Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Subgroup Identification for Time-to-event Data Based on AFT Model

Sheng-li An (asl0418@126.com)

Pei Kang

Department of Biostatistics, School of Public Health, Southern Medical University

Ying-xin Liu

Deparment of Biostatistics, School of Public Health, Southern Medical University

Fu-qiang Huang

Deparment of Biostatistics, School of Public Health, Southern Medical University

Research article

Keywords: accelerated failure time model, adaptive design, change-point algorithm, false discovery rate, precision medicine, subgroup identification

Posted Date: December 17th, 2019

DOI: https://doi.org/10.21203/rs.2.18979/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Subgroup Identification for Time-to-event Data Based on AFT Model

Sheng-li An[1#], Pei Kang[1#], Ying-xin Liu[1], Fu-qiang Huang[1]

Authors' Affiliations: [1]State Key Laboratory of Organ Failure Research, National Clinical Research Center for Kidney Disease, Guangzhou, China; Department of Biostatistics, School of Public Health, Southern Medical University, Guangzhou, China

[#] **Sheng-li An and Pei Kang contributed equally to this work.**

Corresponding Authors: Sheng-li An, State Key Laboratory of Organ Failure Research, National Clinical Research Center for Kidney Disease, Department of Biostatistics, School of Public Health, Southern Medical University, 1838 North Guangzhou Avenue, Guangzhou 510515, China. Phone: 020-61649465; Email: asl0418@126.com

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The data analyzed during this study are included in these published article.

AIDS Clinical Trials Group Protocol 175 (ACTG175).

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by Grant from School of Public Health of Southern Medical University, China (GW201821).

Acknowledgements

Not applicable.

SUMMARYS

Considering the problem of identifying subgroup in a randomized clinical trial with respect to survival time, we present an analysis strategy to find subgroup of enhanced treatment effect. We fit univariate accelerated failure time (AFT) models with covariate-treatment interactions to identify predictive covariates. The false discovery rate is controlled by Benjamini-Hochberg procedure. Then a composite score conversion is employed to transform the set of identified covariates for each patient into a univariate score. To classify patient subgroups, a change-point algorithm is applied to searching for the threshold cutoff instead of using the median. Moreover, we adopted a biomarker adaptive design to check whether the treatment effect exists within certain subgroup. The simulation results show that the change-point method is remarkably superior to the median cutoff particularly when the subgroup sizes vary considerably. Furthermore, the 2-stage adaptive design has good power properties in detecting treatment effect while the type I error is generally controlled. As an illustration, we apply the proposed methods to an AIDS study. In conclusion, when the sample size is sufficient and the censoring rate is mild, the AFT model combined with change-point algorithm performs well in identifying subgroup.

Keywords: accelerated failure time model; adaptive design; change-point algorithm; false discovery rate; precision medicine; subgroup identification

1. BACKGROUND

Personalized medicine, which tailors to a patient's genetic and other unique characteristic, is a rapidly emerging field of health care. The ultimate goal of personalized medicine is to optimize the benefit of treatment by prescribing the right drugs for the right patients with minimal side effects [1]. Today, many cancer treatments are being developed for targeted therapies [2-4], in which only a subpopulation is anticipated to benefit from the therapy. To ensure the effect of personalized medicine, it is essential to identify the subgroup who beneficially responses to the targeted treatment than others based on each patient's characteristic. For this reason, the subgroup analysis, if properly used, can contribute to more informed clinical decisions, improved efficiency of the treatment, and reduced cost and side effects.

The subgroup analysis has been studied by a number of scholars [5-10]. But almost all of these studies focus on continuous or binary outcome and relatively few methods pertain to right-censored survival endpoints, while it is common that the outcome of interest is a time to an event in clinical data.

For survival outcomes, development of predictive covariates for treatment selection mainly consists of [11]: predictive covariates identification, and subgroup detection. Predictive covariates identification refers to establish a mathematic model to screen the candidate covariates which contribute to diverse treatment effect of patients. Subgroup detection means to identify the potential subgroup with higher treatment response through a classification model.

When interest lies in the relationship between covariates and duration time, the most popular regression model in survival analysis is Cox proportional hazards model. Chen's research [12] used Cox models to identify predictive covariates. But it only evaluates the performance of classifiers and models when the subgroup proportion is 0.2 and 0.5. It is not clear whether and how the sample size, predictive covariate size and censoring rate affect its performance. Besides, it does not explain why the significant level is directly set to be 0.001 and 0.0025 in the subgroup identification when multiple univariate models are used. An alternative to the Cox regression model is the AFT model

which doesn't rely on the proportional hazard assumption. The AFT model simply regresses the logarithm of the survival time over the covariates and has an intuitive interpretation. Thus, our work extends that of Chen's [12] in some ways. Our study will focus on the AFT model with the Benjamini-Hochberg procedure to adjust the significance level. Additionally, we will assess its performance when the sample size, censoring rate, subgroup proportion or the number of predictive covariates changes, while they were not provided in Chen's study.

The remainder of the paper is organized as follows. Section 2 presents the proposed analysis strategy for subgroup identification. In Section 3, details are demonstrated for a simulation study. In Section 4, an application to an AIDS study is described. The findings of the study are discussed in Section 5.

2. ANALYSIS STRATEGY FOR SUBGROUP IDENTIFICATION

2.1. Predictive Covariates Identification

For a given patient, let *TR* denote the arm indicator (*TR* = 0 for control and *TR* = 1 for treatment), and z_{ij} represent the measurement for the *j*-th covariate in the *i*-th patient, and t_i denote the observed survival time and y_i is $\log(t_i)$ ($i = 1,\ldots,n$), where *n* is total number of patients. Patients are randomly assigned to two arms.

Assume there exists two subgroups in the sampled patients: responder ($g+$) and non-responder ($g-$). Let $u_{S\cdot TR}$ denote $\mathrm{E}(y_{i|S\cdot TR})$, where the subscript *S* is subgroup indicator (*S* = 0 for $g-$ and *S* = 1 for $g+$). That is, the expected mean of the logarithm of survival time for the four subgroups is $u_{S\cdot TR}$. In our study, we only consider the situation where treatment is effective among $g+$ patients (i.e., $u_{00} = u_{10} = u_{01} < u_{11}$). The outcome y_i for patient *i* in the treatment arm (*TR* = 1) relies on the underlying value of the predictive covariates z_{ij}'s. Traditionally, univariate regression models are usually fitted to identify predictive covariates, in which, main effect of the *j*-th variable z_{ij} and treatment *TR* and their interaction (z_{ij}·*TR*) are included. In our study, we fit a univariate Weibull AFT model：

$$y_i = \log(t_i) = \alpha_j + \beta_{1j} z_{ij} + \beta_{2j} TR_i + \beta_{3j} z_{ij} \cdot TR + \varepsilon_{ij}, \quad (1)$$

where a significant interaction coefficient β_{3j} indicates different treatment responses between the $g+$ and $g-$ patients defined by the covariate z_{ij}; that is, the different z_{ij} values of the $g+$ and $g-$ patients might contribute to diverse treatment effects. And the set of variables z_{ij}, for which the coefficient β_{3j} is significant, are regarded as the so-called predictive covariates. Using Equation 1, however, the power to test for interaction effect β_{3j} is often poor [13].

In contrast, Freidlin and Simon [14] proposed excluding the main effect of covariate. In our study, the univariate Weibull AFT model without term $\beta_{1j} z_{ij}$ is as follows.

$$y_i = \log(t_i) = \alpha_j + \beta_{2j} TR_i + \beta_{3j} z_{ij} \cdot TR + \varepsilon_{ij}. \quad (2)$$

Equation 2 may possess a higher power to detect an interaction effect than Equation 1. Detailed evaluation and comparisons between Equation 1 and Equation 2 will be evaluated in our study.

As mentioned above, either Equation 1 or Equation 2 must be performed for each covariate in the study so that all potential predictive covariates are identified. As it involves multiple hypotheses testing, the significance threshold should be adjusted. There are adjustment approaches based on the concept of false discovery rate (FDR) , which is the proportion of significant results that are false positives[15]. In our study, the Benjamini-Hochberg procedure is employed to control the FDR. The method sorts the m P-values in increasing order and finds the largest k for which $P_k \leq k \div m \times q$, where q is the false discovery rate desired. All hypotheses with index at most k are rejected.

Let $U = \{x_1, x_2, \ldots, x_k\}$ denote the set of predictive covariates at the pre-specified false discovery rate of q=0.05 in the Benjamini-Hochberg procedure, where k is the number of significant x's. Based on the set U, a classification model will be developed for subgroup detection.

2.2. Subgroup Detection

Subgroup detection refers to selecting the responders in the patient population who tend to benefit from a particular treatment. It can generally be divided into two steps. The first step is to acquire each patient's response (score), which predicts each patient's treatment response on the basis

of the values of his/her predictive covariates x's. The second step is to dichotomize the patients by finding a cutoff-point for the predictive scores.

In the first step, let $w_j (j = 1,\ldots,k)$ be the weight assigned to the j-th predictive covariate. According to the composite score proposed by Matsui *et al.* [16], the weights can be the estimated regression coefficient β_{3j} ($j = 1,\ldots,k$) from Equation 1 or Equation 2. Thus, the predictive score for a patient with a set of covariate values is $l(x) = \Sigma_j w_j x_j = \Sigma_j \beta_{3j} x_j$.

In the second step, it may be an intuitive and simple choice to serve a percentile of the predictive scores as the cutoff-point based on the subgroup proportion, while the median is the most shared option [17-21]. Unfortunately, this ratio remains a mystery in reality. Besides, Freidlin and Simon [22] proposed a voting-based classifier (VBC). But 2 pre-determined tuning parameters R and G are necessitated and their specifications are often subjective and require additional analyses. Alternatively, we apply a change-point algorithm [23], which is a likelihood-based method and has been widely used to detect single or multiple change locations and divide data into a consecutive subset.

In our study, for a given ordered sequence of quantitative scores $l=\{l_{(1)},l_{(2)},\ldots,l_{(n)}\}$, a change-point will occur within this set when there exists an integer τ between 1 and n-1, such that the statistical properties of $\{l_{(1)},l_{(2)},\ldots,l_{(\tau)}\}$ and $\{l_{(\tau+1)},l_{(\tau+2)},\ldots,l_{(n)}\}$ are different in some way, such as mean and variance. And the change-point $l_{(\tau)}$ is to maximize the log-likelihood function,

$$ML(\tau) = log\, P\left(l_{(1):(\tau)}\middle|\hat{\theta}_1\right) + log\, P\left(l_{(\tau+1):(n)}\middle|\hat{\theta}_2\right), \quad (3)$$

where $P(.)$ is the probability density function associated with the distribution of the data, and $\hat{\theta}_1$, $\hat{\theta}_2$are the maximum likelihood estimates of the parameters. Once the cutoff-point $l_{(\tau)}$ is obtained, it is available to divide patients into responder and non-responder subgroup depending on their predictive score above or under $l_{(\tau)}$. The R package `changepoint` [24] is adopted to implement the change-point algorithm. Besides, Chen et al. [12] proposed applying the change-point method to the VBC method so that there is no need to specify 2 tuning parameters.

Besides, it is also crucial to assess whether there exists subgroups prior to the subgroup detection. On the basis of the predictive covariates, we also apply a bootstrap likelihood ratio test

(LRT) proposed by Gail and Simon [25] to evaluate the heterogeneity of patients. Provided that the test is significant at a pre-specified α level, then we proceed to subgroup selection. Otherwise, we stop the process and conclude that the sampled patients are homogeneous. The R package `mclust` [26] is adopted to perform the LRT.

2.3. *Biomarker Adaptive Design*

When a clinical trial involves one or more subgroups, it is crucial to assess the treatment effect on the overall patients as well as the $g+$ subgroup [11]. Biomarker adaptive designs have been developed to evaluate the treatment effect on the $g+$ subgroup [14, 27, 28]. Simply, biomarker adaptive design (presented in Figure 1) consists of an overall test at significance level α_1 and a subgroup test at α_2 ($\alpha_1+\alpha_2=\alpha$, typically, $\alpha = 0.05$). Firstly, a test of overall treatment effect is carried out at α_1 using the whole patients. Secondly, comparison of control and treatment arms in the selected subgroup $g+$ is carried out at α_2.

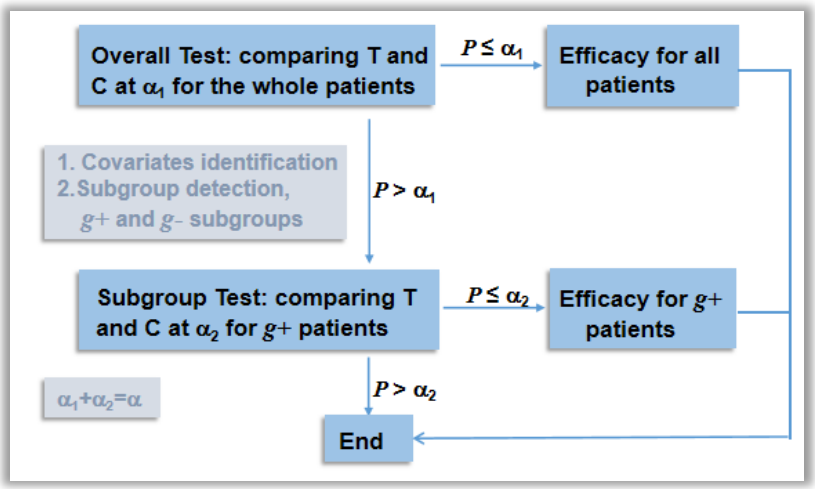

Figure 1. Biomarker adaptive design. T and C denote treatment group and control group, respectively.

3. SIMULATION

The pipeline of the proposed method is shown in Figure 2. Simulations are used to provide an evaluation of both the accuracy of subgroup detection procedure and the efficacy of treatment effect.

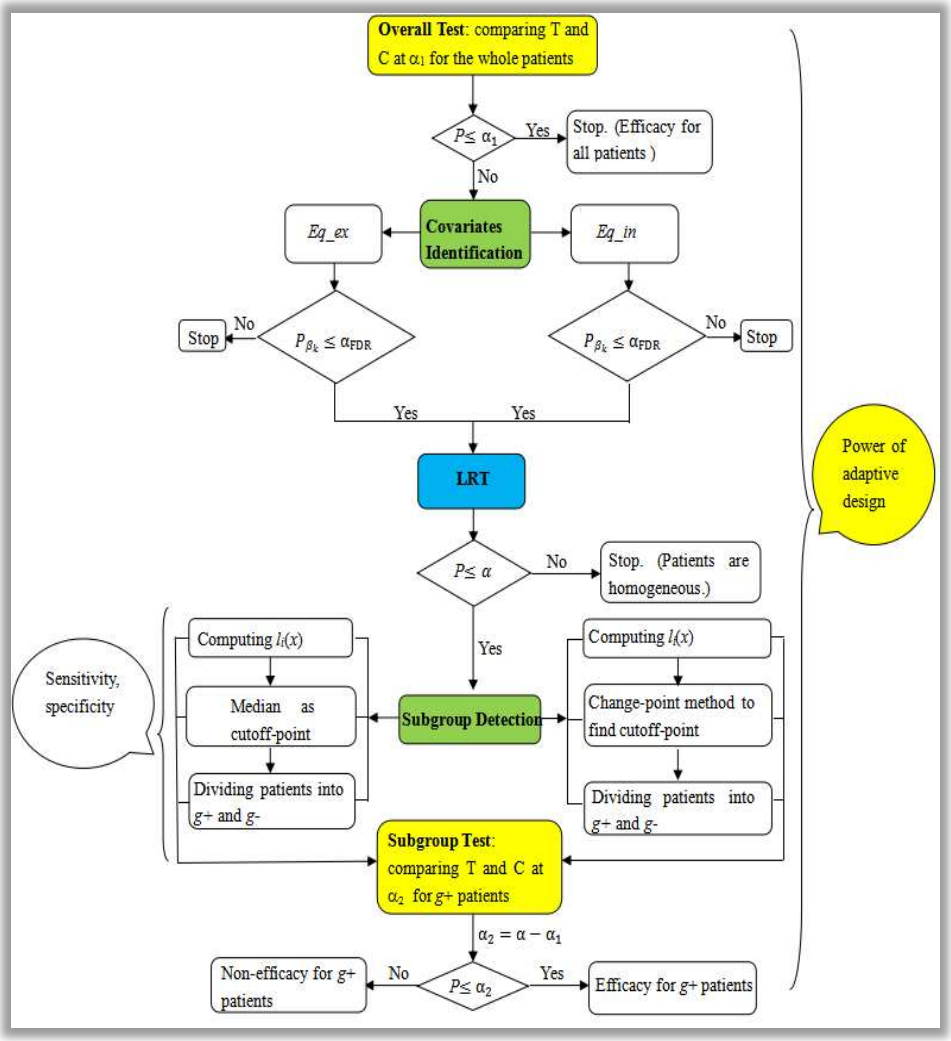

Figure 2. The pipeline of the proposed method. T and C denote treatment group and control group respectively. Eq_ex is the univariate Weibull AFT model excluding main effect of covariate, and Eq_in is the univariate Weibull AFT model including main effect of covariate.

3.1. Accuracy of Subgroup Detection Procedure

We consider 100 variables divided into two blocks: the first k variables serve as predictive covariates and are generated from $N(\mu_1, \sigma_1^2)$*and* $N(\mu_2, \sigma_2^2)$for the covariates values in $g+$ and $g-$ patients respectively. The remaining covariates are all generated from $N(\mu_2, \sigma_2^2)$. We generate data sets of different sample sizes with equal number of subjects in both arms, various predictive

covariates size, or diverse subgroup proportion. Censoring time is simulated from the exponential distribution so that the censoring rate ranges from 10% to 65%. The settings for μ_1, μ_2, σ_1, and σ_2 are to be 6, 8, 0.5, and 0.25, respectively.

Let y_i be the logarithm of failure time and x_{ij} be the covariate for the i^{th} individual in a random sample of size n. The survival time is simulated from the AFT model

$$y_i = \log(t_i) = \mu + \beta_1 TR_i + \sum_{j=1}^{k} \beta_j x_{ij} TR_i + \sigma\varepsilon_i, \; i = 1, \cdots, n, \qquad (4)$$

with the intercept term μ, the true regression coefficient vector β, the random error ε_i which follows the extreme value distribution and yields a Weibull regression model, and the scale parameter σ which equals 1. As aforementioned, we only consider the situation where treatment is only effective among $g+$ patients (i.e., $u_{00} = u_{10} = u_{01} < u_{11}$). According to Equation 3, y_i is equivalent to β_0 for all control arm patients, and equals to $\beta_0 + \beta_1 + \sum_j^k \beta_j \mu_1$ and $\beta_0 + \beta_1 + \sum_j^k \beta_j \mu_2$ for responders and non-responders respectively in the treatment arm. To satisfy the situation considered, we have $\beta_1 + \sum_j^k \beta_j \mu_2 = 0$. As μ_2 is set to be 8, so $\beta_1 + \sum_j^k \beta_j \times 8 = 0$.

Two predictive-covariate selection models are considered: the univariate Weibull AFT model excluding main effect of covariate (*Eq_ex*), and the univariate Weibull AFT model including main effect (*Eq_in*). Four subgroup selection classifiers are evaluated: composite score with change-point algorithm (*Composite C*), composite score with median (*Composite M*), VBC employed with change-point algorithm (*VBC C*), and VBC with median (*VBC M*). Responder detection is performed using 10-fold cross validation without performing LRT. Efficacy of the responder detection is evaluated in terms of sensitivity (the percentage of correct identification of $g+$ patients out of the true positive patients), specificity (the percentage of correct identification of $g-$ patients out of the true negative patients) and Youden index computed over 1000 simulated data sets.

In Figure 3, the performances of two predictive-covariate detection models and 4 classifiers are illustrated when the sample size, censoring rate, or subgroup proportion changes.

According to the results, model *Eq_ex* generally outperforms model *Eq_in*. An exception occurs in the model combined *Composite M* or *VBC M*, where *Eq_ex* yields higher sensitivity but lower specificity relative to *Eq_in*. In different situations, the sensitivity of four classifiers performs

similarly within the same model. As the sample size increases (shown in Figure 3.A), sensitivity of 4 classifiers shows dramatic increase. With growing censoring rate (presented in Figure 3.D), sensitivity of 4 classifiers declines dramatically. As the subgroup proportion rises (shown in Figure 3.G), sensitivity of 4 classifiers increases by about 10% and then maintains stable. In terms of specificity, the differences between classifiers with change-point and median become apparent. High specificity of the classifiers with change-point (*Composite C*, *VBC C*) maintains stable while that of the classifiers with median (*Composite M*, *VBC M*) drops by 4% to 26% when the sample size increases (presented in Figure 3.B), and rises by 13% to 26% with growing censoring rate or subgroup proportion (shown in Figure 3.E, Figure 3.H), respectively. In generally, the change-point algorithm has an advantage in subgroup selection particularly when the sizes of the 2 subgroups differ substantially. As the subgroup proportion comes up to 0.5, the advantage fade away. When the enhanced treatment effect is fixed, the performance of four classifiers diverges little under different predictive covariates sizes (figure not shown). In a word, with the increase of sample size or decrease of censoring rate, *Eq_ex* combined with *Composite C* gradually shows its advantage in subgroup detection.

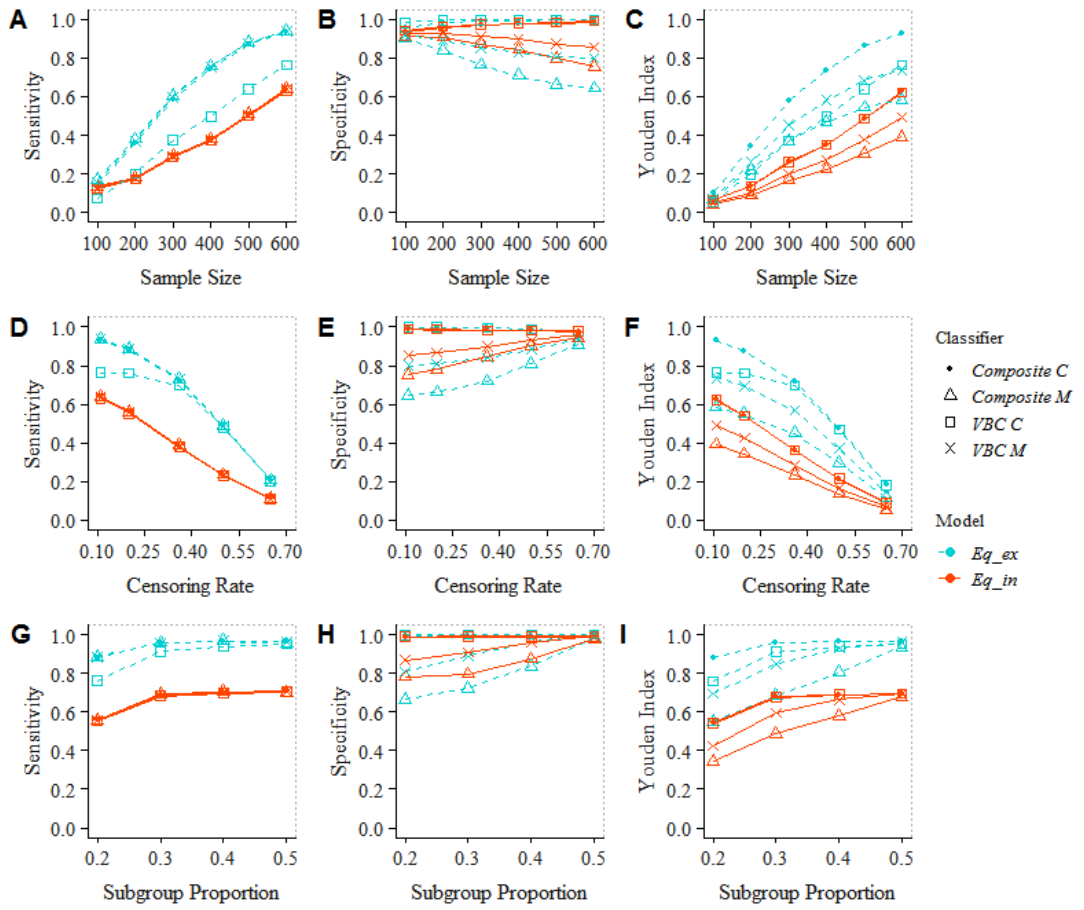

Figure 3. Performance of subgroup selection in different scenarios. Sensitivity is the percentage of correct identification of g+ patients out of the true positive patients, specificity is the percentage of correct identification of g- patients out of the true negative patients. The number of predictive covariates k is 4. In panels A, B, C, the censoring rate is 0.10 and the subgroup proportion is 0.20. In panels D, E, F, the sample size is 600 and the subgroup proportion is 0.20. In panels G, H, I, the sample size is 600 and the Diff is 0.10 (Diff=censoring rate - subgroup proportion/2).

3.2. The Efficacy of Treatment Effect

As mentioned above, we propose using the 2-stage biomarker adaptive design to increase the power of detecting treatment effect.

The powers of 4 classifiers with or without LRT are obtained. As LRT only slightly affects the power (result not shown here), only powers of classifiers without LRT in different situations are

presented in Figure 4, where the dotted line with inverted triangle (reference line) represents the empirical powers of the overall arm comparison at significant level 0.05 (conventional design). The other lines represent the adaptive powers in different situations. In addition, the power of the adaptive design was calculated as the percentage of replications with either positive overall test result (α_1=0.03) or subgroup test result (α_2=0.02). The simulation times is set to be 1000.

Comparing the performance of the conventional design to that of the adaptive design, it is obvious that the adaptive design enhances the power of detecting treatment effect. There appears to be a large improvement where the difference is more than 26% and 19.4% for *Eq_ex* and *Eq_in*, respectively, when the sizes of the 2 subgroups being substantially different. The powers of 4 classifiers within the same model show similar results. As a whole, *Eq_ex* combined with *Composite C* is powerful in detecting the treatment effect.

We also simulate 1000 null datasets to evaluate the type Ⅰ error of the procedure with or without LRT where there is no treatment effect and patients are homogenous. The results (figure not shown) suggest that the type Ⅰ errors of *Composite C*, *VBC C* and *VBC M* based on *Eq_ex* are controlled within the range of (0.02, 0.05) with or without LRT. Those of others are too large without LRT, but descend substantially to (0.02, 0.06) with the LRT.

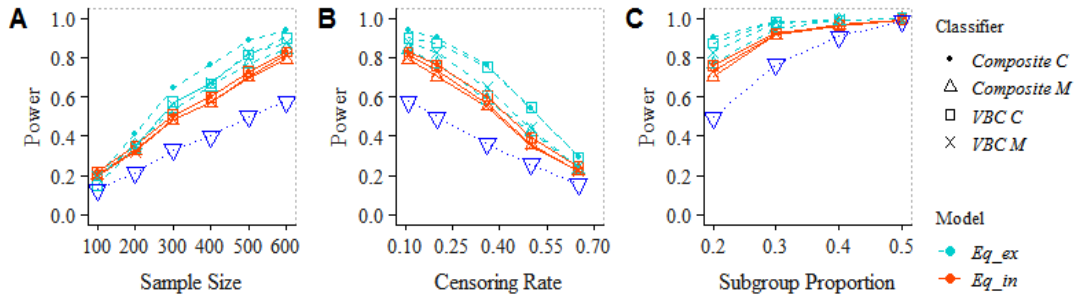

Figure 4. Power of 4 classifiers without LRT in different scenarios. The number of predictive covariates k is 4. In panel A, the censoring rate is 0.10 and the subgroup proportion is 0.20. In panel B, the sample size is 600 and the subgroup proportion is 0.20. In panel C, the sample size is 600 and the Diff is 0.10 (Diff=censoring rate - subgroup proportion/2). The dotted line with inverted triangle (reference line) represents the powers of conventional design.

4. APPLICATION TO AIDS STUDY

We illustrate the proposed method to data from AIDS Clinical Trials Group Protocol 175 (ACTG175). In this study, 2139 HIV-infected subjects were randomized to four different treatment groups, including zidovudine (ZDV) monotherapy, ZDV + didanosine (ddI), ZDV + zalcitabine (zal) and ddI monotherapy. We focus on patients receiving treatments: ZDV+ddI (denoted as 1) and ddI monotherapy (denoted as 0). Among them, there are 522 receiving treatment 1 and 559 receiving treatment 0. And the censoring rate is 80.27% and 77.10%, respectively. We choose the days of observing the event ((i) a decline in CD4 T cell count of at least 50 (ii) an event indicating progression to AIDS, or (iii) death) as the response. The covariates include patient's age and weight, the CD4 and CD8 counts (coded as CD40 and CD80 respectively), Karnofsky score (karnof), days of previously received antiretroviral therapy (preanti), hemophilia (hemo), homosexual activity (homo), history of intravenous drug use (drug), previous non-zidovudine antiretroviral therapy (oprior), previous zidovudine use in the 30 days (z30), previous zidovudine use (zprior), race, gender, antiretroviral history (str2), and symptomatic status (symptom). The first six variables are continuous while others are binary.

Our objective is to detect whether there is a subgroup with an enhanced treatment effect. In predictive covariate identification, we used the univariate AFT model excluding the main effect of covariates (*Eq_ex*) with the FDR controlled by Benjamini-Hochberg procedure. Two subgroup selection classifiers were considered: composite score with change-point algorithm (*Composite C*), composite score with median (*Composite M*). And 10-fold cross-validation is used.

Survival curves of the combination and separate medication groups is shown as Figure 5. The summary result is shown in Table 1 where the median survival time and group size (or subgroup size) for each arm are presented in Row 3 to Row 8. The corresponding *P* values for the overall test at α_1=0.03 and arm comparison in the selected subgroup at α_2=0.02 are presented in Row 1 and Row 2. As the 1st test in the 2-stage adaptive design is not significant (P = 0.181), we conclude that the treatment is not effective for all patients. In the 2nd test, the corresponding *P* values for the subgroup effect with the 2 classifiers are 0.369, and 0.356 respectively. And the survival curves for the 4 subgroups identified by the *Composite C* and *Composite M* are presented in Figure 6. It indicated that responders (T+ and C+) can be distinguished from non-responders (T- and C-),

although the difference between responders in ZDV+ddI group (T+) and responders in ddI group (C+) is non-significant. It means that the responders (or non-responders) in the two arms have the same treatment effect, while the responders have more enhanced treatment effect than the non-responders (P<0.001) (shown in Figure 7). The number of g+ patients identified by *Composite C* and *Composite M* was 280+304=584 (54.02%) and 259+273=532 (49.21%), respectively. The performance of *Composite C* and *Composite M* is similar, which may be due to g+ subgroup proportion near 0.5.

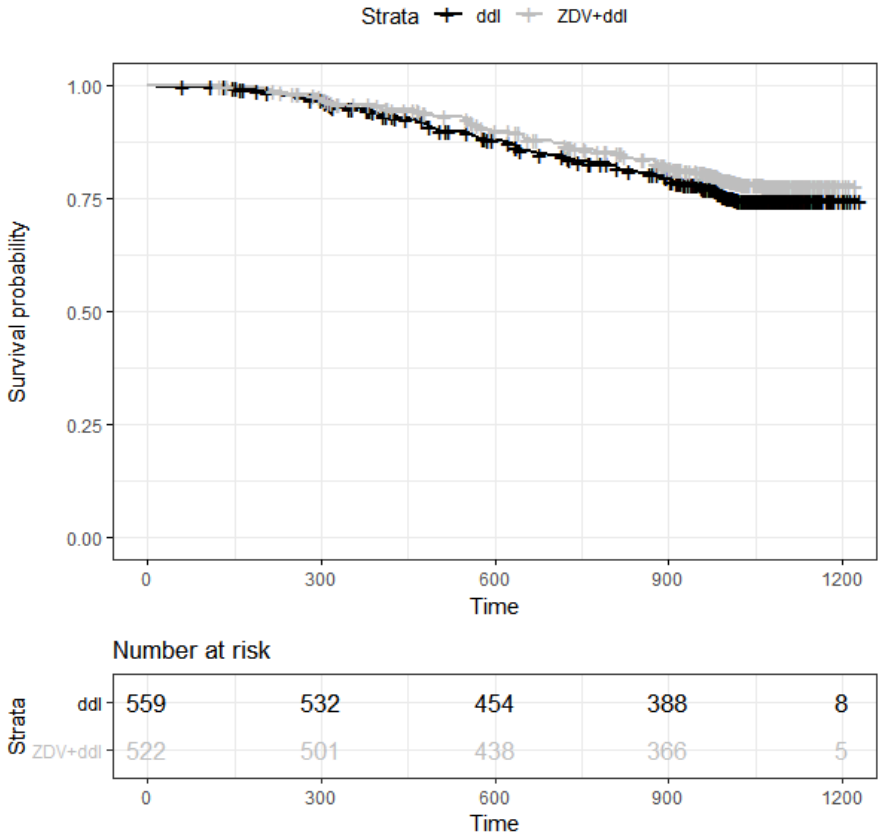

Figure 5. Survival curves of the combination and separate medication groups. ZDV+ddI indicates the zidovudine + didanosine group, and ddI indicates didanosine monotherapy.

Table 1. Summary result of ACTG175

	Composite C	*Composite M*

Overall test *	0.181	0.181
Subgroup test *	0.369	0.356
ZDV+ddI §	1014 (522)	1014 (522)
ddI§	1000 (559)	1000 (559)
ZDV+ddI *g*+§	1042 (280)	1048(259)
ZDV+ddI *g*-§	991 (242)	993 (263)
ddI *g*+§	1022 (304)	1029 (273)
ddI *g*-§	989(255)	990.5 (286)

Overall test: the 1st stage of biomarker adaptive design;

Subgroup test: the 2st stage of biomarker adaptive design;

** P value of Log-rank test;*

§ median survival time (group size or subgroup size).

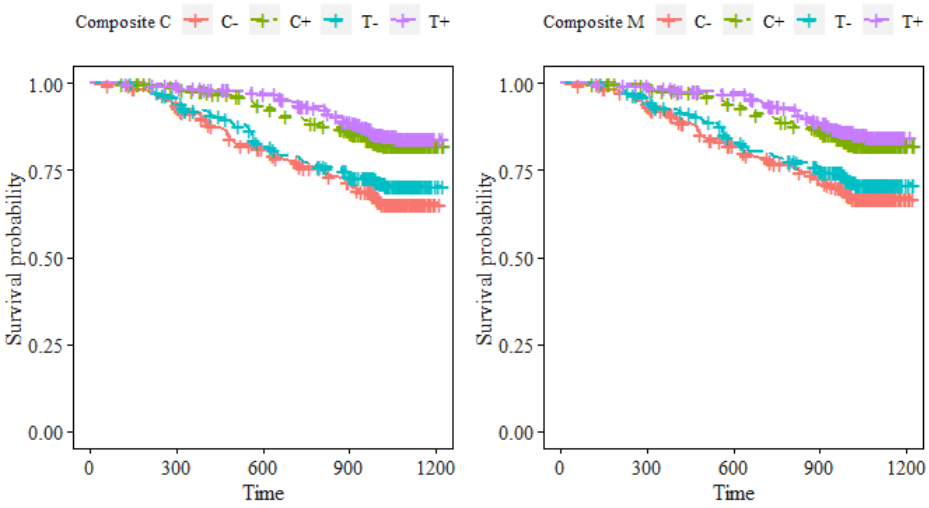

Figure 6. Survival curves for subgroups identified by Composite C (left) and Composite M (right). The four subgroups correspond to responder in ZDV+ddI group (T+), non-responder in ZDV+ddI group (T-), responder in ddI group (C+) and non-responder in ddI group (C-).

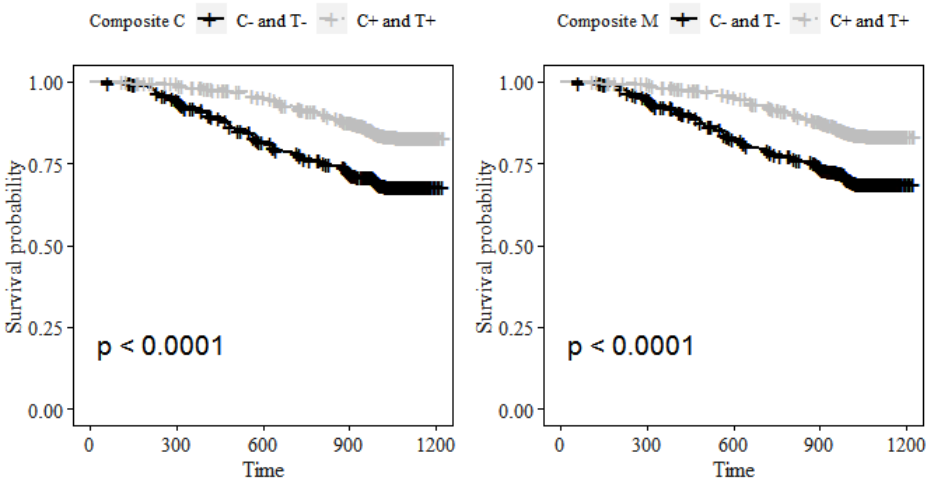

Figure 7. Survival curves for subgroups identified by Composite C (left) and Composite M (right). The two subgroups respectively correspond to responders in ZDV+ddI group (T+) and in ddI group (C+), non-responders in ZDV+ddI group (T-) and in ddI group (C-).

5. CONCLUSION AND DISCCUSION

Traditional methods often deal with the average effect of a new treatment. It is likely that in the new treatment group, some patients with certain characters (the so-called responders) may have a different treatment effect from the other ones when the two treatment arms do not differ in overall survival. There has been study focusing on the model-based selection for survival outcomes using Cox model combined with a change-point algorithm [12]. This study has evaluated whether the strategy proposed by Chen [12] can be used to subgroup identification in the AFT model with Benjamini-Hochberg procedure to control the FDR. These evaluations are made through simulations conducted across different sample size, censoring rate, subgroup proportion or predictive covariate size. In our study, relatively high-sensitivity and high-specificity are found in subgroup detection and largely improved power is presented in 2-stage biomarker adaptive design across different situations for the univariate AFT model excluding the main effect (*Eq_ex*) combined with change-point algorithm. Besides, its corresponding type Ⅰ error is controlled well no matter whether LRT is conducted whereas that of Chen's [12] (approximately 0.14) is large without performing the LRT.

This is mainly due to that the use of Benjamini-Hochberg procedure that can adjust the significant level in the predictive covariate identification, while it is not clear why the significant level of the Cox model one is directly set to be 0.001 and 0.0025 in Chen's study [12]. In addition, it is found that with the increase of sample size or decrease of censoring rate, *Eq_ex* combined with *Composite C* gradually shows its advantage in subgroup selection, but that is unknown in Chen's study [12].

In the real data analysis, classifier based on median and classifier based on change-point algorithm show similar effects, which is probably related to the high censoring rate (78.6%), and the small difference between *g*+ and *g*- subgroup sizes. Remarkably, it turns out that the identified responders enjoy a significantly beneficial treatment effect than the non-responders, irrespective of which therapy they received. Besides, the responders in both arms have the same therapeutic effect.

The above simulations and results are based on continuous variables. A future simulation to evaluate the performance with binary variables in our proposed analysis strategy is anticipated. Besides, the R package `mclust` is used in the LRT test, where categorical variables are not allowed. In the future, R package `mixtools` can be applied to data with both continuous and categorical variables. In addition, the univariate composite model only takes into account individual covariate contributions to the treatment response and it disregards the correlation among the covariates and their possible interaction effects. Possible alternative needs to be developed.

In predictive covariate identification with high dimensional survival data, in which the number of covariates is larger compared to sample size, a penalized AFT model based on the lasso, ridge or elastic net method may be used to screen out less useful covariates in future research. Furthermore, with the development of medical studies, more and more fatal diseases are now curable. Thus, there is an urgency to develop a method to identify cured subgroup patients in treatment group. The mixture cure model is a special type of survival models and it assumes that the studied population is a mixture of susceptible individuals who may experience the event of interest, and cure/non-susceptible individuals who will never experience the event. It is anticipated to apply the mixture cure model to find out the cure subgroup in future study.

References:

1. Kang S, Lu W, Song R: Subgroup detection and sample size calculation with proportional hazards regression for survival data. *STAT MED* 2017, 36(29).

2. Balis FM: Evolution of anticancer drug discovery and the role of cell-based screening. *J Natl Cancer Inst* 2002, 94(2):78-79.

3. Rothenberg ML, Carbone DP, Johnson DH: Improving the evaluation of new cancer treatments: challenges and opportunities. *NAT REV CANCER* 2003, 3(4):303-309.

4. Schilsky RL: End points in cancer clinical trials and the drug approval process. *CLIN CANCER RES* 2002, 8(4):935-938.

5. Tianxi C, Lu T, Peggy H W, L J W: Analysis of randomized comparative clinical trial data for personalized treatment selections. *BIOSTATISTICS* 2011, 12(2):270.

6. Zhao L, Tian L, Cai T, Claggett B, Wei LJ: Effectively Selecting a Target Population for a Future Comparative Study. *J AM STAT ASSOC* 2013, 108(502):527-539.

7. Song X, Pepe MS: Evaluating markers for selecting a patient's treatment. *BIOMETRICS* 2004, 60(4):874-883.

8. Shen J, He X: Inference for Subgroup Analysis With a Structured Logistic-Normal Mixture Model. *J AM STAT ASSOC* 2015, 110(509):303-312.

9. Marco B, Gelber RD: Patterns of treatment effects in subsets of patients in clinical trials. *BIOSTATISTICS* 2004, 5(3):465-481.

10. Foster JC, Taylor JMG, Ruberg SJ: Subgroup identification from randomized clinical trial data. *STAT MED* 2011, 30(24):2867-2880.

11. Chen JJ, Lu TP, Chen YC, Lin WJ: Predictive biomarkers for treatment selection: statistical considerations. *BIOMARK MED* 2015, 9(11):1121-1135.

12. Chen YC, Lee UJ, Tsai CA, Chen JJ: Development of predictive signatures for treatment selection in precision medicine with survival outcomes. *PHARM STAT* 2018, 17(2):105-116.

13. Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ: Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *J CLIN EPIDEMIOL* 2004, 57(3):229-236.

14. Freidlin B, Simon R: Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *CLIN CANCER RES* 2005, 11(21):7872-7878.

15. Tsai CA, Hsueh HM, Chen JJ: Estimation of False Discovery Rates in Multiple Testing: Application to Gene Microarray Data. *BIOMETRICS* 2003, 59(4):1071-1081.

16. Matsui S, Simon R, Qu P, Shaughnessy JJ, Barlogie B, Crowley J: Developing and validating continuous genomic signatures in randomized clinical trials for predictive medicine. *CLIN CANCER RES* 2012, 18(21):6065-6073.

17. Chen DT, Davis-Yadley AH, Huang PY, Husain K, Centeno BA, Permuth-Wey J, Pimiento JM, Malafa M: Prognostic Fifteen-Gene Signature for Early Stage Pancreatic Ductal Adenocarcinoma. *PLOS ONE* 2015, 10(8).

18. Chen DT, Hsu YL, Fulp WJ, Coppola D, Haura EB, Yeatman TJ, Cress WD: Prognostic and Predictive Value of a Malignancy-Risk Gene Signature in Early-Stage Non‐Small Cell Lung Cancer. *Journal of the National Cancer Institute* 2011, 103(24):1859-1870.

19. CQ Z, K D, D S, BA W, M M, N P, RK T, K N, C L, N L: Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. *J CLIN ONCOL* 2010, 28(29):4417-4424.

20. Tang H, Xiao G, Behrens C, Schiller J, Allen J, Chow CW, Suraokar M, Corvalan A, Mao J, White MA *et al*: A 12-gene set predicts survival benefits from adjuvant chemotherapy in non-small cell lung cancer patients. *CLIN CANCER RES* 2013, 19(6):1577-1586.

21. Chen HC, Kodell RL, Cheng KF, Chen JJ: Assessment of performance of survival prediction models for cancer prognosis. *BMC MED RES METHODOL* 2012, 12:102.

22. Freidlin B, Simon R: Adaptive Signature Design: An Adaptive Clinical Trial Design for Generating and Prospectively Testing A Gene Expression Signature for Sensitive Patients. *CLIN CANCER RES* 2005, 11(21):7872-7878.

23. Hinkley DV: Inference about the change-point in a sequence of random variables. *BIOMETRIKA* 1970, 57(1):1-17.

24. Killick R, Eckley IA: changepoint: An R Package for Changepoint Analysis. *J STAT SOFTW* 2015, 58(58):1-19.

25. Gail M, Simon R: Testing for qualitative interactions between treatment effects and patient subsets. *BIOMETRICS* 1985, 41(2):361-372.

26. Fraley C, Raftery AE: Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Publications of the American Statistical Association* 2002, 97(458):611-631.

27. Jiang W, Freidlin B, Simon R: Biomarker-Adaptive Threshold Design: A Procedure for Evaluating Treatment With Possible Biomarker-Defined Subset Effect. *J Natl Cancer Inst* 2007, 99(13):1036-1043.

28. Scher HI, Nasso SF, Rubin EH, Simon R: Adaptive clinical trial designs for simultaneous testing of matched diagnostics and therapeutics. *Clinical Cancer Research An Official Journal of the American*

Association for Cancer Research 2011, 17(21):6634.