus 79.1%). This is due to the fact that the treatment selection rule employed in the adaptive design is not perfect and Doses L and M are occasionally selected for the final analysis, which leads to some power loss compared to this particular two-arm design.

References

Dmitrienko, A., D'Agostino, R.B. (2018). Multiplicity considerations in clinical trials. New England Journal of Medicine. 378, 2115-2122.

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

Table 1. Futility stopping in the traditional and adaptive designs

Treatment effect scenario	Parameter	Value
Scenario 1	Probability of dropping Dose L for futility	13.4%
	Probability of dropping Dose M for futility	13.6%
	Probability of dropping Dose H for futility	13.1%
Scenario 2	Probability of dropping Dose L for futility	26.0%
	Probability of dropping Dose M for futility	25.6%
	Probability of dropping Dose H for futility	13.6%

Scenario 1: The common effect size is 0.4. Scenario 2: The effect size at Doses L and M is 0.3 and the effect size at Dose H is 0.4.

Table 2. Treatment selection in the adaptive design

Treatment effect scenario	Parameter	Value
Scenario 1	Probability of selecting Dose L as the best performing treatment	33.6%
	Probability of selecting Dose M as the best performing treatment	32.9%
	Probability of selecting Dose H as the best performing treatment	32.8%
Scenario 2	Probability of selecting Dose L as the best performing treatment	19.6%
	Probability of selecting Dose M as the best performing treatment	20.4%
	Probability of selecting Dose H as the best performing treatment	54.7%

Scenario 1: The common effect size is 0.4. Scenario 2: The effect size at Doses L and M is 0.3 and the effect size at Dose H is 0.4.

Table 3. Power calculations in the traditional and adaptive designs

Treatment effect scenario	Parameter	Value
Scenario 1	Power for Dose L (traditional design)	81.5%
	Power for Dose M (traditional design)	81.2%
	Power for Dose H (traditional design)	81.3%
	Power for the best dose (adaptive design)	90.0%
Scenario 2	Power for Dose L (traditional design)	58.1%
	Power for Dose M (traditional design)	58.2%
	Power for Dose H (traditional design)	81.2%
	Power for the best dose (adaptive design)	79.1%

Scenario 1: The common effect size is 0.4. Scenario 2: The effect size at Doses L and M is 0.3 and the effect size at Dose H is 0.4.