




A three-stage approach to identify biomarker signatures for cancer genetic data with survival endpoints

Xue Wu¹ · Chixiang Chen² · Zheng Li³ · Lijun Zhang⁴ · Vernon M. Chinchilli¹ · Ming Wang⁴ 

Accepted: 11 February 2024
© The Author(s) 2024

Abstract

The identification of prognostic and predictive biomarker signatures is crucial for drug development and providing personalized treatment to cancer patients. However, the discovery process often involves high-dimensional candidate biomarkers, leading to inflated family-wise error rates (FWERs) due to multiple hypothesis testing. This is an understudied area, particularly under the survival framework. To address this issue, we propose a novel three-stage approach for identifying significant biomarker signatures, including prognostic biomarkers (main effects) and predictive biomarkers (biomarker-by-treatment interactions), using Cox proportional hazard regression with high-dimensional covariates. To control the FWER, we adopt an adaptive group LASSO for variable screening and selection. We then derive adjusted p -values through multi-splitting and bootstrapping to overcome invalid p values caused by the penalized approach's restrictions. Our extensive simulations provide empirical evaluation of the FWER and model selection accuracy, demonstrating that our proposed three-stage approach outperforms existing alternatives. Furthermore, we provide detailed proofs and software implementation in R to support our theoretical contributions. Finally, we apply our method to real data from cancer genetic studies.

Keywords Survival outcomes · High-dimensional data · Cox proportional hazard · Group LASSO · Biomarker selection · Family-wise error rate

✉ Ming Wang
mxw827@case.edu

¹ Division of Biostatistics and Bioinformatics, Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA, USA

² Division of Biostatistics and Bioinformatics, Department of Epidemiology and Public Health University of Maryland School of Medicine, Baltimore, MD, USA

³ Novartis Pharmaceuticals, East Hanover, NJ, USA

⁴ Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, USA

1 Introduction

Personalized medicine has gained escalating importance in contemporary clinical practice, as the potential for tailored treatments designed for individual patients holds promise for augmenting the effectiveness of interventions (Hamburg and Collins 2010; Chin 2011). However, in the case of diseases such as cancer, significant heterogeneity exists among patients, impacting disease progression and responses to specific treatments. Consequently, identifying cancer subgroups and disparities in diagnoses can be immensely valuable in tailoring optimized therapies for each patient, ultimately improving healthcare outcomes, including survival rates. To achieve this goal, the discovery of both prognostic markers (primary biomarkers) and predictive biomarkers (biomarker-treatment interactions) is crucial in cancer drug development and clinical practice. These biomarkers have the potential to personalize therapies for individual patients and enhance treatment effectiveness.

Recent advances in biotechnology have led to the generation of vast amounts of complex biological and molecular data. Modern high-throughput technologies can simultaneously measure the expression levels of thousands of genes. Databases like the Gene Expression Omnibus (GEO) and Array Express provide extensive resources for cancer genetic research (Barrett 2010). However, a common challenge in these datasets is that the number of genes (p) is often equal to or even greater than the number of samples (n). This situation becomes more complex when researchers aim to identify both prognostic biomarkers (genes) and predictive biomarkers (gene-treatment interactions), increasing the dimensionality of the data. In such cases, it is crucial to assess the significance of each variable using p values and to correct for multiple testing, especially in clinical applications. Of note, controlling the false positive rate, specifically the family-wise error rates (FWER), is a paramount consideration. False positives can lead to erroneous conclusions, resource wastage, the diversion of research efforts towards unproductive avenues, among others. Hence, our commitment to controlling false positives is rooted in the need to uphold the integrity of scientific and medical research. The central focus of this paper resides in the identification of biomarker signatures, encompassing both prognostic and predictive biomarkers, through the assignment of valid p values within the context of the survival framework, all while effectively controlling FWER.

Over the past two decades, a multitude of regularization techniques have emerged to facilitate feature selection and yield sparse parameter estimates when grappling with high-dimensional data. One prominent method is the least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996), a preeminent method known for furnishing sparse estimates through an L_1 penalty, encompassing optimal tuning parameters for linear models. This work has led to the development of various extensions (Fan and Li 2001; Zou 2006; Ghosh 2007; Yuan and Lin 2006; Wang and Leng 2008). To deal with ultra-high-dimensional data, Fan and Lv (2008a) introduced a correlation screening technique termed Sure Independence Screening (SIS). This method effectively reduces the

dimensionality from an extremely large scale to a more manageable one, making it easier to use LASSO for variable selection. This combination is referred to as SIS-Lasso. These techniques, tailored for variable selection and screening, have been expanded to accommodate survival outcomes within the context of Cox proportional hazards (PH) models (Tibshirani 1997; Fan and Li 2002; Zhang and Lu 2007; Simon 2011; He 2019; Fan 2010; Zhao and Li 2012).

While regularization and screening techniques are effective in producing sparse and interpretable estimates, they face challenges in maintaining control over type I error rates, primarily due to issues with p values obtained from penalized likelihood. Recent advancements have addressed these challenges in high-dimensional linear models. For example, Wasserman and Roeder (2009) introduced a "screen and clean" procedure, involving data division into a training set for variable screening (using LASSO) and a testing set for significance testing. Several other methods exhibit screening properties, such as the adaptive Lasso (Zou 2006) and the smoothly clipped absolute deviation (SCAD) Fan and Li (2001). Later, Meinshausen (2009) enhanced the approach by iteratively repeating the split-and-fit procedure, computing p values for each split, and aggregating them to establish a collective p -value for the purpose of controlling FWER. Related works include Meinshausen and Yu (2009); Bühlmann (2013); Zhang and Zhang (2014); Dezeure (2015). Recent breakthroughs in this domain include the work of Zuo et al. (2021), who introduced a groundbreaking variable selection approach termed "penalized regression with second-generation p values" (ProSGPV). This method combines an L_1 penalization scheme with second-generation p values (SGPV) to identify variables suitable for inclusion in the model. While these methods have proven effective in generalized linear models for high-dimensional data, their application within the survival framework is an emerging area that requires further development and exploration.

In this paper, we present a novel method for detecting biomarker signatures that considers both main effects and biomarker-by-interaction effects within the survival framework while effectively managing the FWER. To achieve this, we extend the concept introduced by Meinshausen (2009) to the Cox survival model, employing a three-stage process that enables the identification of prognostic and predictive biomarkers while assigning valid p values. Our contributions include: (1) Application to high-dimensional datasets using a penalized technique for variable selection, facilitating the identification of biomarker signatures that encompass both prognostic and predictive biomarkers; (2) Addressing the challenge of multiple testing by obtaining p values from randomized multi-split data, ensuring robust control of the FWER; (3) Providing a user-friendly R implementation of our algorithm, available at <https://github.com/aliviawu/Biomarker-Paper/tree/main>. Additionally, we offer comprehensive theoretical properties in the Supplementary Materials. By integrating main effects and biomarker-by-interactions within the survival framework and ensuring strict control over the FWER, our approach makes a valuable contribution to the field of biomarker identification and statistical inference in high-dimensional data scenarios.

The remainder of this paper is organized as follows. In Sect. 2, we provide a detailed description of our proposed three-stage approach within the Cox PH model framework. This approach considers main effects and biomarker-by-interaction effects while

effectively controlling the FWER. Section 3 presents the results of our simulation studies, including comparisons with existing methods. In Sect. 4, we apply our proposed method to multiple real-world datasets. Finally, in Sect. 5, we summarize our findings and discuss potential directions for future research.

2 Materials and methods

To achieve FWER control in biomarker selection, we propose a three-stage approach that builds on the penalized-likelihood approach using adaptive gLASSO (Wang and Leng 2008). Our approach extends the idea of p-value adjustment developed by Meinshausen (2009) to the Cox proportional hazards framework. Additionally, we provide a description of several existing alternatives for comparison purposes.

2.1 Notation

Consider a study with n subjects and p potential biomarkers. Let T_i denote the event time for the i^{th} subject and C_i denote the censoring time. The follow-up time is defined as $Y_i = \min(T_i, C_i)$, and the event indicator is $\delta_i = I(T_i \leq C_i)$, where $I(\cdot)$ is an indicator function. We focus on right-censored data. The p -dimensional candidate biomarkers are denoted by $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{ip})^T$. The treatment status for the i^{th} patient is denoted by H_i . We assume that T_i and C_i are conditionally independent given \mathbf{X}_i . The hazard function of the i^{th} patient under the Cox PH model can be expressed as:

$$h(t|\mathbf{X}_i, H_i) = h_0(t) \exp(\alpha_0 H_i + \alpha_1 X_{i1} + \dots + \alpha_p X_{ip} + \gamma_1 X_{i1} H_i + \dots + \gamma_p X_{ip} H_i), \quad (1)$$

with the total number of regression parameters as $2p + 1$. We denote the number of main biomarkers (prognostic) and its interaction with treatment (predictive) to be $\tilde{p} = 2p$. Our objective is to identify prognostic biomarkers ($\alpha_j \neq 0, j = 1, \dots, p$) and predictive biomarkers ($\gamma_j \neq 0, j = 1, \dots, p$) in situations where $n \ll \tilde{p}$.

2.2 Adaptive gLASSO for variable selection

Under the Cox PH framework, let D denote the indices of the subjects who experienced the event of interest, and for each $r \in D$, the observed failure time is denoted by t_r . The set $R_r = \{i : Y_i \geq t_r\}$ includes the indices of the individuals who are at risk of experiencing the event at time t_r . Let $\theta = \{\alpha_0, \alpha_j, \gamma_j, j = 1, \dots, p\}$ be a vector of parameters with dimensionality $\tilde{p} + 1$, and let the semi-parametric partial likelihood function for parameter estimation be denoted by

$$L(\theta) = \prod_{r \in D} \frac{h(t|\mathbf{X}_i, H_i)}{\sum_{i \in R_r} h(t|\mathbf{X}_i, H_i)}. \quad (2)$$

Let $\ell(\theta)$ denote the logarithm of the partial likelihood function $\log L(\theta)$. The estimates of the parameters θ can be obtained by maximizing $\ell(\theta)$. However, when the

number of parameters \tilde{p} is larger than the sample size n ($n \ll \tilde{p}$), the semi-parametric likelihood estimator in Eq. (2) may not be feasible due to the difficulty in finding the global maximum as the number of biomarkers increases.

To select prognostic and predictive biomarkers with the oracle property, one commonly used method with a weighted adaptive gLASSO penalty can be considered. The objective function that the adaptive gLASSO minimizes is:

$$\ell^{aGL}(\theta) = \ell(\theta) + \lambda_1 \left(\sum_{g=1}^p \frac{1}{\hat{\omega}_g} \|\theta_{G_g}\|_1 \right), \tag{3}$$

where $\|\cdot\|_1$ denotes the L_1 norm, λ_1 is the shrinkage parameter, G_g is the index set belonging to the g -th pair of prognostic and predictive biomarkers ($g = 1, \dots, p$), and $\theta_{G_g} = \{\alpha_g, \gamma_g\}$ is the g -th pair of estimated coefficients belonging to G_g . Additionally, $\hat{\omega}_g$ is an adaptive weight vector obtained by performing L_2 regularization for each coefficient of α and γ , which is defined as follows:

$$\hat{\omega}_g = \frac{1}{\sqrt{(\hat{\alpha}_g^{ini})^2 + (\hat{\gamma}_g^{ini})^2}}, \tag{4}$$

where $\hat{\alpha}_g^{ini}$ and $\hat{\gamma}_g^{ini}$ are initial estimators obtained from a ridge regression (Hoerl and Kennard 1970), $g = 1, \dots, p$.

Subsequently, the group-based estimates can be obtained by maximizing $\ell^{aGL}(\theta)$. Biomarkers with non-zero coefficients ($\alpha_g \neq 0$ or $\gamma_g \neq 0$, where $g = 1, \dots, p$) will be selected. However, it should be noted that the FWER is not well controlled by this method.

2.3 The proposed three-stage approach

To control the FWER, we propose a three-stage strategy based on the penalized likelihood approach. This strategy extends the concept of p-value adjustment introduced by (Meinshausen 2009) to the Cox proportional hazards model. The algorithm for our proposed three-stage approach is described in detail below.

2.3.1 Stage I: conduct feature screening and obtain (nonaggregated) p values

In the first stage, we perform feature screening to reduce the dimensionality from \tilde{p} to a more manageable scale d with $d < n$. We then use the remaining $d/2$ pairs of prognostic and predictive biomarkers for variable selection based on penalized techniques (i.e., adaptive gLASSO) as well as p-value adjustment. Both feature screening and p-value adjustment are performed through a bootstrapping procedure.

For $b = 1 \dots B$,

- (i) Randomly split the data into two sets, a “training” set and a “testing” set, with some allocation rate of m (e.g., $m = 0.5$ indicates an equal sample size). Of note,

the selection of m depends on various factors, including the dataset sample size, the study's objectives, and other relevant considerations.;

- (ii) Perform feature screening via joint hypotheses using the training data. To identify the significant predictors among the main biomarkers and their interactions with treatment, the likelihood ratio test (LRT) can be utilized. For each j in $1, \dots, p$, the log-likelihood under the null hypothesis (H_0) and the alternative hypothesis (H_A) can be compared:

$$H_0 : \alpha_j = 0, \gamma_j = 0, H_A : \text{at least one of } \alpha_j \text{ and } \gamma_j \text{ is not equal to } 0. \quad (5)$$

The LRT statistic is computed as the difference between the partial log-likelihood statistics of two models: H_0 , which only includes the treatment variable (H_j), and H_A , which includes the treatment variable (H_j), a main biomarker (X_j), and its interaction with treatment ($X_j H$). In general, the LRT statistic is expressed as $LRT = -2 \ln(\frac{\ell_0}{\ell_A}) \sim \chi_2^2$. To screen the predictors efficiently, existing screening procedures often require the specification of a threshold. Here, we adopt the conventional threshold of $[n/\log n]$ (Fan and Lv 2008b). Specifically, the screening process retains the top $[n/\log n]$ pairs based on the rank of the Chi-square statistics of the joint hypothesis testing for each biomarker and its interaction with treatment.

- (iii) Apply adaptive gLASSO on the training set to select pairs using the pre-selected pairs of main biomarkers and their interactions with treatment. The selected pairs are denoted by $\tilde{S}^{(b)}$;
- (iv) Obtain the p -values for each selected pair in $\tilde{S}^{(b)}$ by performing LRT using testing dataset. Using only testing data, we employ a LRT hypothesis test for each selected pair in $\tilde{S}^{(b)}$, and calculate the corresponding chi-square based p -values, $\tilde{p}_j^{(b)}$, for $j \in \tilde{S}^{(b)}$. For unselected pairs, we set the corresponding p -values to 1.
- (v) Adjusted p -values based on the Bonferroni correction for $\tilde{p}_j^{(b)}$, denoted by $p_j^{(b)}$ ($j = 1, \dots, p$). For $j = 1, 2, \dots, p$, we have

$$p_j^{(b)} = \min \left\{ \frac{1}{2} \tilde{p}_j^{(b)} |\tilde{S}^{(b)}|, 1 \right\}, \quad (6)$$

where $|\tilde{S}^{(b)}|$ denotes the total number of selected pairs in $\tilde{S}^{(b)}$.

2.3.2 Stage II: obtain aggregated p values

In the first stage, we obtain a total of B p -values for each pair of prognostic and predictive biomarkers. To aggregate these p -values, we use quantiles introduced by (Meinshausen 2009).

Specifically, we define

$$Q_j(\eta) = \min \left\{ 1, q_\eta \left(\left\{ p_j^{(b)} / \eta; b = 1, \dots, B \right\} \right) \right\},$$

where q_η is the η^{th} quantile for the set $\left\{ p_j^{(b)} / \eta; b = 1, \dots, B \right\}$. The aggregated p -value is denoted as p_j^* and is defined as

$$p_j^* = \min \left\{ 1, (1 - \log \eta_{\min}) \inf_{\eta \in (\eta_{\min}, 1)} Q_j(\eta) \right\}, \quad (7)$$

where $\eta_{\min} \in (0, 1)$, and recommended choice is 0.05 as suggested by (Meinshausen 2009). Note that $H_{0j} : \alpha_j = \gamma_j = 0$ is rejected if $p_j^* \leq \alpha$, where α is the pre-specified FWER to be preserved ($j = 1, \dots, p$).

2.3.3 Stage III: biomarker signature identification

In the last stage, we perform further regression analysis on the pairs of prognostic and predictive biomarkers selected in Stage II. We fit a Cox PH model using the entire dataset by maximizing the partial likelihood. We identify biomarkers with p -values less than the pre-specified significance level α . The significant biomarkers, along with their interactions with treatment, form the biomarker signature for predicting patient outcomes in response to treatment.

Identification of prognostic and predictive biomarkers related to disease progression and treatment response in a survival framework is crucial for personalized medicine. However, existing penalized approaches for high-dimensional data often suffer from a lack of control over the FWER. Our proposed three-stage strategy that combines penalized likelihood techniques with adjusted p values obtained through random data splitting provides a reliable and interpretable approach for identifying biomarker signatures with strong prognostic and predictive power, while effectively controlling FWER.

In Stage I, we use a bootstrapping procedure to reduce the dimensionality of the training dataset to a moderate scale and select active pairs of prognostic and predictive biomarkers via an adaptive group LASSO technique. We then accumulate corrected p -values $p_j^{(b)}$ for each active pair via a likelihood ratio test based on the testing dataset. In Stage II, we summarize these non-aggregated p -values to p_j^* using an adaptive empirical quantile function and select the pairs based on p_j^* . Finally, in Stage III, we identify the final biomarkers by fitting the selected pairs from Stage II with a Cox PH model based on the entire dataset.

2.4 Other existing methods

To empirically evaluate the performance of our proposed approach, we compare it with several existing methods. In the literature, there are various methods for variable selection from biomarker main effects and biomarker-by-treatment interactions. Ternès (2016) conducted a comprehensive summary of possible approaches for high-dimensional Cox PH regression. They compared these methods through simulations with different numbers of biomarkers and varying effects of main biomarkers and interactions with treatment, and evaluated their selection abilities in null (i.e., no interactions with treatment) and alternative scenarios (i.e., at least one interaction with treatment). In the null scenarios, group LASSO and gradient boosting methods performed poorly in the presence of non-null main effects but performed well in alternative scenarios with high interaction strengths. Adaptive LASSO with grouped

weights was found to be too conservative. Principal component analysis (PCA) combined with LASSO performed moderately. Both LASSO and adaptive LASSO performed well, although LASSO was relatively poor in the presence of only non-null main effects. Here, we describe several competing methods that we consider for comparison.

2.4.1 LASSO

In the Cox PH framework, variable selection is typically performed by minimizing the log partial likelihood subject to a penalty on the parameters, as proposed by Tibshirani (1997). We use the LASSO penalty for both the main effects α_j and their interaction effects with treatment γ_j ($j = 1, \dots, p$) in Eq. (1) to perform variable selection, enabling us to identify both prognostic and predictive biomarkers. With the semi-parametric partial likelihood function defined in Eq. (2), let $\theta = \{\alpha_0, \alpha_j, \gamma_j\}_{j=1}^p$, where α_0 denotes the treatment effect, α_j denotes the j^{th} prognostic biomarker effect and γ_j denotes the j^{th} predictive biomarker effect. The partial log-likelihood with the LASSO penalty is

$$\ell^L(\theta) = \ell(\theta) + \lambda \left(\sum_{j=1}^p |\alpha_j| + \sum_{j=1}^p |\gamma_j| \right), \quad (8)$$

where prognostic biomarkers and predictive biomarkers are equally penalized with the shrinkage parameter λ . This tuning parameter λ is chosen by fivefold cross-validation. The LASSO-based coefficient estimators can then be obtained by maximizing $\ell^L(\theta)$, and the predictive and prognostic biomarkers ($\alpha_j \neq 0, \gamma_j \neq 0$) are selected.

2.4.2 Adaptive LASSO with grouped weights

Adaptive LASSO is a penalization method that assigns different penalty weights to the main effects and interaction effects, with larger coefficients penalized less than smaller ones to highlight their differences (Zou 2006; Zhang and Lu 2007). In the initial stage, this method estimates the weights by including the treatment and all biomarker main effects and interactions with the treatment, and applies a ridge penalty (Hoerl and Kennard 1970). Let α_j and γ_j ($j = 1, \dots, p$) be the main effects and interaction effects of the biomarkers, respectively. The penalty term with the shrinkage parameter λ_2 to control the magnitude of α_j and γ_j is

$$\lambda_2 \left(\sum_{j=1}^p \alpha_j^2 + \sum_{j=1}^p \gamma_j^2 \right). \quad (9)$$

In the second stage, a common grouped weight is estimated for all α_j and a single weight is assigned to all γ_j as the average of α_j and γ_j from the preliminary stage, i.e., $\alpha_R = \frac{1}{p} \sum_{j=1}^p |\alpha_j|$, $\gamma_R = \frac{1}{p} \sum_{j=1}^p |\gamma_j|$. The penalized log-likelihood with $\theta = \{\alpha_0, \alpha_j, \gamma_j\}_{j=1}^p$ is

$$\ell^{aL}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + \lambda \left(\frac{1}{\alpha_R} \sum_{j=1}^p |\alpha_j| + \frac{1}{\gamma_R} \sum_{j=1}^p |\gamma_j| \right). \tag{10}$$

2.4.3 Gradient boosting

Boosting algorithms are designed to enhance prediction accuracy by training a sequence of weak models, each correcting the errors of its predecessors. In high-dimensional settings, the process starts from the null model and updates a single coefficient at each step. This iterative process stops when the model achieves a balance between bias and variance. Gradient boosting reformulates this approach as a numerical optimization problem, where the objective is to minimize the model’s loss function by adding weak learners using gradient descent (Friedman 2001). Bühlmann and Yu (2003) proposed L_2 -Boost with a novel component-wise smoothing spline learner, providing an effective procedure for carrying out boosting for high-dimensional regression problems with continuous predictors. In our study, we first estimate the treatment effect preliminarily and then fix it as an offset.

2.4.4 PCA+LASSO

In the first stage, we use PCA (Hastie 2017) to reduce the dimensionality of the main effect matrix. The second stage applies the LASSO penalty to the interactions, which allows for the identification of predictive biomarkers based on the first K principal components of the main effects. In the final stage, we fit a Cox PH model by maximizing the partial likelihood based on all biomarkers and selected biomarker-treatment interactions. We then select prognostic and predictive biomarkers with p -values less than α .

3 Simulation studies

3.1 Simulation setup

We make the following assumptions regarding the true hazard function for the i -th patient:

$$h(t|\mathbf{X}_i) = h_0(t) \exp \left(\alpha_0 H_i + \alpha_4 X_{i4} + \gamma_4 X_{i4} H_i + \alpha_5 X_{i5} \right). \tag{11}$$

Here, α_0 denotes the impact of treatment H , α_4 and α_5 signify the prognostic effects of biomarkers X_4 and X_5 , respectively, and γ_4 represents the predictive effect of biomarker X_4 . We investigate four distinct scenarios, each characterized by distinct values of $(\alpha_0, \alpha_4, \gamma_4, \alpha_5)$: Scenario 1 (S1): (1, 1, 1, 1); Scenario 2 (S2): (1, 0.5, 1, 1); Scenario 3 (S3): (1, 0.5, 1.5, 1); Scenario 4 (S4): (1, 0.5, 2.5, 1.5). To simulate individual participants, we generate a treatment indicator variable for each using a $Binom(1, 0.5)$ distribution. The expression level of the primary j -th biomarker X_{ij} follows a standard normal distribution. The pairwise correlation between prognostic

biomarkers is set to $\rho = 0.15$ in order to mimic our real data applications. The baseline hazard function $h_0(t)$ is set up as a constant $\frac{1}{100}$ and $\frac{1}{270}$, resulting in mean censoring rates of around 40% or 60%, respectively. Censoring times C_i are generated from a uniform distribution $Unif(0, 150)$. For the estimation of survival times, we employ the method introduced by (Bender 2005). This involves drawing from a $Unif(0, 1)$ random variable and subsequently applying the transformations:

$$t_i = -\log(U)/h(t|\mathbf{X}_i).$$

The observed time to an event is then computed as $\min(t_i, C_i)$, along with the event status denoted as $I(t_i < C_i)$. To provide a comprehensive evaluation of our results, we generate 1,000 Monte Carlo datasets for each scenario. Furthermore, we consider varying sample sizes of 300, 500, or 1000, with the allocation rate $m = 0.5$ and the total number of candidate prognostic and predictive biomarkers \tilde{p} set to 1000, 2000, or 4000.

We evaluate four primary performance metrics: selection accuracy, mean squared error (MSE), relative bias of regression coefficient estimates, and control of FWER. Selection accuracy measures the percentage of times the true biomarker is selected out of 1,000 replicates. We estimate MSE as the mean of the squared difference between the true and estimated parameters across 1,000 replicates. Relative bias is evaluated as the mean difference divided by the true parameter value across 1,000 replicates. FWER control measures the proportion of times in 1,000 replicates that at least one biomarker, which is not one of the three candidate biomarkers, is selected while controlling the FWER at the nominal level of $\alpha = 0.05$. We investigate the impact of sample size, number of biomarkers, and censoring rates on these four metrics across various scenarios. Additionally, we compare the proposed method with five other methods, namely LASSO, gLASSO, adaptive LASSO, PCA+LASSO, and gradient boosting.

3.2 Simulation results

3.2.1 FWER control

Our proposed method effectively controls the FWER at a nominal level of 0.05 for selecting prognostic and predictive biomarkers, as demonstrated in Fig. 1. We evaluated the actual FWER across 1,000 replicates for four different scenarios with varying sample sizes (n), the total number of biomarkers (\tilde{p}), and censoring rates, except for the case where $\tilde{p} = n = 1000$ because it does not represent a high-dimensional scenario. Our method shows effective FWER control at approximately 0.05 for all four scenarios, particularly when the sample size is 1000. Although the FWERs were inflated for sample sizes of 500 or 800, we observed a return to 0.05 with a sample size of 1000. Furthermore, we conducted additional simulations using a different allocation rate, i.e., $m = 0.7$ (allocating 70% of the data for training and 30% for testing). We observed minimal differences in FWERs. For example, with sample sizes of $n = 300, 500$ in S1 ($\tilde{p} = 2000$), the FWERs were 0.062 and 0.048, respectively, and also similar trends emerged

across varying sample sizes (not shown due to space limit). Moreover, we also considered a higher value of the correlation coefficient, $\rho = 0.3$, for S1. Across different combinations of sample sizes and the number of biomarkers denoted as $(n, \tilde{p}) = (300, 1000), (300, 2000), (500, 1000), (500, 2000), (1000, 2000), (1000, 4000)$ we obtained satisfactory results with the FWERs of 0.055, 0.056, 0.051, 0.045, 0.059, and 0.048, respectively. In addition, We also incorporated a non-randomized clinical trial with an 80% treatment proportion, mirroring our second data application. Thus, in Scenario 1 ($\alpha_4 = \alpha_5 = \gamma_4 = 1$), with sample sizes of $n = (300, 500)$ and $\tilde{p} = 2000$, the obtained FWERs are 0.045 and 0.040, respectively.

3.2.2 Estimates of effects

The accurate estimation of prognostic and predictive biomarker effects is crucial for predicting hazard and survival rates. Boxplots of the estimates of α_4 , γ_4 , and α_5 are presented in Fig. 2 for scenarios S1 and S4, both with a 60% censoring rate. Additional scenarios are presented in Figure S5 and S6 in the Supplementary Materials. The average coefficient estimates of α_4 and α_5 from 1,000 simulated datasets closely match the true effects of (1, 1) and (0.5, 2.5) for scenarios S1 and S4, respectively. As the sample size increases from 300 to 1000, the dispersion of the estimated γ_4 gradually decreases, and the estimates tend to center around the true effects. Moreover, the coefficient estimates for the 40% censoring rate exhibit similar trends in terms of deviation from the true values, as shown in Figure S5 in the Supplementary Materials.

The MSEs and biases of the estimates are presented in Table 2 and Table S1 in the Supplementary Materials, respectively. The results indicate that, with a censoring rate of 40%, the MSEs and biases for all estimates are generally smaller compared to those with a 60% censoring rate. Moreover, increasing the sample size leads to a reduction in MSEs and relative bias, which is consistent with our expectations. Regarding the interaction effect of γ_4 , although the MSEs are relatively higher compared to the main effects (e.g., α_4 and α_5), the bias decreases as the true interaction effect increases from 1 to 1.5. Additionally, the underlying true effect strength of the primary biomarkers influences the estimation of their interaction effect. For example, as the true effect value of α_4 decreases from 1 (S1) to 0.5 (S2), its bias decreases, while the bias of its interaction effect with γ_4 increases.

3.2.3 Selection accuracy

The results of the selection accuracy analysis are summarized in Table 1. The findings indicate that selection accuracy improves with larger sample sizes, lower censoring rates, and greater biomarker effects. When the sample size is sufficiently large (i.e., >800), the true biomarker effects and censoring rate have minimal effects on selection accuracy. Both the biomarker and interaction term selections are influenced by the underlying true effect strength, and the accuracy of the interaction term selection tends to increase as the main effect decreases. When α_4 decreases from 1 (S1) to 0.5 (S2), the selection accuracy of X_4 decreases, but the accuracy of X_4H increases, particularly with small sample sizes and high censoring rates.

Table 1 Selection accuracy (%) of biomarker effect estimators for proposed three-stage method; four scenarios: $S1(\alpha_4 = 1, \gamma_4 = 1, \alpha_5 = 1)$; $S2(\alpha_4 = 0.5, \gamma_4 = 1, \alpha_5 = 1)$; $S3(\alpha_4 = 0.5, \gamma_4 = 1.5, \alpha_5 = 1)$; $S4(\alpha_4 = 0.5, \gamma_4 = 2.5, \alpha_5 = 1.5)$

Censor	n	\tilde{p}	S1			S2			S3			S4		
			X_4	X_4H	X_5	X_4	X_4H	X_5	X_4	X_4H	X_5	X_4	X_4H	X_5
40%	300	1000	1.000	1.000	0.998	0.983	1.000	1.000	0.978	1.000	1.000	0.981	1.000	1.000
		2000	1.000	1.000	0.997	0.979	1.000	1.000	0.981	1.000	1.000	0.985	1.000	1.000
		4000	1.000	1.000	0.998	0.984	1.000	1.000	0.992	1.000	1.000	0.981	1.000	1.000
		1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	500	2000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000
		4000	1.000	1.000	1.000	0.998	1.000	1.000	0.999	1.000	1.000	0.998	1.000	1.000
		1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		2000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	800	2000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		4000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		2000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
60%	300	1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		2000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		4000	1.000	0.987	1.000	0.898	0.990	1.000	0.897	1.000	1.000	0.916	1.000	1.000
		1000	1.000	0.985	1.000	0.892	0.991	1.000	0.896	1.000	1.000	0.920	1.000	1.000
	500	2000	1.000	0.986	1.000	0.891	0.986	1.000	0.893	1.000	1.000	0.912	1.000	1.000
		4000	1.000	0.999	1.000	0.989	1.000	0.987	1.000	1.000	1.000	0.996	1.000	1.000
		1000	1.000	0.999	1.000	0.975	1.000	0.975	1.000	1.000	1.000	0.978	1.000	1.000
		2000	1.000	1.000	1.000	0.981	1.000	0.984	1.000	0.984	1.000	0.986	1.000	1.000
	800	2000	1.000	1.000	1.000	0.997	1.000	1.000	0.997	1.000	1.000	0.999	1.000	1.000
		4000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1000	1.000	1.000	1.000	0.999	1.000	1.000	0.999	1.000	1.000	0.998	1.000	1.000
		2000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1000	2000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
	4000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
	1000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	
	2000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	

Furthermore, sample size is a critical factor in determining selection accuracy. As the sample size increases from 300 to 500, the selection percentages for all biomarkers are consistently around 100%.

In conclusion, higher selection percentages are observed when the sample size is large, the censoring rate is small, and the true biomarker effect is strong. The proposed method is not significantly affected by the number of candidate biomarkers, as the selection accuracy remains relatively stable as \tilde{p} increases.

3.2.4 Method comparison

We compared our three-stage strategy with five other methods: LASSO, adaptive LASSO, gLASSO, PCA+LASSO, and gradient boosting. The comparison was based on 11 setups combining various sample sizes and biomarker numbers under high censoring rate, and additional scenarios are available in the Supplementary Materials. Since LASSO and adaptive LASSO had similar performance, we excluded the results of adaptive LASSO from the analysis. As shown in Fig. 3, our proposed three-stage method effectively controls the FWERs, whereas the other four methods fail to achieve this goal. For all alternative methods, the FWERs for all scenarios are close to 1. These results agree with our expectation. Notably, methods such as LASSO, boosting, PCA+LASSO, and adaptive LASSO do not take into account the correlation between the primary biomarker and its interaction with treatment or impose the hierarchy constraint, despite incorporating various regularization strategies for variable selection. Group LASSO does attempt to tackle these concerns through its regularization approach; however, the p values obtained for the selected variables do not effectively control the FWER. In contrast, our method is designed to provide valid asymptotic control over variable inclusion at the nominal level, which is made possible by integrating the multiple sample-splitting approach and incorporating features such as correlation and hierarchy (e.g., group lasso) into the regularization techniques after initial feature screening.

For scenarios S1 and S2 (with details available in the Supplementary Materials), we evaluated the selection accuracy of prognostic and predictive biomarkers using five methods under four scenarios, with sample sizes of 300, 500, 800, and 1000, 2000 biomarkers, and 40% and 60% censoring rates. Overall, gLASSO and PCA+LASSO showed relatively high selection accuracy of the interaction term X_4H , but performed poorly in selecting the main effects. On the other hand, LASSO and gradient boosting achieved selection accuracies close to 1 and were insensitive to the censoring rate, sample size, number of biomarkers, and scenario. However, our proposed method controls the selection accuracy with respect to increases in the sample size or underlying biomarker effects or a decrease in the censoring rate. Furthermore, scenarios S3 and S4 (details available in the Supplementary Materials) compared the coefficient estimates based on different methods. In comparison with the four existing methods, our model provided unbiased estimates of the effects with improved efficiency throughout. The simulations were carried out using R 4.1.2 on a high-performance computing cluster. For each dataset with 50 iterations of the bootstrapping procedure, the computation time for the proposed method is approximate 6 min, considering the number of biomarkers as 1,000 and sample size of 300,

Table 2 Mean square errors (MSEs) of biomarker effect estimators for three-stage method; four scenarios: $S1(\alpha_4 = 1, \gamma_4 = 1, \alpha_5 = 1)$; $S2(\alpha_4 = 0.5, \gamma_4 = 1, \alpha_5 = 1)$; $S3(\alpha_4 = 0.5, \gamma_4 = 1.5, \alpha_5 = 1)$; $S4(\alpha_4 = 0.5, \gamma_4 = 2.5, \alpha_5 = 1.5)$

Censor	n	\tilde{p}	S1			S2			S3			S4		
			α_4	γ_4	α_5	α_4	γ_4	α_5	α_4	γ_4	α_5	α_4	γ_4	α_5
40%	300	1000	0.020	0.040	0.021	0.016	0.042	0.022	0.016	0.035	0.021	0.016	0.067	0.030
		2000	0.022	0.043	0.022	0.015	0.046	0.022	0.015	0.038	0.022	0.016	0.064	0.028
		4000	0.022	0.043	0.023	0.016	0.045	0.024	0.017	0.037	0.021	0.016	0.064	0.027
	500	1000	0.011	0.024	0.012	0.009	0.025	0.013	0.009	0.020	0.013	0.009	0.038	0.016
		2000	0.011	0.024	0.012	0.009	0.025	0.012	0.009	0.021	0.012	0.010	0.039	0.017
		4000	0.012	0.023	0.012	0.010	0.024	0.012	0.010	0.020	0.012	0.009	0.037	0.016
	800	1000	0.008	0.016	0.007	0.006	0.016	0.007	0.006	0.013	0.008	0.006	0.021	0.009
		2000	0.008	0.015	0.007	0.006	0.015	0.008	0.006	0.012	0.008	0.006	0.022	0.011
		4000	0.007	0.014	0.008	0.006	0.015	0.008	0.006	0.013	0.008	0.006	0.022	0.010
	1000	2000	0.006	0.013	0.006	0.005	0.011	0.006	0.005	0.010	0.006	0.005	0.018	0.008
		4000	0.006	0.012	0.006	0.005	0.012	0.006	0.005	0.010	0.006	0.005	0.017	0.007
		1000	0.033	0.056	0.036	0.025	0.050	0.038	0.025	0.062	0.037	0.023	0.093	0.046
60%	300	2000	0.032	0.057	0.032	0.024	0.050	0.034	0.025	0.066	0.034	0.022	0.092	0.041
		4000	0.030	0.056	0.030	0.025	0.051	0.036	0.025	0.066	0.036	0.024	0.088	0.042
		1000	0.019	0.036	0.019	0.016	0.031	0.019	0.016	0.038	0.019	0.014	0.051	0.025
	500	2000	0.020	0.039	0.019	0.016	0.034	0.020	0.016	0.040	0.021	0.015	0.052	0.025
		4000	0.018	0.036	0.018	0.015	0.031	0.019	0.015	0.037	0.019	0.013	0.052	0.022
		1000	0.011	0.020	0.011	0.009	0.018	0.011	0.009	0.021	0.011	0.009	0.030	0.013
	800	2000	0.011	0.022	0.011	0.010	0.019	0.013	0.010	0.023	0.013	0.009	0.030	0.015
		4000	0.010	0.022	0.012	0.009	0.019	0.012	0.009	0.023	0.012	0.008	0.031	0.015
		1000	0.009	0.017	0.009	0.009	0.016	0.009	0.009	0.019	0.009	0.008	0.024	0.010
	1000	2000	0.009	0.016	0.008	0.008	0.015	0.009	0.008	0.017	0.009	0.007	0.024	0.010
		4000	0.009	0.016	0.008	0.008	0.015	0.009	0.008	0.017	0.009	0.007	0.024	0.010
		1000	0.009	0.016	0.008	0.008	0.015	0.009	0.008	0.017	0.009	0.007	0.024	0.010

however, this time may increase (up to 26 min) with larger number of biomarkers and sample sizes. Other alternative methods need less time due to the lack of bootstrap procedures (i.e., 3 min for the case with 1,000 biomarkers and sample size of 300).

4 Applications

In the first example, we present an application of our method using an existing breast cancer study that includes patients with estrogen receptor (ER)-negative tumors (Desmedt 2011; Hatzis 2011). The gene expression data associated with the study is publicly available from the Gene Expression Omnibus database (refer to GSE16446 and GSE25066). The study comprised 614 patients, with 507 receiving only anthracycline-based adjuvant chemotherapy (coded as 0) (Desmedt 2011), and 107 receiving anthracycline with taxane-based chemotherapy (coded as 1) (Hatzis 2011). The gene expression data has been pre-processed (e.g., normalization, filtering out low-expression genes), resulting in 1,689 gene variables for direct analysis. The primary outcome of interest is the distant recurrence-free survival, with a censoring rate of approximately 78% for both groups.

In the second application, we examine the effect of Tamoxifen treatment on patients with ER-positive breast tumors and evaluate gene expression biomarkers and their interactions with treatment (Loi 2007). The original dataset comprises 414 patients from the cohort GSE6532, collected by Loi (2007), to identify ER-positive subtypes with gene expression profiles. Our analysis focuses on the primary outcome of distant metastasis-free survival (DMFS). After excluding 34 patients, who lack any records of time-to-event data (no follow-up or dropout information) for survival outcomes, we are left with 255 patients who received Tamoxifen treatment and 125 patients who did not. The censoring rates for the two groups are 73.3% and 77.6%, respectively, and there are 44,916 gene expression measurements for each patient.

We applied our approach to identify prognostic and predictive gene biomarkers in the two applications and compared it to existing methods. To implement our proposed method, we opted for 50 iterations in the bootstrapping procedure, and we utilized a 70% allocation rate to partition the data into training and testing datasets. The nominal level was set at 0.05, and for each existing method, a feature screening was constructed based on the training set. After obtaining a total of 50 p values for each pair of prognostic and predictive biomarkers, we calculated the aggregated p values for each pair via the quantile function introduced by Meinshausen (2009) with 0.05 as the η_{min} .

As shown in Table 3, our proposed method did not identify any prognostic or predictive biomarkers for the first dataset. However, for the second dataset, the interaction of the gene HYPK ('218,680_x_at') with the treatment indicator was selected with significance, indicating that HYPK is a predictive biomarker for Tamoxifen treatment regarding the outcome of DMFS. This finding is consistent with the literature (Hans-Dieter and RoyerMatthias 2017), where HYPK is suggested as a novel predictive biomarker for breast cancer. LASSO selected the largest

number of biomarkers, followed by boosting, adaptive LASSO, and gLASSO, while PCA+LASSO led to the fewest selections. This performance is similar to what we observed in our simulation studies. Additionally, the HYPK gene selected by our proposed method was also identified by the gradient boosting methods. We further listed the gene symbols of selected prognostic and predictive biomarkers based on the methods in Tables S.4-S.5 of the Supplementary Materials.

Per reviewers' suggestion, we present another data application to further explore our method, as provided in the Supplementary Material. In particular, we analyzed a microarray dataset (refer to GSE22762) (Herold 2011) comprising 151 chronic lymphocytic leukemia (CLL) patients. The primary objective was to identify prognostic and predictive biomarkers associated with overall survival (OS) and salvage chemotherapy. Our proposed method successfully identified 2 predictive biomarkers and 3 prognostic biomarkers. For additional details, please refer to Section D of the Supplementary Material.

5 Discussion

In this paper, we presented a three-stage strategy for identifying prognostic and predictive biomarker signatures, which can be extended to higher-order terms, such as pairwise interactions among the biomarkers, depending on clinical interest or practical necessity. Our work builds upon the concept of multi-splitting for p -value adjustment to identify prognostic and predictive biomarkers under the survival framework, with a focus on Cox PH regressions. Specifically, we extend the approach proposed by (Meinshausen 2009) by generating pairwise p values through joint hypothesis testing. However, we note that if we were to generate p values via individual hypothesis testing, as in the approach proposed by (Meinshausen 2009), the resulting family-wise error rates (FWERs) would be overly conservative when identifying prognostic and predictive biomarkers in our case.

We conducted extensive simulation studies, which demonstrated that our proposed approach can control the FWER well around the nominal level, whereas existing methods such as LASSO, gLASSO, PCA+LASSO, and gradient boosting fail to control FWER. For example, LASSO produces FWERs close to 1 for all scenarios, while boosting, gLASSO, and PCA+LASSO have unstable FWERs across different scenarios. Controlling the FWER in cancer studies can improve the sensitivity of biomarker selection and testing during screening. Additionally, compared with existing methods, our proposed method provides accurate estimates of the effects of selected biomarkers with centers closer to true effects and lower dispersion across a variety of scenarios. The mean square errors and relative bias of the estimates produced by our proposed method are consistently lower across a variety of scenarios. However, gLASSO and PCA+LASSO have the largest variability of estimates, and the boosting method underestimates the effects in most scenarios. Furthermore, our method can control the selection accuracy of prognostic and predictive biomarkers close to 100% with an increase in sample size or underlying biomarker effects or a decrease in the censoring rate.

Table 3 Number of selected predictive and prognostic biomarkers

Methods	Breast cancer: taxane		Breast cancer: tamoxifen	
	#predictive bio-markers	#prognostic bio-markers	#predictive bio-markers	#prognostic biomarkers
Three-stage method	0	0	1	0
Group LASSO	4	5	1	5
LASSO	5	18	11	22
Adaptive LASSO	1	21	0	17
Boosting	9	14	8	15
PCA+LASSO	0	0	2	1

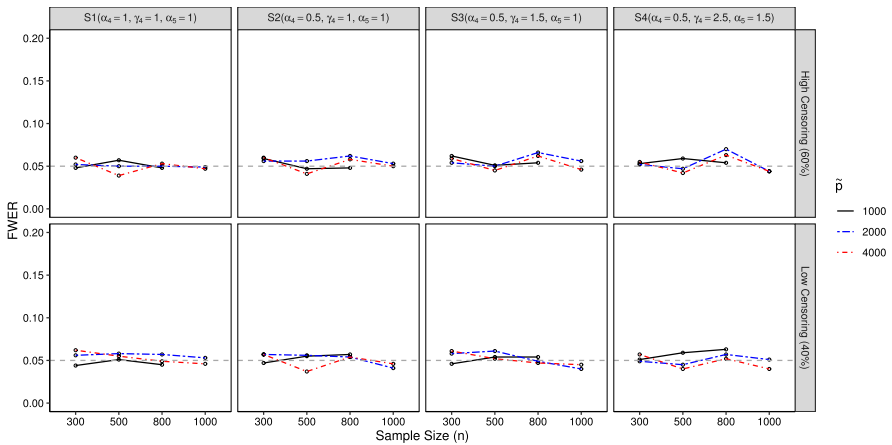


Fig. 1 A matrix of panels for family-wise error rates calculated via our proposed model based on simulation studies; rows represent high or low censoring rates; columns represent four simulation scenarios; x-axis is the sample size and y-axis is FWERs; three types of lines represent different total number of biomarkers \tilde{p}

In contrast, existing methods (e.g., gLASSO and PCA+LASSO) have relatively lower and inconsistent selection accuracy of the main effects, regardless of censoring rates.

Furthermore, we applied our proposed method and other existing methods to analyze two breast cancer datasets and a chronic lymphocytic leukemia dataset, as detailed in the supplementary material. While there is no literature revealing true prognostic and predictive biomarkers based on these three gene expression datasets, our proposed method yielded intriguing findings. In the second data example, our proposed method identified the gene HYPK as a predictive biomarker for Tamoxifen treatment regarding the outcome of DMFS. This aligns with a finding from (Hans-Dieter and Royer-Matthias 2017), suggesting HYPK as a novel predictive biomarker for breast cancer. Notably, the gradient boosting method also identified the HYPK gene. For the CLL

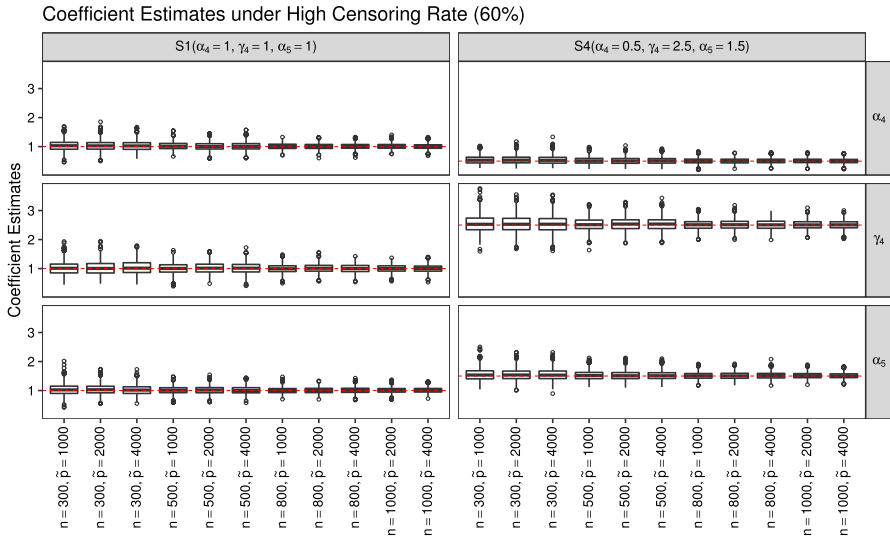


Fig. 2 A matrix of panels for coefficient estimates for the high censoring rate; rows represent three biomarker effects; columns represent four simulation scenarios; x-axis represents 11 setups of sample sizes and the numbers of biomarkers; y-axis represents the estimated coefficients; red dotted lines represent the strength of biomarker effects

data application shown in the Supplementary Material, our proposed method successfully identified 2 predictive biomarkers and 3 prognostic biomarkers associated with OS and chemotherapy. Among the selected biomarkers, TCF7 stood out as both a prognostic and predictive biomarker, indicating a significant impact on OS in CLL and providing insights into the effect of chemotherapy on CLL patients. This finding aligns with the observations in the paper by (Herold 2011).

In summary, our proposed approach provides a robust tool for identifying prognostic and predictive biomarkers in cancer studies, demonstrating superior performance in simulation studies compared to existing methods, particularly in terms of controlling the FWER. Noted that the FWER control represents a conservative approach, emphasizing the importance of avoiding any false discoveries within a family of tests, making it more restrictive compared to less stringent methods. Of note, in real-world applications, various factors, such as non-randomized clinical trial, sample size, the distribution of biomarker values, the magnitude of biomarker effects and varied pairwise correlations among biomarkers, may influence the selection results. In our data applications, we present several examples for illustration, with promise findings and the genes selected by our method verified by the existing literature.

Regarding the choice of quantiles for aggregating p values, we have adopted a strategy akin to (Meinshausen 2009), considering a lower bound value of $\eta_{\min} = 0.05$ that has been both suggested and exclusively explored in prior studies (Meinshausen 2009; Renaux 2020; Shi et al. 2023; Buzdugan 2016). While there are alternative methods for p -value aggregation (Mitchell 2015), our context is unique because the p values for each variable are generated through repeated data random-splits,

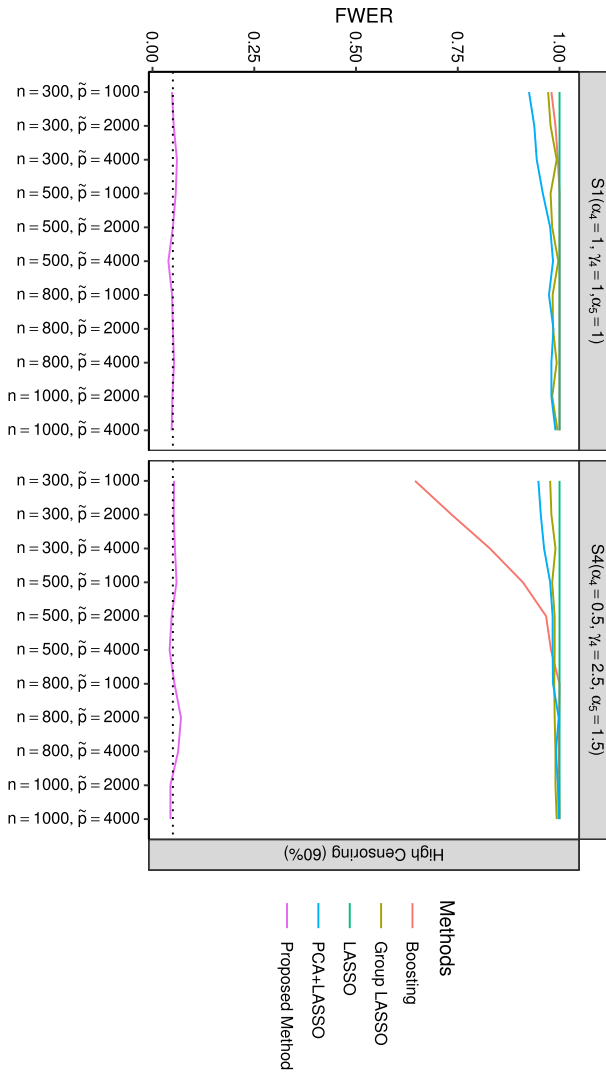


Fig. 3 A matrix of panels for FWERs comparisons for different methods based on simulation studies; rows represent high or low censoring rates; columns represent four simulation scenarios; x-axis represents different sample sizes and the numbers of biomarkers, y-axis represents FWERs; five colored lines represent different methods

resulting in empirical distributions. Using quantiles to combine and aggregate p values provides a flexible means of error rate control, with the advantage of subjective quartile selection. The challenge of selecting the quartile parameter, denoted as η_{\min} ($0 < \eta_{\min} < 1$), has been acknowledged Meinshausen (2009), and there is no universally accepted value for η_{\min} that guarantees error control. However, the outcomes from the chosen value have proven satisfactory, but further exploration can be pursued in this regard.

In terms of other future work, it may be worthwhile to investigate the issue of multicollinearity between gene expression levels, particularly when the correlation among biomarkers is relatively high. Additionally, the accelerated failure time model can be easily adapted into our three-stage framework to derive the p values and identify biomarker signatures when Cox PH models are not appropriate due to violations of the PH assumption. Overall, our proposed method has wide-ranging potential for application in cancer genetics studies and can be readily extended to other areas as necessary.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10260-024-00748-y>.

Acknowledgements The authors gratefully acknowledge Dr. Hao Feng and Ms. Wen Tang for their assistance with the data processing of the microarray dataset (GSE22762) (Herold et al., 2011) during the paper revision. Additionally, they appreciate the valuable feedback provided by the reviewers, which has contributed to enhancing the quality of the paper.

Funding Wang's work received partial support from the start-up funding provided by the Department of Population and Quantitative Health Sciences at Case Western Reserve University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data availability The data analysed during the current study are publicly available from the Gene Expression Omnibus database. Please see below: GSE16446: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16446>; GSE25066: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25066>; GSE6532: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse6532>. GSE22762: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22762>.

Declarations

Conflict of interest The authors declare no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Barrett T et al (2010) Ncbi geo: archive for functional genomics data sets-10 years on. *Nucleic Acids Res* 39(suppl-1):D1005–D1010
- Bender R et al (2005) Generating survival times to simulate cox proportional hazards models. *Stat Med* 24(11):1713–1723
- Bühlmann P (2013) Statistical significance in high-dimensional linear models. *Bernoulli* 19(4):1212–1242
- Buzdugan L et al (2016) Assessing statistical significance in multivariable genome wide association analysis. *Bioinformatics* 32(13):1990–2000
- Bühlmann P, Yu B (2003) Boosting with the l_2 loss. *J Am Stat Assoc* 98(462):324–339
- Chin L et al (2011) Cancer genomics: from discovery science to personalized medicine. *Nat Med* 17(3):297–303
- Desmedt C et al (2011) Multifactorial approach to predicting resistance to anthracyclines. *J Clin Oncol* 29(12):1578–1586

- Dezeure R et al (2015) High-dimensional inference: confidence intervals, p -values and R-software hdi. *Stat Sci* 30(4):533–558
- Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 96(456):1348–1360
- Fan J, Li R (2002) Variable selection for coxs proportional hazards model and frailty model. *Ann Stat* 30(1):74–99
- Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc: Ser B (Stat Methodol)* 70(5):849–911
- Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc: Ser B (Stat Methodol)* 70(5):849–911
- Fan J et al (2010) High-dimensional variable selection for cox's proportional hazards model. *Theory powering applications - a festschrift for Lawrence D. Brown, Institute of Mathematical Statistics Collections Borrowing Strength*, pp 70–86
- Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Ann Stat* 29(5):1189
- Ghosh S (2007) Adaptive elastic net: an improvement of elastic net to achieve oracle properties. Preprint, p 1
- Hamburg MA, Collins FS (2010) The path to personalized medicine. *N Engl J Med* 2010(363):301–304
- Hans-Dieter, RoyerMatthias, KHR-P (2017) Novel prognostic and predictive biomarkers (tumor markers) for human breast cancer. EP2669682B1
- Hastie T et al (2017) *The elements of statistical learning: data mining, inference, and prediction*. Springer
- Hatzis C et al (2011) A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA* 305(18):1873–1881
- He K et al (2019) An improved variable selection procedure for adaptive lasso in high-dimensional survival analysis. *Lifetime Data Anal* 25(3):569–585
- Herold T et al (2011) An eight-gene expression signature for the prediction of survival and time to treatment in chronic lymphocytic leukemia. *Leukemia* 25(10):1639–1645
- Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12(1):55–67
- Loi S et al (2007) Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J Clin Oncol* 25(10):1239–1246
- Meinshausen N, Yu B (2009) Lasso-type recovery of sparse representations for high-dimensional data. *Ann Stat* 37(1):246–270
- Meinshausen N et al (2009) p values for high-dimensional regression. *J Am Stat Assoc* 104(488):1671–1681
- Mitchell MW (2015) A comparison of aggregate p value methods and multivariate statistics for self-contained tests of metabolic pathway analysis. *PLoS One* 10(4):e0125081
- Renaux C et al (2020) Hierarchical inference for genome-wide association studies: a view on methodology with software. *Comput Stat* 35(1):1–40
- Shi H et al (2023) Tests for ultrahigh-dimensional partially linear regression models
- Simon N et al (2011) Regularization paths for cox's proportional hazards model via coordinate descent. *J Stat Softw* 39(5):1–13
- Ternès N et al (2016) Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces. *Biom J* 59(4):685–701
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc: Ser B (Methodol)* 58(1):267–288
- Tibshirani R (1997) The lasso method for variable selection in the cox model. *Stat Med* 16(4):385–395
- Wang H, Leng C (2008) A note on adaptive group lasso. *Comput Stat Data Anal* 52(12):5277–5286
- Wasserman L, Roeder K (2009) High dimensional variable selection. *Ann Stat* 37(5A):2178
- Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. *J R Stat Soc: Ser B (Stat Methodol)* 68(1):49–67
- Zhang C-H, Zhang SS (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *J R Stat Soc Ser B (Stat Methodol)* 76(1):217–242
- Zhang HH, Lu W (2007) Adaptive lasso for cox's proportional hazards model. *Biometrika* 94(3):691–703
- Zhao SD, Li Y (2012) Principled sure independence screening for cox models with ultra-high-dimensional covariates. *J Multivar Anal* 105(1):397–411
- Zou H (2006) The adaptive lasso and its oracle properties. *J Am Stat Assoc* 101(476):1418–1429
- Zuo Y et al (2021) Variable selection with second-generation p values. *The American Statistician*, pp 1–11