

Adaptive designs with data-driven population selection

MedianaDesigner package

1. Introduction

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

2.1. Trial design

Adaptive designs with data-driven population selection are utilized in trials with two or more patient populations (these trials are known as multi-population trials). Consider a two-arm Phase III trial for evaluating the efficacy profile of an experimental treatment compared to a control, e.g., placebo. The primary efficacy endpoint in this trial could be a continuous, binary or time-to-event endpoint. The efficacy assessment will be performed in two populations, namely, the overall population of patients with the condition of interest as well as a prospectively specified subset of patients. The subset is defined using a binary classifier based on a baseline patient characteristic. Since biomarkers are most commonly used for this purpose, patients in the selected subset will be referred to as biomarker-positive patients and patients in the complementary subset will be referred to as biomarker-negative patients. Two interim analyses will be utilized in the trial. The first analysis will support a futility assessment and the second analysis will focus on identifying the best population or populations for the final analysis.

The unblinded interim analyses will support the following decision rules:

- Futility stopping rule at the first interim analysis: A futility assessment will be carried out to assess the efficacy profile in the overall population. The trial will be terminated for futility if the overall population effect is unlikely to be significant at the final analysis.
- Population selection rule at the second interim analysis: The most promising population or populations will be selected for evaluating the treatment effect at the final analysis.

It is important to note that the decision rules (futility stopping and population selection rules) are non-binding and could be overridden by the trial's sponsor or data monitoring committee. As in the adaptive design module with treatment selection (ADTreatSel function), the adaptive design considered in this module can be extended by incorporating an efficacy stopping rule at the first or second interim analysis (to stop the trial early due to strong evidence of effectiveness in the overall or biomarker-positive populations) or a sample size/event count re-estimation rule (to increase the target number of patients or events if the predicted probability of success in either patient population is lower than expected).

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

2.2. Futility stopping rule

As in the other adaptive design modules (ADSSMod and ADTreatSel functions), the futility stopping rule at the first interim analysis will be set up using conditional power, see, for example, Wassmer and Brannath (2016, Chapter 7). Conditional power will be computed in the overall patient population as the probability of a significant overall effect with respect to the primary efficacy endpoint at the final analysis conditional upon the interim data in the overall population.

Let CP denote the conditional power in the overall population evaluated at the first interim analysis (the derivation of conditional power is provided in the documentation for the ADSSMod function and will be omitted). The trial will be terminated due to futility at this interim look if the conditional power does not exceed a pre-defined threshold denoted by c , where $0 < c < 1$, i.e.,

$$CP \leq c.$$

As pointed out in the other adaptive design modules, the futility threshold is typically set to a fairly low value, e.g., this threshold rarely exceeds 0.3.

2.2. Population selection rule

The population selection rule to be applied at the second interim analysis will be aimed at choosing the most relevant populations for the final analysis. This decision rule will be defined using the influence and interaction conditions (see, for example, Millen et al., 2012). These conditions were introduced to facilitate the process of formulating meaningful regulatory claims in multi-population trials.

To define the decision rules based on the influence and interaction conditions, let $\hat{\theta}_+$ and $\hat{\theta}_-$ denote the interim estimates of the effect sizes in the populations of biomarker-positive and

biomarker-negative patients, respectively. As in the other adaptive design modules, the effect size for time-to-event endpoints is defined as the negative log-hazard ratio.

The influence and interaction conditions are applied sequentially. The influence condition states that the final analysis will be performed in the overall population (and potentially in the biomarker-positive population) if there is evidence of a meaningful treatment effect within the biomarker-negative population, which means that a beneficial treatment effect is not restricted to the biomarker-positive subset. The influence condition is satisfied if the interim effect size in the biomarker-negative population is greater than a clinically relevant threshold, i.e.,

$$\hat{\theta}_- \geq c_{inf},$$

where $c_{inf} \geq 0$ is known as the influence threshold. If the influence condition is not met, the overall effect is likely to be driven by a strong treatment effect in the biomarker-positive subset and thus the final analysis will be restricted to the biomarker-positive population. If the patient enrollment is not completed by the time the second interim analysis is conducted, only biomarker-positive patients will be enrolled after this interim look.

Secondly, if the influence condition is satisfied, the interaction condition is applied to determine if the final analysis will be performed in both populations or only in the overall population. This condition states that the final evaluation will include both patient populations if a differential treatment effect is present, i.e., the treatment benefit in the biomarker-positive population is much stronger than that in the biomarker-negative population. Mathematically, this means that

$$\hat{\theta}_+/\hat{\theta}_- \geq c_{int},$$

where $c_{int} > 1$ is known as the interaction threshold. If this condition is not met, the treatment effect is likely to be homogeneous across the biomarker-positive and biomarker-negative populations and only the overall effect will be examined at the final analysis. For more information on the selection of the influence and interaction thresholds, see Dmitrienko and Paux (2017).

2.3. Adaptive design methodology

As in the other adaptive design modules, an appropriate adjustment needs to be applied at the final analysis to account for the data-driven decision rule (population selection rule) at the second interim look. This adjustment guarantees that the Type I error rate in the trial is preserved at the nominal level, i.e., a one-sided $\alpha = 0.025$.

To define the adjustment, assume first that the influence and interaction conditions introduced in Section 2.2 are both met at the second interim analysis, which means that the treatment effect will be evaluated in the overall and biomarker-positive populations at the final analysis. Let p_0 and p_+ denote the one-sided treatment effect p-values in the overall and biomarker-positive

populations, respectively. Since there are two opportunities to claim that the trial's outcome is positive, a multiplicity adjustment will be used to control the Type I error rate. For example, if a multiplicity adjustment based on the Hochberg test is chosen, a significant effect will be established in the overall population if

$$\min(2\min(p_0, p_+), \max(p_0, p_+)) \leq \alpha \text{ and } p_0 \leq \alpha.$$

Similarly, a significant effect will be established in the biomarker-positive population if

$$\min(2\min(p_0, p_+), \max(p_0, p_+)) \leq \alpha \text{ and } p_+ \leq \alpha.$$

If the interim data at the second look support a decision to select a single population for the final analysis, only one of the two treatment effect p-values will be defined. In particular, if the influence condition is not met, the final analysis will focus on the biomarker-positive population, which means that p_0 will be undefined. This p-value will be set to 1 and thus the treatment effect in the biomarker-positive population will be significant if

$$p_+ \leq \alpha/2.$$

Also, if the influence condition is satisfied but the interaction condition is not satisfied at the second interim analysis, the final evaluation will be performed in the overall population. This implies that p_+ will be set to 1 and a significant effect will be concluded in the overall population if

$$p_0 \leq \alpha/2.$$

3. Case study

To illustrate the process of designing Phase III clinical trials with population selection rules, a multi-population Phase III oncology trial will be used. This trial will be conducted to investigate the effectiveness of an experimental treatment versus control, e.g., best supportive care. The efficacy profile of this treatment will be examined in the overall population of patients as well as a pre-defined subset of patients with a positive biomarker status. An enhanced treatment effect is anticipated within the biomarker-positive population compared to the complementary subset of biomarker-negative patients. The primary efficacy analysis is formulated in terms of overall survival.

The trial's sponsor would like to evaluate the potential benefits of an adaptive design with two interim analyses that support futility assessment and population selection, respectively, and compare it to a traditional design that focuses on the efficacy evaluation in the overall population.

The adaptive and traditional designs will be event-driven in the sense that the interim and final analyses will be performed after a pre-defined number of events. The adaptive design will utilize the following decision rules:

- A futility stopping rule based on conditional power will be applied at the first interim analysis. This analysis will be carried out after accruing 40% of the total number of events in the overall population. The futility threshold (c) is set to 20%. This value was derived based on the approach implemented in the futility module (FutRule function). An optimal value for the threshold in the overall population was chosen assuming the median survival times of 7.5 and 11 months in the control and treatment arms. With this futility threshold, the specificity rate was close to 80% and the sensitivity rate approached 90%.
- A population selection rule will be applied at the second interim analysis, which will be conducted after 60% of the total number of events in the overall population. This rule will be based on the influence and interaction conditions with the following thresholds:
 - Influence threshold $c_{inf} = 0$.
 - Interaction threshold $c_{int} = 1.3$.

The resulting influence condition will be fairly liberal in the sense that a decision to proceed with the overall population analysis will be supported unless a negative treatment effect is observed in the biomarker-negative population at the second interim analysis.

The interaction condition based on the selected threshold will conclude that a differential treatment effect is present and thus recommend that the treatment effect should be evaluated in both populations at the final analysis if the effect size in the biomarker-positive subset is greater than that in the complementary subset by more than 30%.

- A multiplicity adjustment at the final analysis will be based on the Hochberg test.

The traditional design will employ a single interim analysis, which will rely on the same futility stopping rule as in the adaptive design. As stated above, under the traditional design, only the overall effect will be evaluated at the final analysis.

An initial power calculation was performed in this trial under the assumption that the median survival times in the control and treatment arms are equal to 7.5 and 11 months, respectively. These median survival times correspond to the hazard ratio of 0.68. Assuming 90% power and no interim futility assessment, the final analysis should be performed after 290 events. The sample size in the trial was chosen using a 1:2 randomization ratio, namely, it is assumed that 140 patients will be assigned to the control arm and 280 patients will be assigned to the treatment arm.

To assess the performance of the proposed adaptive design, the total number of events in the overall population was set to 290, which means that the first and second interim looks will be taken after 114 and 174 events have been accrued. The target number of events within the

biomarker-positive population was set to 190. This event count is required in the adaptive design if the biomarker-positive population is the only population chosen for the final analysis. This target number of events results in 90% power under an optimistic assumption that the median survival times are equal to 7.5 and 12 months in the control and treatment arms, i.e., the hazard ratio is 0.63. The population prevalence of biomarker-positive patients was assumed to be 50% and the annual dropout rate was set to 5%.

Operating characteristics of the traditional and adaptive designs were computed under three treatment effect scenarios. Under all scenarios, a common median survival time of 7.5 months was considered in the control arm regardless of the biomarker status. The assumed median survival times in the treatment arm are summarized in Table 1. As shown in this table, a consistently strong treatment effect, which corresponds to the hazard ratio of 0.63, is assumed within the biomarker-positive population. By contrast, the treatment effect is getting weaker within the complementary population, i.e., the hazard ratio ranges between 0.75 under Scenario 1 and 0.94 under Scenario 3.

Table 1. Treatment effect assumptions

Treatment effect scenario	Parameter	Value
Scenario 1	Median survival time in the treatment arm (biomarker-negative subset)	10 months
	Median survival time in the treatment arm (biomarker-positive subset)	12 months
Scenario 2	Median survival time in the treatment arm (biomarker-negative subset)	9 months
	Median survival time in the treatment arm (biomarker-positive subset)	12 months
Scenario 3	Median survival time in the treatment arm (biomarker-negative subset)	8 months
	Median survival time in the treatment arm (biomarker-positive subset)	12 months

The key operating characteristics of the two trial designs are presented in Tables 2 and 3.

Table 2 provides an overall comparison of the traditional and adaptive designs and presents the common probability of stopping due to futility at the first interim look as well as the probability of success for each design. As expected, the probability of stopping for futility increases as the assumed hazard ratio in the biomarker-negative population approaches 1. A weaker efficacy signal in this population directly influences the overall treatment effect and the futility stopping rule is triggered more frequently under Scenario 3 compared to Scenario 1. Next, when comparing the power values for the two designs in Table 2, it is helpful to note that, if the adaptive design selects two populations for the final analysis, the probability of success is defined as the probability of establishing a significant treatment effect in either population. It follows from Table 2 that the two designs guarantee virtually the same level of power (about 80%) if a strong treatment effect is expected in biomarker-negative patients (Scenario 1). Under the other scenarios, the treatment effect within the biomarker-negative population is assumed to be weaker and the adaptive approach demonstrates a power advantage over the traditional design. This power gain is due to the flexible population selection rule employed by the adaptive design. If the experimental treatment appears to benefit only biomarker-positive patients, the adaptive design will most likely select this population for the final analysis. The ability to identify the best patient population in a data-driven manner results in a higher probability of success compared to the traditional approach that always evaluates the overall effect in the trial. Of course, both designs are underpowered under Scenarios 2 and 3 and the target number of events needs to be increased if a weaker efficacy signal is anticipated within the biomarker-negative population.

The flexible population selection rule used in the adaptive design is illustrated in Table 3. This table presents the probabilities of individual outcomes at the second interim analysis under the three treatment effect scenarios. Beginning with Scenario 1, the treatment effect is expected to be strong in the biomarker-positive population as well as its complement and, to improve the probability of success in the trial, the decision rule selects the overall population or both populations for the final evaluation most of the time. The probability of choosing only the biomarker-positive population is quite low under this scenario (about 20%). With a weaker treatment effect, the decision rule is more likely to identify the biomarker-positive population (with or without the overall population) as the best population to be examined at the final analysis. For example, under Scenario 3, the beneficial effect of the experimental treatment is limited to the subset of biomarker-positive patients and this population is chosen over 50% of the time and, in addition, both populations are selected about 30% of the time. In this case it will counter-intuitive to evaluate the treatment effect only in the overall population and, as expected, the population selection rule rarely recommends this population (only about 13% of the time).

References

Dmitrienko, A., Paux, G. (2017). Subgroup analysis in clinical trials. Clinical Trial Optimization Using R. Dmitrienko, A., Pulkstenis, E. (editors). Chapman and Hall/CRC Press, New York.

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

Millen, B., Dmitrienko, A., Ruberg, S., Shen, L. (2012). A statistical framework for decision making in confirmatory multi-population tailoring clinical trials. *Drug Information Journal*. 46, 647-656.

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

Table 2. 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	12.8%
	Traditional design: Power	79.8%
	Adaptive design: Power	79.2%
Scenario 2	Probability of stopping for futility at the first interim analysis	20.7%
	Traditional design: Power	66.8%
	Adaptive design: Power	68.7%
Scenario 3	Probability of stopping for futility at the first interim analysis	31.1%
	Traditional design: Power	49.1%
	Adaptive design: Power	58.4%

Table 3. Population selection in the adaptive design

Treatment effect scenario	Parameter	Value
Scenario 1	Probability of selecting the overall population only for the final analysis	39.9%
	Probability of selecting the biomarker-positive population only for the final analysis	19.9%
	Probability of selecting both populations for the final analysis	40.2%
Scenario 2	Probability of selecting the overall population only for the final analysis	25.6%
	Probability of selecting the biomarker-positive population only for the final analysis	34.1%
	Probability of selecting both populations for the final analysis	40.3%
Scenario 3	Probability of selecting the overall population only for the final analysis	12.8%
	Probability of selecting the biomarker-positive population only for the final analysis	57.3%
	Probability of selecting both populations for the final analysis	29.9 %