Published in final edited form as:

J Biopharm Stat. 2020 November 01; 30(6): 1060–1076. doi:10.1080/10543406.2020.1832110.

Design and Analysis of Biomarker-Integrated Clinical Trials with Adaptive Threshold Detection and Flexible Patient Enrichment

Ting Wang¹, Xiaofei Wang^{2,*}, Stephen L George², Haibo Zhou¹

¹Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, U.S.A.

²Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, U.S.A.

Summary:

We propose a new adaptive threshold detection and enrichment design in which the biomarker threshold is adaptively estimated and updated by optimizing a trade-off between the size of the biomarker positive population and the magnitude of the treatment effect in that population. Enrichment is based on an enrollment criterion that accounts for the uncertainty in estimation of the threshold. Early termination for futility is allowed based on predictive success probability. Valid testing and estimation techniques for the treatment effect overall and in patient subgroups are studied. Simulations and an example demonstrate advantages of the proposed design over existing designs.

Keywords

biomarker-integrated clinical trials; adaptive threshold estimation; continuous biomarker; multi-stage design; patient enrichment; predictive probability of success; testing and estimation of treatment effects

1. Introduction

To improve the efficiency of clinical trials in precision medicine, biomarker-integrated designs have been proposed to facilitate the evaluation of optimal treatments for patients with different biomarker levels (Mandrekar and Sargent, 2009; Freidlin et al., 2010; Wang et al., 2018, 2019). These designs assume that at baseline a binary biomarker or a continuous biomarker with a fixed cutoff dichotomizes patients into biomarker-positives and biomarker-negatives and patients are randomized to different treatments, stratified by the biomarker category. When biomarkers are continuous, however, there often exists no validated “optimal” thresholds at randomization. For example, in the JAVELIN Lung 200 study, Avelumab did not improve overall survival over Docetaxel in patients with platinum-treated PD-L1-positive non-small-cell lung cancer (NSCLC) (Barlesi et al., 2018). Exploratory analyses showed that Avelumab was associated with improved overall survival relative to

* xiaofei.wang@duke.edu.

Conflict of interest

The authors declare no potential conflict of interests.

Docetaxel in patients whose tumors had high levels of PD-L1 expression (e.g., at 50% and 80% cutoffs). However, the finding on the significant treatment effect in the patients with high PD-L1 levels is considered ad-hoc, as the original trial was not designed to simultaneously identify the optimal threshold, define the biomarker-positive patient subgroup and test and estimate the treatment effect of this patient subgroup. Thus, there exists a need for study designs that build a prospective procedure to co-determine the biomarker threshold and the treatment effect of the identified patient subpopulation during the trial itself. One early approach was the adaptive signature design (Freidlin and Simon, 2005; Freidlin et al., 2010). More recent designs combine adaptive threshold detection and patient enrichment (Simon and Simon, 2013; Lai et al., 2014; Ohwada and Morita, 2016; Simon and Simon, 2017; Zhang et al., 2017; Diao et al., 2018; Xu et al., 2020) in a batchwise fashion. In these designs, the target population is adaptively learned, the biomarker threshold is adaptively identified, and enrollment of future patients is restricted to the patients who are identified as “biomarker-positives”.

Simon and Simon (2013) described a design with adaptive enrichment for a single continuous biomarker and a binary outcome. They defined the biomarker-positive subgroup as a subgroup in which each patient benefits from the treatment, i.e., the treatment effect is positive. In the first block, the patient are enrolled without restriction and the biomarker threshold is searched sequentially in the possible range of the biomarker values, and from the second block onward, only the patients that satisfy the positive subgroup criterion are enrolled. However, their design assumes that there are a discrete set of candidate cutoff points, rather than allowing for more general estimation of the biomarker threshold. Simon and Simon (2017) further applied a Bayesian method to Simon and Simon (2013). The posterior probability that the treatment effect is positive is used as a measure of the potentiality of the biomarker and the optimal threshold of this probability is chosen to maximize a pre-defined utility – expected future patient outcome (EFPO) penalized by accrual time. This utility averages the better effect of treatment and control, both of which change with respect to the biomarker, among the whole range of the biomarker (i.e., all patients). At last, the treatment effect of “biomarker-positives” is tested based on the data from all enrolled patients. Diao et al. (2018) extended the idea of Simon and Simon (2013) to survival endpoints with a comparison of different criteria for determination of the biomarker threshold. Xu et al. (2020) proposed a similar design using a Bayesian model with the enrollment criteria adaptively updated, with interim quantitative decision-making rules for enrichment. They considered a partition of the patients according to the biomarker, which categorizes the biomarker instead of treating it as continuous. Xia et al. (2020) proposed a design which updates the biomarker signature as the trial proceeds, utilizes enrichment strategies for patient selection, and uses Bayesian response-adaptive randomization for treatment assignment. However, these designs described above have largely ignored investigation of the estimation accuracy of the biomarker threshold and the estimation of the treatment effect of the target patient population defined by the biomarker threshold. They also focus on testing the treatment effect in biomarker-positive patients, aggressively excluding biomarker-negative patients in early stage, thereby preventing accurate estimation of the biomarker threshold and the treatment effect in biomarker-negative patients. In many cases, estimation of treatment effects in biomarker-negatives and

in all patients are also of interest. The inclusion of a proportion of biomarker-negative patients with biomarker values close to the estimated cutoff is expected to increase the accuracy of the cutoff estimation.

Several papers considered the optimal threshold as the biomarker value that maximizes a utility involving the subgroup treatment effect and the subgroup size. Lai et al. (2014) defined the best subgroup by the biomarker threshold that maximizes the Kullback-Leibler information number, which is the same as maximizing the product of the squared root of the prevalence of the subgroup and the treatment effect in that subgroup under the assumption of equal variances of the treatment effect in all subgroups, and is also the same as maximizing the power of the test of the treatment effect in that subgroup. Zhang et al. (2017) proposed a utility equal to the product of the power of the test of the treatment effect in a subgroup and the prevalence of that subgroup. Defining the best subgroup based on such utility allows for a trade-off between the size of the subgroup and the treatment effect in that subgroup. For example, for a change-point model with a single biomarker, selecting a small subgroup with a higher average treatment effect may not be the best choice. Some patients with biomarker values below the change-point should also be considered. Equally important, a good utility should have a natural interpretation and be easily understood by clinicians. For example, the biomarker value that maximizes the utility of the product of the prevalence and the treatment effect of the target population simply corresponds to the intersection point of the biomarker-response curves of the two treatment groups. Therefore, one advantage of choosing the intersection point as the biomarker threshold is that the expected treatment effect will be positive for each biomarker-positive patient and will be negative for each biomarker-negative patient.

In this paper, we first describe the Biomarker Enrichment and Adaptive Threshold (BEAT) design, in which the estimated optimal biomarker threshold is updated in blocks by maximizing a utility that involves a balanced trade off between the size of the biomarker-positive population and the treatment effect in that population. This leads to a biomarker threshold that maximizes the difference in average treatment effect between treatment groups in the target population. It also coincides with the intersection point of the biomarker-response curves. We allow flexible patient enrichment by allowing enrollment strategies for biomarker-positive and biomarker-negative subgroups to depend on the estimation accuracy of the threshold. Patients with a positive biomarker defined by the updated biomarker threshold are all enrolled in the next block and patients with a negative biomarker are enrolled or excluded based on a flexible strategy that combines the estimated biomarker threshold, the estimated precision and the predictive probability of failure for such patients. Early termination for futility is also considered based on the predictive probability for biomarker-positive patients. The proposed design provides methods for biomarker threshold estimation and its precision, and the estimation of treatment effects of biomarker-positive patients, biomarker-negative patients and all patients. In comparison with the all-comers design, the proposed design reduces trial cost and time, especially when the prevalence rate of the biomarker-positives is low. In comparison with the existing adaptive threshold detection and enrichment designs such as Simon and Simon (2017), the proposed design has the following advantages: (1) in each block, we allow unbiased estimation for the biomarker threshold as well as its precision; (2) we build in early stopping of the entire trial for futility

based on predictive probability (Yin et al., 2012); (3) we allow flexible enrichment for biomarker-negative patients to reduce the false negatives due to the uncertainty of the threshold estimation in early blocks; and (4) we classify all enrolled patients as positives or negatives based on the estimated optimal threshold at the end of the trial, and provide valid hypothesis testing and estimation of treatment effects for all patients and all patient subgroups (positives and negatives). In a contrast, Simon and Simon (2017) test the treatment effect of all enrolled patients without re-classifying them using the estimated threshold at the end of the trial.

The remainder of this paper is organized as follows. In Section 2, we provide the details of the proposed design. In Section 3, we present the estimation methods for biomarker threshold and decision strategies for the proposed design. In Section 4, we present the methods for estimating and testing the treatment effects at the end of clinical trial. In Section 5, we present simulation results comparing the proposed design with some existing designs. In Section 6 we illustrate the proposed design using the JAVELIN Lung 200 study as an example. Concluding remarks are given in Section 7.

2. Design

2.1 Notation and assumptions

We make the following assumptions:

- i. Let $X_i \in [0, 1]$ denote a biomarker measured on patient i . The X_i values are assumed to be drawn independently from some probability distribution G . When X is a bounded continuous marker with range $[a, b]$, where $b > a > 0$, the proposed design is applicable when X can be transformed to the range $[0, 1]$ by taking its percentiles. Let Y_i be the outcome for patient i , where $Y_i = 1$ for responder and $Y_i = 0$ for non-responder, and Z_i be the treatment assignment, where $Z_i = 0$ for control C and $Z_i = 1$ for treatment T.
- ii. Assume that $p_T(x)$ and $p_C(x)$ are response probabilities of patients in treatment and control respectively at biomarker value x and follow a logistic regression model as below:

$$\text{logit}[p_z(x)] = \gamma_0 + \gamma_1 x + \gamma_2 z + \gamma_3 zx \quad (1)$$

where γ_1, γ_2 and γ_3 represent biomarker effect, treatment effect and interaction effect.

- iii. Assume that response rates in control and treatment, and the treatment effect are all positively related to the biomarker value, which is equivalent to $\gamma_1 > 0$ and $\gamma_3 > 0$ (thus we also have $\gamma_1 + \gamma_3 > 0$). If there exists an intersection x^* so that $p_T(x^*) = p_C(x^*)$, then $p_T(x) < p_C(x)$ for $x < x^*$ and $p_T(x) > p_C(x)$ for $x > x^*$. In section 3.1, we demonstrate that this intersection is the optimal threshold that maximizes a utility involving a trade-off between the prevalence of the

biomarker-positive patients and the treatment effect in the biomarker-positive patients.

2.2 Design procedures

We accrue patients sequentially in K blocks of size $2n_k$ patients with n_k patients assigned to control and n_k patients assigned to treatment in block k . The number of blocks K and the block sizes n_k are pre-specified. We observe the responses y_i for patients in the k th block before accruing patients for block $k + 1$. For accrual to the k th block we define an enrollment cutoff, $\tau_k \in [0, 1]$, and only enroll patients with $x > \tau_k$. We allow stopping for futility after a minimum of K^* blocks, where K^* is pre-specified. That is, if the estimated optimal biomarker threshold is close to 1 in or after the k^* th block, we stop for futility.

As visualized in Figure 1, the proposed design proceeds as follows:

- i. Pre-specify K , K^* and $n_1, \dots, n_k$. Set τ_1 , the enrollment cutoff of the first block, as 0.
- ii. For blocks $k = 1, \dots, K$ repeat:
 - a. Enroll $2n_k$ patients with enrollment cutoff τ_k and randomize to treatment and control.
 - b. Estimate the optimal biomarker threshold $\hat{c}_k$ and its estimated standard error $\widehat{sd}(\hat{c}_k)$ based on all available outcomes. If $k = K$, go to (iii). Otherwise, calculate new enrollment cutoff $\tau_{k+1} = \max\{0, \hat{c}_k - t \times \widehat{sd}(\hat{c}_k)\}$, where $t \geq 0$ is a user-specified parameter that is used to flexibly define the enrollment cutoff.
 - c. If $\tau_{k+1} > 0$, calculate $PPoF\{-\}$, the predictive probability of failure, the predictive probability that at the end of the trial, the treatment will be worse than control for the biomarker-negative patients whose biomarkers are smaller than the enrollment cutoff τ_{k+1} . See Section 3.2 for details. Update $\tau_{k+1} = 0$ if $PPoF\{-\} \leq \eta_1$ (i.e., control is not efficacious in the biomarker-negatives whose biomarkers are smaller than the enrollment cutoff and we can continue to enroll these patients for further investigation).
 - d. If $\hat{c}_k = 1$, when $k < K^*$, continue the trial, otherwise, stop. If $\hat{c}_k < 1$, calculate $PPoS\{+\}$, the predictive probability of success, the predictive probability that at the end of the trial, the treatment will be superior to control for biomarker-positive patients $[\hat{c}_k, 1]$. See Section 3.2 for details. Stop the trial for futility if $PPoS\{+\} < \eta_2$.
- iii. At the final analysis, based on the final optimal threshold $\hat{c}_K$ estimate, test and estimate the treatment effects in biomarker-positives, biomarker-negatives and overall patients.

For the parameter $t \geq 0$, a larger t means that we enroll more biomarker-negative patients with biomarker values less than the currently estimated biomarker threshold to facilitate the future estimation of the biomarker threshold and reduce the possible loss of biomarker-positive patients. $t = 0$ means we are very confident about the estimation of the biomarker threshold and use it as the enrollment cutoff.

3. Methods

3.1 Estimator of biomarker threshold and its asymptotic variance

The optimal biomarker threshold c_0 is the biomarker value of the intersection of treatment and control response curves if the intersection is within the range of the biomarker X , i.e., $[0, 1]$, and is the boundary value if the intersection is out of the range. Specifically,

$$c_0 = \min\{\max\{0, x^*\}, 1\} \quad (2)$$

where $x^* = -\gamma_2/\gamma_3$ is the biomarker value of the intersection of two biomarker-response curves.

Note that $x = c_0$ also maximizes $\int_x^1 (p_T(x) - p_C(x))d\pi(x) = \Pi(x)[\mu_T(x) - \mu_C(x)]$, where $\pi(x)$ is the density function of the biomarker x , $\Pi(x) = P[X > x]$ is the positive biomarker prevalence, and $\mu_T(x) = E[Y|X > x, Z = 1]$ and $\mu_C(x) = E[Y|X > x, Z = 0]$ are the average treatment effect in patients with $X > x$ receiving treatment and control respectively. Thus, the optimal biomarker threshold we consider involves a trade-off between the prevalence of the subgroup and the treatment effect in biomarker-positive patients.

We can calculate $\hat{x}^* = -\hat{\gamma}_2/\hat{\gamma}_3$ by fitting the logistic regression in equation (1), then by equation (2), we have $\hat{c}_0 = \min\{\max\{0, \hat{x}^*\}, 1\}$. We can use the delta method to estimate $\text{Var}(\hat{x}^*)$ using $\text{Cov}(\hat{\gamma})$ with $\gamma = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)'$ as in the equation (1), which can be obtained by fitting the logistic regression in (1).

Let $h(\gamma) = -\gamma_2/\gamma_3$, then $x^* = h(\gamma)$. Then,

$$\text{Var}(\hat{x}^*) = \text{Var}(h(\hat{\gamma})) = \nabla h(\gamma)^T \text{Cov}(\gamma) \nabla h(\gamma) \quad (3)$$

which can be estimated by $\nabla h(\hat{\gamma})^T \widehat{\text{Cov}}(\hat{\gamma}) \nabla h(\hat{\gamma})$, i.e.,

$$\widehat{\text{Var}}(\hat{x}^*) = \begin{pmatrix} 0 & 0 & -\frac{1}{\hat{\gamma}_3} & \frac{\hat{\gamma}_2}{\hat{\gamma}_3^2} \end{pmatrix} \widehat{\text{Cov}}(\hat{\gamma}) \begin{pmatrix} 0 \\ 0 \\ -\frac{1}{\hat{\gamma}_3} \\ \frac{\hat{\gamma}_2}{\hat{\gamma}_3^2} \end{pmatrix} = \frac{1}{\hat{\gamma}_3^2} \widehat{\text{Var}}(\hat{\gamma}_2) - \frac{2\hat{\gamma}_2}{\hat{\gamma}_3^3} \widehat{\text{Cov}}(\hat{\gamma}_2, \hat{\gamma}_3) + \frac{\hat{\gamma}_2^2}{\hat{\gamma}_3^4} \widehat{\text{Var}}(\hat{\gamma}_3)$$

Since $\hat{c}_0$ is the truncated value of $\hat{x}^*$, we approximate $\text{Var}(\hat{c}_0)$ using the idea of the truncated normal distribution (Barr and Sherrill, 1999) with $\hat{x}^*$ as the mean and $\text{Var}(\hat{x}^*)$ as the variance of the original normal distribution. Specifically, after each simulation, when calculating $\text{Var}(\hat{c}_0)$, we assume $\hat{x}^*$ is the mean of the intersection with a variance of $\text{Var}(\hat{x}^*)$. Thus, we can expect that the possible intersection will spread around $\hat{x}^*$. Since $c_0 = \min\{\max\{0, x^*\}, 1\}$, for all the possible intersection values smaller than 0, we should truncate it to 0 for estimating c_0 , and for all the possible intersection values larger than 1, we should truncate it to 1 for estimating c_0 . Let $\tilde{x} \sim N(\hat{x}^*, \text{Var}(\hat{x}^*))$, when $\hat{x}^* \in (0, 1)$, we have

$$\widehat{\text{Var}}(\hat{c}_0) = \int_{-\infty}^0 (0 - \hat{x}^*)^2 f(\tilde{x}) d\tilde{x} + \int_0^1 (\tilde{x} - \hat{x}^*)^2 f(\tilde{x}) d\tilde{x} + \int_1^{\infty} (1 - \hat{x}^*)^2 f(\tilde{x}) d\tilde{x}$$

At the end of k th block, the same procedure is followed to estimate the optimal biomarker threshold $\hat{c}_k$ and its estimated standard error $\widehat{sd}(\hat{c}_k)$.

3.2 Calculation of the Predictive Probability of Success / Failure

In the new design, we use predictive probability to determine futility in the biomarker-positives and check efficacy of control in the biomarker-negatives.

A general procedure to calculate the predictive probability for a two-arm randomized trial is as follows (Yin et al., 2012). Let p_1 and p_2 be the response rate of treatment 1 and 2, and assign p_s a prior distribution of $\text{Beta}(\alpha_s, \beta_s)$ for $s = 1, 2$. If among o_s subjects treated in arm s , we observe w_s responses, and the maximum sample size in arm s is O_s , then assuming Y_s is the number of responses in the future patients, the predictive probability that p_2 is superior to p_1 is:

$$\begin{aligned} PP(p_1, p_2, \delta) &= E_{Y_1, Y_2} [I\{P(p_2 - p_1 > \eta | W_s = w_s, Y_s, s = 1, 2) \geq \delta\}] \\ &= \sum_{y_1=0}^{O_1-o_1} \sum_{y_2=0}^{O_2-o_2} P(Y_1 = y_1 | W_1 = w_1) P(Y_2 = y_2 | W_2 = w_2) \\ &\quad \times I\{P(p_2 - p_1 > \eta | W_s = w_s, Y_s = y_s, s = 1, 2) \geq \delta\} \end{aligned} \quad (4)$$

where η is a clinically meaningful treatment difference, δ is a threshold probability, and $I\{\}$ is the indicator function.

In the proposed design, we set $\eta = 0$, $PPoS\{+\} = PP(p_C, p_T, \delta\{+\})$ on biomarker-positive patients and $PPoF\{-\} = PP(p_T, p_C, \delta\{-\})$ on the biomarker-negative patients whose biomarkers are smaller than the enrollment cutoff.

In the k th block of the proposed design, we enroll and randomize $2n_k$ patients, estimate optimal biomarker threshold $\hat{c}_k$ and its estimated standard error $\widehat{sd}(\hat{c}_k)$ based on all available outcomes, and calculate new enrollment cutoff $\tau_{k+1} = \max\{0, \hat{c}_k - t \times \widehat{sd}(\hat{c}_k)\}$.

If $\tau_{k+1} > 0$, we calculate $PPoF_{\{-\}}^{(k)}$ for the biomarker-negative patients whose biomarkers are smaller than the enrollment cutoff τ_{k+1} by following these steps:

i. We order the available enrollment cutoffs: $\{\tau_1, \dots, \tau_{k+1}\}$ together with 0 as an increasing unique order, $\{\delta_{(1)}^{(k-)}, \dots, \delta_{(J^{(k-)})}^{(k-)}\}$, in which “ $k-$ ” represents the negative patients in the k th block and $J^{(k-)}$ is the total number of unique enrollment cutoffs. Assume τ_{k+1} ranks $L^{(k-)}$.

ii. In $(\delta_{(j)}^{(k-)}, \delta_{(j+1)}^{(k-)}]$, we estimate response rate $\hat{p}_{i,j}^{(k-)}$, number of enrolled patients $n_{i,j}^{(k-)}$ and number of screened patients $m_{i,j}^{(k-)}$ for $i = T, C$.

iii. Thus, for biomarker-negative patients whose biomarkers are smaller than the enrollment cutoff, we estimate the average response rate as

$$\hat{p}_i^{(k-)} = \sum_{j=1}^{L^{(k-)}-1} \hat{p}_{i,j}^{(k-)} \frac{m_{i,j}^{(k-)}}{\sum_{j=1}^{L^{(k-)}-1} m_{i,j}^{(k-)}}, \text{ and the expected future}$$

$$\text{observation number as } O_i^{(k-)} - o_i^{(k-)} = (\sum_{k'=k+1}^K n_{k'}) \times \frac{\sum_{j=1}^{L^{(k-)}-1} m_{i,j}^{(k-)}}{\sum_{j=1}^{J^{(k-)}-1} m_{i,j}^{(k-)}}, \text{ and}$$

calculate the observed number $o_i^{(k-)} = \sum_{j=1}^{L^{(k-)}-1} n_{i,j}^{(k-)}$, and the response number $w_i^{(k-)} = o_i^{(k-)} \hat{p}_i^{(k-)}$ for $i = T, C$.

iv. Therefore, in the k th block,

$$\begin{aligned} PPoF_{\{-\}}^{(k)} &= PP(p_T^k, p_C^k, \delta_{\{-\}}) = \sum_{y_T=0}^{O_T^{(k-)} - o_T^{(k+)}} \sum_{y_C=0}^{O_C^{(k-)} - o_C^{(k-)}} P(Y_T = y_T | W_T \\ &= w_T^{(k-)}) \\ &P(Y_C = y_C | W_C = w_C^{(k-)}) I\{P(p_C > p_T | W_i = w_i^{(k-)}, Y_i = y_i, i = T, C) \geq \delta_{\{-\}}\} \end{aligned} \quad (5)$$

The calculation of $PPoS_{\{+\}}^{(k)}$ for biomarker-positive patients, $[\hat{c}_k, 1]$ is similar.

4. Estimation and Test of Treatment Effects

At the end of the trial, an immediate need is to estimate and test the treatment effects of various patient subgroups and overall patients. Simon and Simon (2017) test $H_0: p_T \leq p_C$ vs $H_A: p_T > p_C$ for all enrolled patients, where p_T and p_C are the average response rates of the treatment group and the control group. However, based on the final estimated threshold, these enrolled patients may include both biomarker-positive and biomarker-negative

patients. Therefore, at the end of the trial, we first classify all enrolled patients as biomarker-positives or biomarker-negatives based on the final estimated optimal threshold, and test and estimate the treatment effects for overall patients and all patient subgroups. Specifically, for overall patients, we test $H_0: p_T \leq p_C$ vs $H_A: p_T > p_C$; for biomarker-positive patients, we test $H_0^+: p_T^+ \leq p_C^+$ vs $H_A^+: p_T^+ > p_C^+$; but for biomarker-negative patients, for whom the treatment may not be better than control, usually, we are interested in whether the treatment is significantly worse than control. Therefore, here we test $H_0^-: p_T^- \geq p_C^-$ vs $H_A^-: p_T^- < p_C^-$. However, the direction can be reversed if the interest is whether the treatment is significantly better than control.

In our design, we have available enrollment cutoffs: $\{\tau_1, \dots, \tau_K\}$, which can be ordered together with $\hat{c}_K$ and 1 as an increasing unique order, $\{\delta_{(1)}, \dots, \delta_{(J)}\}$. Let L represent the rank of $\hat{c}_K$ in the order $\{\delta_{(1)}, \dots, \delta_{(J)}\}$. In each $(\delta_{(j)}, \delta_{(j+1)})$, we have response rate $\hat{p}_{i,j}$, number of enrolled patients $n_{i,j}$, number of screened patients $m_{i,j}$ for $i = T, C$. Thus, we can estimate the average response rate in biomarker-positives, biomarker-negatives and all patients for $i = T, C$ as

$$\hat{p}_i^+ = \sum_{j=L}^{J-1} p_{i,j} \frac{m_{i,j}}{\sum_{j=L}^{J-1} m_{i,j}}, \quad \hat{p}_i^- = \sum_{j=1}^{L-1} p_{i,j} \frac{m_{i,j}}{\sum_{j=1}^{L-1} m_{i,j}}, \quad \hat{p}_i = \sum_{j=1}^{J-1} p_{i,j} \frac{m_{i,j}}{\sum_{j=1}^{J-1} m_{i,j}}$$

Therefore, we can estimate treatment effect in biomarker-positives, biomarker-negatives and overall as $\hat{p}_T^+ - \hat{p}_C^+$, $\hat{p}_T^- - \hat{p}_C^-$ and $\hat{p}_T - \hat{p}_C$. To test the treatment effect in positive with

$$\text{hypotheses } H_0^+: p_T^+ \leq p_C^+ \text{ vs } H_A^+: p_T^+ > p_C^+, \text{ given } \text{Var}(\hat{p}_i^+) = \sum_{j=L}^{J-1} \frac{p_{i,j}(1-p_{i,j})}{n_{i,j}} \left(\frac{m_{i,j}}{\sum_{j=L}^{J-1} m_{i,j}} \right)^2$$

for $i = C, T$, the test statistic that follows a standard normal distribution under H_0^+ is

$$S^+ = (\hat{p}_T^+ - \hat{p}_C^+) / \sqrt{\text{Var}(\hat{p}_T^+) + \text{Var}(\hat{p}_C^+)}. \text{ The test procedures for the treatment effects in}$$

biomarker-negatives and overall are similar with hypotheses $H_0^-: p_T^- \geq p_C^-$ vs $H_A^-: p_T^- < p_C^-$ and $H_0: p_T \leq p_C$ vs $H_A: p_T > p_C$.

5. Simulation Studies

We compare the proposed design, BEAT, with the biomarker all-comers design and the design described in Simon and Simon (2017), referred to hereafter as the Simon design. The choice of the Simon design as a focus of performance evaluation of the proposed design is motivated by the fact that the Simon design is one of the most representative designs with features of adaptive threshold detection and patient enrichment and it is directly relevant to the proposed design. The biomarker all-comers design enrolls all eligible patients regardless of the biomarker value and randomizes them to treatment and control groups. After the trial, one can fit two biomarker-response curves based on the patients' response and biomarker and find the optimal biomarker threshold. The Simon design enrolls patients and identifies the biomarker threshold sequentially with enrichment after the first block by only enrolling the patients that satisfy the positive subgroup's criterion. The posterior probability that the treatment effect is positive is used as a measure of the positive potentiality of the biomarker and choose the optimal threshold for this probability, which maximizes the utility, defined as

the expected future patient outcome (EFPO) penalized by accrual time, as the criterion for the positive subgroup. As there is no estimation methods of treatment effects proposed for the Simon design, we apply our methods described in the previous section for testing and estimation of the treatment effects in these designs.

We show the results of the proposed design with different $t = 0, 0.5, 1$. The primary comparison measures include mean estimate and empirical standard error of the biomarker threshold, average screening number, average enrolled and “lost” number of biomarker-positive patients (i.e., biomarker-positive patients who are not enrolled), rejection rates and treatment effects coverage. We set block number $K = 6$ with enrolled number per arm $n_k = 100$ for $k = 1, \dots, K$. We set $\delta_{\{+\}} = 0.5$ for $PPoS_{\{+\}}$ with $PPoS_{\{+\}} < \eta_+ = 0.5$ considered as futility of treatment in the biomarker-positives, and set $\delta_{\{-\}} = 0.7$ for $PPoF_{\{-\}}$ with $PPoF_{\{-\}} > \eta_- = 0.7$ considered as the treatment is much worse than the control in the selected biomarker-negatives. We consider $c_0 = 0.3, 0.5$ and 0.7 , and for $c_0 = 0.7$, we show the results when two biomarker-response curves are distant and closer separately for comparison, as seen in Figure 2.

Table 1 shows the threshold estimation, futility rate and means of other important design operating parameters. From Table 1, empirical standard errors of the proposed design are always smaller than those of the Simon design, and when $c_0 = 0.7$, even smaller than the all-comers design. This advantage helps the proposed design perform better in several other aspects. Besides, the proposed design has much smaller screening numbers compared to the Simon design, significant enrichment of biomarker-positive patients compared to the all-comers design, and much fewer lost in biomarker-positive patients for enrollment compared to the Simon design.

Our enrolled biomarker-positives are a little fewer than the Simon design, but our results support that it does not affect much in our further testing and estimation of treatment effects. Tables 2, 3 and 4 show that the testing and estimation of different treatment effects (biomarker-positives, biomarker-negatives and overall). From Table 2, the rejection rates in biomarker-positives of the proposed design are always higher than the all-comers design and are similar to or even higher than the Simon design. Therefore, though we enroll a little fewer biomarker-positive patients than the Simon design, we can achieve higher power by estimating the biomarker threshold more accurately and losing fewer biomarker-positive patients. The estimate of the treatment effect in biomarker-positives of all designs are close to the true value while the proposed design and the Simon design have smaller standard errors due to enrichment of biomarker-positives. All coverage rates are close to the nominal level of 95%.

From Table 3, the rejection rates in biomarker-negatives of the proposed method are smaller than the all-comers design but in most cases are larger than the Simon design. Moreover, for the proposed design, the rejection rate in biomarker-negatives increases with respect to t , thus the proposed design with $t = 1$ has the largest rejection rate in biomarker-negatives among all enriched designs and is close to the rejection rate in biomarker-negatives of the all-comers design. The proposed design and the Simon design have larger standard errors of

the biomarker-negatives due to enrichment of biomarker-positives while the Simon design always has the largest since it enrolls much fewer biomarker-negatives, which also causes the lower coverage of the treatment effect in biomarker-negatives.

From Table 4, standard errors of the overall treatment effect have similar trend as the treatment effect in biomarker-negatives: the proposed design and the Simon design have larger standard errors while the Simon design always has the largest. All designs have good estimate and coverage of the overall treatment effect.

We also show the simulation results of a special case in Table 5 and Table 6 when the response rates of treatment and control are always the same, i.e. $p_T(x) = p_C(x)$ for all $x \in [0, 1]$, represented as 'na'. In this case, the null hypotheses, $H_0^+: p_T^+ \leq p_C^+$, $H_0^-: p_T^- \geq p_C^-$ and $H_0: p_T \leq p_C$ are all true. Type I error control of the proposed design is evaluated with this setting.

From Table 5, note that there is no true biomarker threshold. The proposed design with stop for futility has a very high futility rate, indicating the futility strategy works well. We also present the results of the proposed design without stopping for futility for comparison. The Simon design has a much larger screening number compared to the proposed design and the all-comers design.

In Table 6, since for $c_0=na$, only overall treatment effect is meaningful, we only show testing and estimation of the overall treatment effect. We can see the proposed design with no stop for futility and other design clearly control the type I errors for testing the overall treatment effect. For the proposed design with stop for futility, by considering the futility simulations as no rejection, we also have type I error controlled by the proposed design with stop for futility.

Finally, we investigate the hyper-parameters. In the proposed design, we have hyper-parameters for enrollment cutoff selection and futility decision, including t for the enrollment cutoff, δ_+ and η_+ for the *PPoS*, and δ_- and η_- for the *PPoF*. A reasonable t will help avoid wrongly excluding true biomarker-positive patients in the early enrollment of the trial. In previous simulations, we set $\delta_+ = 0.5$, $\eta_+ = 0.5$, $\delta_- = 0.7$ and $\eta_- = 0.7$. Table 7 shows the impact of the choice of hyper-parameters on the performance of the proposed design when $c_0 = 0.5$ and $t = 0.5$ with one of the hyper-parameters changed at a time.

From Table 7, we can see that the change of δ_+ and η_+ mainly affects the futility rate. Higher δ_+ or η_+ results in higher futility rate. We also observed that the change of δ_- and η_- mainly affects the rejection rate in biomarker-negative patients. Higher δ_- or η_- results in higher rejection rate in biomarker-negative patients. The change of these hyper-parameters does not affect other important outcomes, such as estimate of the biomarker threshold, standard error, estimate of the treatment effect, etc. For a typical biomarker-integrated clinical trial, reasonable initial values of these hyper-parameters could be $\delta_+ = 0.5$, $\eta_+ = 0.5$, $\delta_- = 0.7$ and $\eta_- = 0.7$. As the impact of these hyper-parameters on the design performance is significant,

optimal choice of hyper-parameters for a specific clinical trial should be carefully investigated through simulation.

6. Redesigning JAVELIN Lung 200 – A hypothetical example

We illustrate the proposed design using the JAVELIN Lung 200 data from Barlesi et al.. (Barlesi et al., 2018), which provides the enrolled numbers and response numbers of patients with biomarker (PD-L1) known in certain intervals (percentages of PD-L1) in treatment (Avelumab) and control (Docetaxel) groups separately, as in Table 8.

Based on the available information as Table 8, we assume the true biomarker-response relationship for treatment and control groups are:

$$\begin{aligned} \text{logit}(p_T(x)) &= -4 + 3x \\ \text{logit}(p_C(x)) &= -2.2 + 0x \quad (p_C(x) \equiv 10\%) \end{aligned}$$

Suppose that we require the BEAT design to have the following operating characteristics: (1) estimation of the optimal threshold with rejection of 95%CI half-width less than 0.20, i.e. the standard error of the estimated threshold is less than 0.1; (2) rejection of the null hypothesis for positives, $H_0^+: p_T^+ \leq p_C^+$ at 1-sided significance level of 0.05 and power of 90%; (3) power of at least 50% in testing the treatment effect for negatives in favor of the control; and (4) an early stopping probability for futility of less than 10% under the alternative hypothesis.

The 4-block design with $t = 0.5$ or 1, enrolled number per arm $n_1 = n_2 = 111$ and $n_3 = n_4 = 110$ and hyperparameters $\delta\{+\} = 0.5$, $n_+ = 0.4$, $\delta\{-\} = \eta_- = 0.7$ has all of the above operating characteristics. The performance of the biomarker all-comers design and the Simon design are compared with the BEAT design for the same sample size. Table 9 shows the threshold estimation, futility rate and mean of important numbers of different design and Table 10 shows testing and estimation of different treatment effects.

From Table 9, both the proposed design and the Simon design heavily enrich the biomarker-positive patients compared to the all-comers design. However, the proposed design has smaller standard error (< 0.1) of the estimated biomarker threshold and much smaller screening number with much fewer biomarker-positive patients lost compared to the Simon design.

From Table 10, it clearly shows that all enriched designs increase the power of treatment effect in biomarker-positives from the all-comers design's 69% to around 90%. Moreover, the proposed design with $t = 0.5$ or 1 has larger power for testing the treatment effect in biomarker-negatives compared to the Simon design. Since two response curves cross, for testing the overall treatment effect, we have a very small power. All designs have good estimate and coverage of the treatment effects.

7. Concluding Remarks

We propose a new adaptive threshold detection and enrichment design, Biomarker Enrichment and Adaptive Threshold (BEAT) design, to sequentially enrich the biomarker-positive patients with an adaptively updated estimate of the biomarker threshold. The proposed design has several important advantages: unbiased estimation for the biomarker threshold, flexible enrichment for biomarker-negative patients to reduce the false negatives due to the uncertainty of the threshold estimation in early blocks, and valid hypothesis testing and estimation of treatment effects for all patients and all patient subgroups which are defined by the “chosen” optimal threshold at the end of the trial.

Compared to the existing designs that have no clear definition of the optimal biomarker threshold, we define the optimal biomarker threshold as the biomarker value that maximizes the product of the prevalence of the biomarker-positive population and the treatment effect in that population, which is equal to the biomarker value corresponding to the intersection of two biomarker-response curves if the intersection is within the range of the biomarker. We further provide an efficient and unbiased estimator of the optimal biomarker threshold, and importantly, provide its estimation accuracy. Most existing designs aggressively enrich the biomarker-positive patients with a large screening number and result in a large number of lost biomarker-positive patients. In contrast, the BEAT design enriches the patients with a flexible and efficient enrichment strategy, in which the enrollment threshold depends on the estimation accuracy of the biomarker threshold. The parameter $t \geq 0$ is user-specified to flexibly define the enrollment cutoff, with a smaller value (e.g., $t = 0.5$) indicating more aggressive screening (i.e., more patients excluded) and a larger value (e.g., $t = 1$) indicating more conservative screening (i.e., fewer patients excluded). As a result, the new design requires screening fewer patients and loses fewer biomarker-positives. Finally, we provide valid hypothesis testing and estimation of treatment effects for all patients and all patient subgroups, defined by the “chosen” optimal threshold, in the patient population at the end of the trial. Simon and Simon (2017) proposed the inverse normal combination test statistic to test the treatment effect among the “biomarker-positives” that has often been used in adaptive trial designs (Bauer et al., 2016). The test allows type I error control under the null hypothesis that there is no treatment effect for the “biomarker-positives” defined by different cutoff points, which is a strong assumption. The power of this test is not properly defined as there is no single treatment effect for different “biomarker-positives” and the power is not defined and therefore calculated at the time of study design. Our simulations show the new design has smaller standard error for the estimation of treatment effects than the Simon design if the same estimation method is applied to the data collected in the Simon design. We also employ predictive probability for futility of treatment in biomarker-positive patients and efficacy of control in selected biomarker-negative patients. The performance of the new design could be significantly influenced by the choice of hyper-parameters and initial values that can be used at trial design are given in Section 5. We have assumed that the biomarker-positives are those with larger biomarker values. Our approach can be easily modified to fit the scenario that biomarker-positives have smaller biomarker values. Although our design is especially relevant when the control therapy is an “active control”, it is applicable to any case where the new treatment may perform worse than the control therapy for some

subgroup of patients. For example, adding a new cancer drug to an effective control treatment might add toxicity with no efficacy benefit, thereby causing some patients to not receive a full course of the effective control therapy, resulting in reduced effectiveness for the new treatment. An R package is available for free download to facilitate the use of the proposed design.

In this paper, we considered a binary outcome and logistic regression is used to model the relation between the outcome with biomarker and treatment. Based on specific needs, we can extend logistic model to flexible and possible more complicated, such as spline regression models. In the simulation, we pre-specified the number of blocks K and the block sizes n_k to provide reasonable accuracy for estimating the threshold. How to choose optimal K and n_k for a specific trial requires further investigation. The idea of the BEAT design can be extended to biomarker-integrated trial with continuous or time-to-event outcomes. The extension to the case of continuous outcome is straightforward, substituting a linear regression model for logistic regression to characterize the relationship between the outcome, biomarker and treatment. Although conceptually straightforward, the extension to the case of time-to-event outcome will require significant additional work. Parametric and semiparametric survival models can be used to estimate the relationship between survival time, biomarker and treatment, which then can be used to estimate the biomarker threshold and the enrollment criteria for patient enrichment. Analogous to the difference of response rates between treatment arms, the difference in restricted mean survival time between treatments as a function of biomarker value can be used to estimate the optimal biomarker threshold. Any time-to-event outcome such as overall survival may take more time to observe compared to a binary or continuous outcome. While the trial is ongoing, an intermediate survival outcome, such as progression-free survival, can be used to estimate the intersection points which define the biomarker-positive patients for the purpose of patient enrichment.

Additional work can also be done for the situation when multiple biomarkers or even high-dimensional biomarkers are implicated in treatment selection. A functional of these biomarkers and its interaction with treatment can be estimated through a model related them to the outcome. The functional and model parameters could be estimated with accumulatively observed data. Subsequently the estimation of the biomarker hyperplane and the definition of biomarker-positive patients can be updated adaptively. Innovative methods for estimating the biomarker threshold and optimizing the enrichment strategies in those settings can be further studied. In practice, the goal of a biomarker-integrated trial is to determine the optimal biomarker threshold defining biomarker-positive versus biomarker-negative patients, and test multiple hypotheses, for example, the treatment effect of the biomarker-positives and the overall treatment of all patients. An optimal enrichment strategy, specially, the enrollment probability for a coming patient with specific biomarker value, will be optimized to achieving adequate power for testing each hypothesis with minimum overall sample size and study cost or time.

Acknowledgments

The research work was partly supported by National Cancer Institute grant P01CA142538 (Ting Wang, Xiaofei Wang and Haibo Zhou), National Institute of Environmental Health Sciences grant P30ES010126 (Ting Wang and Haibo Zhou), and National Institute on Aging grant R01AG066883 (Xiaofei Wang).

References

- Barlesi F, Vansteenkiste J, Spigel D, Ishii H, Garassino M, de Marinis F, Özgüro lu M, Szczesna A, Polychronis A, Uslu R, et al. (2018). Avelumab versus docetaxel in patients with platinum-treated advanced non-small-cell lung cancer (JAVELIN Lung 200): an open-label, randomised, phase 3 study. *The Lancet Oncology* 19, 1468–1479. [PubMed: 30262187]
- Barr DR and Sherrill ET (1999). Mean and variance of truncated normal distributions. *The American Statistician* 53, 357–361.
- Bauer P, Bretz F, Dragalin V, König F, and Wassmer G (2016). Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Statistics in Medicine* 35, 325–347. [PubMed: 25778935]
- Diao G, Dong J, Zeng D, Ke C, Rong A, and Ibrahim JG (2018). Biomarker threshold adaptive designs for survival endpoints. *Journal of Biopharmaceutical Statistics* 28, 1038–1054. [PubMed: 29436940]
- Freidlin B, Jiang W, and Simon R (2010). The cross-validated adaptive signature design. *Clinical Cancer Research* 16, 691–698. [PubMed: 20068112]
- Freidlin B, McShane LM, and Korn EL (2010). Randomized clinical trials with biomarkers: Design issues. *Journal of the National Cancer Institute* 102, 152–160. [PubMed: 20075367]
- Freidlin B and Simon R (2005). Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research* 11, 7872–7878. [PubMed: 16278411]
- Lai TL, Lavori PW, and Liao OY-W (2014). Adaptive choice of patient subgroup for comparing two treatments. *Contemporary Clinical Trials* 39, 191–200. [PubMed: 25205644]
- Mandrekar SJ and Sargent DJ (2009). Clinical trial designs for predictive biomarker validation: Theoretical considerations and practical challenges. *Journal of Clinical Oncology* 27, 4027–4034. [PubMed: 19597023]
- Ohwada S and Morita S (2016). Bayesian adaptive patient enrollment restriction to identify a sensitive subpopulation using a continuous biomarker in a randomized phase 2 trial. *Pharmaceutical Statistics* 15, 420–429. [PubMed: 27485377]
- Simon N and Simon R (2013). Adaptive enrichment designs for clinical trials. *Biostatistics* 14, 613–625. [PubMed: 23525452]
- Simon N and Simon R (2017). Using Bayesian modeling in frequentist adaptive enrichment designs. *Biostatistics* 19, 27–41.
- Wang X, Cai J, and George SL (2019). Enriched biomarker-driven trials. *Wiley Stats Ref: Statistics Reference Online* 10.1002/9781118445112.stat08153.
- Wang X, Zhou J, Wang T, and George SL (2018). On enrichment strategies for biomarker stratified clinical trials. *Journal of Biopharmaceutical Statistics* 28, 292–308. [PubMed: 28933670]
- Xia F, George SL, Ning J, Li L, and Huang X (2020). A signature enrichment design with Bayesian adaptive randomization. *Journal of Applied Statistics* 10.1080/02664763.2020.1757048.
- Xu Y, Constantine F, Yuan Y, and Pritchett YL (2020). Asied: a Bayesian adaptive subgroup-identification enrichment design. *Journal of Biopharmaceutical Statistics* 30, 623–638. [PubMed: 31782938]
- Yin G, Chen N, and Jack Lee J (2012). Phase II trial design with Bayesian adaptive randomization and predictive probability. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 61, 219–235.
- Zhang Z, Li M, Lin M, Soon G, Greene T, and Shen C (2017). Subgroup selection in adaptive signature designs of confirmatory clinical trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 66, 345–361.

Figure 1:
Flow chart of the Biomarker Enrichment and Adaptive Threshold (BEAT) design.

Figure 2:
Biomarker-response curves of treatment and control (a) when $c_0 = 0.3$, (b) when $c_0 = 0.5$, (c) when $c_0 = 0.7$ and two biomarker-response curves are distant, (d) when $c_0 = 0.7$ and two biomarker-response curves are closer compared to (c).

Table 1:

Threshold estimation, futility rate and other operating characteristics

True c_0	Design	Threshold Est.		% of	Screened	Enrolled	Excluded
		Mean	$SD/\widehat{SD}$	Futility	Patients	Positives	Positives
0.3	All-comers	0.283	0.130 / 0.136	0.000	1200 / 1200	863	0
	Simon	0.310	0.157 / <i>na</i>	0.000	1902 / 1902	1084	178
	BEAT _{$\tau=0$}	0.287	0.140 / 0.143	0.116	1473 / 1373	995	40
	BEAT _{$\tau=0.5$}	0.292	0.145 / 0.138	0.124	1453 / 1320	982	18
	BEAT _{$\tau=1$}	0.295	0.139 / 0.136	0.110	1388 / 1272	950	7
0.5	All-comers	0.497	0.064 / 0.065	0.000	1200 / 1200	600	0
	Simon	0.511	0.087 / <i>na</i>	0.000	2547 / 2547	1039	173
	BEAT _{$\tau=0$}	0.495	0.077 / 0.078	0.155	2064 / 1870	979	44
	BEAT _{$\tau=0.5$}	0.502	0.080 / 0.078	0.131	1998 / 1742	948	16
	BEAT _{$\tau=1$}	0.501	0.075 / 0.073	0.128	1851 / 1644	902	3
0.7	All-comers	0.709	0.082 / 0.078	0.000	1200 / 1200	353	0
	Simon	0.692	0.070 / <i>na</i>	0.000	4409 / 4409	970	296
	BEAT _{$\tau=0$}	0.701	0.063 / 0.067	0.378	3282 / 2649	942	76
	BEAT _{$\tau=0.5$}	0.703	0.062 / 0.065	0.323	3087 / 2017	877	28
	BEAT _{$\tau=1$}	0.701	0.061 / 0.062	0.364	2784 / 1867	800	3
0.7	All-comers	0.710	0.152 / 0.164	0.000	1200 / 1200	348	0
(closer)	Simon	0.689	0.127 / <i>na</i>	0.000	4756 / 4756	905	409
	BEAT _{$\tau=0$}	0.689	0.111 / 0.113	0.587	3159 / 2498	857	76
	BEAT _{$\tau=0.5$}	0.697	0.117 / 0.113	0.600	2965 / 1980	770	26
	BEAT _{$\tau=1$}	0.688	0.108 / 0.109	0.621	2305 / 1556	673	2

¹The Simon design does not provide a method to estimate the standard error of the biomarker threshold.

²For futility simulations, we only record its futility index and screening number.

³For screening number, the first one refers to the average screening number of simulations without futility, the second one refers to the average screening number of all simulations, including futility simulations.

Table 2:

Testing and estimation of treatment effect in biomarker-positives

True c_0	Design	Rej % of Pos	Trt Effect of Pos		Cover % for Pos effect ($\hat{c}_0/c_0$)
			Mean	$SD / \widehat{SD}$	
0.3	True		0.111		
	All-comers	0.962	0.111	0.030 / 0.034	0.956 / 0.969
	Simon	0.966	0.115	0.031 / 0.032	0.949 / 0.954
	BEAT _{$\varepsilon=0$}	0.977	0.112	0.028 / 0.032	0.940 / 0.975
	BEAT _{$\varepsilon=0.5$}	0.978	0.115	0.030 / 0.032	0.940 / 0.966
	BEAT _{$\varepsilon=1$}	0.983	0.116	0.029 / 0.032	0.951 / 0.960
0.5	True		0.114		
	All-comers	0.920	0.114	0.032 / 0.040	0.945 / 0.978
	Simon	0.987	0.118	0.028 / 0.032	0.950 / 0.966
	BEAT _{$\varepsilon=0$}	0.977	0.114	0.027 / 0.032	0.934 / 0.969
	BEAT _{$\varepsilon=0.5$}	0.989	0.117	0.025 / 0.032	0.956 / 0.979
	BEAT _{$\varepsilon=1$}	0.994	0.119	0.025 / 0.033	0.961 / 0.980
0.7	True		0.070		
	All-comers	0.396	0.071	0.044 / 0.054	0.952 / 0.986
	Simon	0.734	0.074	0.029 / 0.034	0.942 / 0.980
	BEAT _{$\varepsilon=0$}	0.782	0.072	0.024 / 0.033	0.940 / 0.984
	BEAT _{$\varepsilon=0.5$}	0.803	0.075	0.024 / 0.033	0.952 / 0.994
	BEAT _{$\varepsilon=1$}	0.800	0.077	0.025 / 0.035	0.956 / 0.993
0.7 (closer)	True		0.036		
	All-comers	0.170	0.043	0.052 / 0.058	0.949 / 0.986
	Simon	0.302	0.043	0.030 / 0.037	0.958 / 0.983
	BEAT _{$\varepsilon=0$}	0.408	0.051	0.022 / 0.035	0.934 / 0.982
	BEAT _{$\varepsilon=0.5$}	0.352	0.048	0.025 / 0.037	0.949 / 0.984
	BEAT _{$\varepsilon=1$}	0.371	0.053	0.026 / 0.039	0.947 / 0.986

¹. Cover % is the rate for the 95% confidence intervals of the estimated treatment effect in biomarker-positives covering the true treatment effect in biomarker-positives based on estimated threshold $\hat{c}_0$ and true threshold c_0 separately.

Table 3:

Testing and estimation of treatment effect in biomarker-negatives

True c_0	Design	Rej % of Neg	Trt Effect of Neg		Cover % for
			Mean	$SD/\widehat{SD}$	Neg effect ($\hat{c}_0/c_0$)
0.3	True		-0.037		
	All-comers	0.194	-0.039	0.048 / 0.052	0.964 / 0.984
	Simon	0.171	-0.070	0.132 / 0.084	0.867 / 0.884
	BEAT _{$\varepsilon=0$}	0.145	-0.057	0.062 / 0.067	0.956 / 0.977
	BEAT _{$\varepsilon=0.5$}	0.179	-0.064	0.066 / 0.067	0.954 / 0.973
	BEAT _{$\varepsilon=1$}	0.236	-0.066	0.068 / 0.064	0.938 / 0.964
0.5	True		-0.106		
	All-comers	0.894	-0.106	0.032 / 0.039	0.953 / 0.988
	Simon	0.401	-0.121	0.083 / 0.084	0.941 / 0.971
	BEAT _{$\varepsilon=0$}	0.418	-0.111	0.048 / 0.073	0.972 / 0.992
	BEAT _{$\varepsilon=0.5$}	0.512	-0.123	0.051 / 0.071	0.969 / 0.984
	BEAT _{$\varepsilon=1$}	0.653	-0.128	0.052 / 0.068	0.953 / 0.976
0.7	True		-0.144		
	All-comers	0.998	-0.145	0.030 / 0.033	0.930 / 0.969
	Simon	0.641	-0.145	0.063 / 0.074	0.953 / 0.972
	BEAT _{$\varepsilon=0$}	0.729	-0.154	0.056 / 0.071	0.958 / 0.970
	BEAT _{$\varepsilon=0.5$}	0.812	-0.156	0.052 / 0.069	0.969 / 0.985
	BEAT _{$\varepsilon=1$}	0.867	-0.159	0.053 / 0.065	0.956 / 0.972
0.7 (closer)	True		-0.079		
	All-comers	0.796	-0.081	0.030 / 0.034	0.946 / 0.979
	Simon	0.345	-0.095	0.053 / 0.069	0.966 / 0.977
	BEAT _{$\varepsilon=0$}	0.419	-0.105	0.049 / 0.064	0.946 / 0.977
	BEAT _{$\varepsilon=0.5$}	0.521	-0.111	0.049 / 0.062	0.955 / 0.968
	BEAT _{$\varepsilon=1$}	0.622	-0.107	0.043 / 0.057	0.969 / 0.983

¹. Cover % is the rate for the 95% confidence intervals of the estimated treatment effect in biomarker-negatives covering the true treatment effect in biomarker-negatives based on estimated threshold $\hat{c}_0$ and true threshold c_0 separately.

Table 4:

Testing and estimation of overall treatment effect

True ϵ_0	Design	Rej %	Overall Trt Effect		Cover %
			Mean	SD / $\widehat{SD}$	
0.3	True		0.066		
	All-comers	0.748	0.066	0.029 / 0.028	0.954
	Simon	0.494	0.054	0.040 / 0.036	0.938
	BEAT _{$\epsilon=0$}	0.659	0.062	0.033 / 0.030	0.960
	BEAT _{$\epsilon=0.5$}	0.664	0.060	0.035 / 0.030	0.944
	BEAT _{$\epsilon=1$}	0.644	0.059	0.033 / 0.030	0.943
0.5	True		0.003		
	All-comers	0.064	0.004	0.028 / 0.029	0.946
	Simon	0.037	-0.007	0.046 / 0.046	0.948
	BEAT _{$\epsilon=0$}	0.025	0.002	0.035 / 0.038	0.972
	BEAT _{$\epsilon=0.5$}	0.050	-0.005	0.038 / 0.039	0.967
	BEAT _{$\epsilon=1$}	0.045	-0.005	0.038 / 0.038	0.958
0.7	True		-0.080		
	All-comers	0.000	-0.080	0.029 / 0.029	0.951
	Simon	0.000	-0.081	0.052 / 0.052	0.946
	BEAT _{$\epsilon=0$}	0.000	-0.085	0.046 / 0.051	0.962
	BEAT _{$\epsilon=0.5$}	0.000	-0.084	0.044 / 0.049	0.968
	BEAT _{$\epsilon=1$}	0.000	-0.088	0.044 / 0.047	0.951
0.7 (closer)	True		-0.044		
	All-comers	0.000	-0.044	0.030 / 0.029	0.937
	Simon	0.000	-0.054	0.043 / 0.050	0.966
	BEAT _{$\epsilon=0$}	0.005	-0.055	0.045 / 0.045	0.951
	BEAT _{$\epsilon=0.5$}	0.002	-0.064	0.045 / 0.045	0.953
	BEAT _{$\epsilon=1$}	0.004	-0.058	0.038 / 0.041	0.972

¹. Cover % refers to the rate for the 95% confidence intervals of the estimated overall treatment effect covering the true overall treatment effect.

Table 5:

Threshold estimation, futility rate and other operating characteristics for the special case

True c_0	Design	Threshold Est.		% of	Screened
		Mean	$SD / \widehat{SD}$	Futility	Patients
<i>na</i>	All-comers	0.507	0.340 / 0.404	0.000	1200 / 1200
	Simon	0.274	0.343 / <i>na</i>	0.000	4020 / 4020
	BEAT _{$t=0$}	0.469	0.302 / 0.307	0.811	1927 / 1305
	BEAT _{$t=0.5$}	0.506	0.323 / 0.312	0.819	1983 / 1390
	BEAT _{$t=1$}	0.456	0.301 / 0.312	0.808	1518 / 1110
	BEAT _{$t=0$} *	0.531	0.328 / 0.369	0.000	2035 / 2035
	BEAT _{$t=0.5$} *	0.570	0.342 / 0.359	0.000	1917 / 1917
	BEAT _{$t=1$} *	0.540	0.339 / 0.368	0.000	1499 / 1499

¹The Simon design does not provide a method to estimate the standard error of the biomarker threshold.

²For futility simulations, we only record its futility index and screening number.

³For screening number, the first one refers to the average screening number of simulations without futility, the second one refers to the average screening number of all simulations, including futility simulations.

⁴The results with * are from the proposed design with no stop for futility.

Table 6:

Testing and estimation of treatment effects for a special case

True ϵ_0	Design	Rej %	Rej %	Rej %	Overall Trt Effect		Cover %
(True x^*)		of Pos	of Neg	overall	Mean	SD / $\widehat{SD}$	
<i>na</i>	True				0		
	All-comers	0.062	0.048	0.048	0.000	0.028 / 0.029	0.951
	Simon	0.097	0.178	0.060	-0.003	0.037 / 0.035	0.940
	BEAT _{$t=0$}	0.189	0.109	0.105	-0.004	0.043 / 0.034	0.898
	BEAT _{$t=0.5$}	0.158	0.165	0.079	-0.011	0.049 / 0.035	0.912
	BEAT _{$t=1$}	0.179	0.163	0.108	-0.001	0.041 / 0.032	0.920
	BEAT _{$t=0$} *	0.061	0.050	0.052	-0.002	0.035 / 0.031	0.950
	BEAT _{$t=0.5$}	0.064	0.119	0.045	-0.014	0.043 / 0.033	0.932
	BEAT _{$t=1$} *	0.056	0.126	0.045	-0.011	0.040 / 0.031	0.924

¹ Cover % refers to the rate for the 95% confidence intervals of the estimated overall treatment effect covering the true overall treatment effect.

Table 7:Impact of hyper-parameters on the performance of the proposed design when $c_0 = 0.5$ and $t = 0.5$

Cases	Changed	Threshold Est.		% of	Screened	Enrolled/Excluded	Rej Rate
	Value	Mean	$SD / \widehat{SD}$	Futility	Patients	Positives	Negatives
Original		0.502	0.080 / 0.078	0.155	1998 / 1742	948/16	0.512
δ_+	0.2	0.501	0.075 / 0.076	0.082	1977 / 1839	946/15	0.525
	0.8	0.509	0.078 / 0.075	0.216	2003 / 1645	950/11	0.566
η_+	0.2	0.502	0.076 / 0.080	0.025	1950 / 1906	933/17	0.535
	0.8	0.502	0.076 / 0.076	0.418	1987 / 1305	953/11	0.555
δ_-	0.4	0.499	0.075 / 0.078	0.152	1967 / 1434	942/14	0.410
	0.9	0.498	0.082 / 0.078	0.127	1889 / 1773	916/12	0.708
η_-	0.4	0.501	0.077 / 0.079	0.137	2007 / 1782	964/17	0.385
	0.9	0.505	0.074 / 0.072	0.153	1888 / 1663	902/11	0.760

Table 8:

Response rates reported by JAVELIN Lung 200

PD-L1 %	Avelumab	Docetaxel
	Responses / Enrolled (Rate)	Responses / Enrolled (Rate)
< 1%	9 / 132 (6.8%)	13 / 131 (9.9%)
[1%, 50%)	8 / 96 (8.3%)	16 / 118 (13.6%)
[50%, 80 %)	5 / 48 (10.4%)	4 / 41 (9.8%)
≥ 80%	37 / 120 (30.8%)	11 / 106 (10.4%)

¹. Avelumab is the experimental treatment and Docetaxel is the standard treatment for control.

Table 9:

Threshold estimation, futility rate and other operating characteristics of the hypothetical example

Design	Threshold Est.		% of	Screened	Enrolled	Excluded
	Mean	$SD / \widehat{SD}$	Futility	Patients	Positives	Positives
True	0.600					
All-comers	0.598	0.095 / 0.092	0.000	884 / 884	356	0
Simon	0.595	0.117 / <i>na</i>	0.000	2474 / 2474	712	230
BEAT _{$t=0$}	0.597	0.091 / 0.092	0.131	1859 / 1846	650	58
BEAT _{$t=0.5$}	0.602	0.093 / 0.091	0.102	1688 / 1562	618	22
BEAT _{$t=1$}	0.603	0.088 / 0.086	0.092	1508 / 1402	570	4

¹The Simon design does not provide a method to estimate the standard error of the biomarker threshold estimate.

²For futility simulations, we only record its futility index and screening number.

³For screened patients, the first one refers to the average screening number of simulations without futility, the second one refers to the average screening number of all simulations, including futility simulations.

Table 10:

Testing and estimation of treatment effects of the hypothetical example

Treatment effect	Design	Rej %	Mean	$SD / \widehat{SD}$	Cover %
Positives	True		0.074		
	All-comers	0.691	0.074	0.029 / 0.037	0.952 / 0.987
	Simon	0.896	0.076	0.025 / 0.028	0.932 / 0.975
	BEAT _{<i>t</i>=0}	0.911	0.078	0.023 / 0.028	0.954 / 0.975
	BEAT _{<i>t</i>=0.5}	0.907	0.078	0.022 / 0.028	0.947 / 0.983
	BEAT _{<i>t</i>=1}	0.901	0.079	0.023 / 0.029	0.955 / 0.982
Negatives	True		-0.051		
	All-comers	0.775	-0.052	0.019 / 0.023	0.941 / 0.981
	Simon	0.340	-0.054	0.035 / 0.041	0.944 / 0.975
	BEAT _{<i>t</i>=0}	0.442	-0.061	0.029 / 0.037	0.960 / 0.983
	BEAT _{<i>t</i>=0.5}	0.516	-0.063	0.028 / 0.036	0.962 / 0.984
	BEAT _{<i>t</i>=1}	0.607	-0.063	0.027 / 0.034	0.962 / 0.984
Overall	True		-0.001		
	All-comers	0.046	-0.001	0.020 / 0.020	0.950
	Simon	0.056	-0.004	0.027 / 0.027	0.944
	BEAT _{<i>t</i>=0}	0.041	-0.006	0.024 / 0.025	0.960
	BEAT _{<i>t</i>=0.5}	0.030	-0.007	0.023 / 0.025	0.967
	BEAT _{<i>t</i>=1}	0.030	-0.008	0.023 / 0.024	0.961

J. For treatment effects in biomarker-positives and biomarker-negatives, Cover % refers to the rate for the 95% confidence intervals of the estimated treatment effect covering the true treatment effect based on estimated threshold $\hat{c}_0$ / true threshold c_0 separately. For overall treatment effect, Cover % refers to the rate for the 95% confidence intervals of the estimated overall treatment effect covering the true overall treatment effect.