HHS Public Access
Author manuscript
J Biopharm Stat. Author manuscript; available in PMC 2019 January 22.

Published in final edited form as:
J Biopharm Stat. 2018 ; 28(6): 1038–1054. doi:10.1080/10543406.2018.1434191.

Biomarker threshold adaptive designs for survival endpoints

Guoqing Diao[a], **Jun Dong**[b], **Donglin Zeng**[c], **Chunlei Ke**[b], **Alan Rong**[d], and **Joseph G Ibrahim**[c]

[a]Department of Statistics, George Mason University, Fairfax, Virginia, USA

[b]Amgen Inc., Thousand Oaks, California, USA

[c]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

[d]Astellas Pharma US, Inc., Los Angeles, California, USA

Abstract

Due to the importance of precision medicine, it is essential to identify the right patients for the right treatment. Biomarkers, which have been commonly used in clinical research as well as in clinical practice, can facilitate selection of patients with a good response to the treatment. In this paper, we describe a biomarker threshold adaptive design with survival endpoints. In the first stage, we determine subgroups for one or more biomarkers such that patients in these subgroups benefit the most from the new treatment. The analysis in this stage can be based on historical or pilot studies. In the second stage, we sample subjects from the subgroups determined in the first stage and randomly allocate them to the treatment or control group. Extensive simulation studies are conducted to examine the performance of the proposed design. Application to a real data example is provided for implementation of the first-stage algorithms.

Keywords

Adaptive enrichment design; predictive biomarker; survival endpoint; two-stage design

Introduction

Many new anticancer agents are molecularly targeted and therefore may only benefit a subgroup of a histologically defined population. Predictive biomarkers have been utilized to identify the sensitive subset to optimize treatment efficacy and safety in clinical practice. For example, in recurrent or metastatic squamous-cell carcinoma of the head and neck (SCCHN), loss of phosphate and tensin homolog (PTEN) expression had a negative effect on tumor response to epidermal growth factor receptor (EGFR) monoclonal antibodies (Mao et al., 2010); and EGFR expression level has been negatively associated with overall survival (Ang et al., 2002; Grandis et al., 1998; Nicholson et al., 2001; Reimers et al., 2007), which makes EGFR a potential positive prognostic factor for anti-EGFR antibodies. The

CONTACT Guoqing Diao, gdiao@gmu.edu, Department of Statistics, George Mason University, Fairfax, Virginia, USA. .
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lbps.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

SPECTRUM study is a phase-3 trial of chemotherapy with or without panitumumab (an anti-EGFR antibody) in patients with SCCHN (Vermorken et al., 2013). EGFR and PTEN data had been collected retrospectively using stored samples. It is of interest to prospectively evaluate the predictive value of EGFR and PTEN for panitumumab in SCCHN, while preliminary information may be derived from the SPECTRUM study. The benefits and challenges of incorporating biomarkers into the development of anticancer agents have been increasingly discussed (Renfro et al., 2016). Many design options have been proposed in the literature.

A biomarker is typically measured as a continuous or semi-continuous variable. The target subset is defined based on a chosen cutpoint of the biomarker. If the cutpoint is known but it is not known whether the biomarker is predictive, a biomarker-by-treatment interaction design (all-comer) can be employed where the study is stratified by the biomarker status and the hypotheses can be set up based on the overall population or the biomarker positive subset. If there exists strong evidence that the treatment is only effective in the positive biomarker subgroup, e.g., from early phase studies, an enrichment design can be utilized to limit the study population only to the positive subset. In addition, a subgroup enrichment design can include an interim analysis to decide whether or not the treatment is only effective in the positive subset and thus discontinue enrollment of patients in the non-sensitive subset. Freidlin et al. (2010) provided detailed discussions on phase-3 clinical trials that integrate treatment and biomarker evaluation.

In practice, there may be limited data available to identify a single biomarker or to determine a known cutpoint to define a sensitive subset before initiation of the phase-3 study. Then the selection of the cutpoint will be based on the study itself and will be part of the study objective. A number of methods have been suggested to prospectively including selection of the cutpoint into the study design (Renfro et al., 2016). For example, adaptive signature design (ASD) enrolls all comers into two stages (Freidlin and Simon, 2005). A classifier (i.e., cutpoint) is developed using data from stage- 1 patients only, and the classifier is not used to restrict enrollment in stage 2 but to define a subset of sensitive patients. Note that when there are multiple biomarkers or the treatment effect is not monotonic as a function of the biomarker under the alternative, there may not exist a simple cutpoint. For clarification, we emphasize here that we use “cutpoint” only when there is only one biomarker and the treatment effect is monotonic as a function of the biomarker under the alternative. In the final analysis, the comparison can be made based on the overall population using data from all patients enrolled in both stages or on the sensitive subset accrued during stage 2 together with some multiplicity adjustment procedure. Freidlin et al. (2010) extended ASD to cross-validated ASD where the stage-1 data is also used to test the sensitive subset to improve efficiency. A subgroup enrichment design can also be developed to limit enrollment of the second-stage patients to the sensitive subset.

Recently, Simon and Simon (2013) proposed a different adaptive enrichment phase-3 design (AED). The design in the beginning enrolls all patients, and sequentially restricts entry in an adaptive manner. It gives much of the efficiency of the “enrichment” approach without the need to choose a subset beforehand. The objective of the design is to improve the chance of a positive study by progressively limiting enrollment to patients who respond to the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

treatment. The appropriate primary null hypothesis of this design is that no subpopulation benefits from treatment over control. Therefore, this design may not be able to draw a conclusion about which subset of patients will benefit the most from the treatment. Consequently, the AED of Simon and Simon (2013) essentially reduces to a non-adaptive design when all patients respond to the treatment, even though the treatment effect may be different on subjects with different biomarker values. In Simon and Simon (2013), there was little detailed discussion regarding the algorithm to enrich the study particularly for survival endpoints, which are commonly used in clinical trials for anticancer agents. Additionally, little was discussed on the data used in the final analysis. More recently, Renfro et al. (2014) proposed an adaptive design by selecting the biomarker threshold for which the interaction effect between the biomarker and the treatment is the most significant. However, the design assumes that the experimental treatment is hypothesized to work better for patients with higher biomarker values than patients with lower biomarker values. This assumption may not be true in practice and consequently may lead to loss of power.

In this manuscript, we will propose a new AED design, called the biomarker threshold adaptive design (BTAD). Similar to the AED design in Simon and Simon (2013), our design does not adaptively adjust the total sample size after stage 1 or the sample size in stage 2. The stage-1 analysis can be based on historical or pilot studies. The enrichment in stage 2 is expected to increase power for hypothesis testing using either data from stage 2 alone or combined data from both stages. The Cox regression model for survival endpoints is employed for the AED. However, the proposed methods can be easily generalized to any other applications where a regression model is mainly used for inference. Different criteria for determination of the biomarker cutpoint based on stage-1 data are proposed. Our algorithm can potentially provide better enrichment in stage 2 than that of Simon and Simon (2013), and thus result in a more powerful procedure. Furthermore, while motivated by cancer studies, the proposed design is general and can be readily applied to clinical trials on other diseases.

This paper is organized as follows. In the next section, we use a group sequential framework and propose an optimal algorithm to determine, at a given interim analysis, which patient population will be enrolled next to enrich the study. Factors that may impact the algorithm are investigated. Furthermore, we discuss valid statistical methods for the final data analysis. We then evaluate the performance of the proposed design through extensive simulations. A real data example from the SPECTRUM study is used to demonstrate the determination of the biomarker cutpoint and to evaluate the performance of the proposed design compared with the non-adaptive design as well. We conclude the paper with some discussions.

Biomarker threshold adaptive design

We consider a BTAD to identify one or more predictive biomarkers. Specifically, our goal is to identify a subset of patients according to a biomarker of interest such that the treatment achieves the maximum beneficial effect compared to the control in this subset. One can then oversample patients in that subset to improve the overall efficiency of the design. In the simplest case, there exists a threshold for a single biomarker, such that the treatment effect is larger (or smaller) in one group than the other. Let T, X and A denote the survival endpoint,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the biomarker of interest and the treatment indicator, respectively. The treatment indicator A takes value 1 for the new treatment and 0 for the control. We propose two ways to determine the threshold. In this paper, we refer to the subject subset with either $X \le c$ or $X > c$ identified by our methods with better treatment effect as the "biomarker positive subgroup", and the complement as the "biomarker negative subgroup."

BTAD1.

For a given cutpoint c, we first consider the following Cox models for the two subgroups $X \le c$ and $X > c$, respectively

$$\lambda(t \mid X \le c, A) = \lambda_{1c}(t)\exp\left(\beta_{1c}A\right)$$

and

$$\lambda(t \mid X > c, A) = \lambda_{2c}(t)\exp\left(\beta_{2c}A\right).$$

Note that the smaller the regression coefficients β_{1c} and β_{2c}, the better the effect of the new treatment compared to the control. Based on the observed data in the first stage $\left\{\left(Y_i = \min\left(T_i, C_i\right),\ \Delta_i\ = I\left(T_i \le C_i\right), X_i, A_i\right),\ i = 1, \ldots, n_1\right\}$, where Ci, and Δ_i are the censoring time and event indicator, respectively, we can obtain the estimators of β_{1c} and β_{2c}, denoted by $\hat{\beta}_{1c}$ and $\hat{\beta}_{2c}$. We denote the total sample size and that of stages 1 and 2 as n, n_1 and n_2, respectively. We select the threshold $\hat{c}$ to minimize min $\left\{\hat{\beta}_{1c}, \hat{\beta}_{2c}\right\}$ for $c \in \mathscr{X}$, where $\mathscr{X}$ is the support of X. In the second stage, we then sample the remaining $n_2 = n - n_1$ subjects only from subgroup $X \le \hat{c}$ if $\hat{\beta}_{1\hat{c}} < \hat{\beta}_{2\hat{c}}$, and subgroup $X > \hat{c}$ otherwise.

BTAD2.

Alternatively, we can fit a Cox model including both the main effects of X and A and their interaction effect

$$\lambda(t|X, A) = \lambda_c(t)\exp\left\{\gamma_{1c}I(X > c) + \gamma_{2c}A + \gamma_{3c}I(X > c)A\right\}.$$

Based on the observed data from the n_1 subjects in the first stage, we can obtain the estimate of the treatment by biomarker interaction effect γ_{3c}, denoted by $\tilde{\gamma}_{3c}$. We select the threshold $\tilde{c}$ to maximize $\left|\tilde{\gamma}_{3c}\right|$ for $c \in \mathscr{X}$. In the second stage, we sample the remaining $n_2 = n - n_1$ subjects from subgroup $X \le \tilde{c}$ if $\tilde{\gamma}_{3\tilde{c}} > 0$ and $X > \tilde{c}$ otherwise.

It is worth clarifying that by "biomarker negative subgroup", we do not mean that in this group the treatment is not promising. Instead, we mean that the treatment effect is better in the "biomarker positive subgroup" than in the "biomarker negative subgroup." Therefore, it is possible that the treatment is promising for the overall population, but the proposed designs intend to identify the subpopulation such that the treatment is more promising than

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the other. One limitation is that the proposed designs may fail to recruit patients for which there is also treatment effect but the treatment is not as effective as in the other subgroup.

There are a few differences between *BTAD1* and *BTAD2*. First of all, *BTAD1* aims to identify the subgroup that responds the best to the treatment whereas *BTAD2* aims to identify a subgroup that responds to the treatment better than the other subgroup and the threshold is chosen such that the difference between treatment effects in these two subgroups is the greatest. Second, compared to *BTAD1*, *BTAD2* has an additional assumption that the hazard functions in the control group are proportional between the two biomarker subgroups. On the other hand, when the proportional hazards assumption is valid, *BTAD2* tends to yield more efficient parameter estimators and is numerically more stable especially when the sample size is small.

We can use a grid search method, for example, at certain sample percentiles of X, to select the threshold. To ensure reliable estimates of the unknown parameters, especially when the sample size is small or moderately large, we also suggest selecting the threshold such that there are at least 30% of subjects in each biomarker subgroup. A computer program in C language implementing *BTAD1* and *BTAD2* is available freely at http://mason.gmu.edu/gdiao/software/BTAD.

Once we collect all the data, it is of interest to test the hypothesis

$$H_0: \lambda(t|X = x, A = 0) = \lambda(t|X = x, A = 1), t \in [0, \tau], x \in \mathscr{X}, \quad (1)$$

where $\mathscr{X}$ is the support of X and τ is the end of study. That is, under the null hypothesis, there is no difference between the hazard functions in the treatment group and control group for any biomarker value. Furthermore, one may be interested in estimating the treatment effect. A natural question is which dataset to use in the final analysis after collecting all the data from the first and second stages. One can consider three types of datasets: (a) data from the second stage only; (b) all the data including both stages; and (c) data with subjects from the biomarker positive group only from both stages, that is, data including subjects selected according to the determined threshold from the first stage and all subjects from the second stage. We emphasize here that while using the first two types of data can preserve the type-I error rate, using the third type of data will lead to an inflated type-I error rate. When the null hypothesis is true, regardless of the threshold selected in the first stage, the data in the second stage are still collected under the null hypothesis; therefore, using the first two types of data can still preserve the type-I error rate. However, since we determine the threshold by selecting a subgroup in the first stage in which the treatment effect is better than the other subgroup, biased sampling arises and leads to an inflated type-I error rate if we include only the biomarker positive group in the final analysis. This observation is evident in the simulation studies in Section 3.

Note that the null hypothesis defined in (1) is general. In practice, to test this null hypothesis, one has to impose certain model assumptions. For example, by testing $\beta = 0$ in the Cox model with only the treatment indicator A as the covariate, i.e.,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\lambda(t|A) = \lambda_0(t)\exp(\beta A),$$

besides the Cox model assumption, we assume that patients within each biomarker subgroup have equal chances to receive either treatment or that there is no biomarker effect on the survival endpoint. On the other hand, if we assume the model

$$\lambda(t|X, A) = \lambda_0(t)\exp\left(\beta_1 A + \beta_2 X + \beta_3 X \times A\right), \quad (2)$$

then testing the null hypothesis defined in (1) is equivalent to testing $\beta_1 = \beta_3 = 0$. Under the null hypothesis defined in (1), all these tests using the appropriate data sets with correct model specification can control the type-I error; however, they will lead to different powers given different alternative hypothesis (for example, see Figure 2 in the Simulation Studies Section).

Suppose the true model is (2). We consider three models in the final analysis. The first model includes the treatment effect only,

$$\lambda^*(t|X, A) = \lambda_0^*(t)\exp\left(\beta_1^* A\right).$$

This is a misspecified model if $\beta_2 \neq 0$ or $\beta_3 \neq 0$. Therefore $\hat{\beta}_1^*$ is not a consistent estimator of β_1 or β_3. Instead, β_1^* may have a complicated form involving all the parameters in the true model. The second model we consider includes both treatment and biomarker main effects

$$\lambda^*(t|X, A) = \lambda_0^*(t)\exp\left(\beta_1^* A + \beta_2^* X\right).$$

In this case, $\beta_1^* \approx \beta_1 + \beta_3 E(X)$. If $\beta_1 = 0$ and $E(X) = 0$, we may fail to detect the true treatment effect. Hence, one needs to use caution when interpreting the results under this model. We may also fit the correctly specified model in the final analysis including both main effects and the interaction effect between the treatment and biomarker. We will still have consistent parameter estimates under this model using both the first-stage and second-stage data, as is evident by the simulation studies in Section 3. In all three models, we propose to use a Wald-type test statistic to test the null hypothesis H_0 of no treatment effect.

Simulation studies

We conducted extensive simulation studies to examine the performance of BTAD and we compare it to that of the standard non-adaptive design. We consider a two-arm trial with one biomarker X. We generate data from the exponential distribution under the following two scenarios:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\lambda(t|X,A) = \exp\left(\beta_1 A + \beta_2 X + \beta_3 X \times A\right) \quad (3)$$

and

$$\lambda(t|X,A) = \exp\left\{\beta_1 A + \beta_2 I(X > c) + \beta_3 I(X > c)A\right\}, \quad (4)$$

where c is the population median of the distribution of X. Under both scenarios, the baseline hazard is equal to 1. We considered three different distributions for X: (1) Uniform(0,1); (2) N(0; 1); and (3) EXP(1). The regression parameters are set to be

$$\beta_1 = 0,\ \beta_2 = -0.5,\ \beta_3 = 0,\ -0.1, \ldots, -0.6.$$

Under model (3), the hazard ratio between the treatment group and control group given $X = x$ is $\exp(\beta_1 + \beta_3 x)$. Therefore, for a positive number x, the above hazard ratio decreases as β_3 decreases. Consequently the smaller the value of β_3, the better the treatment effect for positive values of X. For example when $X = 0.5$, the hazard ratio ranges from 1 to 0.741 as j_3 varies from 0 to –0.6. On the other hand, when $X = -0.5$, the hazard ratio ranges from 1 to 1.350. For each simulation, we generate 1,000 replicates with $n = 200$ or $n = 300$. In all simulations, we fix the censoring time at 10.

In the first set of simulation studies, we compare the performance of the non-adaptive design, *BTAD1*, and *BTAD2*. Specifically, we sample $n_1 = n/2$ subjects from the general population. Based on these $n/2$ subjects, we determine the threshold c and biomarker positive group. We use a grid search method for selecting the threshold from the 30th, 40th,..., and 70th sample percentiles of the biomarker data. We then sample subjects from the biomarker positive group only in the second stage. In the final analysis, we evaluate the treatment effect based on all n subjects from both the first stage and the second stage with or without adjusting for the biomarker.

Table 1 presents the frequencies of selected groups $X \leq c$ and $X > c$ with $n = 200$ and $n = 300$ under different scenarios with β_3 fixed at –0.6. This table suggests that increasing the sample size can improve the frequency of a correctly selected biomarker subgroup. Histograms of the selected cutpoints under data generation model (3) with $\beta_3 = -0.6$ for $n = 300$ are presented in Figure 1. Similar results are obtained under data generation model (4) (data not shown).

Using all the data in both stages, we compare the powers of detecting the treatment effect of the following 10 methods with different designs and different analyses:

1. non-adaptive design and X adjusted;
2. non-adaptive design and X not adjusted;
3. adaptive design using *BTAD1*, X adjusted;

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

4. adaptive design using *BTAD1*, X not adjusted;
5. adaptive design using *BTAD2*, X adjusted;
6. adaptive design using *BTAD2*, X not adjusted;
    - a. adaptive design using *BTAD1* under the constraint such that only the subgroup $X > c$ is selected, X adjusted;
    - b. adaptive design using *BTAD1* under the constraint such that only the subgroup $X > c$ is selected, X not adjusted;
    - c. adaptive design using *BTAD2* under the constraint such that only the subgroup $X > c$ is selected, X adjusted;
    - d. adaptive design using *BTAD2* under the constraint such that only the subgroup $X > c$ is selected, X not adjusted.

Note that methods (7)-(10) assume that subjects with larger values of the biomarker respond better to the new treatment as in Simon and Simon (2013). This assumption may not be true in practice. Figures 2 and 3 present the type-I error rates and powers for testing the null hypothesis H_0 under data generation models (3) and (4), respectively. The tests under the proposed adaptive design have correct control of the type-I error rate and are consistently more powerful than the tests under the non-adaptive design. The analyses with and without adjusting for the biomarker effect yield comparable results in most cases; however, the test of the treatment effect without adjusting for the biomarker effect can be substantially more powerful than the one adjusting for the biomarker effect when the mean biomarker value and the main treatment effect are close to 0. For example, as shown in Figure 2 when $X \sim N(0, 1)$, since the treatment effect β_1^* under the model adjusting for X is approximately $\beta_1 + \beta_3 E(X) = 0$, there is essentially no power to detect treatment effect under the non-adaptive design. Methods 7–10 show that if the assumption on the constraint is correct, then the power is higher, as expected. *BTAD1* and *BTAD2* show similar performance in all simulations in Figures 2 and 3.

In the second set of simulations, we investigate which types of datasets are suitable for the final analysis under the proposed BTAD. In particular, we compare the type-I error rates and powers of the tests for testing H_0 at the nominal significance level of 0.05 using the three different types of data in the final analysis as described in Section 2. Figure 4 displays the results based on the adaptive design using *BTAD1*. It is obvious that using the biomarker positive group only can lead to an inflated type-I error rate, while using the second-stage data only or all the data including both first and second-stage data preserves the type-I error rate. The tests using all the data are more powerful than those using the second-stage data only with the exception when $X \sim N(0, 1)$. As discussed in the end of Section 2, the treatment effect β_1^* under the fitted models is approximately $\beta_1 + \beta_3 E(X)$. Therefore, the power of the test depends on both the sample size in the final analysis and the approximate effect size $\beta_1 + \beta_3 E(X)$. Although the sample size in the combined data set including both stages is larger than the sample size in the second stage only, the expectation of X in the second-stage

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

data set can be much larger than that in the combined data set leading to a larger effect size and consequently a larger power.

Finally, we conducted simulation studies to examine the estimates of the regression coefficients by fitting the true model including both the main effects of biomarker and treatment and their interaction effect. We considered a non-adaptive design, *BTAD1* and *BTAD2* using the second-stage data only, using both the first-stage and second-stage data, and using data from biomarker positive subgroup only. The results for data generation model (3) are presented in Table 2–4. By fitting the full model with interaction effects using all the data, we can still estimate the parameters consistently, although there is some efficiency loss compared to the non-adaptive design. Using the biomarker positive subgroup only will lead to a biased estimate of the treatment effect. Furthermore, using the second-stage data only or the biomarker positive subgroup data only leads to substantial efficiency loss of the parameter estimates. *BTAD1* and *BTAD2* have similar performance.

SPECTRUM study

We consider a head and neck cancer clinical trial as a case study to demonstrate how to determine the cutpoint using our algorithms for biomarkers with different patterns of biomarker-outcome relationship as described in Section 2. The SPECTRUM study is a multicenter, randomized, open- label, phase-3 trial of chemotherapy with or without panitumumab in patients with recurrent or metastatic SCCHN (Vermorken et al., 2013). The primary endpoint was overall survival. For our analysis, 526 (80.0%) subjects were randomly selected from the 657 subjects in the SPECTRUM trial for demonstration purpose. The dataset with 526 subjects is treated as the data collected in stage 1. There are no stage-2 data; therefore, the inference after stage 2 is not available in this analysis.

In a literature review of more than 200 tumor studies by Nicholson et al. (2001), EGFR is reported to act as a prognostic factor for OS in solid tumors with increased EGFR expression generally associated with a reduced OS rate. In SCCHN, a correlation between higher EGFR expression and poorer OS was claimed by Ang et al. (2002). The OS rates for subjects with high EGFR expressing SCCHN (> median mean absorbance) compared to those with low EGFR expressing SCCHN were significantly lower ($p = 0.0006$). These studies (reporting mean absorbance) were performed using automated image analysis to quantify expression levels. EGFR-negative tumors have shown a tendency toward a better prognosis in OS (70% vs. 45%, $p = 0.10$) by Reimers et al. (2007). In addition, when tumor levels of EGFR expression (intensity and tumor cell extent) were analyzed as continuous variables, cause-specific survival was reduced among subjects with higher levels of EGFR (p = 0.0001) (Grandis et al., 1998).

In SCCHN, PTEN has been shown to be frequently altered at the genetic and biochemical level (Pedrero et al., 2005). In metastatic colorectal cancer, the prevailing hypothesis is that lack of PTEN expression predicts resistance to EGFR antibody therapy. There have been four small studies that have evaluated the association of PTEN expression and patient response to EGFR antibodies and a meta-analysis of these data showed that loss of PTEN

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

expression had a negative effect on tumor response to EGFR monoclonal antibodies (pooled risk ratio, 0.22; 95% confidence interval, 0.10–0.50; *P* = 0.001) (Mao et al., 2010).

We first perform an exploratory analysis on two biomarkers MEMASPE (tumor cell percentage with membrane staining total 1±3% for EGFR) and CYTOSPPT (Tumor cell percentage with cytoplasmic staining total 1±3% for PTEN). Among the total number of 526 subjects, 163 subjects have non-missing MEMASPE values and 290 subjects have non-missing CYTOSPPT values. Subjects with missing biomarker values were excluded from the analysis. For each biomarker, we first plot the histogram. Additionally, we fit the following Cox model allowing for biomarker-dependent treatment effects,

$$\lambda(t|X = x, A) = \lambda_0(t)\exp\left\{g_0(x) + d(x)A\right\}, \quad (5)$$

where $g_0(x)$ and $d(x)$ are estimated nonparametrically by using B-splines. The results are presented in Figure 5. Both biomarkers are left skewed with a spike around a value of 100. It appears that the treatment effect, as a function of MEMASPE, is monotonically increasing; that is, patients with larger values of MEMASPE respond to the treatment better. In contrast, the treatment effect as a function of CYTOSPPT, is not monotone. The treatment doesn’t appear to have a beneficial effect on patients with CYTOSPPT values around 20 whereas patients with CYTOSPPT values around 80 respond to the treatment the best.

We used a grid search approach ranging from the 30th sample percentile to the 70th sample percentile with a step size of 2% to determine the biomarker positive subgroup. For MEMASPE, both *BTAD1* and *BTAD2* identify the positive subgroup with MEMASPE > 94.6, which is the 66th sample percentile. This result agrees well with the results presented in Figure 5, and the proposed methods correctly identify the biomarker subgroup which responds to the treatment the best. For CYTOSPPT, *BTAD1* identifies the positive subgroup with CYTOSPPT > 92.5 (sample median) and *BTAD2* identifies the positive subgroup with CYTOSPPT > 76.05 (30th sample percentile). Although the proposed methods do not identify the subgroup in which the treatment effect is the largest, they still identify a subgroup with beneficial treatment effect compared to a majority of the other patients. More discussion on this topic is provided in Section 5.

Finally, we conducted a simulation study based on the SPECTRUM study to examine the performance of the proposed BTAD using the biomarker MEMASPE. In the first stage, we randomly sampled 60 patients from the entire dataset and then determine the cutpoint using *BTAD1* and *BTAD2.* In the second stage, we randomly sampled up to 30 patients with positive biomarkers from the remaining 103 patients. We then repeated this procedure 1,000 times and estimated the treatment effect using all the data from both the first stage and the second stage. With the biomarker adjusted, the empirical powers for testing the treatment effect at the significance level of 0.05 are 91.6% and 89.6% based on *BTAD1* and *BTAD2,* respectively, compared to a power of 62.2% based on the non-adaptive design. Without adjusting for the biomarker, the empirical powers are 85.9%, 83.0%, and 50.7% based on *BTAD1, BTAD2,* and the non-adaptive design, respectively. In about 25% of the 1000 simulated datasets, there are fewer than 30 patients with positive biomarkers in the second

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

stage. Even though the sample sizes are smaller on average in the BTAD than the nonadaptive design, there is substantial power increase of the BTAD over the non-adaptive design.

Discussion

In this paper, we proposed a method to determine the threshold based on the estimates of the treatment effect in each biomarker subgroup. Naturally, one would want to take into account the variation of the parameter estimates in the decision process. For example, one may use a statistic for testing the treatment effect in each subgroup to select the threshold. However, such an approach may lead to selecting the biomarker negative group with a larger sample size in the first stage.

Renfro et al. (2014) considered the same model as in *BTAD2* including the interaction effect between treatment and biomarker. Assuming that patients with higher biomarker values respond to the treatment better, the threshold was selected such that the treatment-by-biomarker interaction is most significant. To relax the assumption on the direction of the interaction effect, we can extend the approach of Renfro et al. (2014) such that the threshold $\tilde{c}$ is selected the same way but in the second stage we sample subjects from subgroup with $X \le \tilde{c}$ if $\tilde{\gamma}_{3c} > 0$ and $X > \tilde{c}$ otherwise. Our limited simulation studies suggest that in general the performance of this extension is comparable with *BTAD2;* however, compared to *BTAD1* and *BTAD2,* this approach tends to select threshold near the sample median of the biomarker value and can have reduced power, particularly when *X* is generated from a normal distribution.

We have assumed that treatment effect is monotone as a function of the biomarker, in particular, a model with a change point, at which the treatment effect changes. In practical applications, it is possible that the treatment effect curve is not monotone. Violation of the monotonicity assumption may lead to inaccurate identification of the biomarker positive subgroup. In this case, we may consider the general model (5) as described in Section 4. In particular, when *d(x)* is a step function with a change-point at *c*, the above model reduces to the model under consideration in this paper. The general model (5) also includes settings such that the marginal treatment effect is not 0 and/or the interaction effect is not linear on the biomarker scale. In general, we can estimate $g_0(x)$ and *d(x)* nonparametrically, e.g., using smoothing splines. We can then determine the biomarker positive group as

$$\mathscr{R}_c = \{x : \hat{d}(x) < c\},$$

for a constant *c*. The constant *c* can be chosen to minimize

$$\frac{\int_{\mathscr{R}_c} \hat{d}(x) dF_X(x)}{\int_{\mathscr{R}_c} dF_X(x)} - \frac{\int_{\mathscr{R}_c^C} \hat{d}(x) dF_X(x)}{\int_{\mathscr{R}_c^C} dF_X(x)},$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

subject to

$$\min\left\{\int_{\mathscr{R}_c} dF_X(x), \int_{\mathscr{R}_c^C} dF_X(x)\right\} \geq q, \text{ for a pre } - \text{ specified value } q,$$

where $\mathscr{R}_c^C$ is the complement set of $\mathscr{R}_c$, and $F_X(x)$ is the cumulative distribution function of X. Future research is warranted along this direction.

When there is no interaction effect between the biomarker of interest and treatment, this biomarker is not a predictive biomarker. Therefore, such a biomarker is not an appropriate candidate for identifying subpopulations of patients who are most likely to respond to the treatment. In view of this, we can first perform hypothesis testing to test $H_0 : \beta_{1c} = \beta_{2c}$ for *BTAD1* and $H_0 : \gamma_{3c} = 0$ for *BTAD2*. Then, we continue the algorithm of the adaptive design given that there is some level of evidence of an interaction effect between the biomarker and the treatment.

Simon and Simon (2013) proposed to sample all subjects from the positive biomarker group. Such a design may not differ much from the non-adaptive design if the treatment is better than the control in every biomarker group. On the other hand, our proposed methodology is based on sampling subjects who will benefit the most from the treatment. Even if the treatment is better than the control in every biomarker group, our proposed design aims to select the group who respond to the treatment the best. Therefore, our proposed methods tend to be more powerful than the method of Simon and Simon (2013) because of larger effect sizes.

We can extend our proposed methods to two or more biomarkers **X**. For example, for p biomarkers, we can have a p – dimensional search for the thresholds $(C_1,\ldots,C_P)$. However, the computational burden increases geometrically as the number of biomarkers increase. An alternative approach is to first construct a composite score or risk factor $\beta^{\mathrm{T}}\mathbf{X}$, which can be obtained by fitting a Cox model

$$\lambda(t|\ \mathbf{X}\ , A) = \lambda(t)\exp\left(\beta^T\ \mathbf{X}\ + \gamma A\right).$$

We then treat the composite score as a new biomarker and apply the aforementioned methods to determine a threshold for this composite score. This is a current area of research.

When comparing the non-adaptive and adaptive approaches, we assume that the sample sizes used are the same. Like many other enrichment designs, the proposed adaptive approaches can be more costly compared to a non-adaptive design as the adaptive designs typically require additional screening and more patients need to be recruited in order to have sufficient number of biomarker positive patients in the second stage. It would be interesting to conduct cost-benefit analysis to compare the non-adaptive and adaptive designs.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

In this paper, we focus on survival endpoints and assume a Cox model for the relationship between treatment and the survival outcome. However, the proposed methods can be generalized to regression models for different types of outcome data, such as binary, count, and normal outcomes. Second, when the proportional hazards assumption does not hold, the proposed methods can still preserve type-I error under the null hypothesis but may not be as powerful as seen now. Alternatively, other regression models for survival endpoints such as the proportional odds model, accelerated failure time model, or additive hazard model can be considered.

References

Ang KK, Berkey BA, Tu X, Zhang H-Z, Katz R, Hammond EH, Fu KK, Milas L (2002). Impact of epidermal growth factor receptor expression on survival and pattern of relapse in patients with advanced head and neck carcinoma. Cancer Research 62:7350–7356 [PubMed: 12499279]

Freidlin B, Jiang W, Simon R (2010). The cross-validated adaptive signature design. Clinical Cancer Research 16:691–698. doi:10.1158/1078-0432.CCR-09-1357. [PubMed: 20068112]

Freidlin B, 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. doi:10.1158/1078-0432.CCR-05-0605. [PubMed: 16278411]

Grandis JR, Tweardy DJ, Melhem MF (1998). Asynchronous modulation of transforming growth factor alpha and epidermal growth factor receptor protein expression in progression of premalignant lesions to head and neck squamous cell carcinoma. Clinical Cancer Research 4:13–20 [PubMed: 9516947]

Mao C, Liao R, Chen Q, et al. (2010). Loss of PTEN expression predicts resistance to EGFR-targeted monoclonal antibodies in patients with metastatic colorectal cancer. British Journal of Cancer 102:940. doi:10.1038/sj.bjc.6605575. [PubMed: 20160728]

Nicholson RI, Gee JMW, Harper ME (2001). EGFR and cancer prognosis. European Journal of Cancer 37:9–15. doi:10.1016/S0959-8049(01)00231-3. [PubMed: 11165124]

Pedrero JMG, Carracedo DG, Pinto CM, Zapatero AH, Rodrigo JP, Nieto CS, Gonzalez MV (2005). Frequent genetic and biochemical alterations of the pi 3-k/AKT/PTEN pathway in head and neck squamous cell carcinoma. International Journal of Cancer 114:242–248. doi:10.1002/ijc.20711. [PubMed: 15543611]

Reimers N, Kasper HU, Weissenborn SJ, Stutzer H, Preuss SF, Hoffmann TK, Speel EJM, Dienes HP, Pfister HJ, Guntinas-Lichius O, et al. (2007). Combined analysis of HPV-DNA, p16 and EGFR expression to predict prognosis in oropharyngeal cancer. International Journal of Cancer 120:1731–1738. doi:10.1002/ijc.22355. [PubMed: 17236202]

Renfro LA, Coughlin CM, Grothey AM, Sargent DJ (2014). Adaptive randomized phase ii design for biomarker threshold selection and independent evaluation. Chinese Clinical Oncology 3.

Renfro LA, Mallick H, An M-W, Sargent DJ, Mandrekar SJ (2016). Clinical trial designs incorporating predictive biomarkers. Cancer Treatment Reviews 43:74–82. doi:10.1016/j.ctrv.2015.12.008. [PubMed: 26827695]

Simon N, Simon R (2013). Adaptive enrichment designs for clinical trials. Biostatistics 14:613–625. doi:10.1093/biostatistics/kxt010. [PubMed: 23525452]

Vermorken JB, Stohlmacher-Williams J, Davidenko I, Licitra L, Winquist E, Villanueva C, Foa P, Rottey S, Skladowski K, Tahara M, et al. (2013). Cisplatin and fluorouracil with or without panitumumab in patients with recurrent or metastatic squamous-cell carcinoma of the head and neck (spectrum): An open-label phase 3 randomised trial. The Lancet Oncology 14:697–710. doi:10.1016/S1470-2045(13)70181-5. [PubMed: 23746666]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

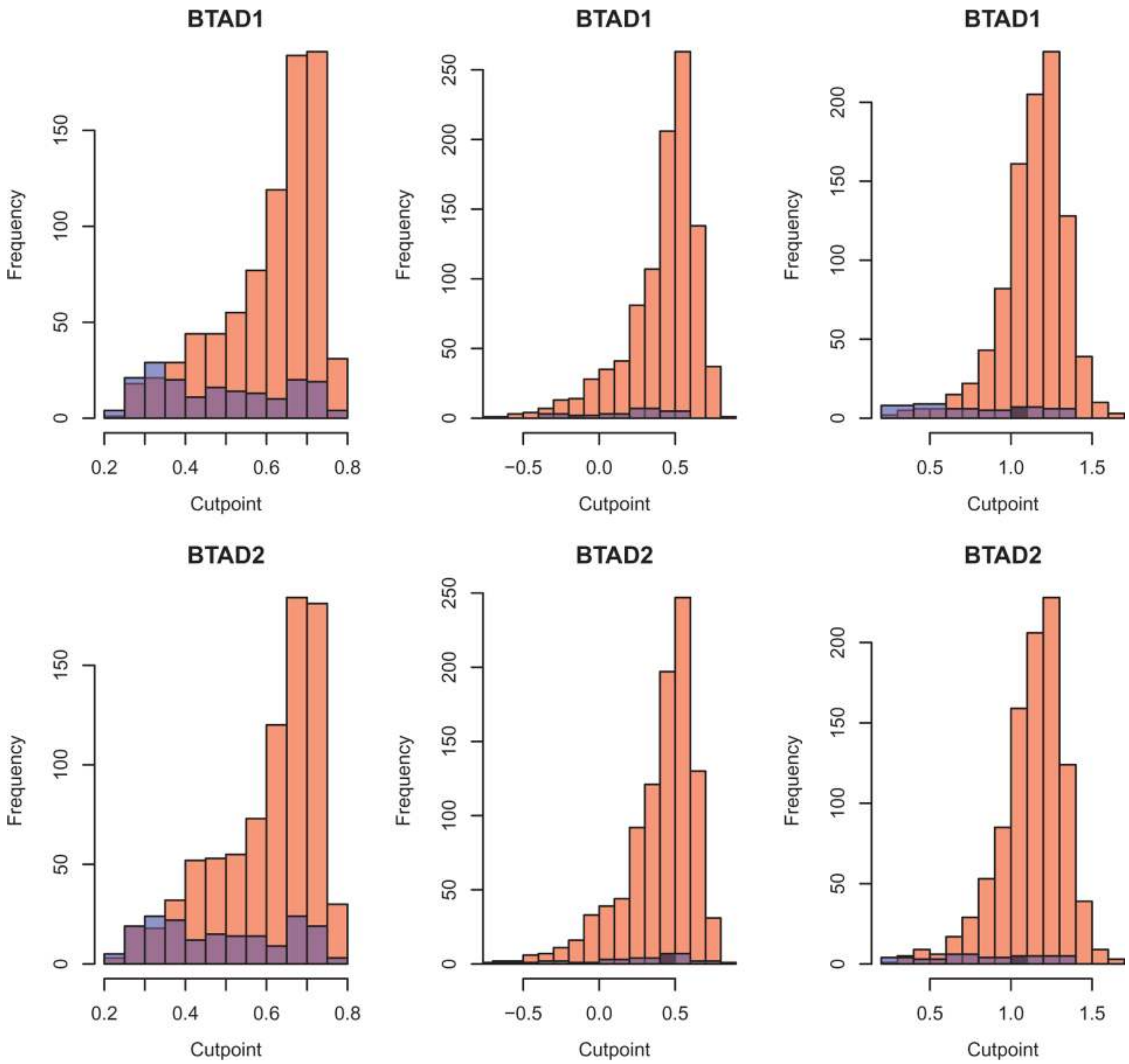

Figure 1.

Histograms of selected cutpoints under data generation model (3) with $\beta_3 = -0.6$. From the left panel to the right panel, the distributions of X are Uniform(0,1), N(0,1), and EXP(1), respectively. The upper and lower panels correspond to the results from *BTAD1* and *BTAD2,* respectively. The magenta bars correspond to selected groups $X>c$, and the blue bars for $X \leq c$.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

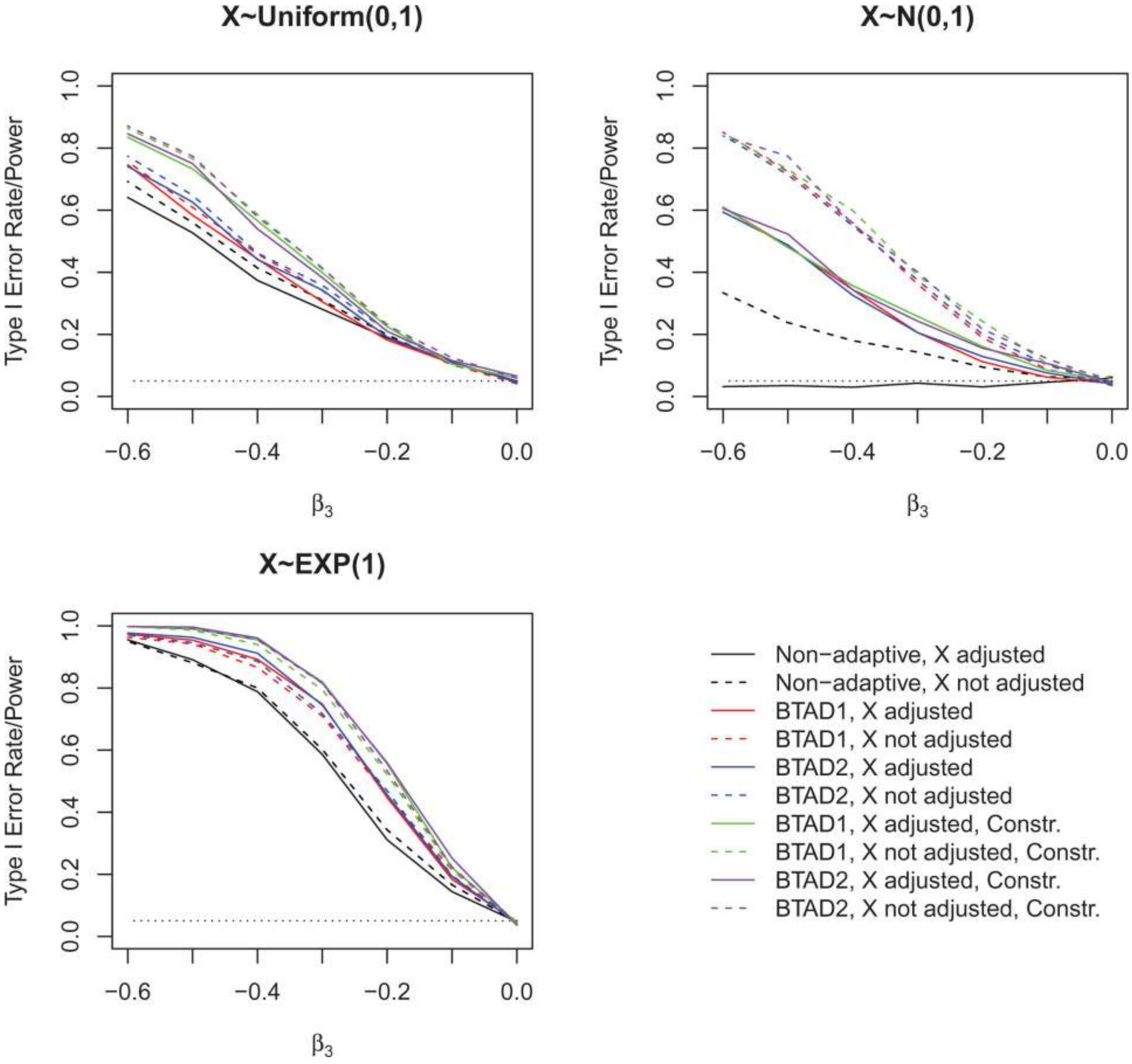

Figure 2.

Type-I error rates and powers under data generation model (3). All data from both stages are included in the analysis. The dotted reference line corresponds to the *y*-axis at 0.05.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

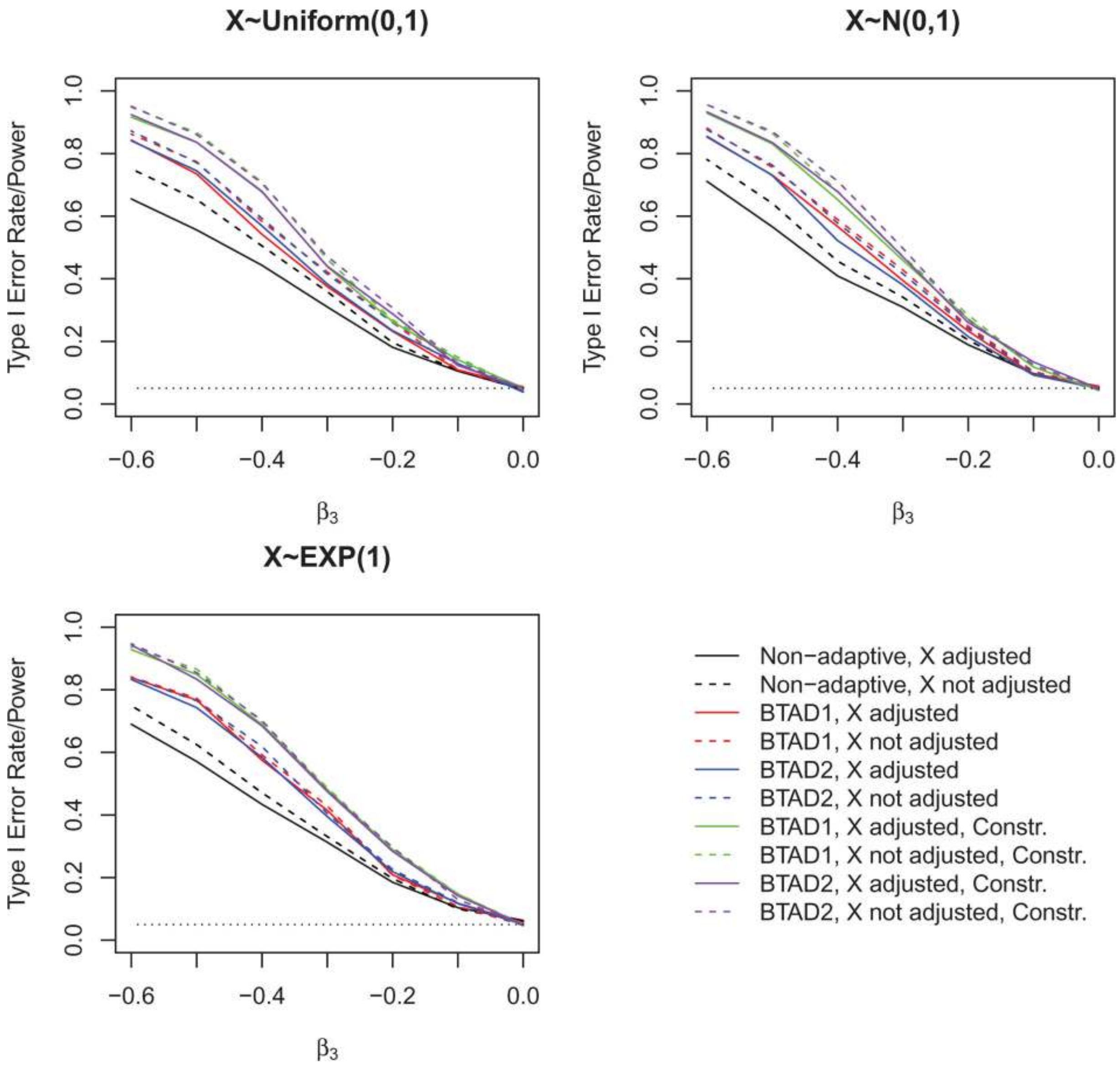

Figure 3.

Type-I error rates and powers under data generation model (4). All data from both stages are included in the analysis. The dotted reference line corresponds to the *y*-axis at 0.05.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

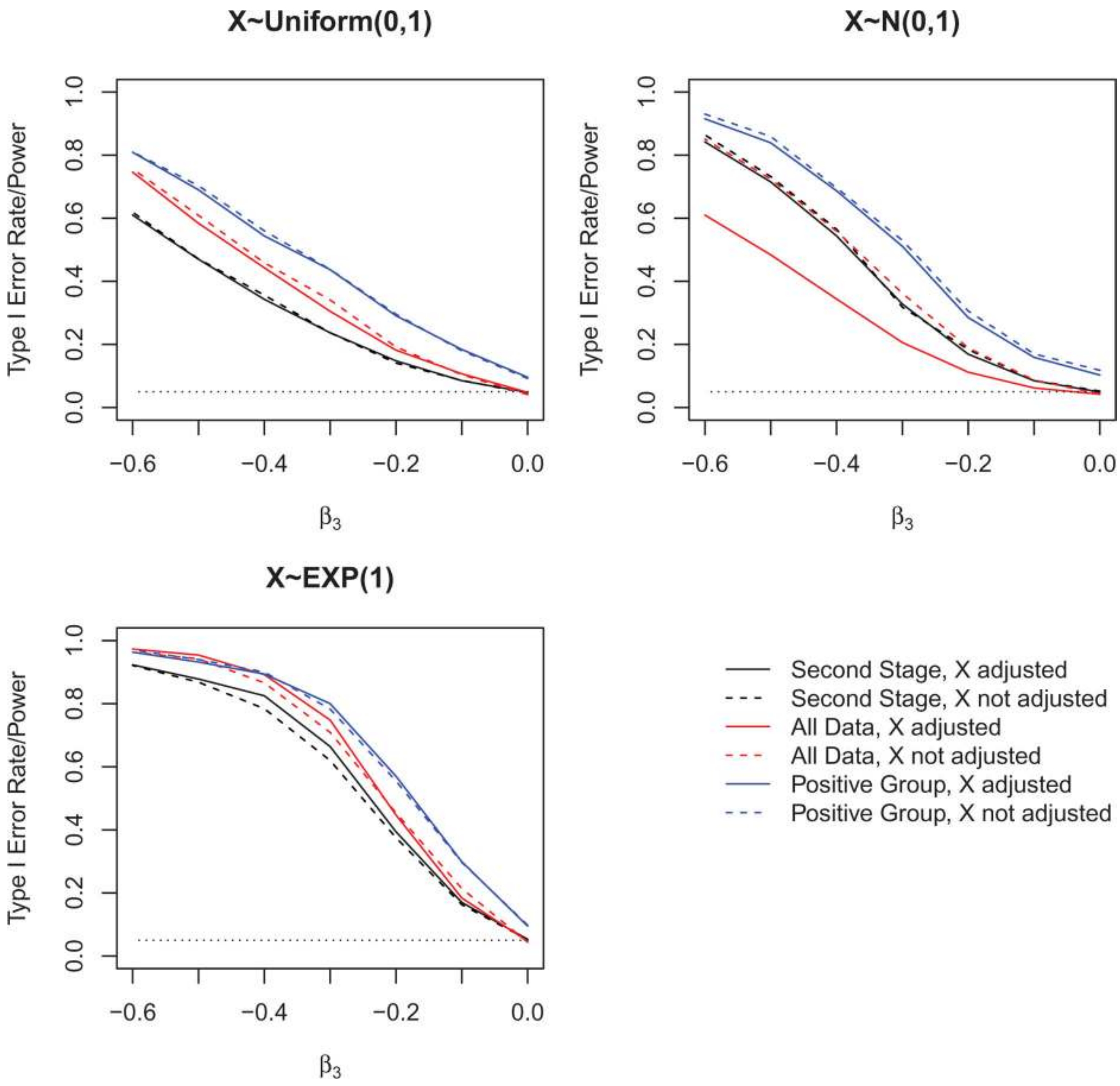

Figure 4.

Type-I error rates and powers for *BTAD1* under data generation model (3) using different types of data in the final data analysis. The dotted reference line corresponds to the *y*-axis at 0.05.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

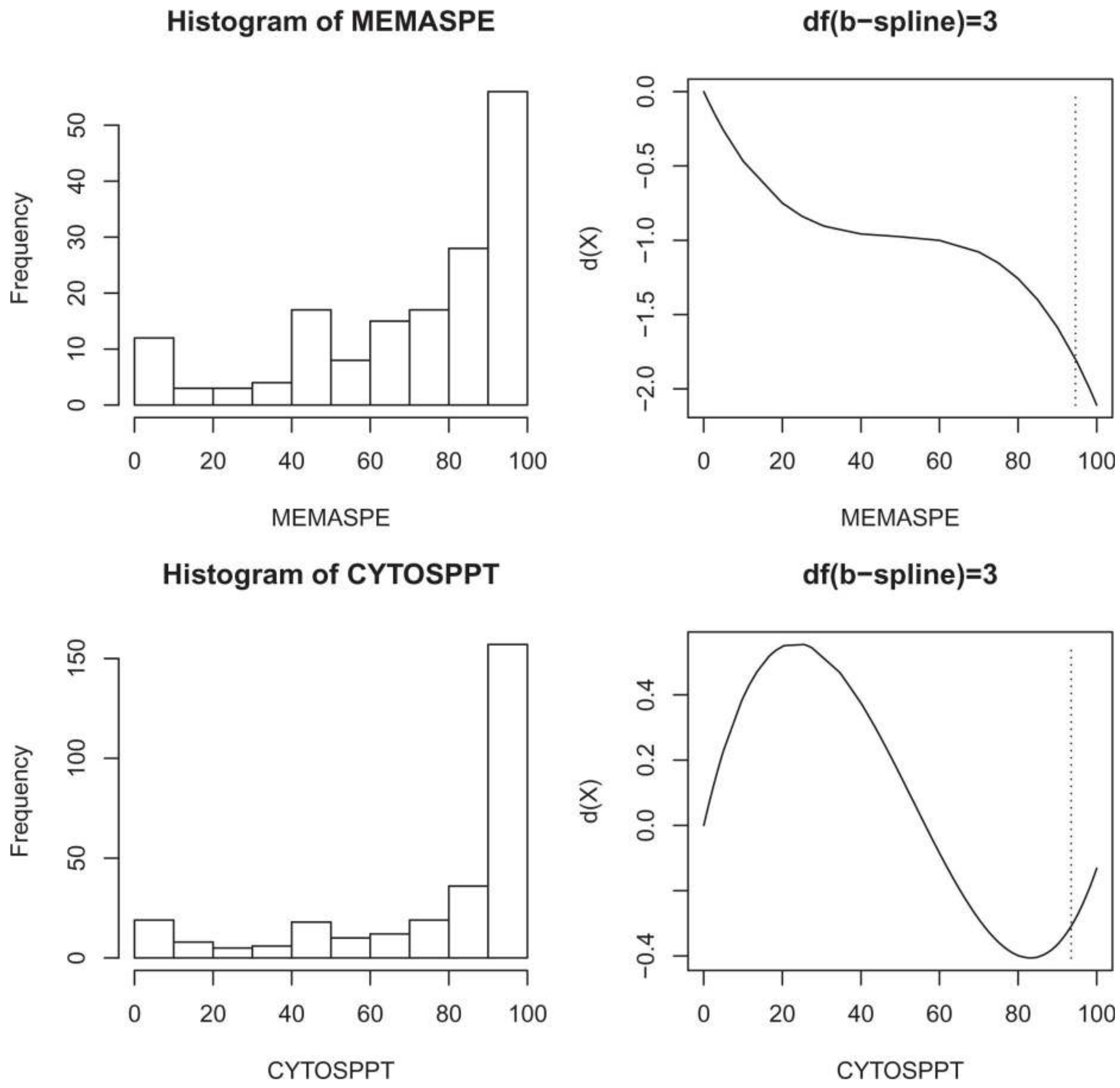

Figure 5.

Results for MEMASPE (tumor cell percentage with membrane staining total 1±3% for EGFR) and CYTOSPPT (tumor cell percentage with cytoplasmic staining total 1±3% for PTEN). Among the total number of 526 subjects, 163 subjects have non-missing MEMASPE values and 290 subjects have non-missing CYTOSPPT values. The dotted vertical lines represent the selected thresholds by using *BTAD1* (94.6 for MEMASPE and 92.5 for CYTOSPPT).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1.

Relative frequencies of selected subgroups. Under the true simulation models, the subgroup (X > c) is expected to be selected.

	BTAD1		***BTAD2***	
Data generation model	$X \le c$	$X > c$	$X \le c$	$X > c$
		$n = 200$		
(3), X~*Uniform(0*,1)	0.242	0.758	0.239	0.761
(4), X~*Uniform(0*,1)	0.137	0.863	0.142	0.858
(3), X~N(0,1)	0.031	0.969	0.031	0.969
(4), X~N(0,1)	0.142	0.858	0.145	0.855
(3), X~*EXP(1)*	0.090	0.910	0.066	0.934
(4), X~*EXP(1)*	0.151	0.849	0.152	0.848
		$n = 300$		
(3), X~*Uniform(0*,1)	0.181	0.819	0.180	0.820
(4), X~*Uniform(0*,1)	0.072	0.928	0.088	0.912
(3), X~N(0,1)	0.021	0.979	0.021	0.979
(4), X~N(0,1)	0.094	0.906	0.093	0.907
(3), X~*EXP(1)*	0.041	0.959	0.027	0.973
(4), X~EXP(1)	0.084	0.916	0.082	0.918

Note: the true regression parameters ($\beta_1, \beta_2, \beta_3$) are set to be (0, –0.5, –0.6).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2.

Parameter estimates based on the full model under various designs for data generation model (3) when *X*~*Uniform*(0,1)

			BTAD1						***BTAD2***					
	Non-adapt		**Second stage**		**All**		**Positive**		**Second stage**		**All**		**Positive**	
Parameter	**Est**	**SE**	**Est**	**SE**	**Est**	**SE**	**Est**	**SE**	**Est**	**SE**	**Est**	**SE**	**Est**	**SE**
$(\beta_1,\beta_2,\beta_3) = (-0.5,0,0)$														
β_1	−0.490	0.371	−0.533	1.449	−0.508	0.370	−0.494	1.206	−0.486	1.419	−0.500	0.374	−0.462	1.231
β_2	0.015	0.283	−0.058	1.264	0.045	0.306	−0.176	1.075	−0.012	1.137	0.044	0.309	−0.134	1.028
β_3	−0.025	0.516	0.118	2.120	0.012	0.530	0.034	1.764	−0.039	1.938	0.006	0.546	−0.097	1.730
$(\beta_1,\beta_2,\beta_3) = (-0.5,0,-0.3)$														
β_1	−0.498	0.350	−0.543	1.461	−0.503	0.375	−0.531	1.277	−0.575	1.488	−0.512	0.366	−0.560	1.278
β_2	0.007	0.294	−0.060	1.396	0.051	0.317	−0.203	1.198	−0.023	1.422	0.041	0.330	−0.183	1.189
β_3	−0.326	0.502	−0.260	2.012	−0.326	0.535	−0.259	1.761	−0.281	2.057	−0.305	0.553	−0.266	1.765
$(\beta_1,\beta_2,\beta_3) = (-0.5,0,-0.6)$														
β_1	−0.507	0.350	−0.477	1.458	−0.501	0.396	−0.560	1.217	−0.503	1.505	−0.506	0.388	−0.556	1.239
β_2	0.005	0.281	−0.014	1.552	0.037	0.336	−0.229	1.300	0.064	1.589	0.045	0.332	−0.174	1.322
β_3	−0.606	0.499	−0.638	2.050	−0.605	0.544	−0.469	1.724	−0.712	2.156	−0.613	0.539	−0.533	1.828

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3.

Parameter estimates based on the full model under various designs for data generation model (3) when $X \sim N(0,1)$

			BTAD1						*BTAD2*					
	Non-adapt		Second stage		All		Positive		Second stage		All		Positive	
Parameter	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE	Est	SE
							$(\beta_1,\beta_2,\beta_3) = (-0.5.0.0)$							
β_1	−0.508	0.107	−0.519	0.291	−0.503	0.113	−0.491	0.240	−0.530	0.311	−0.507	0.116	−0.499	0.254
β_2	−0.001	0.144	−0.005	0.462	0.034	0.154	−0.124	0.378	−0.001	0.468	0.035	0.158	−0.124	0.381
β_3	0.003	0.148	−0.002	0.415	−0.002	0.160	−0.045	0.343	0.006	0.415	0.002	0.160	−0.041	0.347
							$(\beta_1,\beta_2,\beta_3) = (-0.5.0,-0.3)$							
β_1	−0.506	0.110	−0.512	0.306	−0.509	0.112	−−0.509	0.261	−0.507	0.296	−0.510	0.114	−0.507	0.255
β_2	0.000	0.146	0.029	0.494	0.027	0.162	−0.068	0.421	0.013	0.510	0.025	0.163	−0.079	0.432
β_3	−0.305	0.150	−0.312	0.443	−0.301	0.161	−0.303	0.365	−0.325	0.443	−0.302	0.162	−0.308	0.365
							$(\beta_1,\beta_2,\beta_3) = (-0.5.0,-0.6)$							
β_1	−0.505	0.108	−0.495	0.306	−0.508	0.111	−0.515	0.250	−0.473	0.326	−0.504	0.112	−0.494	0.264
β_2	−0.007	0.148	0.037	0.533	0.003	0.166	−0.025	0.453	0.044	0.570	0.005	0.162	−0.016	0.478
β_3	−0.609	0.150	−0.659	0.476	−0.608	0.158	−0.613	0.397	−0.660	0.497	−0.608	0.160	−0.613	0.407

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4.

Parameter estimates based on the full model under various designs for data generation model (3) when X~E(0,1)

			BTAD1						***BTAD2***					
	Non-adapt		**Second stage**		**All**		**Positive**		**Second stage**		**All**		**Positive**	
Parameter	**Est**	**SE**	**Est**	**SE**	**Est**	**SE**	**Est**	**SE**	**Est**	**SE**	**Est**	**SE**	**Est**	**SE**
							$(\beta_1,\beta_2,\beta_3) = (-0.5.0.0)$							
β_1	−0.510	0.125	−0.574	0.886	−0.525	0.138	−0.486	0.778	−0.530	0.907	−0.527	0.135	−0.446	0.751
β_2	−0.006	0.213	−0.010	0.466	0.001	0.223	−0.048	0.385	−0.019	0.490	−0.003	0.219	−0.056	0.408
β_3	0.009	0.169	0.066	1.193	0.040	0.184	−0.077	1.018	0.019	1.313	0.045	0.184	−0.110	1.104
							$(\beta_1,\beta_2,\beta_3) = (-0.5.0,-0.3)$							
β_1	−0.503	0.122	−0.535	0.576	−0.518	0.123	−0.495	0.518	−0.529	0.561	−0.519	0.122	−0.502	0.484
β_2	−0.003	0.221	0.005	0.612	0.008	0.258	−0.058	0.516	0.015	0.619	0.001	0.255	−0.058	0.520
β_3	−0.308	0.186	−0.307	0.812	−0.285	0.193	−0.325	0.711	−0.338	0.84	−0.279	0.189	−0.343	0.683
							$(\beta_1,\beta_2,\beta_3) = (-0.5.0,-0.6)$							
β_1	−0.505	0.126	−0.523	0.429	−0.509	0.119	−0.505	0.375	−0.502	0.328	−0.510	0.119	−0.499	0.279
β_2	0.023	0.224	0.049	0.718	0.024	0.272	−0.013	0.621	0.096	0.707	0.026	0.281	0.017	0.632
β_3	−0.623	0.208	−0.649	0.683	−0.607	0.208	−0.645	0.627	−0.670	0.525	−0.604	0.206	−0.643	0.450