Tests and Classifications in Adaptive Designs

Qiusheng Chen[*] Xu-Feng Niu[†]

Abstract

Statistical tests for biomarker identification and classification methods for patients grouping are two important topics in adaptive designs of clinical trials. In this article, we evaluate three test methods for biomarker identification: a model-based identification method, the popular t-test, and the nonparametric Wilcoxon Rank Sum test. For selecting the best classification methods in Stage 2 of an Adaptive Signature Design, we examine classification methods including the recent developed machine learning approaches such as Random Forest, Lasso and Elastic-Net Regularized Generalized Linear Models (Glmnet), Support Vector Machine(SVM), Gradient Boosting Machine (GBM), and Extreme Gradient Boosting(XGBoost). Statistical simulations are carried out in our study to assess the performance of biomarker identification methods and the classification methods. The best identification method and the classification technique will be selected based on the True Positive Rate(TPR,also called Sensitivity) and the True Negative Rate(TNR,also called Specificity).

Key Words: Adaptive Signature Designs, Boosting and Optimization, Classification trees, Sensitive Genes, Sensitive Patients, Targeted Agent.

[*] Qiusheng Chen is graduate student, Department of Statistics, Florida State University, Tallahassee, FL 32306

[†] Xu-Feng Niu is Professor, Department of Statistics, Florida State University, Tallahassee, FL, 32306

1. Introduction

Clinical trials play an important role in medical research, in which participants (usually human volunteers) receive specific interventions based on the protocol designed by the researchers. The interventions in a clinical trial could be different medical products, such as new drugs, new devices, or new procedures that are compared with a placebo. For example, in a clinical trial evaluating the effect of a new drug to reduce blood pressure, patients with high blood pressure may receive a certain dose of the drug to see whether their blood pressure decreases.

Different statistical designs are widely used in clinical trials. For example, in a randomized controlled clinical design with at least one control group, participants are randomly assigned to the treatments and the effects of different treatments are evaluated compared. In a group sequential design, patients are divided into a number of equal-sized groups, receive treatments sequentially, and the decision to stop the trial or not is based on repeated significance tests of the accumulated data after each group is evaluated. In a clinical trial usually the Double-Blind method is employed to avoid potential bias introduced by human factors, in which both participants and investigators are unaware who will get which specific treatment.

Adaptive designs in clinical trials were proposed in the late 1970s when Efron (1971) discussed how to balance a sequential experiment. Wei (1978) introduced a class of designs for sequential clinical trials, the biased-coin design, for the purpose of reducing experimental bias and increasing the precision of inference about treatment effects. The main idea of an adaptive design in clinical trials is that the investigator may modify trial and/or statistical procedures based on the review of data from different stages during the experimental process, which may identify clinical benefits of the treatments more efficiently and increase the success probability of the clinical development without undermining the validity and integrity of the trial. Chow et al. (2005) presented some statistical consideration of adaptive methods in clinical development, in which the authors mentioned that statistical procedures in a clinical trial including randomization, study design, study objectives/hypotheses, sample size, data monitoring and interim analysis, statistical analysis plan, and/or methods for data analysis. Group sequential designs in clinical trials were discussed by many authors, including Lan and DeMets (1978), Posch and Bauer (1999), Jennison and Turnbull (1999), and Liu et al. (1999). Chow and Chang (2008) provided a review on adaptive design methods in clinical trials, in which they pointed out the popularity of adaptive designs is mainly due to three reasons: reflecting the medical practice in real world, ethical with respect to both efficacy and safety (toxicity) of the test treatments under investigation, and flexible also efficient in the early and late phase of clinical development.

Adaptive Signature Design for clinical trials of targeted agent was proposed in last ten or more years. For instance, due to the heterogeneous feature of tumor types in an ontology study, a new generation of agents under development is molecularly targeted. When these agents enter the definitive stage of clinical evaluation, researchers ideally wish to use reliable assays to select sensitive patients, and restrict eligibility to patients with sensitive tumors perform specific evaluation on the subset. Freidlin and Simon (2005) proposed an Adaptive Signature Design (ASD) for generating and prospectively testing a gene expression signature for sensitive patients. In this proposed design, a signature to identify sensitive patients is not available. The design combines the prospective development of a pharmacogenomics diagnostic test (signature) to select sensitive patients with properly powered test for overall effect. The ASD consists of two steps, signature development and validation on mutually exclusive subgroups of patients (e.g., half of the population is used to develop a signature and another half to validate it). In the first step, a set of candidate predictive biomarkers (genes) are identified using the training data set. The response variable used in Freidlin and Simon (2005) was binary with as the probability of response for the patient. A logit model with each gene as the predictor was fit to the patients in the training set and genes with a significant coefficient were selected as predictive biomarkers based on a pre-specified type I error cutoff threshold. In the second step, the predictive biomarkers (or sensitive genes) identified in stage one were used to classify the patients in the test data set as sensitive patients and non-sensitive patients. Specifically, the ASD in Freidlin and Simon (2005) used a machine learning voting(MLV) method to identify the sensitive patient subgroup in the test data set, which involves two pre-specified tuning parameters R and G. After the patient classification, a test of treatment effect was performed on the sensitive patient subgroup.

The development of biomarker-adaptive designs including ASD generally involves three main components: biomarker identification, classifier development, and performance assessment. In this article we propose to study statistical tests and methods in the first two components of biomarker-adaptive designs: 1) compare three test methods for biomarker identification, i.e., the model-based identification method, the popular t-test, and the nonparametric Wilcoxon Rank Sum test; and 2) extend the classification method comparison performed by Lee et al. (2005) and compare classification methods including the recent developed machine learning approaches such as Random Forest, the Lasso and Elastic-Net Regularized Generalized Linear Models(Glmnet), Support Vector Machine(SVM), Gradient Boosting Machine(GBM) and the Extreme Gradient Boosting(XGBoost). The best identification method and the classification method will be selected based on the True Positive Rate (TPR) and the False Positive Rate (FPR).

Rest of this article will be organized as the follows. In Section 2, statistical structures of classification methods will be studied. Specifically, Gradient Boosting Machine (GBM) defined by Friedman (2001) will be discussed briefly, including function estimation and numerical optimization in function space. A comparison procedure for evaluating different test methods in biomarker identification and for selecting the best classification methods in an adaptive design will be proposed. Some basic terminologies used in the procedure will be defined, such as True Positive Rate(TPR,also called Sensitivity) and the True Negative Rate(TNR,also called Specificity).

In Section 3, Statistical simulations will be carried out to assess the performance of biomarker identification methods and the classification procedures, in which training data and testing data will be generated in different situations for the comparison of different methods. Specifically, binary response variable in a clinical trial will be generated by different logistic regression models and using different response rate. For predictors of the response variable, a large number of genes is generated based on normal distributions with a uniform random noise term. Subjects in the training data and testing data are grouped as sensitive patients and non-sensitive patients. The biomarker test methods and the classification methods will be applied to the simulated data. The best identification method and the best classification technique will be selected using the procedure proposed in Section 2.

Discussion and future study related to tests and classification methods in clinical trials will be presented in Section 4. Instead of a binary response variable, continuous response variables such as survival time of patients after treatments or time to recurrence of an event will be considered. Corresponding models for a continuous response variable such as Cox Proportional Hazard Model or the General Hazard Rate Model that extend the time-varying covariates and time-dependent effects models will be investigated in adaptive signature designs and subgroup identification. Applications of these models and techniques will be discussed too. Furthermore, some other subgroup identification methods will be investigated, including the Virtual Twins method proposed by Foster et al. (2011), the subgroup identification based on differential effect search (SIDES) method introduced by Lipkovich et al. (2011), and mining data to find subsets of high activity (ARF) by Dhammika and Javier (2004).

2. Classification Methods and Comparison Procedure

In this Section, we first introduce the statistical structures of boosting-based classification methods. In particular, we discuss the basic idea and statistical definition of the Gradient Boosting Machine (GBM) that was defined by Friedman (2001). The main techniques include function estimation and numerical optimization in function space. For evaluating different test methods in biomarker identification and for selecting the best classification methods in an adaptive design, we propose a comparison procedure to achieve this purpose.

2.1 Gradient Boosting Machine

In statistical classification analysis, a common task is to build a non-parametric regression model such as a classification tree based on a training data set, which are used to split new subjects from the testing data set into different groups. In order to increase the accuracy of classification, researchers in the later 1990s proposed some machine learning methods that generate many trees and aggregate results from a sequence of trees. Gradient boosting machines (GBM) are actually a family of powerful machine-learning techniques that is based on boosting and optimization, starting at some weak classifiers and generating the next classifier to improve the already trained ensemble classifier. The final predictions in GBM are generated by minimizing an arbitrary differentiable loss function.

Friedman (2001) developed explicit regression gradient boosting algorithm and discussed the relationship between boosting and optimization of function estimation. In function estimation, we usually have a random "response" variable y and a set of random "explanatory" variables denoted by $\mathbf{x} = \{x_1, \dots, x_n\}$. Suppose that the real relationship between y and $\mathbf{x}$ is $y = F^*(\mathbf{x}) + \epsilon$ where ϵ is a random error term. Based on an observed "training" sample $\{y_i, \mathbf{x}_i\}_{i=1}^N$, we want to find an estimate or approximation $\hat{F}(\mathbf{x})$ of the function $F^*(\mathbf{x})$ that minimizes the expected value of some given loss function $L(y, F(\mathbf{x}))$ over the joint distribution of all $(y, \mathbf{x})$-values,

$$F^* = \underset{F}{argmin}\, E_{y,\mathbf{x}} L(y, F(\mathbf{X})) = \underset{F}{argmin}\, E_{\mathbf{x}}[E_y(L(y, F(\mathbf{x})))|\mathbf{x}]. \tag{1}$$

Friedman (2001) pointed out that frequently employed loss functions in practice include squared-error $L(y, F) = (y - F)^2$ and absolute error $L(y, F) = |y - F|$ for $y \in \mathcal{R}^1$ for regression models and negative binomial log-likelihood, $L(y, F) = log(1 + e^{-2yF})$, when $y \in \{-1, 1\}$ for classification.

For the type of functions to be estimated, a common procedure in statistical analysis is to restrict $F(\mathbf{x})$ as a member of a parameterized class of functions $F(\mathbf{x}; \mathbf{P})$, where $\mathbf{P} = \{P_1, P_2, \dots\}$ is a finite set of parameters whose joint values identify individual class members. Friedman (2001) focused on "additive" expansions of functions with the form:

$$F(\mathbf{x}; \{\beta_m, a_m\}_1^M) \quad = \quad \sum_{m=1}^{M} \beta_m h(\mathbf{x}; a_m), \tag{2}$$

where the function $h(\mathbf{x}; a)$ in (2) is usually a simple parameterized function of the input variables $\mathbf{x}$, characterized by parameters $a = \{a_1, a_2, \dots\}$. The individual terms differ in the joint values a_m chosen for these parameters. For example, in the Classification and Regression Tree (CART) method introduced by Breiman et al. (1984), each of the functions $h(\mathbf{x}; a_m)$ is a small regression tree, in which the parameters a_m's are the splitting variables, split locations and the terminal node means of the individual trees.

When a parameterized model $F(\mathbf{x}; \mathbf{P})$ is chosen instead of a general non-parametric function $F(\mathbf{x})$, the function optimization problem becomes the following parameter optimization:

$$\mathbf{P}^* \quad = \quad \underset{P}{argmin}\, \Phi(\mathbf{P}), \tag{3}$$

where

$$\Phi(\mathbf{P}) = E_{y,x} L(y, F(\mathbf{x}; \mathbf{P}))$$

and

$$F^*(\mathbf{x}) = F(\mathbf{x}; \mathbf{P}^*).$$

For most $F(\mathbf{x};\mathbf{P})$ and loss function $L(y,F)$, usually there is no explicit solution for estimation the function $F(\mathbf{x};\mathbf{P})$. Some numerical optimization methods have to be used to solve (3), which often involves expressing the solution for the parameters in the form:

$$\mathbf{P}^* \quad = \quad \sum_{m=0}^{M} \mathbf{p}_m, \tag{4}$$

where $\mathbf{p}_0$ is an initial guess for the parameters and $\{\mathbf{p}_m\}_1^M$ are successive increments ("steps or boosts"), each based on the sequence of preceding steps. The prescription for computing each step $\mathbf{p}_m$ is defined by a given optimization method.

More details of this approach and algorithms were presented in Friedman (2001).

2.2 Method Comparison Procedure

In genetics study of human being and animals, usually a large number of genes are involved. For instance, recent studies estimated the number of genes in human to be between 19,000 and 20,000. For a given biological state or disease such as cancer, researchers often need to decide which gene and how many genes are related to the disease. As we mentioned in Section 1, in this study the following three tests will be used to screen and identify sensitive genes: a model-base test, the well-know t test, and the nonparametric Wilcoxon Rank Sum Test. For simplicity of the study, we consider only one treatment such as a new drug in a clinical trial versus a control group. The response variable is assumed to be binary with two possibilities: the patient has a response to the treatment such as a tumor size reduced, or the patient has no response to the treatment.

It should point out that many other statistical test methods could be used to detect sensitive genes among a large group of genes, such as the popular stepwise method and LASSO. However, when several thousand genes are involved in a research, using the stepwise method is time consuming and very slow. LASSO is relatively quicker but sometimes still select too many sensitive genes (predictors).

A simulation study will be carried out to compare the performance of the three gene-screening methods in Section 3. Similar to the criteria used in Troyanskaya et al. (2002), true positive rate (P^g_{sen}) and true specificity rate (P^g_{spec}) will be calculated for each of the three methods:

$$\begin{aligned} P^g_{sen} &= \frac{\text{Number of sensitive genes identified}}{\text{Number of known sensitive genes}} \\ P^g_{spec} &= \frac{\text{Number of non-sensitive genes identified}}{\text{Number of known non-sensitive genes}} \end{aligned} \tag{5}$$

In the simulation study, patients with sensitive genes (called sensitive patients) and without sensitive genes (called non-sensitive patients) will be generated. Six classification methods will be used on a training data set first then applied to a test data set for patient identification. The six classification methods are: Logistic regression, Support Vector Machine (SVM), Random Forest, the Lasso and Elastic-Net Regularized Generalized Linear Models (Glmnet), the Gradient Boosting Machine (GBM), and the Extreme Gradient Boosting (XGBoost).

Similar to the gene screening procedure, the true positive rate (P_{sen}^{p}) and true specificity rate (P_{spec}^{p}) defined below will be calculated for each of the six methods and used to evaluate the performance of the method:

$$\begin{aligned} P_{sen}^{p} &= \frac{\text{Number of sensitive patients identified}}{\text{Number of known sensitive patients}} \\ P_{spec}^{p} &= \frac{\text{Number of non-sensitive patients identified}}{\text{Number of known non-sensitive patients}} \end{aligned} \tag{6}$$

3. Simulation Study

In Section one we have discussed three test methods applicable for biomarker-identification (i.e. sensitive genes in this study) and six methods that are available for classification in the second stage of an adaptive design. In this section, simulation Study will be conducted to evaluate the performance of these methods. The best biomarker-identification method and the best subgroup classification method will be determined based on the results of simulation study. All the procedures, from data generation to the model fitting, are done by using the statistical computing package R.

3.1 Simulation Design

In this simulation design we assume that characterized patients with tumors are randomly assigned to treatment group or control group in a clinical trial. Data are simulated to describe expression levels of different microarray genes: the higher mean of gene expression value, the more sensitive the gene is. The number of sensitive genes in a patient determines the sensitivity of this patient. Therefore patients' response rates to treatment in the simulated data are determined in advance.

Specifically, assume that there are N patients, each patient has L evaluated genes, and K of the L genes are "sensitive". For the ith patient, let π_i denote its response rate and let t_i be the treatment that patient receives ($t_i = 0$ for patient with standard treatment or placebo, $t_i = 1$ for patient with novel treatment).

First gene expressions are generated from normal distributions with different means, and uniform random noises are added to the gene expressions for the purpose of approximation the true distributions of gene expressions observed in practice (see, e.g., Troyanskaya et al. (2002)). Similar to the simulation study conducted in Freidlin and Simon (2005), in this simulation the response probabilities of patients will be generated by the following logistic regression model:

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \lambda t_i + \gamma_1 t_i x_{i1} + \cdots + \gamma_K t_i x_{il},\ for\ i = 1, \ldots, n; \tag{7}$$

where π_i denote the probability of response for the ith patient, λ is the base level of treatment main effect regardless of gene expression of different patients, and the γ's are the coefficients of the interaction terms between treatment and sensitive gene expressions. To simplify the simulation, all gene main effects and the treatment-expression interactions for the nonsensitivity genes are assumed to be 0.

For further simplicity, in this simulation we assume $\gamma_1 = \cdots = \gamma_k$ in (7). Then the response rates for patients generated from (7) will be

$$\pi_i = \frac{e^{(\beta_0 + \lambda t_i + \gamma \cdot \sum_{k=1}^{K} t_i x_{ik})}}{1 + e^{(\beta_0 + \lambda t_i + \gamma \cdot \sum_{k=1}^{K} t_i x_{ik})}} \tag{8}$$

3.1.1 Simulation setup

In this simulation study, gene expressions for 400 patients will be generated first, with 200 patients in treatment group and the rest 200 patients in control group. Because of computing power limitations, we assume that each patient has $L = 1000$ genes. I had tried $L = 10,000$ and $L = 5,000$, but it would take several days on my copmputer to get results for one simulation run due to the slow looping in R.

Among the $L = 1000$ genes, we assume each sensitive patient has $K = 10$ sensitive genes, which are generated from normal distributions with positive mean μ_1: the larger of the mean μ_1 is, the higher response rate of the patient. All generated genes data are blurred by a set of randomly generated uniform noise. Non-sensitive patients are defines as these patients with nonsensitive genes that are generated from normal distributions with mean 0. Uniform random noises are also added to nonsensitive gene expressions.

Among the 400 patients, we assume 40 patients (10% of the total number of patients) have sensitive genes thus are sensitive to the treatment, while patients in the control group are assumed to be all non-sensitive, i.e. with mean 0 of gene expression.

The simulation setup in this study can be summarized as

- For representing different scenarios, gene expression levels are generated as follows:

  (a) sensitive genes in sensitive patients are generated from multivariate normal distribution in four different scenarios with different mean μ_1 ($\mu_1 = \{1.3, 1, 0.8, 0.6\}$) and variance σ_1 ($\sigma_1 = 1$).

  (b) non-sensitive genes in both sensitive patients and non sensitive patients are generated from multivariate normal distribution with mean 0 and variance σ_2 ($\sigma_2 = 1$).

- Random noises are generated from the uniform distribution. Different noise levels are applied, such as : $U(-0.01, 0.01)$, $U(-0.5, 0.5)$, and $U(-1, 1)$.

Table 1: Response rates π_i for sensitive patients and nonsensitive patients

Sensitive patients π_1	Mean of gene expression m
0.9661	1.3
0.9206	1.0
0.8641	0.8
0.7773	0.6
Nonsensitive patients π_0	Mean of gene expression m
0.2497	0
0.2497	0
0.2497	0
0.2497	0

- Binary response Y_i for patient i is generated from the Bernoulli distribution with probability π_i calculated from Model (7).

Set $\beta_0 = -1.1, \lambda = 0.55, \gamma_1 = \cdots = \gamma_k = 0.3$, the response rates for the four mean scenarios are listed in Table 1.

3.1.2 Simulation Procedure

Step 1: Use logistic model, Wilcoxon-test, t-test to select sensitive genes (p-value < 0.05 as a sensitive gene).

Logistic model based test: for each gene j fit the single gene logistic model $logit(\pi_i) = \mu + \lambda t_i + \beta_j t_i x_{ij}$ with treatment-expression interaction term or $logit(\pi_i) = \mu + \lambda_j t_i + \beta_j x_{ij}$ without interaction term. Decide a gene j to be sensitive if the p-value for β_j is significant at a specified level. If a sensitive gene is correctly selected as sensitive, then save the total counts in 500 loops for $L = 1,000$ genes as N^g_{l1}, if a nonsensitive gene is correctly select as nonsensitive then save the counts as N^g_{l2}.

Wilcoxon-test: for each gene j, test its expression by treatment group and control group. Decide gene j to be sensitive if the p-value for Wilcoxon test statistic "W_g" less than a specified level. If a sensitive gene is correctly selected as sensitive, then save the total counts iin 500 loops for $L = 1,000$ genes as N^g_{w1}, if a nonsensitive gene is correctly identified as nonsensitive then save the total counts in 500 loops for $L = 1,000$ N^g_{w2}.

t-test: for each gene j, test its expression by treatment group and control group. Decide gene j to be sensitive if the p-value for Student test statistic "T_g" is less than a specified level. If a sensitive gene is correctly selected as sensitive, save the total counts in 500 loops for $L = 1,000$ genes as N_{t1}^{g}, if a nonsensitive gene is correctly identified as nonsensitive then save the total counts in 500 loops for $L = 1,000$ genes N_{t2}^{g}.

Step 2: Combine response Y, TRT variable and the selected genes into new data set and then divide it into two halves, "`trainDF`" and "`testDF`". E.x. after performing t-test, the selected genes together with response and treatment variables form six training data sets: "`trainDF_t_lg`", "`trainDF_t_rf`", "`trainDF_t_gbm`", "`trainDF_t_svm`", "`trainDF_t_glm`", "`trainDF_t_xgb`"; and six testing data sets: "`testDF_t_lg`", "`testDF_t_rf`", "`testDF_t_gbm`", "`testDF_t_svm`", "`testDF_t_glm`", "`testDF_t_xgb`". In total we setup 18 "`trainDF`" data sets and 18 "`testDF`" data sets. $(6 * 3 = 18$ models).

Step 3: For each classification method, after getting 3 training set in the last step, the predicting step on "`testDF`"can be carried out. The response rates of all patients in "`testDF`" can be calculated. Patient with response rate of at least 0.5 is defined as sensitive patient, otherwise the patient is non sensitive. So far, if an originally sensitive patient is correctly selected by those classification models as sensitive, then count the correct identification time and save total counts. E.x. in one single classification step by using Support vector machine, number of sensitive patients correctly detected in test set "`testDF_t_svm`" is $N_{t\_svm_1}^{p}$ out of total 20 sensitive patients; number of nonsensitive patients correctly selected as nonsensitive in test set "`testDF_t_svm`" is $N_{t\_svm_2}^{p}$ out of total 180 non sensitive patients.

3.2 Simulation Results

The two criteria that will be used to assess efficiency of each test method and each classification method are Sensitivity and Specificity.

(i)Sensitivity of gene (P_{sen}^{g}) is the estimated probability that a sensitive gene is identified as sensitive. E.x. gene sensitivity for t-test is: $P_{sen}^{g} = N_{t1}^{g}/(1000 * 1\%)$

(ii)Sensitivity of patient (P_{sen}^{p}) is the estimated probability that a sensitive patient is classified as sensitive. E.x. patient sensitivity for t-test is: $P_{sen}^{p} = N_{t1}^{p}/(100 * 20\%)$.

(iii)Specificity of gene(P_{spec}^{g}) is the estimated probability that a nonsensitive gene is identified as nonsensitive. E.x. specificity for t-test is: $P_{sen}^{g} = N_{t2}^{g}/(1000 * 99\%)$

(iv)Specificity of patient(P_{spec}^{p}) is the estimated probability that a nonsensitive patient is classified as nonsensitive. E.x. patient specificity for t-test is: $P_{spec}^{p} = N_{t2}^{p}/(100 * 80\%)$.

3.2.1 Results on comparing three gene identification methods

Table 2 is a summary of the average of gene sensitivity and specificity in 1000 repeats under Model 7. Columns are results by the three methods of LR, t-test, Wilcoxon RST. Rows indicate four scenarios that the mean values of gene expression are of different levels $\{1.3, 1.0, 0.8, 0.6\}$. Uniform noises 0, $U(-0.01, 0.01)$, $U(-0.5, 0.5)$, and $U(-1, 1)$ are added to the generated covariate data. The simulation results show that the logistic model-based method outperformed the Wilcoxon Rank Sum test and the popular t-test, which is expected since the response variable was generated by the logistic regression model.

Table 2: Simulation results of Gene sensitivity and specificity by the three identification methods under Model 7 with uniform noises added to the generated covariate data

Mean of	Logistic		Wilcoxon		t-test	
Gene expression	regression		Rank Sum			
	P^g_{sen}	P^g_{spec}	P^g_{sen}	P^g_{spec}	P^g_{sen}	P^g_{spec}
U=0						
$m = 1.3$	0.9975	0.9520	0.9288	0.8622	0.9524	0.9496
$m = 1.0$	0.9715	0.9518	0.9129	0.8625	0.8167	0.9500
$m = 0.8$	0.9021	0.9518	0.8736	0.8629	0.6418	0.9499
$m = 0.6$	0.7433	0.9519	0.7701	0.8623	0.4264	0.9497
U(-0.01,0.01)						
$m = 1.3$	0.9977	0.9521	0.9291	0.8626	0.9529	0.9498
$m = 1.0$	0.9734	0.9518	0.9122	0.8625	0.8152	0.9499
$m = 0.8$	0.8998	0.9516	0.8715	0.8624	0.6388	0.9501
$m = 0.6$	0.7462	0.9518	0.7697	0.8628	0.4279	0.9501
U(-0.5,0.5)						
$m = 1.3$	0.9967	0.9517	0.9239	0.8629	0.9375	0.9501
$m = 1.0$	0.9686	0.9519	0.8895	0.8624	0.7838	0.9501
$m = 0.8$	0.8908	0.9519	0.8289	0.8629	0.6091	0.9502
$m = 0.6$	0.7420	0.9517	0.6929	0.8627	0.3973	0.9502
U(-1,1)						
$m = 1.3$	0.9935	0.9519	0.8898	0.8624	0.8766	0.9499
$m = 1.0$	0.9575	0.9516	0.7982	0.8627	0.6886	0.9499
$m = 0.8$	0.8769	0.9518	0.6853	0.8625	0.5128	0.9501
$m = 0.6$	0.7471	0.9517	0.5208	0.8627	0.3331	0.9501

3.2.2 Results on comparing six patients classification methods

Table 3 shows the average sensitivity and specificity of patient classifications in 500 replicates under Model 7 with uniform noises $U(-0.01, 0.01)$ added to the generated covariate data. Columns are the results by the six classification methods of Glmnet, Random Forest, GBM, LR, XGBoost and SVM. Rows indicate the four scenarios that the mean values of gene expression are of different levels $\{1.3, 1.0, 0.8, 0.6\}$. In each scenario, results from the three identification methods are presented.

Table 3: Simulation results of patients' sensitivity and specificity by the six classification methods under Model 7 with uniform noises $U(-0.01, 0.01)$ added to the generated covariate data

.

Mean of Gene expression	Gene test method	Sensitivity and Specificity	Glmnet	Random Forest	GBM	Logistic Regression	XGBoost	SVM
$m = 1.3$	LG	P^p_{sen}	0.8415	0.9894	0.9151	0.6951	0.9796	0.9196
		P^p_{spec}	0.8311	0.8382	0.8317	0.7970	0.8421	0.8542
	WRT	P^p_{sen}	0.9030	0.9870	0.9390	0.5327	0.9722	0.7184
		P^p_{spec}	0.8691	0.8489	0.8464	0.7830	0.8419	0.8616
	t-test	P^p_{sen}	0.9437	0.9891	0.9662	0.7568	0.9857	0.8988
		P^p_{spec}	0.8780	0.8609	0.9068	0.8140	0.8451	0.8770
$m = 1.0$	LG	P^p_{sen}	0.7411	0.9140	0.8313	0.6393	0.9262	0.7629
		P^p_{spec}	0.8305	0.8331	0.8305	0.8019	0.8379	0.8585
	WRT	P^p_{sen}	0.7940	0.9078	0.8417	0.5228	0.911	0.5346
		P^p_{spec}	0.8657	0.8415	0.8369	0.7791	0.8350	0.8828
	t-test	P^p_{sen}	0.7871	0.8816	0.8406	0.6629	0.9261	0.6224
		P^p_{spec}	0.8647	0.8462	0.8953	0.8149	0.8380	0.8928
$m = 0.8$	LG	P^p_{sen}	0.6543	0.7295	0.6812	0.6048	0.8425	0.6172
		P^p_{spec}	0.8280	0.8376	0.8318	0.8023	0.8364	0.8548
	WRT	P^p_{sen}	0.6770	0.7245	0.7016	0.5285	0.8030	0.3548
		P^p_{spec}	0.8538	0.8556	0.8427	0.7796	0.8314	0.9069
	t-test	P^p_{sen}	0.6057	0.6869	0.6882	0.5994	0.7974	0.4232
		P^p_{spec}	0.8562	0.8501	0.8964	0.8111	0.8305	0.8970
$m = 0.6$	LG	P^p_{sen}	0.5235	0.5610	0.5591	0.5518	0.7115	0.4142
		P^p_{spec}	0.8306	0.8365	0.8347	0.8063	0.8369	0.8778
	WRT	P^p_{sen}	0.5232	0.5073	0.5356	0.5134	0.6364	0.1659
		P^p_{spec}	0.8454	0.8578	0.8432	0.7788	0.8328	0.9464
	t-test	P^p_{sen}	0.4439	0.4820	0.4993	0.5205	0.6117	0.1924
		P^p_{spec}	0.8504	0.8482	0.8923	0.8044	0.8232	0.9382

Figure 1 presents the same results in Table 3 from the three test methods and the six classification methods, where different shapes and colors are used for the average sensitivity and specificity of patient classifications in 500 replicates under Model 7 with uniform noises $U(-0.01, 0.01)$ added to the generated covariate data. From the plots it is very clear that XGBoost outperformed other classification methods in terms of higher sensitivity, with the GBM and Random Forest the close second. SVM has higher average specificity than other five methods but with quite low sensitivity, especially for the cases $m = 0.8$ and $m = 0.6$.

Figure 1: Simulation results of patients' sensitivity and specificity by the six classification methods under Model 7 with uniform noises $U(-0.01, 0.01)$

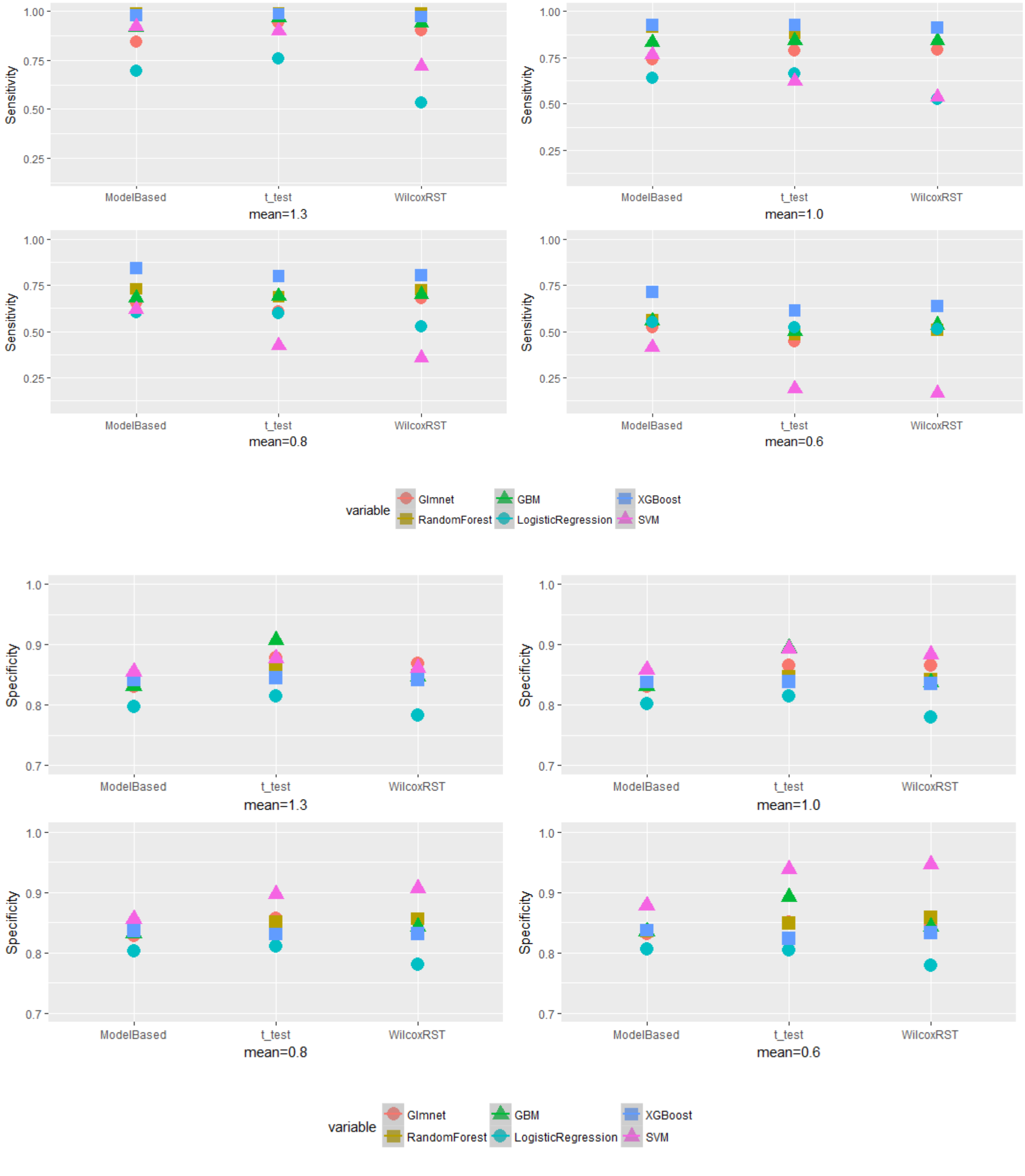


Table 4 lists the average sensitivity and specificity of patient classifications in 500 replicates under Model 7 with uniform noises $U(-0.5, 0.5)$ added to the generated covariate data. Columns are the results by the six classification methods of Glmnet, Random Forest, GBM, LR, XGBoost and SVM. Rows indicate the four scenarios that the mean values of gene expression are of different levels $\{1.3, 1.0, 0.8, 0.6\}$. Again in each scenario, results from the three identification methods are presented.

Table 4: Simulation results of patients' sensitivity and specificity by the six classification methods under Model 7 with uniform noises $U(-0.5, 0.5)$ added to covariates

.

Mean of Gene expression	Gene test method	Sensitivity and Specificity	Glmnet	Random Forest	GBM	Logistic Regression	XGBoost	SVM
$m = 1.3$	LG	P^p_{sen}	0.8415	0.9871	0.9059	0.6963	0.9734	0.9027
		P^p_{spec}	0.8311	0.8396	0.8340	0.8031	0.8420	0.8513
	WRT	P^p_{sen}	0.9030	0.9778	0.9199	0.5406	0.9616	0.6784
		P^p_{spec}	0.8691	0.8473	0.8446	0.7831	0.8426	0.8671
	T-test	P^p_{sen}	0.9437	0.9787	0.9458	0.7648	0.9807	0.8533
		P^p_{spec}	0.8780	0.8587	0.9033	0.8112	0.8447	0.8789
$m = 1.0$	LG	P^p_{sen}	0.7411	0.8881	0.8154	0.6334	0.9154	0.7608
		P^p_{spec}	0.8303	0.8288	0.8218	0.8069	0.8348	0.8508
	WRT	P^p_{sen}	0.7940	0.8897	0.8230	0.5295	0.8808	0.5050
		P^p_{spec}	0.8657	0.8374	0.8345	0.7822	0.8317	0.8813
	t-test	P^p_{sen}	0.7871	0.8583	0.8140	0.6711	0.8993	0.5964
		P^p_{spec}	0.8647	0.8386	0.8928	0.8112	0.8353	0.8891
$m = 0.8$	LG	P^p_{sen}	0.7411	0.7319	0.6824	0.6013	0.8309	0.5821
		P^p_{spec}	0.8305	0.8275	0.8267	0.8005	0.8321	0.8584
	WRT	P^p_{sen}	0.7940	0.7145	0.6732	0.5169	0.7799	0.3410
		P^p_{spec}	0.8657	0.8380	0.8369	0.7807	0.8287	0.9045
	t-test	P^p_{sen}	0.7871	0.6473	0.6370	0.5861	0.7722	0.4007
		P^p_{spec}	0.8647	0.8414	0.8924	0.8053	0.8255	0.8957
$m = 0.6$	LG	P^p_{sen}	0.5327	0.5271	0.5278	0.5386	0.7037	0.4103
		P^p_{spec}	0.8316	0.8404	0.8329	0.8055	0.8316	0.8762
	WRT	P^p_{sen}	0.5057	0.4934	0.5144	0.5087	0.6083	0.1335
		P^p_{spec}	0.8461	0.8454	0.8414	0.7820	0.8248	0.9540
	t-test	P^p_{sen}	0.4382	0.4794	0.4972	0.5158	0.5891	0.2044
		P^p_{spec}	0.8526	0.8439	0.8903	0.8035	0.8226	0.9320

Figure 2 presents the same results in Table 4 from the three test methods and the six classification methods, where different shapes and colors are used for the average sensitivity and specificity of patient classifications in 500 replicates under Model 7 with uniform noises $U(-0.5, 0.5)$ added to the generated covariate data. Similar to the results in Figure 1, XGBoost outperformed other classification methods in terms of higher sensitivity, with the GBM and Random Forest the close second. SVM has higher average specificity than other five methods but with quite low sensitivity, especially for the cases $m = 0.8$ and $m = 0.6$.

Figure 2: Simulation results of patients' sensitivity and specificity by the six classification methods under Model 7 with uniform noises $U(-0.5, 0.5)$

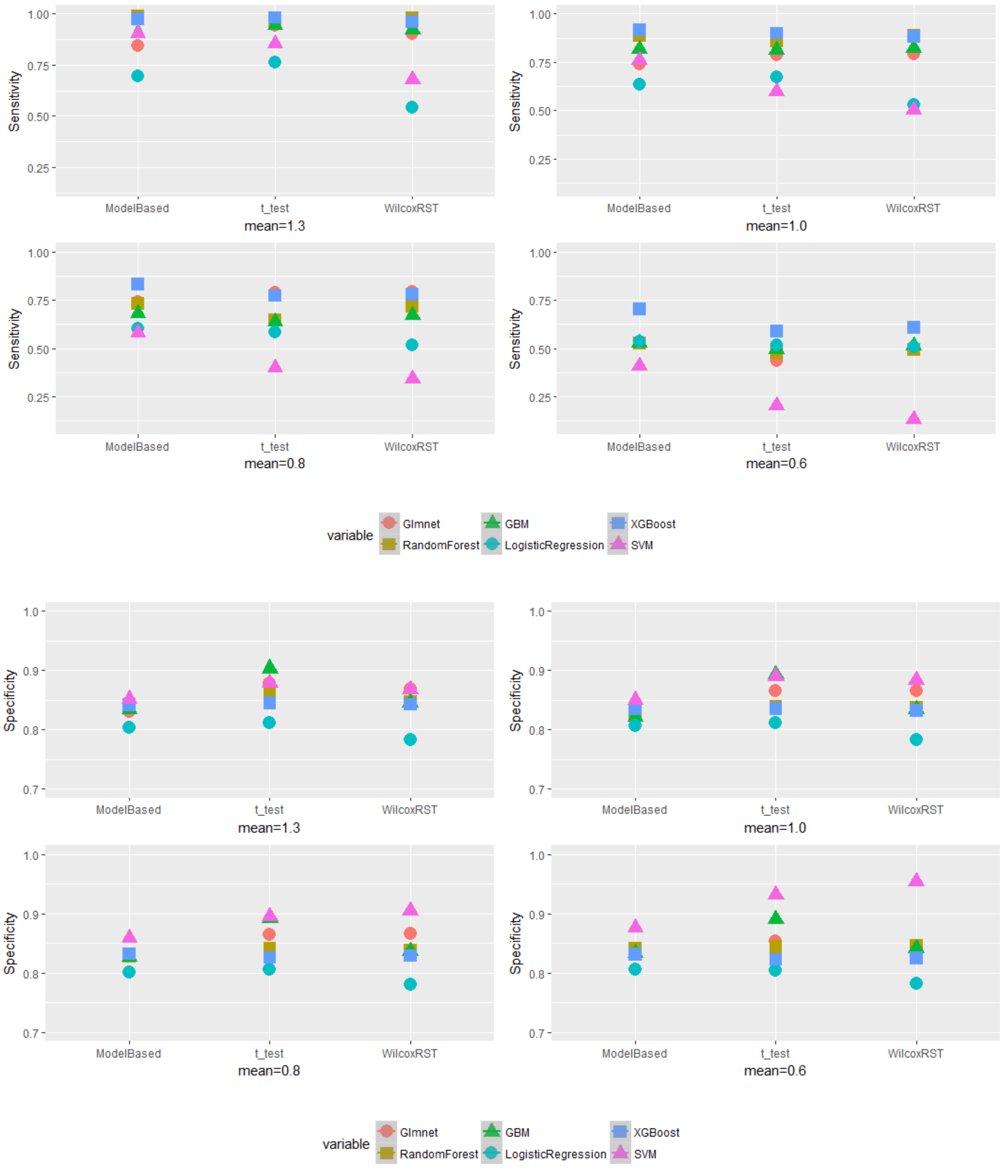


Table 5 presents the average sensitivity and specificity of patient classifications in 500 replicates under Model 7 with uniform noises $U(-1, 1)$ added to the generated covariate data. Columns are the results by the six classification methods of Glmnet, Random Forest, GBM, LR, XGBoost and SVM. Rows indicate the four scenarios that the mean values of gene expression are of different levels $\{1.3, 1.0, 0.8, 0.6\}$. In each scenario, results from the three identification methods are presented.

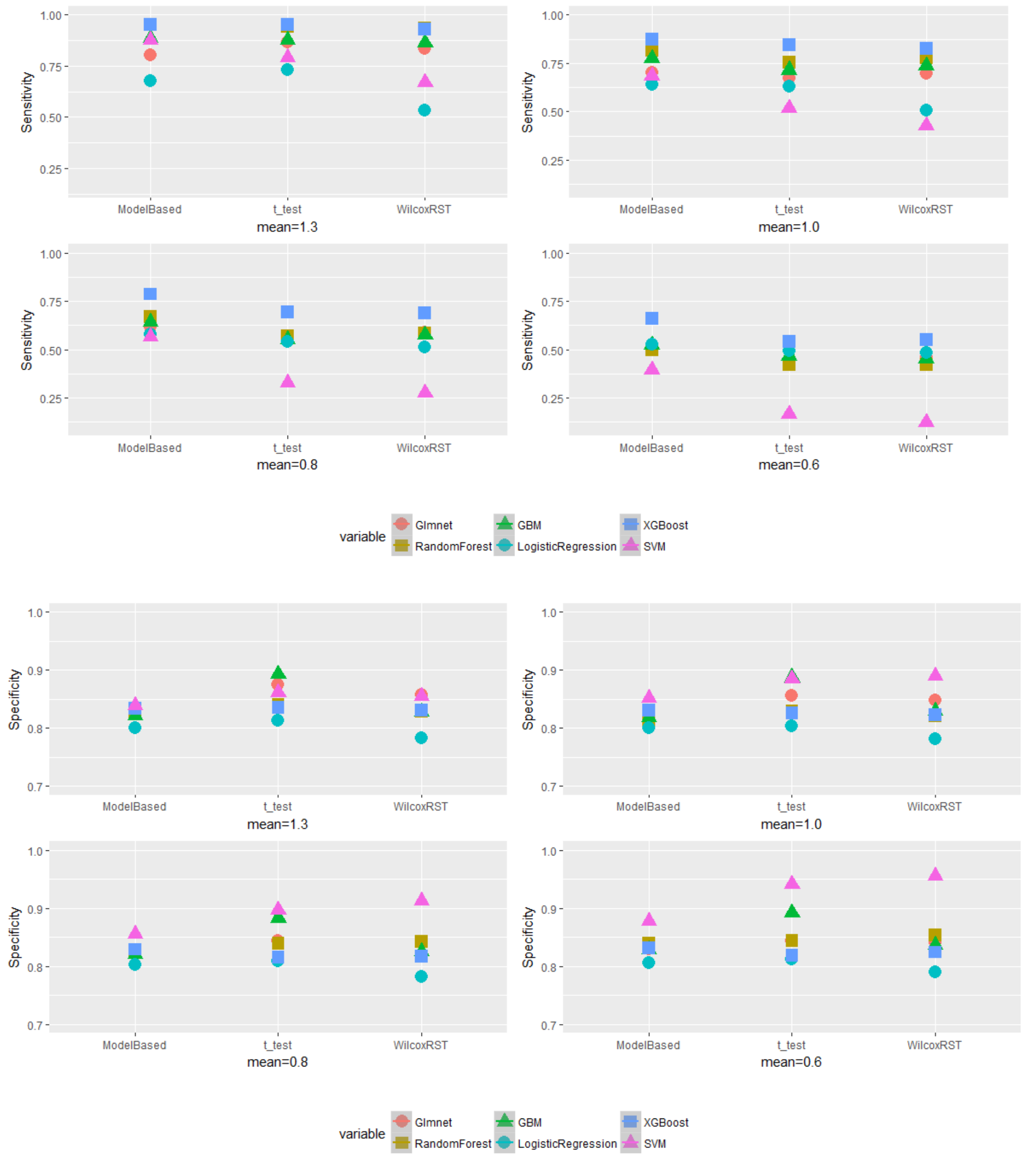
Table 5: Simulation results of patients' sensitivity and specificity by the six classification methods under Model 7 with uniform noises $U(-1, 1)$ added to covariates

.

Mean of Gene expression	Gene test method	Sensitivity and Specificity	Glmnet	Random Forest	GBM	Logistic Regression	XGBoost	SVM
$m = 1.3$	LG	P^p_{sen}	0.8019	0.9506	0.8865	0.6765	0.9526	0.8774
		P^p_{spec}	0.8245	0.8255	0.8217	0.8006	0.8336	0.8383
	WRT	P^p_{sen}	0.8331	0.9334	0.8629	0.5342	0.9308	0.6677
		P^p_{spec}	0.8579	0.8288	0.8283	0.7824	0.8308	0.8548
	t-test	P^p_{sen}	0.8657	0.9407	0.8783	0.7297	0.952	0.7899
		P^p_{spec}	0.8746	0.8403	0.8934	0.8127	0.8354	0.8619
$m = 1.0$	LG	P^p_{sen}	0.7002	0.8156	0.7729	0.6393	0.8724	0.6857
		P^p_{spec}	0.8272	0.8127	0.8191	0.8007	0.8311	0.8516
	WRT	P^p_{sen}	0.6980	0.7759	0.7366	0.5070	0.8240	0.4244
		P^p_{spec}	0.8483	0.8202	0.8300	0.7815	0.8229	0.8897
	t-test	P^p_{sen}	0.6710	0.7523	0.7148	0.6304	0.8434	0.5149
		P^p_{spec}	0.8554	0.8288	0.8874	0.8037	0.8252	0.8849
$m = 0.8$	LG	P^p_{sen}	0.6191	0.6713	0.6436	0.5798	0.7864	0.5655
		P^p_{spec}	0.8279	0.8254	0.8202	0.8019	0.8289	0.8555
	WRT	P^p_{sen}	0.5903	0.5863	0.5776	0.5148	0.6883	0.2771
		P^p_{spec}	0.8421	0.8417	0.8252	0.7813	0.8173	0.9132
	t-test	P^p_{sen}	0.5505	0.5711	0.5508	0.5421	0.6923	0.3273
		P^p_{spec}	0.8445	0.8395	0.8833	0.8094	0.8154	0.8973
$m = 0.6$	LG	P^p_{sen}	0.5163	0.4980	0.5218	0.5233	0.6612	0.3947
		P^p_{spec}	0.8295	0.8392	0.8278	0.8067	0.8308	0.8778
	WRT	P^p_{sen}	0.4584	0.4217	0.4515	0.4809	0.5509	0.1250
		P^p_{spec}	0.8399	0.8537	0.8371	0.7864	0.8240	0.9555
	t-test	P^p_{sen}	0.4282	0.4235	0.4670	0.4808	0.5419	0.1677
		P^p_{spec}	0.8443	0.8443	0.8935	0.8149	0.8185	0.9406

Figure 3 presents the same results in Table 5 from the three test methods and the six classification methods, where different shapes and colors are used for the average sensitivity and specificity of patient classifications in 500 replicates under Model 7 with uniform noises $U(-1, 1)$ added to the generated covariate data. Results in this Figure are similar to the results shown in Figures 1 and 2.

Figure 3: Simulation results of patients' sensitivity and specificity by the six classification methods under Model 7 with uniform noises $U(-1, 1)$



Based on the simulation results, we conclude the model-based test is the best choice for initial covariate selection. For patient classification, XGBoost outperformed other methods in terms of higher sensitivity, performance of GBM and Random Forest is quite robust in both sensitivity and specificity, and SVM has the highest specificity but with quite low sensitivity. These three classification methods, XGBoost, GBM, and Random Forest, should be used and compared further in real data analysis.

4. Future Study

In our simulation study logistic models were assumed for the relationship between binary response and covariates. However in real world the types of covariates, response, and models could vary. So we may also consider continuous response such as survival time of patients after treatments or time to recurrence of an event. For continuous response, Cox Proportional Hazard Model or the General Hazard Rate Model that extends the time-varying covariates and time-dependent effects models could be used in ASD and subgroup identification.

In the comparison procedure discussed in Section 2, we use biomarker identification methods to search on the covariate space of genomic data and use most recent machine learning classification methods to identify subgroup then perform treatment effect analysis on the classified subgroup. There are several exploratory approaches of subgroup identification reported in recent years, which utilize tree-based methods such as the basic CART algorithm or random forest and meanwhile incorporating evaluation of treatment effect during the identification process. These methods, including the Subgroup identification based on differential effect search (SIDES) method, ARF (Activity Region Finder) and Virtual Twins, aim at finding a specific covariate subspace, with which patients would expect to have higher treatment effect than the counter subset. These give us a new insight for our "test + classification" procedure. We would try to borrow their ideas in the second stage of our enriched ASD, compare the results of using our selected methods to evaluate treatment effect and results by replacing the Random Forest, GBM, XGBoost with these three methods in the classification step.

In the future research, we plan to study the above three subgroup identification methods carefully. In the traditional subgroup identification approaches, including the machine learning methods Random Forest, GBM, and XGBoost, high-dimensional covariate space of patients, such as the space spanned by thousands of genes, are participated into subspaces with much lower dimensions, such as subspaces generated by very few sensitive genes or biomarkers. After the covariate space partition, the treatment effect on patients with a specific set of covariates, such as patients with sensitive genes under treatment vs other patients, is evaluated. Subgroup identification procedures such as SIDES take a more localized partition of the covariance space and focus on identification of 'interesting' areas in the covariate space, such the areas specified by $I(x_1 < c1)$ and $I(x_2 > c2)$ where the treatment effect is likely to be large. In my future study, both the traditional subgroup identification methods and the covariate local-focused methods will be apply to real data analysis. Results from the two types of subgroup identification methods will be evaluated and compared.

References

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

Chow, S., Chang, M., and Pong, A. (2005). Statistical consideration of adaptive methods in clinical development. *Journal of Biopharmaceutical Statistics*, 15(4):575–591.

Chow, S.-C. and Chang, M. (2008). Adaptive design methods in clinical trials a review. *Orphanet Journal of Rare Diseases*, pages 3–11.

Dhammika, A. and Javier, C. (2004). Mining data to find subsets of high activity. *Journal of Statistical Planning and Inference*, 122:23–41.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, 58(3):403–417.

Foster, J., Taylor, J., and Ruberg, S. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30:28672880.

Freidlin, B. and Simon, R. (2005). Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Journal of the American Statistical Association*, 11(21):7872–8.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Jennison, C. and Turnbull, B. (1999). *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC, New York, NY.

Lan, K. and DeMets, D. L. (1978). Group sequential procedures: calendar versus information time. *Statistics in Medicine*, 8:1191–1198.

Lee, J., Lee, J., Park, M., and Song, S. (2005). An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, 48(4):869–885.

Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011). Subgroup identification based on differential effect searcha recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, 30(21):2601–21.

Liu, Q., Proschan, M. A., and Pledger, G. W. (1999). A unified theory of two-stage adaptive designs. *Journal of the American Statistical Association*, 97:1034–1041.

Posch, M. and Bauer, P. (1999). Adaptive two-stage designs and the conditional error function. *Biometrical Journal*, 41:689–696.

Troyanskaya, O., Garber, M., Brown, P., Botstein, D., and Altman, R. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18(11):1454–61.

Wei, L. (1978). The adaptive biased-coin design for sequential experiments. *The Annals of Statistics*, 6(1):92–100.