Statistica Sinica Preprint No: SS-2025-0076	
Title	Nonparametric Inference on Treatment-biomarker Interaction Based on Probability Index
Manuscript ID	SS-2025-0076
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0076
Complete List of Authors	Zehui Wang, Yanglei Song, Wenyu Jiang and Dongsheng Tu
Corresponding Authors	Dongsheng Tu
E-mails	dtu@ctg.queensu.ca
Notice: Accepted author version.	

Statistica Sinica

Nonparametric Inference on Treatment-Biomarker Interaction Based on Probability Index

Zehui Wang, Yanglei Song, Wenyu Jiang and Dongsheng Tu

Queen's University

Abstract: In precision medicine, an important task is identifying subgroups of patients who may benefit more from a treatment based on a clinical variable or biomarker. In this paper, we propose a non-parametric treatment-biomarker interaction measure using a probabilistic index to assess differences in treatment effects between subgroups formed by dichotomizing a biomarker at a cutpoint. When the cutpoint is prespecified, null hypothesis of no interaction is tested using Wilcoxon-type statistics. When the cutpoint is not prespecified, the same null hypothesis is tested across a range of cutpoint values by taking the supremum of Wilcoxon-type statistics with p-value calculated by a bootstrap procedure. It is shown that the sizes of the proposed tests converge to the nominal level in both cases. If the null hypothesis is rejected in the case of unspecified cutpoint, a profile estimator for the cutpoint that maximizes the difference in treatment effects is proposed. We show that the estimator has cubic-rate convergence and asymptotically follows a scaled Chernoff's distribution. Furthermore, we introduce an m-out-of-n bootstrap procedure to estimate the unknown scaling factor in the asymptotic distribution. Extensive simulation studies support our theory,

*Dongsheng Tu, ORCID ID: 0000-0003-4842-2184

and the proposed procedures are applied to a dataset from a clinical trial on advanced colorectal cancer.

Key words and phrases: Probability index, Predictive classification, Wilcoxon-type statistics, Bootstrap, U-processes, Chernoff's distribution

1. Introduction

In clinical medicine, after assessing treatment effects among all enrolled patients, subgroup analyses based on clinical variables or biomarkers (such as age and blood pressure) are frequently used to identify subsets of patients who may benefit more from a new treatment than others. Known as predictive classification in the clinical literature (Ballman, 2015), this type of analyses evaluates how biomarkers interact with treatments in terms of some clinical outcome, or in other words, the heterogeneity of treatment effects across subgroups. It plays a crucial role in customizing treatment strategies to optimize clinical benefits of patients. For example, in the CCTG CO.17 trial (Karapetis et al., 2008), cetuximab significantly improved survival for patients with wild-type Kras compared to best supportive care, while there was no significant difference between the two treatments for patients with mutated K-ras. This highly significant interaction led to the recommendation that cetuximab be reserved for those with wild-type K-ras.

Traditional analyses often rely on regression models or mean treatment effects as a measure of comparison (Gavanji et al., 2017; Li et al., 2023a,b). While straightforward, mean-based measures may fail to capture the complexities of ordered outcomes, where numerical differences do not imply proportional changes—for example, a response of 2 on an ordinal scale is not necessarily twice the severity or benefit of a response of 1. Moreover, mean-based metrics are sensitive to outliers and may fail to account for certain aspects of distributional characteristics like variance or skewness, potentially distorting treatment effect interpretation. In this context, the probabilistic index (PI) emerges as a more robust and informative alternative (Birnbaum, 1956; Sen, 1967; Kotz et al., 2003). As pointed out by Patel and Hoel (1973), this measure is easily interpreted and does not require a linear model assumption. Unlike mean treatment effects, the PI can be effectively applied to ordered responses and provides insights into the likelihood of deviations between treatments. In addition, it is resistant to the influence of outliers, making it a valuable tool in clinical studies.

In this paper, we focus on subgroups formed by dichotomizing a biomarker using a cutpoint. Specifically, let $(X_i, Y_i, Z_i), i \in [n] := \{1, ..., n\}$ be independent and identically distributed, where for subject i, Z_i is the biomarker with a distribution function F, and X_i and Y_i are, respectively, the clinical

outcomes under the control and experimental treatments. Given a cutpoint c, it divides the population into biomarker-positive ($Z > c$) and biomarker-negative ($Z \leq c$) subgroups. To assess treatment-biomarker interaction, we propose a nonparametric measure based on the probabilistic index:

$$\theta_c := \Pr\left(X_1 \leq Y_2 | Z_1 > c, Z_2 > c\right) - \Pr\left(X_1 \leq Y_2 | Z_1 \leq c, Z_2 \leq c\right). \quad (1.1)$$

The term $|\theta_c|$ captures the difference in relative treatment effects within each subgroup as defined by c. If $\theta_c = 0$, then there is no difference in treatment effects, whereas a larger $|\theta_c|$ implies stronger predictive effects of the biomarker; see Section 2 for more discussions.

We consider two scenarios: one with a prespecified cutpoint and one without. In the first scenario, the cutpoint, say c_0, is prespecified based on existing subject knowledge or prior medical research; then, our focus is on evaluating the difference in treatment effects between the two subgroups defined by this prespecified cutpoint c_0 by testing the hypothesis:

$$H_0 : \theta_{c_0} = 0 \qquad v.s. \qquad H_1 : \theta_{c_0} \neq 0, \quad (1.2)$$

to which we refer as the problem with prespecified cutpoint.

In practice, however, the cutpoint is often unknown, and in such situations, the initial step would be testing whether there is any difference in the relative treatment effects between biomarker-positive and biomarker-

negative groups across all potential cutpoint values in a range. Specifically, we consider the following hypotheses:

$$H_0 : \theta_c = 0 \text{ for all } c \in [\ell, u] \quad v.s. \quad H_1 : \theta_c \neq 0 \text{ for some } c \in [\ell, u], \quad (1.3)$$

where $\ell < u$ are user-specified constants and we assume $0 < F(\ell) < F(u) < 1$. Under this assumption, the conditional probabilities in (1.1) are well defined and estimable. If the null hypothesis is rejected, we also want to estimate the cutpoint value that maximizes the predictive effects, that is, the quantity c_b defined as follows:

$$c_b := \arg\max_{c \in [\ell, u]} |\theta_c|. \quad (1.4)$$

We refer to the second scenario as the problem with unknown cutpoint.

Our contributions in this paper are as follows. First, we propose using the probabilistic index in predictive classification to compare treatment benefits between two groups defined by a biomarker cutpoint, whether prespecified or not. This nonparametric measure (1.1) captures gradual changes in predictive effects, remains robust to outliers, and offers an intuitive interpretation. While most existing methods focus on mean shifts or regression models for detecting optimal cutoffs, to the best of our knowledge, our work is the first to use a probabilistic-level index for change-point testing. This approach is suitable for both ordinal and continuous outcomes, avoids

restrictive distributional assumptions, and provides a novel perspective in predictive classification.

Second, for both the prespecified and unknown cutpoint scenarios, we develop test statistics based on Wilcoxon-type statistics (Wilcoxon, 1945; Mann and Whitney, 1947) and derive their asymptotic null distributions. For situations with unknown cutpoint, we introduce a bootstrap procedure to calculate the p-value of the test because of the complicated structure in the covariance of the asymptotic null distribution. We establish the consistency of proposed procedures using techniques from the literature on U-processes. Our simulation studies highlight the robustness of our method, in handling extreme values and non-linearly generated data.

Third, we provide an estimate for the optimal cutpoint c_b in (1.4) using the profile method (Faraggi and Simon, 1996; Jonker et al., 2013; Li et al., 2023b) when the null hypothesis of no interaction is rejected by the data in the case of unknown cutpoint. We show that the estimator has cubic-rate convergence and asymptotically follows a scaled Chernoff's distribution (Groeneboom and Wellner, 2001). It is well-known that for Chernoff-type asymptotic distributional approximation, the standard nonparametric bootstrap is invalid, and that one theoretically sound remedy is using subsampling, such as the m-out-of-n bootstrap (Koziol and Jia,

2009; Seijo and Sen, 2011; Cattaneo et al., 2020). Typically, these methods estimate the quantiles of the limiting distribution through subsampling to construct confidence intervals. However, our simulation results indicate that the m-out-of-n bootstrap has a poor size control even for large samples. To address this issue, we introduce an m-out-of-n bootstrap to estimate the scaling factor via the first absolute moment, with a data-driven algorithm to select the optimal size m. The resulting confidence intervals for c_b are demonstrated to achieve coverage probabilities close to nominal levels.

The rest of the paper is structured as follows. We first review related work. In Section 2, we introduce an estimator for the nonparametric measure of treatment-biomarker interaction in (1.1). Sections 3 and 4 address prespecified and unknown cutpoints, respectively. Section 5 presents a profile method for optimal cutpoint estimation and a bootstrap procedure for confidence intervals. Section 6 reports simulation results, and Section 7 applies the method to a colorectal cancer clinical trial dataset. Proofs are in the Appendix.

1.1 Review of Related Work

We next review the literature on predictive classification with a single biomarker cutpoint. Typical approaches use regression models with treat-

ment, biomarker group, and their interaction as covariates. For prespecified cutpoints, tests like the Wald and score tests (Mccullagh and Nelder, 1989) can assess interaction effects. For unknown cutpoints, methods like the minimum p-value (Faraggi and Simon, 1996; Jonker et al., 2013; Li et al., 2023b) and profile methods (Luo and Boyett, 1997; Pons, 2003; Jiang et al., 2007; Gavanji et al., 2017) are used, but treating the estimated cutpoint as exogenous can inflate type I errors due to multiple testing (Hilsenbeck et al., 1992; Faraggi and Simon, 1996; Hilsenbeck and Clark, 1996; Mazumdar and Glassman, 2000). To address these issues, resampling methods like paired bootstrap and multiplier residual bootstrap for linear models (Li et al., 2023b), m-out-of-n bootstrap for generalized linear models (Li et al., 2023a), and permutation and residual bootstrap for Cox models (Jiang et al., 2007; Gavanji et al., 2017) have been proposed. These approaches, while useful, are designed for specific models and abrupt changes. They are prone to model mis-specification and less effective at capturing gradual changes or handling outliers.

We then discuss subsampling methods, specifically the m-out-of-n bootstrap (Bickel et al., 1997), which generates bootstrap samples of size $m = o(n)$ as a modification to the standard nonparametric bootstrap that uses $m = n$. This approach is particularly useful for estimators with a cube-root

convergence rate (Cattaneo et al., 2020). Examples include the isotonic density estimator (Grenander, 1956), the least median of squares estimator (Kim and Pollard, 1990), and the maximum score estimator (Manski, 1975). The asymptotic validity of the m-out-of-n bootstrap has also been established for non-smooth estimation problems, including linear, survival, and generalized linear models with cutpoints (Seijo and Sen, 2011; Xu et al., 2014; Li et al., 2023a).

We should also mention the rich literature on the probabilistic index, which is becoming increasingly common in clinical medicine for its easy interpretation (Moser and McCann, 2008; Jiang et al., 2016; Péron et al., 2016). Various inference methods for this index have been developed, including the Wilcoxon-Mann-Whitney statistic (Wilcoxon, 1945; Mann and Whitney, 1947), kernel smoothing (Baklizi and Eidous, 2006), and density ratio methods (Fokianos and Troendle, 2007). Additionally, Patel and Hoel (1973) introduced a nonparametric measure on the interaction between treatment and a binary covariate B based on $\Pr(X_1 \leq Y_2|B_1 = 1, B_2 = 1) - \Pr(X_1 \leq Y_2|B_1 = 0, B_2 = 0)$, where 0 and 1 represent the two levels of B. In the current paper, we extend this framework to scenarios where the binary covariate is defined by a single cutpoint for a biomarker.

2. Nonparametric Measure of Treatment-Biomarker Interaction

Recall that for subject $i \in [n]$, Z_i is the biomarker with a distribution function F, and X_i and Y_i are, respectively, the clinical outcomes under the control and experimental treatments. Further, different subjects are assumed to be independent and identically distributed (i.i.d.).

In clinical trials, each subject is randomly assigned to one of the treatment groups, and the corresponding outcome is observed. Specifically, for $i \in [n]$, denote by $U_i \in \{0, 1\}$ the group assignment for subject i, and T_i the realized outcome. Thus, $U_i = 0$ (resp. $U_i = 1$) indicates that the subject belongs to the control (resp. experimental) group, and $T_i = (1 - U_i)X_i + U_iY_i$. The observed data is the i.i.d. triplets, $(U_i, T_i, Z_i), i \in [n]$. We denote by (U, T, Z) a generic random vector with the same distribution, and assume that the following holds throughout the paper.

(*C.0*) Assume that $0 < F(\ell) < F(u) < 1$, $\lambda := E[U] \in (0, 1)$, and that U and Z_1 are independent.

Remark 1. Here, $\lambda = E(U) \in (0, 1)$ is constant. We restrict the cutpoint to $[\ell, u]$ to exclude extreme values that would create severe imbalance. In practice, they may be taken as non-extreme empirical quantiles of Z.

Remark 2. For $c \in [\ell, u]$, the probabilistic measure θ_c in (1.1) have the following representation. For $k \in \{0, 1\}$, and $t \in \mathbb{R}$, define

$$\begin{aligned} G^{(+)}(t|c,k) &:= \Pr(T \le t | Z > c, U = k), \\ G^{(-)}(t|c,k) &:= \Pr(T \le t | Z \le c, U = k). \end{aligned} \tag{2.5}$$

Thus $G^{(+)}(t|c,0)$ and $G^{(+)}(t|c,1)$ are, respectively, the distribution functions of X_i and Y_i given that $Z_i > c$, and similarly, $G^{(-)}(t|c,0)$ and $G^{(-)}(t|c,1)$ are the distribution functions of X_i and Y_i given that $Z_i \le c$. Then

$$\theta_c = \int G^{(+)}(t|c,0) dG^{(+)}(t|c,1) - \int G^{(-)}(t|c,0) dG^{(-)}(t|c,1).$$

Based on the treatment group and whether the biomarker is larger than c or not, the population is divided into four sub-populations, and θ_c is a function of their distributions as shown above.

For each $c \in [\ell, u]$, θ_c in (1.1) is identifiable. Specifically, due to assumption $(C.0)$ (in particular, U_1 and Z_1 are independent), we have

$$\Pr\left(X_1 \le Y_2 | Z_1 > c, Z_2 > c\right) = \Pr\left(T_1 \le T_2 | Z_1 > c, Z_2 > c, U_1 = 0, U_2 = 1\right),$$

and the same is true if we replace $Z_1 > c, Z_2 > c$ by $Z_1 \le c, Z_2 \le c$. This observation also motivates the following modified Wilcoxon-Mann-Whitney statistic (Mann and Whitney, 1947):

$$\hat{\theta}_{n,c} = \frac{\sum_{i=1}^n \sum_{j=1}^n T_{i,j}(1-U_i)U_j Z_i^{c+} Z_j^{c+}}{\sum_{i=1}^n (1-U_i) Z_i^{c+} \sum_{i=1}^n U_i Z_i^{c+}} - \frac{\sum_{i=1}^n \sum_{j=1}^n T_{i,j}(1-U_i)U_j Z_i^{c-} Z_j^{c-}}{\sum_{i=1}^n (1-U_i) Z_i^{c-} \sum_{i=1}^n U_i Z_i^{c-}}, \tag{2.6}$$

where $T_{i,j} = I(T_i \leq T_j)$ is the comparison indicator, $Z_i^{c+} = I(Z_i > c)$ and $Z_i^{c-} = I(Z_i \leq c)$ are the subgroup indicators determined by c, and $I(\cdot)$ is the indicator function. In the numerator of the first (resp. second) term on the right-hand-side above, we focus on those patients with a biomarker value $> c$ (resp. $\leq c$), and compare the outcomes from treatment and control within this subgroup. On the other, the terms in the denominators count the number of treatment and control cases within each subgroup.

Next, we develop statistical inference methods for both cases of prespecified and unknown cutpoint based on the above statistics.

3. Testing Interaction with a Prespecified Cutpoint

In this section, we focus on the case in which the cutpoint is prespecified as $c_0 \in [\ell, u]$, and we consider the hypothesis testing problem in (1.2) using $\hat{\theta}_{n,c_0}$ in (2.6) as the test statistic. To quantify the uncertainty of the estimator, we derive the limiting distribution in Theorem 1.

Recall $G^{(+)}(t|c,k)$ and $G^{(-)}(t|c,k)$ in (2.5). Denote by $G^{(\tau)}(t-|c,k) := \lim_{s\uparrow t} G^{(\tau)}(s|c,k)$ the left limit for $t \in \mathbb{R}$ and $\tau \in \{+,-\}$ and by $\mathcal{N}(0,\sigma^2)$ the normal distribution with zero mean and variance σ^2.

Theorem 1. *Suppose that condition (C.0) holds, and that for* $\tau \in \{+, -\}$,

$$Var\left(G^{(\tau)}(X_{-}|c_0, 1)\right) + Var\left(G^{(\tau)}(Y|c_0, 0)\right) > 0. \tag{3.7}$$

For some finite and positive quantity $\sigma^2_{c_0}$ *defined in* (S6.3) *in Appendix S6,*

$$\sqrt{n}(\hat{\theta}_{n,c_0} - \theta_{c_0}) \rightsquigarrow \mathcal{N}(0, \sigma^2_{c_0}),$$

where the symbol "$\rightsquigarrow$*" represents convergence in distribution.*

Remark 3. Condition (3.7) guarantees that the limiting distribution has a positive variance.

The proofs in this section are in Appendix S6. Since $\sigma^2_{c_0}$ has a complicated dependence on the distribution of (T, U, Z), we propose to estimate $\sigma^2_{c_0}$ by the jackknife method (Shao and Tu, 1995). Specifically, define

$$\hat{\sigma}^2_{n,c_0} = \sum_{k=1}^{n} \left\{ \left(\hat{\theta}^{-k}_{n,c_0} - \sum_{j=1}^{n} \hat{\theta}^{-j}_{n,c_0}/n \right)^2 \right\} (n-1), \tag{3.8}$$

where $\hat{\theta}^{-k}_{n,c_0}$ is the estimate of θ_{c_0} based on the sample with the k-th observation left out. The following theorem shows the consistency of the jackknife estimator.

Theorem 2. *Let* $\sigma^2_{c_0}$ *be the finite positive constant in Theorem 1. Assume the conditions in Theorem 1 hold. Then, the jackknife estimator* (3.8) *is*

consistent, that is, $\hat{\sigma}^2_{n,c_0} \xrightarrow{P} \sigma^2_{c_0}$*, where the symbol "*$\xrightarrow{P}$*" indicates convergence in probability. Immediately, it follows that*

$$\frac{\sqrt{n}(\hat{\theta}_{n,c_0} - \theta_{c_0})}{\sqrt{\hat{\sigma}^2_{n,c_0}}} \rightsquigarrow \mathcal{N}(0,1).$$

We can then test whether the treatment effects differ in the two subgroups defined by c_0, i.e., the problem in (1.2). Specifically, the p-value of this test is defined as $2\left[1-\Phi\left(\left|\sqrt{n}\hat{\theta}_{n,c_0}/\sqrt{\hat{\sigma}^2_{n,c_0}}\right|\right)\right]$, where Φ is the distribution function of the standard normal distribution. H_0 is rejected if the p-value is smaller than some pre-specified significant level $\alpha \in (0,1)$. By Theorem 2, this procedure has an asymptotic size α.

4. Testing Interaction with an Unknown Cutpoint

Testing based on a prespecified cutpoint value can only determine whether there is a difference in treatment effects between the subgroups defined by this particular value. However, as previously mentioned, in many practical scenarios when the cutpoint is unknown, it is common to inquire about the existence of predictive effects across all potential cutpoint values. In this section, we focus on the hypothesis testing problem in (1.3).

Recall the definition of c_b in (1.4). Clearly, the null H_0 in (1.3) is equivalent to the statement that $\sup_{c\in[\ell,u]}|\theta_c| = 0 = |\theta_{c_b}|$. Thus we propose

to reject the null when the following test statistic is large:

$$\hat{S}_n := \sqrt{n} \sup_{c \in [\ell, u]} |\hat{\theta}_{n,c}| = \sqrt{n} |\hat{\theta}_{n,\hat{c}_n}|. \tag{4.9}$$

where

$$\hat{c}_n := \arg\max_{c \in [\ell, u]} |\hat{\theta}_{n,c}|, \tag{4.10}$$

and if there are multiple $c \in [\ell, u]$ achieving the same maximum value, we select the minimum one.

Remark 4. In computing $\hat{S}_n$ in (4.9) and $\hat{c}_n$ in (4.10), it suffices to take the sup and arg max over the distinct realized values of $Z_1, \ldots, Z_n$ in $[\ell, u]$, that is, those $c \in C_n := \{Z_1, \ldots, Z_n\} \cap [\ell, u]$, because $|\hat{\theta}_{n,c}|$ is stepwise constant and changes only when c crosses an observed biomarker value.

On one hand, $\hat{S}_n$ may be viewed as the supremum of the test statistics for each single $c \in [\ell, u]$. On the other, it may be viewed as the test statistic corresponding to a single cutpoint value, $\hat{c}_n$, which is an estimator for c_b. One may be tempted to treat $\hat{c}_n$ as deterministic, and calculate a p-value by Theorem 2. However, this approach suffers from serious type I error inflation due to multiple comparisons.

Next, we derive the limiting null distribution of the stochastic process $\{\sqrt{n}\hat{\theta}_{n,c} : c \in [\ell, u]\}$ and the test statistic $\hat{S}_n$. Denote by $\ell^\infty([\ell, u])$ the space of bounded functions on $[\ell, u]$ equipped with the ℓ_∞-norm.

Theorem 3. *Suppose that condition (C.0) holds and that the null* H_0 *in* (1.3) *holds. Then there exists a tight centered Gaussian process* $\mathbf{G} = \{G_c : c \in [\ell, u]\}$ *such that* $\{\sqrt{n}\hat{\theta}_{n,c} : c \in [\ell, u]\}$ *converges weakly in* $\ell^\infty([\ell, u])$ *to* $\mathbf{G}$. *As a result,*

$$\hat{S}_n \rightsquigarrow \sup_{c\in[\ell,u]} |G_c|.$$

The proof of Theorem 3 can be found in Appendix S7, where we decompose $\{\sqrt{n}\hat{\theta}_{n,c} : c \in [\ell, u]\}$ into ratios of U-processes, and apply U-process techniques (Peña and Giné, 1999) and the functional Delta method (Van der Vaart and Wellner, 1996). For more discussions, see Appendix S8.4.

4.1 Bootstrap Procedure

As the Gaussian process $\mathbf{G}$ has an intricate covariance structure, in this subsection, our goal is to develop a valid test by calibrating the distribution of $\hat{S}_n$. To achieve this, we propose the following bootstrap method.

(i) Calculate the maximally selected test statistic $\hat{S}_n$ (defined in (4.9)).

(ii) Let $\{(T_i^*, U_i^*, Z_i^*), i \in [n]\}$ be a bootstrap sample created by sampling with replacement from the data $\{(T_i, U_i, Z_i), i \in [n]\}$. Across $c \in [\ell, u]$, obtain the bootstrap version of the maximal test statistics $\hat{S}_n^*$:

$$\hat{S}_n^* = \sqrt{n} \max_{c\in[\ell,u]} |\hat{\theta}_{n,c}^* - \hat{\theta}_{n,c}|,$$

where $\hat{\theta}^*_{n,c}$ is calculated based on the bootstrap sample.

(iii) Finally, compute the adjusted p-value, that is,

$$p^*_{adjust} = 1 - F^*_n(\hat{S}_n),$$

where F^*_n denotes the distribution function of the bootstrap test statistic $\hat{S}^*_n$, conditional on the data $(T_i, U_i, Z_i), i \in [n]$.

In practice, we approximate the adjusted p-value by Monte Carlo simulations. Specifically, we generate bootstrap samples $\{(T^*_{i,b}, U^*_{i,b}, Z^*_{i,b}), i \in [n], b \in [B]\}$ by sampling with replacement from the data $\{(T_i, U_i, Z_i), i \in [n]\}$ for B times. For the b-th bootstrap sample, we compute the maximal test statistics $\hat{S}^*_{n,b}$ as above. Finally, we approximate the adjusted p-value as the proportion of $\{\hat{S}^*_{n,b} : b \in [B]\}$ greater than $\hat{S}_n$, that is, $B^{-1}\sum_{b=1}^{B} I(\hat{S}^*_{n,b} > \hat{S}_n)$.

The following theorem establishes the consistency of the bootstrap procedure outlined above, whose proof can be found in Appendix S7.1.

Theorem 4. *Suppose that condition (C.0) holds and that the null H_0 in (1.3) holds. Conditional on $(T_i, U_i, Z_i), i \geq 1$, for almost every sequence $(T_i, U_i, Z_i), i \geq 1$, the random process $\{\sqrt{n}(\hat{\theta}^*_{n,c} - \hat{\theta}_{n,c}) : c \in [\ell, u]\}$ converges weakly in $\ell^\infty([\ell, u])$ to the same Gaussian process $\mathbf{G}$ in Theorem 3. There-*

fore, for any significance level $\alpha \in (0,1)$*, we have under the null hypothesis,*

$$\lim_{n\to\infty} \Pr(p^*_{adjust} \le \alpha) = \alpha.$$

Remark 5. The above test procedure is applicable to other test statistics. For instance, we may replace $|\hat{\theta}_{n,c}|$ in (4.9) by $\left|\sqrt{n}\hat{\theta}_{n,c}/\sqrt{\hat{\sigma}^2_{n,c}}\right|$, where $\hat{\sigma}^2_{n,c}$ is defined in (3.8), with c_0 replaced by c. We refer to this approach as the minimum p-value method. Upon rejecting the null hypothesis, the "optimal" cutpoint is estimated by $\tilde{c}_n$, which is the value that yields the lowest among all calculated p-values according to Theorem 2, i.e.,

$$\tilde{c}_n = \arg\min_{c\in[\ell,u]} 2\left[1 - \Phi\left(\left|\frac{\sqrt{n}\hat{\theta}_{n,c}}{\sqrt{\hat{\sigma}^2_{n,c}}}\right|\right)\right].$$

Because the two approaches are similar, we omit detailed discussion and provide simulation results in Appendix S3. Both methods are asymptotically valid, though simulations (Section 6.2) show that the minimum p-value method is less stable and typically has lower power in finite samples.

5. Profile Estimation of the Optimal Cutpoint

If the null hypothesis in (1.3) is rejected, we often need to estimate the optimal cutpoint which maximizes the predictive effects, that is, c_b defined in (1.4). In this section, we assume that the null hypothesis in (1.3) does

not hold, or equivalently $\theta_{c_b} \neq 0$, and aim to estimate and construct a confidence for c_b.

Recall that $|\hat{\theta}_{n,c}|$ represents the absolute difference in relative treatment effects between the biomarker-positive and biomarker-negative groups defined by the value c. A larger value of $|\hat{\theta}_{n,c}|$ indicates a greater potential for benefits in one of these subgroups. Therefore, we propose to estimate c by the profile estimator $\hat{c}_n$ defined in (4.10).

First, we establish the consistency and derive the convergence rate for the profile estimator $\hat{c}_n$ under the following assumptions.

(*C.1*) Assume that c_b in (1.4) is a well-separated maximizer of the function $c \mapsto |\theta_c|$ on $[\ell, u]$, that is, for any $\epsilon > 0$,

$$\inf\{|\theta_{c_b}| - |\theta_c| : |c - c_b| \geq \epsilon,\ c \in [\ell, u]\} > 0.$$

(*C.2*) Assume that the function $c \mapsto \theta_c$ is continuous on $[\ell, u]$. Further, there exist $\zeta > 0$ and $\kappa > 0$ such that if $|c - c_b| \leq \zeta$ and $c \in [\ell, u]$, $|\theta_c| - |\theta_{c_b}| \leq -\kappa(c - c_b)^2$, and that the function $c \mapsto F(c)$ is continuously differentiable on $[c_b - \zeta, c_b + \zeta]$.

Remark 6. Assumption (*C.1*) ensures that the optimal cutpoint c_b is well separated, so ties do not occur asymptotically. In finite samples, tie-breaking (e.g., as in (4.10)) may be used, and when several cutpoints are

nearly optimal, secondary criteria such as interior preference, lower variability, or biological guidance may be incorporated.

Remark 7. If $c \mapsto \theta_c$ is continuous on $[\ell, u]$, the unique maximizer over the compact set is well-separated. Assumption (*C.2*) does not require F to be continuous on $\mathbb{R}$, but only near c_b, permitting discreteness elsewhere. The finitely discrete case will be treated in Appendix S8.6. Further, see Remark 8 when the function $c \mapsto \theta_c$ is not continuous at c_b.

Theorem 5. *Suppose that conditions (C.0) and (C.1) hold. Then* $\hat{c}_n \xrightarrow{P} c_b$ *as* $n \to \infty$. *If, in addition, condition (C.2) holds, then* $n^{1/3}|\hat{c}_n - c_b|$ *is bounded in probability as* $n \to \infty$.

The proof of the above theorem can be found in Appendix S8, where we also explain why this rate differs from the usual parametric rate $n^{-1/2}$. For the cubic convergence rate (i.e., $n^{-1/3}$), we use a modified "peeling device"; see, for example, (Van der Vaart and Wellner, 1996, Theorem 3.2.5).

Remark 8. If $c \mapsto \theta_c$ is twice differentiable, then c_b being a maximizer implies $\theta'_{c_b} = 0$ and $\theta''_{c_b} \leq 0$. If $\theta''_{c_b} < 0$, then for c close to c_b we have $|\theta_c| - |\theta_{c_b}| \leq -\kappa(c - c_b)^2$ for some $\kappa > 0$, which corresponds to the second part of Assumption (*C.2*).

If θ_c is not smooth at c_b, then $\hat{c}_n$ converges to c_b at rate n^{-1}, using

arguments parallel to Seijo and Sen (2011), Xu et al. (2014), and Li et al. (2023a). A formal statement and proof are provided in Appendix S8.

Next, we derive the limiting distribution of the profile estimator $\hat{c}_n$. Recall the definition of $G^{(+)}(\cdot)$ and $G^{(-)}(\cdot)$ in (2.5). Denote by $\mathcal{L}_z^{(1)}$ and $\mathcal{L}_z^{(2)}$ the conditional distribution of X and Y, respectively, given $Z = z$; they are well defined due to the existence of regular conditional distribution (Durrett, 2019, Section 4.1.3). That is, for any Borel $A \subset \mathbb{R}$, $\mathcal{L}_z^{(1)}(A) = \Pr(X \in A | Z = z)$ and $\mathcal{L}_z^{(2)}(A) = \Pr(Y \in A | Z = z)$. We impose the following assumption.

(*C.3*) Assume that $c_b \in (\ell, u)$, and that the function $c \mapsto \theta_c$ is twice differentiable at c_b with the second derivative $\theta''_{c_b} \neq 0$. For any sequence of real numbers $z_n \to c_b$ as $n \to \infty$, $\mathcal{L}_{z_n}^{(i)} \rightsquigarrow \mathcal{L}_{c_b}^{(i)}$ for $i = 1, 2$. Further, for $\tau \in \{+, -\}$, $k \in \{0, 1\}$, the function $t \mapsto G^{(\tau)}(t | c_b, k)$ is continuous.

Theorem 6. *Suppose that conditions (C.0), (C.1), (C.2) and (C.3) hold. Then we have*

$$n^{1/3} (\hat{c}_n - c_b) \rightsquigarrow \left(2\nu / \left|\theta''_{c_b}\right|\right)^{2/3} \mathbb{C},$$

where ν is the scalar defined in (S8.15) *of Appendix S8.3, θ''_{c_b} is the second derivatives of $c \mapsto \theta_c$ at c_b, and the random variable $\mathbb{C}$ follows the Chernoff's distribution.*

The proof of Theorem 6 can be found in Appendix S8.3. Chernoff's distribution (Chernoff, 1964) is defined as the distribution of the maximizer of the two-sided Brownian motion minus a parabola. Specifically, if we denote by $\{\mathbb{Z}(t) : |t| < \infty\}$ a standard two-sided Brownian motion with $\mathbb{Z}(0) = 0$, then $\arg\max_t \{\mathbb{Z}(t) - t^2\}$ has the Chernoff's distribution. For statistical applications and properties of Chernoff's distribution, including its quantiles and the first few absolute moments, see Groeneboom and Wellner (2001).

In the next subsection, we discuss the construction of confidence intervals for c_b based on the above theorem.

5.1 Confidence intervals for the optimal cutpoint

By Theorem 6, the limiting distribution of $n^{1/3}(\hat{c}_n - c_b)$ is a scaled Chernoff's distribution. The scaling factor has a complicated dependence on the distribution of (T, U, Z). In particular, it is challenging to estimate the second derivative θ''_{c_b} of the function $c \mapsto \theta_c$ at c_b. Our experience from simulation studies indicates that even with a large sample size (e.g., when n is in the thousands), the confidence intervals constructed by directly estimating ν and θ''_{c_b} exhibit poor coverage probabilities.

One alternative approach involves using resampling methods. It is well known that the standard empirical bootstrap method would fail in this case

due to nonstandard asymptotics with the cube-root convergence rate (Cattaneo et al., 2020). However, as highlighted by Koziol and Jia (2009) and Seijo and Sen (2011), some subsampling methods, such as the m-out-of-n bootstrap method (Bickel et al., 1997) and subsampling without replacement (Politis and Romano, 1994), remain valid. In general, these methods construct confidence intervals by estimating the $\alpha/2$ and $1-\alpha/2$ quantiles of the limiting distribution through subsampling, where $\alpha \in (0,1)$ is the nominal test level. Although theoretically sound, the m-out-of-n bootstrap suffers from serious under-coverage in our simulation studies, even for a large sample size (see Table 3).

5.1.1 First absolute moment m-out-of-n bootstrap

Our proposal is to estimate the unknown scaling factor by subsampling methods. Specifically, we use the m-out-of-n bootstrap to estimate the first absolute moment of the limiting distribution in Theorem 6, and we denote by $\hat{M}^*$ the estimator. Let $\bar{M}$ be the first absolute moment of Chernoff's distribution. Then we can match the two moments and estimate the unknown scaling factor $\left(\left|2\nu/\theta''_{c_b}\right|\right)^{2/3}$ by $\hat{M}^*/\bar{M}$. Finally, given the estimated scaling factor, we construct confidence intervals for c_b using the quantiles of Chernoff's distribution. The procedure is described below, where m_n and

B are user specified integers and $\alpha \in (0,1)$ is the nominal level of the test.

1. Generate bootstrap data $\{(T^*_{i,b}, U^*_{i,b}, Z^*_{i,b}), i \in [m_n], b \in [B]\}$ by sampling with replacement from the data $\{(T_i, U_i, Z_i), i \in [n]\}$ for B times. Across all c values, obtain the bootstrap version of the estimated optimal cutpoint:

$$\hat{c}^*_{m_n,b} = \arg\max_{c\in[\ell,u]} |\hat{\theta}^*_{m_n,c,b}|, \quad b = 1, 2, ..., B,$$

where $\hat{\theta}^*_{m_n,c,b}$ is calculated based on the b-th bootstrap sample.

2. Compute the bootstrap estimate of the first absolute moment:

$$\hat{M}^* = B^{-1} \sum_{b=1}^{B} m_n^{1/3} \left| \hat{c}^*_{m_n,b} - \hat{c}_n \right|,$$

where $\hat{c}_n$ is the original estimator for c_b defined in (4.10).

3. Let $\bar{M}$, $\mathbb{C}_{\alpha/2}$ and $\mathbb{C}_{1-\alpha/2}$ be the first absolute moment, $\alpha/2$-th quantile, and $(1-\alpha/2)$-th quantile of Chernoff's distribution, which can be found in Groeneboom and Wellner (2001). We construct the $100(1-\alpha)\%$ confidence interval for c_b as follows:

$$\left[\hat{c}_n - \mathbb{C}_{1-\alpha/2} \frac{\hat{M}^*}{\bar{M}} \left(\frac{1}{n}\right)^{1/3}, \hat{c}_n - \mathbb{C}_{\alpha/2} \frac{\hat{M}^*}{\bar{M}} \left(\frac{1}{n}\right)^{1/3} \right].$$

5.1.2 Choice of size m

In the literature, m_n is commonly chosen by a data-driven rule. We adapt the approach of Bickel and Sakov (2008), selecting m from candidate values

as the largest sample size at which the estimated first absolute moment stabilizes. The algorithm is as follows:

Consider a sequence of m's of the form $m_j = \lfloor q^j n \rfloor$ for $j = 1, 2, ...,$ and $q \in (0, 1)$. For each m_j-out-of-n bootstrap, compute $\hat{c}^*_{m_j,b}$ for b-th replication, where $b = 1, 2, ..., B$ and B is the total number of bootstrap replications. Compute the bootstrap estimate of the first absolute moment

$$\hat{M}^*_{m_j} = \frac{1}{B} \sum_{b=1}^{B} m_j^{1/3} |\hat{c}^*_{m_j,b} - \hat{c}_n|,$$

where $\hat{c}_n$ is the profile estimate based on the original dataset. The m will be selected as the value that minimizes the difference between two adjacent first absolute moment estimates:

$$m = \text{argmin}_{m_j} |\hat{M}^*_{m_j} - \hat{M}^*_{m_{j+1}}|,$$

where in the case of ties, we select the largest one. We refer to this algorithm as Method S1 and introduce another one (Method S2), which selects m to achieve a stable empirical distribution, as described in Appendix S1. In Section 6.3, our proposed procedure performs favourably in simulation studies against the standard m-out-of-n method. One possible explanation is that it is easier to estimate the first absolute moment of the limiting distribution in Theorem 6 than its quantiles.

6. Simulation Study in the Case of Unknown Cutpoint

In this simulation study, we focus on the unknown cutpoint problem. In Subsection 6.2, we assess the empirical size and power of the test proposed in Section 4.1 under various scenarios. Further, in Subsection 6.3, we report the coverage probability of the procedure proposed in Subsection 5.1. Additional simulation results on the bias and standard error of the proposed estimator in (4.10), as well as findings related to the problem with prespecified cutpoint, are given in Appendix S1 and Appendix S2, respectively.

6.1 Simulation Setups

We generate the biomarker Z from the uniform distribution on the interval $(0, 1)$, and the group indicator U from a Bernoulli distribution with a success probability 0.5. Further, given the biomarker Z, we generate the outcomes X and Y from the conditional distributions listed in Table 1. Finally, the observed outcome $T = (1 - U)X + UY$.

Cases 1–5 in Table 1 belong to the null hypothesis in (1.3), i.e., there is no predictive effects. Case 1 considers a scenario with no treatment (U) or biomarker (Z) effect. There is only treatment effect in Case 2 and only biomarker effect in Case 3, while there are both treatment and biomarker effects in Case 4 and 5. In Case 5, the biomarker effect has a jump at

0.5, while in Case 4, the biomarker effect is smooth. Further, in Case 5, the potential outcomes have heavy tail distributions. Cases 6–9 in Table 1 consider scenarios under alternative hypothesis. For each case, the optimal cutpoint c_b and the corresponding predictive effect $|\theta_{c_b}|$ are listed defined in (1.4). For each configuration, the results are based on 1000 repetitions.

6.2 Simulation Results: Size and Power of Tests

For the hypothesis testing problem in (1.3), we compare the performance of the following three procedures.

- **(PB)**: Our proposed bootstrap test introduced in Section 4.1;
- **(WB)**: This method treats the estimated optimal cutpoint $\hat{c}_n$ as given and obtains the p-value from the test by Theorem 2 without bootstrap adjustment;
- **(LR)**: The linear regression method proposed by Li et al. (2023b). Specifically, with an unknown cutpoint c_0, they assume

$$E(T) = \alpha_0 + \beta_0 U + \gamma_0 I(Z \le c_0) + \lambda_0 I(Z \le c_0) U, \tag{6.11}$$

where recall that $I(\cdot)$ is the indicator function, and α_0, β_0, γ_0, λ_0 and c_0 are unknown parameters. Under this model, the significance of the differential treatment effect can be assessed by testing the null

Table 1: Simulation setups under the null and alternative hypothesis. $\mathcal{LN}$ denotes the log-normal distribution with the mean and scale parameters, $\mathcal{E}$ denotes the exponential distribution with by the rate parameter, and ϵ has Cauchy distribution with location being 2 and scale 1.

	X	Y	c_b	$\lvert\theta_{c_b}\rvert$
Case 1	$\mathcal{LN}$(1.1,1)	$\mathcal{LN}$(1.1,1)		
Case 2	$\mathcal{LN}$(1.1,1)	$\mathcal{LN}$(1.7,1)		
Case 3	$\mathcal{E}$(exp(10×(Z-0.5)))	$\mathcal{E}$(exp(10×(Z-0.5)))		
Case 4	$\mathcal{E}(Z^2$+0.1)	$\mathcal{E}$(2×(Z^2+0.1))		
Case 5	$2 + I(Z \leq 0.5) + \epsilon$	$3.5 + I(Z \leq 0.5) + \epsilon$		
Case 6a	$\mathcal{E}$(0.6-0.1×I(Z> 0.5))	$\mathcal{E}$(0.4+0.1×I(Z> 0.5))	0.5	0.1
Case 6b	$\mathcal{E}$(0.6-0.1×I(Z> 0.7))	$\mathcal{E}$(0.4+0.1×I(Z> 0.7))	0.7	0.1
Case 7a	$\mathcal{E}$(0.6+0.3×I(Z> 0.5))	$\mathcal{E}$(0.4-0.3×I(Z> 0.5))	0.5	0.3
Case 7b	$\mathcal{E}$(0.6+0.3×I(Z> 0.7))	$\mathcal{E}$(0.4-0.3×I(Z> 0.7))	0.7	0.3
Case 8	$\mathcal{E}$(1)	$\mathcal{E}$(exp(10×(Z-0.5)))	0.5	0.7
Case 9	$\mathcal{E}$(exp(Z))	$\mathcal{E}$(exp(8×(Z-0.45)))	0.52	0.6

hypothesis $H_0 : \lambda_0 = 0$. With c_0 unknown, Li et al. (2023b) first uses the minimum p-value method for the estimation of cutpoint, and then proposes a paired bootstrap method to adjust the p-value.

For methods (PB) and (LR), the number of bootstrap replications is 1000. Two different sample sizes ($n = 300, 500$) are considered. We set $\ell = 0.2, u = 0.8$.

Table 2 summarizes the empirical size and power of the three methods under the null and alternative hypothesis for tests at significance level $\alpha = 0.05$. We observe that for all cases under the null hypothesis (i.e. Case 1–5), the empirical sizes of the proposed method (PB) is close to the nominal 5% level, while there is an almost sevenfold inflation when using method (WB), which treats the estimated $\hat{c}_n$ as deterministic. Moreover, method (LR) exhibits a type I error inflation for Cases 3-4 and is overly conservative for Case 5. Specifically, in Case 3 and 4, the responses are non-linearly generated, and (LR) suffers from model mis-specification. In case 5, the outcomes have heavy tail distributions, which affects the performance of (LR). In comparison, the proposed method (PB) is robust to both model specification and heavy tail distributions.

For empirical power analysis, as summarized in Table 2, all methods demonstrate similar performance, with an overall increasing trend observed as the sample size n and the predictive effect $|\theta_{c_b}|$ increase. In particular, for a sample of size $n = 300$ with $|\theta_{c_b}| \geq 0.3$, the power of the tests exceeds 90% for all methods in Cases 7–9. However, for Case 6a, where $|\theta_{c_b}| = 0.1$ is

Table 2: The empirical size (in percentage) and power under null hypothesis and alternative hypothesis when the cutpoint is not given for significance level $\alpha = 0.05$. Configurations 1-5 focus on size, while configurations 6-9 are for empirical power.

n		1	2	3	4	5	6a	7a	8	9
300	PB	5.1	4.9	5.2	5.7	5.8	20.5	99.7	100	100
	WB	33.3	35.1	30.1	24.4	29.2	43.4	99.9	100	100
	LR	5.8	4.8	7.8	60.7	2.4	29.8	99.9	96.4	97.8
500	PB	5.3	5.1	4.9	6.5	4.4	28.9	100	100	100
	WB	33.3	33.1	29	24.1	24	59.8	100	100	100
	LR	6.3	4.6	7.6	85.5	2.6	48.9	100	99.6	99.5

small, the power does not exceed 60% even for a sample of size $n = 500$. As expected, our procedure is not as powerful as method (LR) when the model in (6.11) is correctly specified, illustrating the tradeoff between robustness and power.

6.3 Simulation Results: Estimation of c_b

For simplicity, we refer to the proposed method in Section 5.1 for constructing confidence intervals for c_b as FAM, which estimates the first absolute

moment of the limiting distribution in Theorem 6 by the m-out-of-n bootstrap method. We compare FAM with the standard empirical bootstrap and m-out-of-n bootstrap, which estimate directly the quantiles of the limiting distribution.

We consider Case 8 and 9 with 95% nominal coverage probability and with different sample sizes: $n = 500, 1000, 2000$, and 4000. Further, for m-out-of-n bootstrap methods, we select m_n to be one of the following: $n^{0.95}$, $n^{0.85}$ and our proposed data-driven procedures (S1 in Subsection 5.1 and S2 in Appendix S1). Each resampling method uses 1000 bootstrap repetitions.

Table 3 provides the estimated coverage probabilities of confidence intervals obtained using different procedures. We observe that our proposed FAM procedure based on m-out-of-n bootstrap have coverage probabilities close to the nominal 95% level, particularly when m is determined by our proposed data-driven algorithms. On the other, the standard empirical bootstrap, with or without FAM, suffers from under-coverage; this is consistent with previous findings (Koziol and Jia, 2009; Seijo and Sen, 2011; Cattaneo et al., 2020). It is interesting that the standard m-out-of-n bootstrap, which is theoretically consistent, have drastic under-coverage (varying between 0.60 and 0.80) even for $n = 4000$. Based on these results, we recommend adopting the FAM procedure, along with the proposed choice

Table 3: Empirical coverage probabilities (in percentage) of the 95% confidence interval constructed by different bootstrap methods under the alternative hypothesis. Moon: m-out-of-n method, FAM: our proposed first absolute moment method.

	Case 8				Case 9			
Sample Size	500	1000	2000	4000	500	1000	2000	4000
Empirical Bootstrap	63.2	65.1	65.1	65.9	60.1	63.4	66.0	67.2
Moon ($m_n = n^{0.95}$)	67.6	71.5	70.6	70.8	62.0	63.9	70.3	70.4
Moon ($m_n = n^{0.85}$)	75.3	76.5	75.3	76.8	63.0	69.3	76.1	77.2
FAM Bootstrap($m_n = n$)	89.3	89.7	89.6	91.0	90.0	90.7	88.9	90.9
FAM Moon ($m_n = n^{0.95}$)	92.0	94.4	92.7	94.2	91.8	93.1	92.6	94.3
FAM Moon ($m_n = n^{0.85}$)	96.2	96.0	96.7	96.6	94.1	95.4	97.0	96.9
FAM Moon (S1)	95.0	95.0	95.8	95.6	93.8	93.9	95.8	94.8
FAM Moon (S2)	95.1	95.0	95.3	95.3	93.5	94.0	95.3	95.1

of m procedure, for constructing confidence intervals for c_b.

7. Real Data Application

The proposed procedures are applied to a dataset from CO.17 trial carried out by the Canadian Cancer Trials Group In this trial, 572 patients diagnosed with advanced colorectal cancers were randomly assigned to two

groups: one receiving Cetuximab in addition to best supportive care (BSC), and the other receiving BSC alone.

In clinical trials of cancer patients, an important outcome of interest is Quality of Life (QoL), which provides information on the impact of treatments in alleviating symptoms and reducing treatment-related side effects from the perspective of patients. In CO.17, QoL was measured using the European Organization for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire (QLQ)-C30, and the primary aim of the QoL analysis was comparing changes in subscales of the questionnaire from baseline to specific time points following randomization between two groups.

In our analysis, we focus on the Physical Function Scale (PFS) with an interest in identifying specific subgroups of patients, based on certain biomarkers, who may experience more favorable changes from baseline at 4, 8, 16 and 24 weeks from randomization, in this subscale. Four biomarkers are considered: mRNA expression of the gene epiregulin (EREG), levels of lactate dehydrogenase (LDH), alkaline phosphatase (ALKPH), and hemoglobin (HGB). For each biomarker, we select ℓ and u as the 20% and 80% empirical quantiles.

We apply both the proposed bootstrap method introduced in Section 4.1 and minimum p-value methods as discussed in Remark 5 to test, sep-

Table 4: Estimated $\hat{c}_n$ and $\tilde{c}_n$ for HGB by profile method and minimum p-value method respectively, and the corresponding p-values based on the bootstrap method. The bootstrap repetition number is B=2000.

HGB									
Week	$\hat{c}_n$	$p_{profile}$	$\tilde{c}_n$	p_{mini}	Week	$\hat{c}_n$	$p_{profile}$	$\tilde{c}_n$	p_{mini}
4	112	0.192	110	0.230	8	110	0.007	110	0.005
16	119	0.662	119	0.646	24	118	0.087	118	0.083

arately for each biomarker, whether there is predictive effect, i.e., the hypotheses in (1.3). Each method uses 2000 bootstrap repetitions. Although estimation of optimal cutpoint is only meaningful when these tests are statistically significant, we report, in Table 4, the optimal cutpoint value estimated by the profile method, i.e., $\hat{c}_n$ in (4.10), and by the minimum p-value method, i.e., $\tilde{c}_n$ in Remark 5, along with the p-values of the tests, for all biomarkers.

For EREG, LDH, and ALKPH, as shown in Table S5 of Appendix S4, none of the p-values are smaller than 0.05, indicating that there was no cutpoint which would define subgroups with statistically significant different relative treatment effects. However, for HGB from baseline at 8 weeks, p-values from both profile and minimum p methods are smaller than 0.05

and the estimated optimal cutpoint value for change in PHS is 110 from both methods. As summarized in Figure 1 and Table S6 of Appendix S4, individuals with HGB lower than 110, when treated with cetuximab plus BSC, had a more positive change in PHS at 8 weeks compared to those treated with BSC alone. Conversely, the change in PHS was similar between the two treatment groups for patients with HGB equal to or higher than 110, suggesting that patients with lower HGB values would have more favorable QoL when treated by cetuximab in comparison with BSC than those with higher values of HGB.

8. Conclusion and Discussions

In this paper, nonparametric procedures for testing treatment-biomarker interaction are proposed based on the probability index in both cases when the cutpoint is prespecified or unknown. Inference procedures for the optimal cutpoint are also developed. Theoretical properties of the proposed procedures are derived and extensive simulation studies validate the robustness and accuracy of our proposed procedures.

The Mann-Whitney-based statistic proposed in this paper offers computational simplicity and robustness to distributional assumptions, though it may be less efficient than certain semiparametric alternatives. A promis-

ing direction is to develop hybrid approaches that balance robustness and efficiency (see Appendix S4.4). Another important extension involves adjusting for confounding variables, which frequently arise in practice. This could be achieved through a regression-based framework for modeling the conditional probability index (Thas et al., 2012; De Schryver and De Neve, 2019), or via machine-learning methods such as double/debiased learning (Chernozhukov et al., 2018; Mi et al., 2021) (see Appendix S4.3). Beyond confounding adjustment, extending our procedures to survival outcomes (Koziol and Jia, 2009; Jiang et al., 2016) and developing strategies for combining multiple biomarkers for subgroup identification (He et al., 2018) represent further promising research directions (see Appendix S4.2).

Supplementary Materials

The proofs and more simulation results are presented in the SMs.

References

Baklizi, A. and O. Eidous (2006). Nonparametric estimation of P(X<Y) using kernel methods. *Metron - International Journal of Statistics LXIV*, 47–60.

Ballman, K. V. (2015). Biomarker: Predictive or Prognostic? *Journal of Clinical Oncology 33*, 3968–3971.

Bickel, P. J., F. Götze, and W. R. van Zwet (1997). Resampling fewer than n observations: Gains,

losses, and remedies for losses. *Statistica Sinica 7*, 1–31.

Bickel, P. J. and A. Sakov (2008, 01). On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica 18*, 967–985.

Birnbaum, Z. W. (1956). On a Use of the Mann-Whitney Statistic. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics 3.1*, 13–18.

Cattaneo, M. D., M. Jansson, and K. Nagasawa (2020). Bootstrap-based inference for cube root asymptotics. *Econometrica 88*, 2203–2219.

Chernoff, H. (1964, 12). Estimation of the mode. *Annals of the Institute of Statistical Mathematics 16*, 31–41.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, 01). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*, C1–C68.

De Schryver, M. and J. De Neve (2019, 08). A tutorial on probabilistic index models: Regression models for the effect size p(y1 ¡ y2). *Psychological Methods 24*, 403–418.

Durrett, R. (2019). *Probability: theory and examples*, Volume 49. Cambridge university press.

Faraggi, D. and R. Simon (1996). A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis. *Statistics in Medicine 15*, 2203–2213.

Fokianos, K. and J. F. Troendle (2007). Inference for the relative treatment effect with the density ratio model. *Statistical Modelling 7*, 155–173.

Gavanji, P., B. E. Chen, and W. Jiang (2017). Residual bootstrap test for interactions in biomarker threshold models with survival data. *Statistics in Biosciences 10*, 202–216.

Grenander, U. (1956). On the theory of mortality measurement. *Scandinavian Actuarial Journal 1956*, 125–153.

Groeneboom, P. and J. A. Wellner (2001). Computing chernoff's distribution. *Journal of Computational*

and Graphical Statistics 10, 388–400.

He, Y., H. Lin, and D. Tu (2018). A single-index threshold cox proportional hazard model for identifying a treatment-sensitive subset based on multiple biomarkers. *Statistics in Medicine 37*, 3267–3279.

Hilsenbeck, S. G. and G. M. Clark (1996). Practical p-value adjustment for optimally selected cutpoints. *Statistics in Medicine 15*, 103–112.

Hilsenbeck, S. G., G. M. Clark, and W. L. McGuire (1992). Why do so many prognostic factors fail to pan out? *Breast Cancer Research and Treatment 22*, 197–206.

Jiang, S., B. E. Chen, and D. Tu (2016). Inference on treatment-covariate interaction based on a nonparametric measure of treatment effects and censored survival data. *Statistics in Medicine 35*, 2715–2725.

Jiang, W., B. Freidlin, and R. Simon (2007). Biomarker-Adaptive Threshold Design: A Procedure for Evaluating Treatment With Possible Biomarker-Defined Subset Effect. *JNCI: Journal of the National Cancer Institute 99*, 1036–1043.

Jonker, D. J., C. S. Karapetis, et al. (2013). Epiregulin gene expression as a biomarker of benefit from cetuximab in the treatment of advanced colorectal cancer. *British Journal of Cancer 110*, 648–655.

Karapetis, C. S. et al. (2008). K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *The New England journal of medicine 359*, 1757–65.

Kim, J. and D. Pollard (1990). Cube root asymptotics. *The Annals of Statistics 18*, 191–219.

Kotz, S., Y. Lumelskii, and M. Pensky (2003). *The Stress–Strength Model And Its Generalizations: Theory and Applications*. WORLD SCIENTIFIC.

Koziol, J. A. and Z. Jia (2009). The concordance index c and the mann-whitney parameter pr(x>y) with randomly censored data. *Biometrical Journal 51*, 467–474.

Li, N., Y. Song, C. D. Lin, and D. Tu (2023a). Bootstrap adjusted predictive classification for identification of subgroups with differential treatment effects under generalized linear models. *Electronic Journal of Statistics 17*(1), 548 – 606.

Li, N., Y. Song, D. Lin, and D. Tu (2023b). Bootstrap adjustment to minimum p-value method for predictive classification. *Statistica Sinica 33*, 2065–2086.

Luo, X. and J. M. Boyett (1997). Estimations of a threshold parameter in cox regression. *Communications in Statistics - Theory and Methods 26*, 2329–2346.

Mann, H. B. and D. R. Whitney (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics 18*, 50–60.

Manski, C. F. (1975, 08). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics 3*, 205–228.

Mazumdar, M. and J. R. Glassman (2000). Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Statistics in Medicine 19*, 113–132.

Mccullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*. CRC Press, Boca Raton.

Mi, X., P. Tighe, F. Zou, and B. Zou (2021). A deep learning semiparametric regression for adjusting complex confounding structures. *The Annals of Applied Statistics 15*, 1086–1100.

Moser, B. K. and M. H. McCann (2008). Reformulating the hazard ratio to enhance communication with clinical investigators. *Clinical Trials: Journal of the Society for Clinical Trials 5*, 248–252.

Patel, K. M. and D. G. Hoel (1973). A nonparametric test for interaction in factorial experiments. *Journal of the American Statistical Association 68*, 615–620.

Peña, V. H. and E. Giné (1999). *Decoupling: From Dependence to Independence*. Springer.

Politis, D. N. and J. P. Romano (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics 22*, 2031–2050.

Pons, O. (2003). Estimation in a cox regression model with a change-point according to a threshold in a covariate. *The Annals of Statistics 31*, 442–463.

Péron, J., P. Roy, B. Ozenne, L. Roche, and M. Buyse (2016). The net chance of a longer survival as a patient-oriented measure of treatment benefit in randomized clinical trials. *JAMA Oncology 2*,

901.

Seijo, E. and B. Sen (2011). Change-point in stochastic design regression and the bootstrap. *The Annals of Statistics 39*, 1580–1607.

Sen, P. K. (1967). A Note on Asymptotically Distribution-Free Confidence Bounds for P{X < Y}, Based on Two Independent Samples. *Sankhyā: The Indian Journal of Statistics, Series A 29*, 95–102.

Shao, J. and D. Tu (1995). *The Jackknife and Bootstrap*. Springer Nature.

Thas, O., J. D. Neve, L. Clement, and J.-P. Ottoy (2012, 07). Probabilistic index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*, 623–671.

Van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes: With Applications to Statistics*. Springer New York.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin 1*, 80–83.

Xu, G., B. Sen, and Z. Ying (2014). Bootstrapping a change-point cox model for survival data. *Electronic Journal of Statistics 8*, 1345–1379.

Department of Mathematics and Statistics, Queen's University, Kingston, ON, K7L 3N6, Canada

E-mail: zehui.wang@queensu.ca

Department of Mathematics and Statistics, Queen's University, Kingston, ON, K7L 3N6, Canada

E-mail: yanglei.song@queensu.ca

Department of Mathematics and Statistics, Queen's University, Kingston, ON, K7L 3N6, Canada

E-mail: wenyu.jiang@queensu.ca

Departments of Public Health Sciences & Mathematics and Statistics and Canadian Cancer Trials Group, Queen's University, Kingston, ON, K7L 3N6, Canada

E-mail: dtu@ctg.queensu.ca