

## Residual Bootstrap Test for Interactions in Biomarker Threshold Models with Survival Data

Parisa Gavanji · Bingshu E. Chen ·  
Wenyu Jiang

Received: date / Accepted: December 4th, 2017

**Abstract** Many new treatments in cancer clinical trials tend to benefit a subset of patients more. To avoid unnecessary therapies and failure to recognize beneficial treatments, biomarker threshold models are often used to identify this subset of patients. We are interested in testing the treatment-biomarker interaction effects in a threshold model with biomarker but an unknown cut point. The unknown cut point causes irregularity in the model, and the traditional likelihood ratio test cannot be applied directly. A test for biomarker-treatment interaction effects is developed using a residual bootstrap method to approximate the distribution of the proposed test statistic. We evaluate the residual bootstrap and the permutation methods through extensive simulation study and find that the residual bootstrap method gives accurate test size, while the permutation method cannot control type  $I$  error sometimes in the presence of main treatment effects. The proposed residual bootstrap test can be used to explore potential treatment-by-biomarker interaction in clinical studies. The findings can be applied to guide the follow-up trial design using biomarker as a stratification factor. We apply the proposed residual bootstrap method to data from Breast International Group (BIG) 1-98 randomized clinical trial and show that patients with high Ki-67 level may benefit more from Letrozole treatment.

**Keywords** Biomarker · Permutation method · Residual Bootstrap · Survival analysis · Biomarker threshold models · Treatment-biomarker interaction.

---

P. Gavanji, W. Jiang  
Department of Mathematics and Statistics, Queen's university, 99 University Ave., Kingston,  
Ontario, Canada K7L 3N6  
B.E. Chen  
Department of Public Health Sciences and Canadian Cancer Trials Group, 10 Stuart Street,  
Kingston, Ontario, Canada K7L 3N6  
E-mail: bechen@ctg.queensu.ca  
Tel.: +613-533-6430

## 1 Introduction

In traditional clinical trials, randomized studies are often conducted to evaluate the treatment effects by including all eligible patients. However, in some subsets of patients with different characteristics, patients may respond differently to the treatments. A patient characteristic affecting a patient's response to a certain treatment is called a predictive biomarker [13] or biomarker henceforth. The study of the interactive impacts of a biomarker on treatment outcomes becomes more and more important for evaluating treatment effects within different biomarker defined patient subsets.

The conventional clinical trial design focuses on the overall treatment effect by including broad eligible patients, which may fail to detect some stronger treatment effects restricted to subsets of patients. Clearly, identifying the subsets of patients who may not benefit, benefit less, or more from the new treatment can help to avoid unnecessary therapy and to make personalized decisions in treating patients. Biomarker threshold models are frequently used [15] to conduct this type of subset analysis. A biomarker is often measured on a continuous scale, such as Ki-67 for breast cancer patients [6]. Royston et al. (2006) suggested that the biomarker should be treated as a continuous variable in order to use all potential information in the data [11]. Sargent et al. (2005) described a predictive (binary) biomarker that splits the patient population into two groups as either good or poor candidates for a specific treatment for optimistic treatment selection [13].

An important aspect of the biomarker study is to test for treatment-biomarker interaction. Existence of the interaction effects implies that the new treatment tend to have different effects on patients with different biomarker values, while no interaction effects implies that the new treatment has the same effects on all patients.

There are various methods in the literature for studying the biomarker-treatment interaction. For example, the traditional Cox proportional hazards model with an interaction term, or the fractional polynomial approach for a continuous biomarker [12]. With a continuous biomarker, it is common in clinical trial analysis to assume that there is a biomarker cut point that categorizes patients into subsets that benefit more, or do not benefit or benefit less from the new treatment. The classical likelihood ratio test for treatment-biomarker interaction also cannot be applied directly for this situation. This is because several regularity conditions are required for the classical asymptotic results for the likelihood ratio test [14], while the biomarker threshold models in the presence of the unknown biomarker cut point do not satisfy these regularity conditions. To address this issue, the problem is often formulated in change point framework for Cox proportional hazard models and profile likelihood methods are often used (Luo and Boyett 1997, Pons 2003) [7], [10]. Luo and Boyett (1997) proposed and justify a maximum profile likelihood method to make statistical inferences about threshold parameter  $c$ . Pons (2003) extended the model and proved the asymptotic consistent for change point problem in

Cox model with interaction terms. However, hypothesis test for interaction in biomarker threshold model remains an open question.

Jiang et al. (2007) proposed a biomarker adaptive threshold design for situations with a known biomarker, but an unknown cut point what defines two patient subsets with different treatment effects. They developed a test method based on permutation for detecting the difference in treatment effects between the two patient subsets [5]. Recently, Chen et al. (2014) proposed a hierarchical Bayesian method to make statistical inference on the biomarker cut point and treatment subset effects at the same time. However, the Bayes approach cannot be used to test the existence of treatment-biomarker interaction effects [4].

Upon a closer inspection on the permutation-based test for treatment-biomarker interaction effects by Jiang et al. (2007), we notice that permutation test is based on an implicit assumption that there are no main treatment effects in the model. In other words, the new treatment only benefits a subset of patients, and has no effects on the rest of the patients. This motivates us to develop a new testing method for the biomarker threshold models with an unknown biomarker cut point, which is very popular in biomarker-aided clinical trial studies. We propose a residual bootstrap method to approximate the distribution of a test statistic for identifying the treatment-biomarker interaction effects. We develop the test procedure based upon the residual bootstrap technique of Loughin (1995), while tackling the modeling challenges caused by the unknown biomarker cut point [8]. The proposed method relaxes the model restriction of the permutation method by allowing main treatment effect and other covariate effects, which is often necessary in common practice for modeling patient subsets that both benefit from the new treatment but at different levels. The proposed residual bootstrap test (RBT) method includes the model of Jiang et al. (2007) as a special case. We then study the finite sample properties of the proposed residual bootstrap method through extensive simulations.

## 2 Model and Methods

### 2.1 Model

In clinical trials, patients are randomly assigned to either a new treatment group or a control group. Let  $Z$  denote the treatment allocation variable, with  $Z = 1$  for the treatment group and  $Z = 0$  for the control group, respectively. These two groups are compared with respect to time to a clinical event, such as death or disease progression. Let  $\tilde{T}_i$  and  $V_i$  be the potential failure time and censoring time for patient  $i$ ,  $i = 1, \dots, n$ . Let  $\delta_i = I(\tilde{T}_i < V_i)$  be a survival status indicator and  $t_i = \min(\tilde{T}_i, V_i)$  be the observed failure or censoring time. Let  $X$  be a biomarker variable. Without loss of generality, we assume that  $0 \leq X \leq 1$ . We further make an assumption that the biomarker is predictive of patient subsets with different treatment benefits. This can be described by

the biomarker threshold model as in [5],

$$\lambda_i(t) = \lambda_0(t) \exp \{ \beta_1 z_i + \beta_2 I(x_i > c) + \beta_3 z_i I(x_i > c) \}, \quad (1)$$

where  $z_i$ ,  $x_i$  are observed values for  $Z$ ,  $X$  for patient  $i$ ,  $\lambda_0(t)$  is the baseline hazard function, and  $\lambda_i(t)$  is the hazard function of patient  $i$ . In this model,  $\beta_1$  is the main treatment effect,  $\beta_2$  is the main biomarker effect, and  $\beta_3$  is the treatment and biomarker interaction effect. The biomarker cut point  $c$  is also an unknown parameter and  $0 \leq c \leq 1$ . If  $c$  is known, model (1) becomes a regular Cox proportional hazards model with an interaction term.

Specifically in model (1), on patients with biomarker values no greater than the cut point  $c$ , the treatment effect in terms of log hazard ratio of treatment versus control groups is expressed by the parameter  $\beta_1$ . On patients with biomarker values greater than  $c$ , the treatment effect (log hazard ratio) is expressed by  $\beta_1 + \beta_3$ . The treatment-biomarker interaction effect  $\beta_3$  is then the difference in treatment effects between the two patient subsets categorized by the biomarker and the cut point. When  $\beta_3 = 0$ , all patients are expected to have the same treatment benefit ( $\beta_1$ ) regardless of their biomarker values.

In this paper, we are interested in studying the treatment-biomarker interaction effect in order to evaluate if patients in different biomarker subsets benefit differently from the new treatment. For this purpose, we focus on testing the existence of the treatment-biomarker interaction effect, with the null hypothesis  $H_0 : \beta_3 = 0$  versus the alternative hypothesis  $H_1 : \beta_3 \neq 0$ . Here we define the null parameter space as  $\Theta_0 = \{(\beta_1, \beta_2, \beta_3, c) : (\beta_1, \beta_2) \in \mathbb{R}^2, \beta_3 = 0, 0 < c < 1\}$ , and the entire parameter space as  $\Theta = \{(\beta_1, \beta_2, \beta_3, c) : (\beta_1, \beta_2, \beta_3) \in \mathbb{R}^3, 0 < c < 1\}$ .

Let  $Y_i(t) = I(t_i \geq t)$  be the indicator function that patient  $i$  is at risk at time  $t$ . For patients  $i = 1, \dots, n$ , the partial likelihood for model (1) can be written as,

$$L(\boldsymbol{\beta}, c) = \prod_{i=1}^n \left\{ \frac{\exp(\mathbf{x}'_i \boldsymbol{\beta})}{\sum_{k=1}^n Y_k(t_i) \exp(\mathbf{x}'_k \boldsymbol{\beta})} \right\}^{\delta_i}, \quad (2)$$

where  $\mathbf{x}'_i = \{z_i, I(x_i > c), z_i I(x_i > c)\}$ , and  $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$ . Let  $\ell$  be the logarithm of the partial likelihood function defined by (2). When the cut point  $c$  is unknown, the partial likelihood function is neither continuous in  $c$  nor differentiable with respect to  $c$ . As a consequence, we are not able to estimate  $c$  directly by maximizing the partial likelihood function, and the classical asymptotic results for the likelihood ratio test statistic is not valid.

However, we can estimate  $c$  by profile likelihood method for  $0 < c < 1$ , and obtain the maximum profile likelihood estimate of  $c$ ,

$$\hat{c} = \arg \max_{0 < c < 1} \ell(\hat{\beta}_{1c}, \hat{\beta}_{2c}, \hat{\beta}_{3c}; c),$$

where  $\hat{\beta}_{1c}$ ,  $\hat{\beta}_{2c}$ , and  $\hat{\beta}_{3c}$  are the maximum partial likelihood estimates of  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  for each given value of  $c$ , under parameter space  $\Theta_c = \{(\beta_1, \beta_2, \beta_3) \in \mathbb{R}^3\}$ . It was shown that  $\hat{c}$  obtained by the maximum profile likelihood method provides a consistent estimate for the threshold parameter  $c$  [4]. The maximum

profile likelihood estimates will be used to obtain the residuals for the residual bootstrap method in Section 2.2.

We now construct the test statistic for treatment-biomarker interaction for unknown threshold parameter  $c$ . For a given value of  $c$ , we denote the corresponding likelihood ratio statistic by

$$LR(c) = 2 \left\{ \ell(\hat{\beta}_{1c}, \hat{\beta}_{2c}, \hat{\beta}_{3c}; c) - \ell(\tilde{\beta}_{1c}, \tilde{\beta}_{2c}, 0; c) \right\},$$

where  $\hat{\beta}_{1c}$ ,  $\hat{\beta}_{2c}$ , and  $\hat{\beta}_{3c}$  are the maximum partial likelihood estimates obtained under the entire parameter space for a given  $c$ ,  $\Theta_c$ , and  $\tilde{\beta}_{1c}$ ,  $\tilde{\beta}_{2c}$  are the maximum partial likelihood estimates under the null parameter space  $\Theta_{0c} = \{(\beta_1, \beta_2) \in \mathbb{R}^2, \beta_3 = 0\}$  for a given  $c$ . We then define a test statistic for testing  $H_0$  in the form of

$$LR = \max_{0 < c < 1} LR(c), \quad (3)$$

which is the maximum of the likelihood ratio statistic  $LR(c)$  over any possible  $c$  value,  $0 < c < 1$ . Note here  $c$  that maximizes  $LR(c)$  is different from  $\hat{c}$  that maximizes the profile likelihood function above.

The distribution of the proposed test statistic  $LR$  is unknown in this setting because it does not follow a conventional chi-square distribution. To conduct statistical inferences for the test statistic  $LR$ , we will use resampling methods to approximate the null distribution of  $LR$ , under the hypothesis  $H_0 : \beta_3 = 0$ . For this purpose, we propose and study the residual bootstrap method in Section 2.2. To compare the proposed method and the existing permutation method, we also briefly review an existing method based on the permutation test [5] in Section 2.3.

## 2.2 Residual Bootstrap Method

Loughin (1995) proposed a residual bootstrap method for the Cox proportional hazards regression model when explanatory variables are non-random constants fixed by the design. We extend the method to the biomarker threshold model (1) in the presence of an unknown cut point, which is beyond the scope of traditional regression models for survival data. Test for the treatment-biomarker interaction effect in model (1), relies on a complicated and unconventional test statistic (3), which needs to be handled carefully in the bootstrap procedure with new strategies.

The proposed sampling scheme [8] is based on the fact that, when the relative risk function is independent of time, the Cox's partial likelihood function defined by (2) remains invariant under monotone increasing transformations of time  $t$ . For a Cox's model, we can write,

$$S(t) = S_0(t)^{\exp(\mathbf{x}'\boldsymbol{\beta})} = \{1 - F_0(t)\}^{\exp(\mathbf{x}'\boldsymbol{\beta})},$$

where  $S_0(t)$  is the baseline survival function, and  $F_0(t) = 1 - S_0(t)$  is the baseline cumulative distribution function. Therefore,

$$F_0(t) = 1 - \{S(t)\}^{\exp(-\mathbf{x}'\boldsymbol{\beta})} \quad (4)$$

is a monotone increasing function of  $t$ . The invariance property of the partial likelihood function implies replacing failure time  $t$  by a monotone transformation  $F_0(t)$  in the data will not change the statistical inference.

Direct application of the method of Loughin (1995) is impossible for the biomarker threshold model (1) because of the unknown cut point  $c$ . Instead, we propose the following residual bootstrap test (RBT) method with four steps for testing the treatment-biomarker interaction effect with  $H_0 : \beta_3 = 0$ , by meticulously incorporating the profile likelihood estimation for  $c$ .

**Step 1.** Use  $\hat{c}$ , the maximum profile likelihood estimate of  $c$ , to fit model (1) as a Cox model based on the original data  $(t_i, \delta_i, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ , and obtain  $\hat{\beta}_1$ ,  $\hat{\beta}_2$ , and  $\hat{\beta}_3$  under the entire parameter space  $\boldsymbol{\Theta}$ . Estimate the survival probabilities  $\hat{u}_i = \hat{S}(t_i)$ ,  $i = 1, \dots, n$ , using the Nelson-Aalen Method [1,9].

**Step 2.** Generate bootstrap data  $(u_i^*, \delta_i^*, \mathbf{x}_i)$  by sampling with replacement from  $(\hat{u}_i, \delta_i, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ . Calculate the probability-scale failure time, under  $H_0$ ,

$$y_i^* = 1 - (u_i^*)^{\exp(-\{z_i \tilde{\beta}_1 + I(x_i > \tilde{c}) \tilde{\beta}_2\})},$$

where  $\tilde{\beta}_1$  and  $\tilde{\beta}_2$  are the maximum partial likelihood estimates, and  $\tilde{c}$  is the maximum profile likelihood estimate of  $c$ , under the null parameter space  $\boldsymbol{\Theta}_0$ . In light of equation (4),  $y_i^*$  is a resampled version of  $F_0(t)$ .

**Step 3.** For each given value of  $c$ , fit two models on the entire parameter space  $\boldsymbol{\Theta}_c$  and the null parameter space  $\boldsymbol{\Theta}_{0c}$  for the bootstrap data  $(y_i^*, \delta_i^*, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ , ( $y_i^*$  as survival time and  $\delta_i^*$  as censoring indicator) to calculate the likelihood ratio statistic  $LR^*(c)$ . Across a grid of  $c$  values,  $0 < c < 1$ , obtain the bootstrap version of the test statistic,

$$LR^* = \max_{0 < c < 1} LR^*(c).$$

**Step 4.** Repeat Steps 2, 3 for  $B$  times, and obtain  $B$  replications of  $LR^*$ . The empirical distribution of  $LR^*$  provides an approximation to the null distribution of the test statistic  $LR$  given by (3). The  $p$ -value of the RBT is the proportion of  $LR^*$  values greater than the observed  $LR$ ,

$$p\text{-value}_{RBT} = \frac{\#\{LR^* > LR\}}{B}, \quad (5)$$

where  $LR$  is the observed test statistic of the form (3) calculated based on the original data.

### 2.3 Permutation Test

Jiang et al. (2007) introduced an adaptive design to investigate if the new treatment benefits all patients or only a subset of patients (treatment-biomarker interaction). In this subsection, we review the permutation test (PT) method of Jiang (2007) as another way to approximate the distribution of the proposed test statistic  $LR$  in (3) for testing the treatment-biomarker interaction.

The PT method consists of  $B$  permutation runs. In each run  $p$ , a permuted version of data is created by randomly assigning patients to the new and control treatments, while keeping other variables the same. On the permuted data of  $p^{th}$  run, fit two Cox models under the null and alternative hypotheses for a grid of  $c$  values, calculate the likelihood ratio statistic  $LR^p(c)$ , and obtain the test statistic,

$$LR^p = \max_{0 < c < 1} LR^p(c).$$

With  $B$  permutation replications,  $p = 1, \dots, B$ , obtain the  $p$ -value

$$p\text{-value}_{PT} = \frac{\#\{LR^p > LR\}}{B}, \quad (6)$$

where  $LR$  of form (3) is obtained based on the original data.

We notice that the PT method for model (1) randomly assigns patients to treatments, hence implicitly assumes that observations are exchangeable under the null hypothesis, i.e., there is no treatment main effect. In other words, the PT method is considered to be correct only when  $\beta_1 = \beta_3 = 0$ . Jiang et al. (2007) has not addressed this restriction, and has not examined the PT method when the assumption of  $\beta_1 = 0$  is violated. The RBT method proposed in this paper has no such restrictions.

### 3 Simulation Study

We conduct simulation studies to evaluate the finite sample performance of the RBT method for testing  $H_0 : \beta_3 = 0$  in model (1) in comparison to the PT method. We generate survival time from a Weibull distribution with hazard function

$$\lambda_i(t) = \nu\gamma(\gamma t)^{\nu-1} \exp\{\beta_1 z_i + \beta_2 I(x_i > c) + \beta_3 z_i I(x_i > c)\}, \quad (7)$$

with parameters  $\gamma$ ,  $\nu$ , and baseline hazard function  $\lambda_0(t) = \nu\gamma(\gamma t)^{\nu-1}$ . The biomarker values,  $x_i$ 's, are generated from a uniform distribution on interval (0,1). Here, we use a censoring time generated from  $U(0, 1.5)$ .

We consider three different designs for assigning patients to treatments in order to compare the performance of the RBT and PT methods in different situations. In design I, we apply an unequal randomization which assigns 80% of the patients to the treatment arm and 20% to the control arm. This type of design is particularly useful when existing evidence suggests that new treatment may have other additional clinical benefit, for example, less toxicity

**Table 1** The empirical size of the test under null hypothesis  $H_0 : \beta_3 = 0$ , for significance level  $\alpha = 0.05$ . Failure time follows Weibull distribution in equation (7) with  $\gamma = 2.0$ . Results are based on  $R = 1000$  replications.

$\nu$	$c_0$	$e^{\beta_1}$	$e^{\beta_2}$	Design I		Design II		Design III	
				PT	RBT	PT	RBT	PT	RBT
1.5	0.25	0.1	0.5	0.058	0.045	0.053	0.042	0.049	0.046
1.5	0.25	0.5	0.3	0.062	0.055	0.063	0.042	0.067	0.051
1.5	0.25	0.9	0.3	0.044	0.037	0.060	0.044	0.042	0.042
1.5	0.25	0.5	0.5	0.054	0.054	0.046	0.041	0.047	0.040
1.5	0.25	1.0	0.5	0.052	0.049	0.048	0.048	0.052	0.051
1.5	0.60	0.1	0.5	0.045	0.036	0.053	0.037	0.043	0.047
1.5	0.60	0.5	0.3	0.057	0.056	0.076	0.036	0.060	0.058
1.5	0.60	0.9	0.3	0.038	0.046	0.135	0.050	0.042	0.049
1.5	0.60	1.0	0.5	0.043	0.045	0.049	0.052	0.046	0.050
1.5	0.75	0.5	0.3	0.054	0.049	0.060	0.047	0.058	0.053
1.5	0.75	0.9	0.3	0.044	0.054	0.075	0.053	0.051	0.047
1.5	0.75	0.5	0.5	0.048	0.045	0.056	0.050	0.044	0.044
1.5	0.75	1.0	0.5	0.042	0.048	0.058	0.046	0.046	0.044
2.0	0.25	0.1	0.5	0.063	0.036	0.056	0.045	0.053	0.049
2.0	0.25	0.5	0.3	0.062	0.049	0.052	0.042	0.054	0.040
2.0	0.25	0.9	0.3	0.047	0.044	0.049	0.041	0.045	0.044
2.0	0.25	0.5	0.5	0.055	0.048	0.048	0.045	0.045	0.045
2.0	0.25	1.0	0.5	0.050	0.047	0.045	0.041	0.045	0.045
2.0	0.60	0.1	0.5	0.045	0.038	0.048	0.039	0.053	0.052
2.0	0.60	0.5	0.3	0.049	0.048	0.088	0.037	0.048	0.042
2.0	0.60	1.0	0.5	0.040	0.040	0.049	0.041	0.047	0.049
2.0	0.75	0.5	0.3	0.052	0.048	0.061	0.046	0.056	0.051
2.0	0.75	0.9	0.3	0.048	0.052	0.074	0.050	0.051	0.050
2.0	0.75	0.5	0.5	0.045	0.047	0.060	0.048	0.046	0.043
2.0	0.75	1.0	0.5	0.045	0.043	0.051	0.043	0.038	0.038

compared with standard treatment. In design II, we assume that patients can be divided into two groups: high risk and low risk groups before randomization [2,16]. We assume 50% of patients are in high risk group and the other 50% in the low risk group, respectively. For the high risk group patient, the new treatment is assigned to the patient with probability 0.75. Otherwise, for the low risk group, the new treatment is assigned to the patient with probability 0.5. Design II is based on the consideration that, for high risk group, the treatment is supposed to benefit this group of patients more. Therefore, it is reasonable to have bigger percentage of high risk patients to receive the new treatment than the standard treatment or control. In design III, we consider a typical randomization and assign two treatments to the patients with equal weights.

To generate simulated data set, we consider different combinations of  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$ , and the true cut point values  $c_0 = 0.25, 0.6$ , and  $0.75$ . We let  $\nu = 1.5, 2.0$ , and  $\gamma = 2.0$  in generating the Weibull failure time. For each set of parameter combination, simulations are repeated  $R = 1000$  times for empirical test size, and  $R = 500$  times for empirical power and empirical bias of  $c$ . For both the bootstrap and permutation methods, we use  $B = 200$  bootstrap/permutation replications, respectively.

Results for the empirical test sizes for both RBT and PT methods for designs I-III are presented in Table 1. The empirical test size is the percentage of the simulation replications that reach the pre-specified level of the statistical significance ( $\alpha = 0.05$ ) when data are generated under the null hypothesis,  $H_0$ . To calculate the empirical test sizes for RBT and PT methods, we calculate p-values from equations (5) and (6), and the empirical test size is the proportion of p-values that are smaller than  $\alpha$ . If the size of the test turns out to be close to the significance level ( $\alpha = 0.05$ ), we conclude that the proposed test procedure controls the type I error-rate, or the test has correct size.

Simulation results in Table 1 indicate that the RBT method provides correct test sizes, while the PT method cannot control the type I error in some cases, especially for the more complicated designs such as design II. For some combinations of parameter settings, the test based on PT method gives wrong test size. For example, when  $\nu = 2.0$ ,  $c_0 = 0.6$ ,  $\beta_1 = \log(0.5)$ ,  $\beta_2 = \log(0.3)$ , using design II, the test size for PT is 0.102, while the test size for RBT is 0.032. We also conducted additional simulations for  $\nu = 0.5, 1.0$  (the results are not shown here), and the test sizes of the RBT method are reasonably close to the significance level  $\alpha = 0.05$  in all situations that we considered, while PT method cannot control type I error in several settings. This is not surprising because the PT method requires the model assumption of no main treatment effect ( $\beta_1 = 0$ ), which apparently does not hold here. The PT method proposed by Jiang et al. (2007) should not be applied unless there is clear prior knowledge that there exists no main treatment effect. However, the proposed RBT method is a valid approach without this kind of model restriction.

Furthermore, we calculate the empirical power of the RBT method under the alternative hypothesis for different values of  $\beta_3 \neq 0$ , and the results are presented in Table 2. The empirical power of the test is defined as the percentage of the simulated replications that reach the pre-specified significance level ( $\alpha = 0.05$ ), when data are generated under the alternative hypothesis. Since the PT method has restrictive model assumptions and cannot control type I error, it is meaningless to further examine its power. The empirical power of RBT method, based on  $R = 500$  replications, as seen in Table 2, is quite satisfactory when testing for the treatment-biomarker interaction effects in various situations. We notice that when  $0 < \exp(\beta_3) < 1$ , a small value of  $\exp(\beta_3)$  is corresponding to a large absolute value of  $\beta_3$ , this leads to a higher power of the test.

We further explore the finite sample properties for the profile likelihood estimate of  $c$ . Table 3 shows the empirical bias and standard error of the profile likelihood estimate of  $c$  for different combinations of the parameters and design I, under the alternative hypothesis. For this table, Weibull failure time is generated from model (7) with  $\nu = 1.5, 2.0$  and  $\gamma = 2.0$  in the baseline hazard function. The profile likelihood method turns out to be quite accurate for estimating  $c$ , with very small bias and standard error. This is consistent with results by Chen et al (2014) that maximum profile likelihood estimate provides unbiased estimate for threshold parameter  $c$  under the alternative hypothesis [4].

**Table 2** The empirical power of RBT,  $H_1 : \beta_3 \neq 0$ , for significance level  $\alpha = 0.05$ . Failure time follows Weibull distribution in equation (7) with  $\gamma = 2.0$ . All results are based on  $R = 500$  replications.

$\nu$	$c_0$	$e^{\beta_1}$	$e^{\beta_2}$	$e^{\beta_3}$	Design I	Design II	Design III
1.5	0.25	0.3	0.5	0.4	0.784	0.874	0.902
1.5	0.25	0.5	0.3	0.5	0.588	0.680	0.710
1.5	0.60	0.3	0.5	0.4	0.792	0.806	0.914
1.5	0.60	0.5	0.3	0.4	0.808	0.748	0.916
1.5	0.75	0.5	0.3	0.2	0.936	0.970	0.994
1.5	0.75	0.5	0.3	0.3	0.776	0.838	0.924
2.0	0.25	0.3	0.5	0.4	0.796	0.912	0.930
2.0	0.25	0.5	0.3	0.5	0.554	0.696	0.692
2.0	0.60	0.3	0.5	0.4	0.856	0.858	0.964
2.0	0.60	0.5	0.3	0.4	0.876	0.824	0.966
2.0	0.75	0.5	0.3	0.2	0.986	0.996	0.998
2.0	0.75	0.5	0.3	0.3	0.808	0.908	0.974

**Table 3** Empirical bias and standard error of the profile likelihood estimate of  $c$ , under  $H_1 : \beta_3 \neq 0$ , for design I. Failure time follows Weibull distribution in equation (7) with  $\gamma = 2.0$ . Results are based on  $R = 500$  replications.

$\nu$	$c_0$	$e^{\beta_1}$	$e^{\beta_2}$	$e^{\beta_3}$	Bias( $\hat{c}$ )	S.E.( $\hat{c}$ )
1.5	0.25	0.3	0.5	0.4	0.0008	0.014
1.5	0.25	0.5	0.3	0.4	-0.0012	0.010
1.5	0.60	0.3	0.5	0.4	-0.0004	0.008
1.5	0.60	0.5	0.3	0.4	-0.0002	0.004
1.5	0.75	0.3	0.5	0.3	-0.0008	0.013
1.5	0.75	0.5	0.3	0.3	-0.0026	0.010
2.0	0.25	0.3	0.5	0.4	-0.0014	0.012
2.0	0.25	0.5	0.3	0.4	-0.0034	0.009
2.0	0.60	0.3	0.5	0.4	0.00004	0.006
2.0	0.60	0.5	0.3	0.4	-0.0001	0.003
2.0	0.75	0.3	0.5	0.3	-0.0025	0.010
2.0	0.75	0.5	0.3	0.3	-0.0028	0.010

It is also of interest to investigate the robustness of the proposed RBT method when the model is mis-specified. Here we conducted simulation studies by introducing an additional biomarker variable  $W \sim N(0, \sigma^2)$ , with  $\sigma = 0.5, 1.0, 2.0$  to the model to generate the data set from

$$\lambda_i(t) = \nu\gamma(\gamma t)^{\nu-1} \exp\{\beta_1 z_i + \beta_2 I(x_i > c) + \beta_3 z_i I(x_i > c) + \beta_4 w_i + \beta_5 w_i z_i\}. \quad (8)$$

Then, we applied the same strategy as before for testing the null hypothesis, while ignoring the effect of the biomarker variable  $W$ . Empirical test sizes based on  $R = 1000$  replications for different combinations of the parameters were reported in Table 4. Since there is no inflation in the empirical test sizes under  $H_0$ , we conclude that the proposed method is robust to model mis-specification.

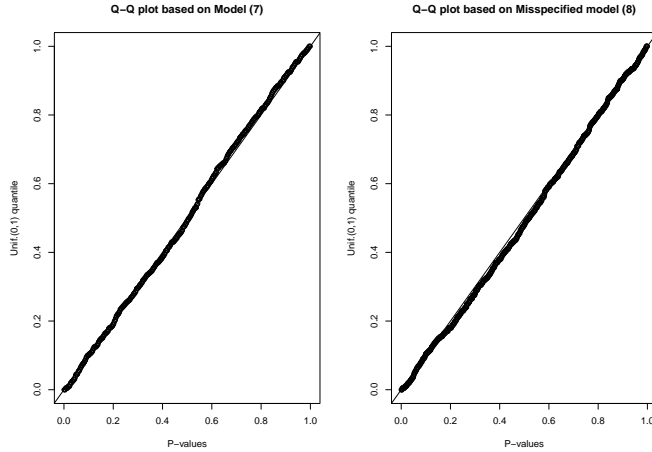
If the null distribution of test statistics  $LR$  is correctly approximated, the p-value shall have a uniform distribution under  $H_0$ . Here we use Q-Q plot to

**Table 4** The empirical test size under null hypothesis  $H_0 : \beta_3 = 0$ , for significance level  $\alpha = 0.05$ . Failure time follows Weibull distribution in equation (8) with  $\gamma = 2.0$ . The model is mis-specified with the biomarker  $W \sim N(0, \sigma^2)$  is left out from the inference. All results are based on  $R = 1000$  replications.

$\nu$	$c_0$	$\sigma^2$	$\exp(\beta_1)$	$\exp(\beta_2)$	$\beta_4$	$\beta_5$	Design I	Design II	Design III
1.5	0.25	0.5	0.3	0.5	0.5	0.0	0.052	0.036	0.055
1.5	0.25	0.5	0.3	0.5	0.5	0.5	0.034	0.051	0.051
1.5	0.25	0.5	0.5	0.3	0.5	0.0	0.068	0.044	0.057
1.5	0.25	0.5	0.5	0.3	0.5	0.5	0.044	0.042	0.057
2.0	0.50	0.5	0.3	0.5	0.5	0.0	0.058	0.064	0.048
2.0	0.50	0.5	0.3	0.5	0.5	0.5	0.050	0.046	0.055
2.0	0.50	0.5	0.5	0.3	0.5	0.0	0.048	0.045	0.045
2.0	0.50	0.5	0.5	0.3	0.5	0.5	0.044	0.032	0.057
1.5	0.25	1.0	0.3	0.5	0.5	0.0	0.061	0.054	0.056
1.5	0.25	1.0	0.3	0.5	0.5	0.5	0.045	0.052	0.062
1.5	0.25	1.0	0.5	0.3	0.5	0.0	0.053	0.037	0.050
1.5	0.25	1.0	0.5	0.3	0.5	0.5	0.044	0.046	0.066
2.0	0.50	1.0	0.3	0.5	0.5	0.0	0.063	0.043	0.061
2.0	0.50	1.0	0.3	0.5	0.5	0.5	0.032	0.049	0.057
2.0	0.50	1.0	0.5	0.3	0.5	0.0	0.052	0.040	0.055
2.0	0.50	1.0	0.5	0.3	0.5	0.5	0.030	0.050	0.049
1.5	0.25	2.0	0.3	0.5	0.5	0.0	0.048	0.041	0.052
1.5	0.25	2.0	0.3	0.5	0.5	0.5	0.037	0.055	0.062
1.5	0.25	2.0	0.5	0.3	0.5	0.0	0.045	0.055	0.050
1.5	0.25	2.0	0.5	0.3	0.5	0.5	0.035	0.060	0.070
2.0	0.50	2.0	0.3	0.5	0.5	0.0	0.062	0.055	0.056
2.0	0.50	2.0	0.3	0.5	0.5	0.5	0.034	0.037	0.064
2.0	0.50	2.0	0.5	0.3	0.5	0.0	0.064	0.040	0.053
2.0	0.50	2.0	0.5	0.3	0.5	0.5	0.031	0.039	0.062

demonstrate that the p-value for RBT method follows a uniform distribution. Figure 1 represents the Q-Q plot where failure time follows Weibull distribution in equation (7) with  $\gamma = 2.0$ , with  $\nu = 1.5$ ,  $c = 0.25$ ,  $\beta_1 = \log(0.5)$ ,  $\beta_2 = \log(0.3)$ , and  $R = 1000$  using Design I (left panel). We further do Q-Q plot for p-values when the model is mis-specified, as data generated from model (8) and tested under model (1). Failure time follows Weibull distribution in misspecified model (8) with above setting with  $\sigma^2 = 1.0$ ,  $\beta_4 = \log(1.65)$ ,  $\beta_5 = \log(1.0)$ , using Design I (right panel). We observed that for both settings, the observations are scatter around the straight line  $y = x$ , which suggests that the p-value follows a uniform distribution very well.

As discussed in Section 1 and Section 2, the classical likelihood ratio test is not valid for testing the treatment-biomarker interaction effects in model (1) in the presence of the unknown biomarker cut point. As an illustration, we simulate data sets based on model (1) with  $c = 0.25$ ,  $\beta_1 = \log(0.3)$ ,  $\beta_2 = \log(0.5)$ ,  $\beta_3 = 0$ , and directly apply the classical likelihood ratio test for  $H_0 : \beta_3 = 0$ ; the empirical test size is given by 0.264, which is obviously far away from the significance level  $\alpha = 0.05$ .



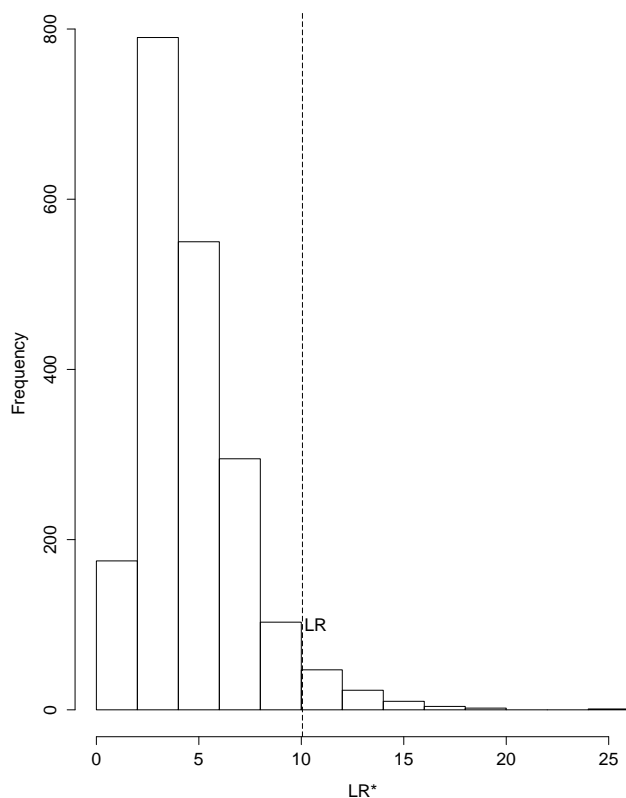
**Fig. 1** Q-Q plots for p-values from RBT vs.  $\text{unif}(0,1)$  quantiles. Failure time follows Weibull distribution in model (7) with  $\gamma = 2.0$ , with  $\nu = 1.5$ ,  $c = 0.25$ ,  $\beta_1 = \log(0.5)$ ,  $\beta_2 = \log(0.3)$ ,  $R = 1000$ , using Design I (left panel). Failure time follows Weibull distribution in misspecified model (8) with above setting with  $\sigma^2 = 1.0$ ,  $\beta_4 = \log(1.65)$ ,  $\beta_5 = \log(1.0)$ , using Design I (right panel)

#### 4 Application

To demonstrate how to apply the RBT method for testing the interaction between the biomarker and the treatment, we apply the method to a data set from Breast International Group (BIG) 1-98 randomized clinical trial [6]. This dataset is available from the *R* software package “*stepp*”. The BIG 1-98 is a Phase III clinical trial of 8010 post menopausal women with hormone-receptor-positive early invasive breast cancer who were randomly assigned adjuvant therapy of Letrozole as new treatment or Tamoxifen as control treatment. Among the 8010 patients, 2685 of them have available Ki-67 biomarker measurements.

The Ki-67 biomarker is a labeling index measurement of cell proliferation. Previous study shows that Letrozole was more effective than Tamoxifen for patients with tumors expressing the highest levels of the Ki-67 labeling index. Lazar et al. (2010) obtained that the  $p$ -value=0.09 ( Fig. 1(c) of [6] ) for the interaction term in hazard rate based on the Subpopulation Treatment Effect Pattern Plot (STEPP) method introduced by Bonetti and Gelber (2000) [3].

To apply the proposed RBT method, we first transform the Ki-67 biomarker value to interval  $(0, 1)$  using empirical percentiles of the original measurements. Then, we apply the RBT method to the data with  $B = 2000$  bootstrap samples. Figure 2 shows the histogram of the bootstrap test statistic samples under the null hypothesis based on the RBT method, which provides an empirical distribution of the proposed likelihood ratio test statistic  $LR$ . As displayed in the graph, the  $p$ -value=0.0505 of the RBT is the proportion of the resampling data with  $LR^*$  greater than the observed  $LR$ .



**Fig. 2** Histogram of the test statistic values based on the residual bootstrap test for data from the Breast International Group BIG 1-98 clinical trial, where  $LR^*$  and  $LR$  are the test statistic of the form (3) based on bootstrapped and original data, respectively. The p-value of the RBT test is 0.0505.

The profile likelihood estimate of cut point value is 0.07 in the transformed scale, which is equivalent to  $\hat{c} = 2.37$  in the original scale. Therefore, we can infer that patients with original Ki-67 values greater than 2.37 may benefit more from the Letrozole treatment.

## 5 Discussion

In clinical trials, testing for treatment-biomarker interaction effects is important when a new treatment tends to benefit a subset of patients more. In this paper, we consider the popular biomarker threshold model (1), which models the biomarker as a binary variable with an unknown cut point, and in this way categorizes patients into two subsets with different treatment effect levels. The classical likelihood ratio test does not work for testing the treatment-

biomarker interaction effects in model (1), because the unknown biomarker cut point causes model irregularity and discontinuous likelihood function. The permutation test method of Jiang et al. (2007) has an implicit model restriction and is only appropriate when no main treatment effects exist ( $\beta_1 = 0$ ), and is not generally applicable. We propose the residual bootstrap test (RBT) method to test for the treatment-biomarker interaction effects. It extends the residual bootstrap for Cox proportional hazards regression model proposed by Loughin (1995) to the biomarker threshold model (1), which is more complex than the conventional Cox model framework.

In the proposed RBT method, we apply the profile likelihood method in the estimation of the biomarker cut point  $c$ . Although the profile likelihood method provides an unbiased estimate for the cut point parameter  $c$  [4, 10], it cannot be used directly for testing the treatment-biomarker interaction effects. In order to deal with the unknown biomarker cut point, we build the RBT method by incorporating the profile likelihood estimation of  $c$  in the bootstrap layers and in the inference procedure whenever necessary.

The proposed RBT method gives correct test size without such restrictions, and it provides good power for detecting departure from the null hypothesis. Based on numerical simulation studies, we also observed that the proposed method is robust to model mis-specification. Ignoring the effect of either a prognostic or a predictive biomarker with normal distribution, the RBT method still controls type  $I$  error of the test. It adds a new and general tool to clinical trial studies with survival outcomes to identify varying treatment effects utilizing patient's biomarker information. We will make the R code for the proposed RBT method available as part of the Biomarker Threshold Models package (<https://CRAN.R-project.org/package=bhm>).

While the proposed residual bootstrap test can be widely used in either clinical trials or epidemiology observation studies, the results shall be interpreted with cautions. Due the fact that biomarker is not considered as a stratification factor during randomization, the detected treatment-biomarker interaction shall not be interpreted the same as those from biomarker-stratified design. We do not suggest using the residual bootstrap test alone to draw conclusion on the biomarker-defined subset treatment effect. On the other hand, we strongly recommend a follow-up biomarker stratified randomization trial to be conducted based on findings from residual bootstrap test.

**Acknowledgements** This work was supported in part by the grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). The computation was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: [www.sharcnet.ca](http://www.sharcnet.ca)) and Compute/Calcul Canada. The authors would like to thank the referees and the Editor for their insightful comments and suggestions.

## References

1. AALEN, O.O., Nonparametric inference for a family of counting processes, *Annals of Statistics*, 6, 701–726 (1978b).

2. BARKER, A.D., SIGMAN, C.C., KELLOFF, G.J., HYLTON, N.M., BERRY, D.A. and ESSERMAN, L.J., I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy, *Clinical pharmacology and Therapeutics*, 86, (1), 97–100 (2009).
3. BONETTI, M. and GELBER, R.D., A graphical method to assess treatment-covariate interactions using the cox model on subsets of the data, *Statistics in Medicine*, 19, (19) 2595–2609 (2000).
4. CHEN, B.E., JIANG, W. and TU, D., A Hierarchical Bayes model for biomarker subset effects in clinical trials, *Computational Statistics and Data Analysis*, 71, 324–334 (2014).
5. JIANG, W., FREIDLIN, B. and SIMON, R., Biomarker-adaptive threshold design: A procedure for evaluating treatment with possible biomarker-defined subset effect, *Journal of the National Cancer Institute*, 99, (13) 1036–1043 (2007).
6. LAZAR, A., COLE, B., BONETTI, M. and GELBER, R., Evaluation of treatment-effect heterogeneity using biomarkers measured on a continuous scale: subpopulation treatment effect pattern plot, *Journal of Clinical Oncology*, 28(29), 4539–44 (2010).
7. LUO, X. and BOYETT, J., Estimation of a threshold parameter in Cox regression. *Communications in Statistics - Theory and Methods*, 26, 2329–2346 (1997).
8. LOUGHIN, T.M., A residual bootstrap for regression parameters in proportional hazards model, *Journal of Statistical Computation and Simulation*, 77, 367–384 (1995).
9. NELSON, W., Theory and applications of hazard plotting for censored failure data, *Technometrics*, 14, 945–965 (1972).
10. PONS, O., Estimation in a Cox regression model with a change-point according to a threshold in a covariate, *The Annals of Statistics*, 31, 2442–2463 (2003).
11. ROYSTON, P., ALTMAN, D.G. and SAUERBREI, W., Dichotomizing continuous predictors in multiple regression: a bad idea, *Statistics in Medicine*, 25, 127–141 (2006).
12. ROYSTON, P. and SAUERBREI, W., A new approach to modeling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials, *Statistics in Medicine*, 23, (16) 2509–2525 (2004).
13. SARGENT, D.J., CONLEY, B.A., ALLEGRA, C. and COLLETTE, L., Clinical Trial Designs for Predictive Marker Validation in Cancer Treatment Trials, *Journal of Clinical Oncology*, 23, (9), 2020–2027 (2005).
14. SERFLING, R.J., *Approximation Theorems of Mathematical Statistics*, Wiley, New York (1980).
15. SIMON, R., Biomarker based clinical trial design, *Chinese Clinical Oncology*, 3, (3):39, 1–8 (2014).
16. ZANG, Y. and LEE, J.J., Adaptive clinical trial designs in oncology, *Chinese Clinical Oncology*, 3, (4) 1–20 (2015).