Received: 28 June 2021 | Revised: 17 July 2022 | Accepted: 21 July 2022

DOI: 10.1002/sim.9545

RESEARCH ARTICLE

Statistics in Medicine WILEY

Fast Lasso-type safe screening for Fine-Gray competing risks model with ultrahigh dimensional covariates

Hong Wang[1] | **Zhenyuan Shen**[1] | **Zhelun Tan**[1] | **Zhuan Zhang**[1] | **Gang Li**[2]

[1]School of Mathematics and Statistics, Central South University, Changsha, Hunan, China

[2]Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, Los Angeles, California, USA

Correspondence
Gang Li, Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, Los Angeles, CA 90095-1772, USA.
Email: vli@ucla.edu

Funding information
National Social Science Foundation of China, Grant/Award Number: 17BTJ019; National Statistical Scientific Research Project of China, Grant/Award Number: 2022LZ28; Changsha Municipal Natural Science Foundation, Grant/Award Number: kq2202080; National Institutes of Health, Grant/Award Numbers: P30 CA-16042, UL1TR000124-02, P01AT003960; Fundamental Research Funds for the Central Universities of Central South University, Grant/Award Number: 2020zzts361

The Fine-Gray proportional sub-distribution hazards (PSH) model is among the most popular regression model for competing risks time-to-event data. This article develops a fast safe feature elimination method, named PSH-SAFE, for fitting the penalized Fine-Gray PSH model with a Lasso (or adaptive Lasso) penalty. Our PSH-SAFE procedure is straightforward to implement, fast, and scales well to ultrahigh dimensional data. We also show that as a feature screening procedure, PSH-SAFE is safe in a sense that the eliminated features are guaranteed to be inactive features in the original Lasso (or adaptive Lasso) estimator for the penalized PSH model. We evaluate the performance of the PSH-SAFE procedure in terms of computational efficiency, screening efficiency and safety, run-time, and prediction accuracy on multiple simulated datasets and a real bladder cancer data. Our empirical results show that the PSH-SAFE procedure possesses desirable screening efficiency and safety properties and can offer substantially improved computational efficiency as well as similar or better prediction performance in comparison to their baseline competitors.

KEYWORDS
adaptive Lasso, competing risks, Fine-Gray model, proportional subdistribution hazards, safe feature screening

1 | INTRODUCTION

Lasso penalization[1] is among the most widely used methodology for high dimensional problems. However, large-scale and high dimensional data can pose substantial computational challenges to solving a Lasso-type penalization problem because its computational cost is in the order of np^2.[2] To improve computational efficiency, a safe feature elimination (SAFE) algorithm has been recently developed for large scale Lasso-type problems.[3] Briefly speaking, SAFE serves as a screening procedure to remove features that are guaranteed to be inactive with zero coefficients in the original Lasso solution. Hence preceding Lasso with SAFE screening can effectively reduce the data dimension and the computational complexity. Due to its superb numerical performance and desirable theoretical properties, SAFE screening has been widely applied to classification and regression problems.[4-7]

The purpose of this article is to develop SAFE algorithms for the Lasso or adaptive Lasso penalized Fine-Gray proportional subdistribution hazards (PSH) model[8] for competing risks survival data with ultrahigh dimensional covariates. Competing risks data arise commonly in many applications when individuals may fail from multiple causes and the

occurrence of one failure event precludes the others from happening.[8-11] The Fine-Gray PSH model directly models the impact of covariates on the marginal probability of failure for a specific cause, namely, the cumulative incidence function (CIF) or subdistribution, and has been commonly used for competing risks data. However, despite of the rich literature on high dimensional methods for the PSH model,[12-18] to the best of our knowledge, no SAFE procedure has been developed for the PSH model.

In this article, we derive a fast safe feature elimination method, named PSH-SAFE, for the Fine-Gray PSH model combined with the Lasso and adaptive Lasso penalties, respectively. The detailed PSH-SAFE rules are described later in (9) and (11). Our PSH-SAFE procedure is straightforward to implement, fast, and scales well to ultrahigh dimensional data. We also show rigorously that as a feature screening procedure, PSH-SAFE is safe in a sense that the eliminated features are guaranteed to be inactive features in the original Lasso-PSH (or adaptive Lasso-PSH) estimator. We conduct extensive simulations to evaluate the performance of the PSH-SAFE procedure in terms of computational efficiency, screening efficiency and safety, and prediction accuracy in multiple scenarios. Our empirical results demonstrate that the PSH-SAFE procedure possesses desirable screening efficiency and safety properties and can offer substantially improved computational efficiency as well as similar or better prediction performance in comparison to their baseline competitors.

The rest of this article is organized as follows. In Section 2.1, we review the PSH model[8] and the pseudo-partial likelihood estimation method. In Section 2.2, we derive the SAFE screening rules for both the Lasso and adaptive Lasso PSH models, with theoretical guarantees. In Section 3, the performance of the proposed method is demonstrated using extensive simulations and a publicly available high-dimensional bladder cancer dataset. Concluding remarks are provided in Section 4.

2 | SAFE SCREENING FOR PENALIZED PROPORTIONAL SUBDISTRIBUTION HAZARDS MODELS

2.1 | Preliminaries

Without loss of generality, we assume that there are two causes of failure for the competing risks outcome, where cause 1 is the event of interest and cause 2 is a competing risk. A competing risks data consists of n iid observations $\{(y_i, \delta_i, \delta_i\varepsilon_i, X_i), i = 1, \dots, n\}$, where $y_i = \min(T_i, C_i)$ is the observed time, T_i, C_i and $\varepsilon_i \in \{1, 2\}$ denote the failure time, the censoring time, and the failure type, $\delta_i = \mathrm{I}(T_i \le C_i)$ is the censoring indicator, and X_i is a vector of p covariates for subject i. Denote the feature matrix by $X = [\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_p}] \in \mathbb{R}^{n\times p}$. Hence, X_i^T denotes the ith row of X and $\boldsymbol{x_i} \in \mathbb{R}^p$ be the jth column of X.

A fundamental quantity in competing risks problems is the CIF for each competing event. The CIF of event type i is

$$F_i(t|X) = P(T \le t, \varepsilon = i|X) = \int_0^t H_i(t|X)S(u|X)\, du,$$

where $H_i(t|X)$ is the cause-specific hazard (CSH) and $S(t|X)$ is the event-free survival. With Fine-Gray's PSH model, one can directly estimate the impact of the covariates on the hazard of the CIF without estimating the individual CSH for the different failure type. The PSH model based on the subdistribution hazard for cause 1 is defined as

$$\begin{aligned} h_1(t|X) &= dF_1(t|X)/\{1 - F_1(t|X)\} \\ &= \lim_{\Delta t\to 0}\frac{1}{\Delta t}Pr\{t \le T \le t + \Delta t, \varepsilon = 1|T \ge t \cup (T \le t \cap \varepsilon \ne 1)|X\}. \end{aligned} \tag{1}$$

The subdistribution hazard of cause 1 is assumed to follow a proportional hazard model, $h_1(t|X) = h_{10}(t)\exp(\boldsymbol{\beta}^T X)$, where $h_{10}(t)$ is an unspecified baseline subdistribution hazard function, and $\boldsymbol{\beta}$ is a p-dimensional vector of regression coefficients.

For right censored competing risk data, the pseudo log-partial likelihood function of the PSH model is defined as[19]

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}\int_0^\infty \left[\boldsymbol{\beta}^T X_i - \log\left\{\sum_j \omega_j(u)Y_j(u)\exp(\boldsymbol{\beta}^T X_j)\right\}\right] \times \omega_i(u)dN_i(u), \tag{2}$$

where $N_i(t) = I(T_i \le t, \varepsilon_i = 1)$, $Y_i(t) = 1 - N_i(t-)$, and $\omega_i(t)$ is a time dependent weight developed based on inverse probability of censoring weighting (IPCW) technique, allowing for dependence between censoring times and covariates. For an individual i at time t, the IPCW weight is defined as $\omega_i(t) = I(C_i \ge T_i \wedge t)\hat{G}(T_i \wedge t)$. Here $G(t) = Pr(C \ge t)$ is the survival function of the censoring time C and $\hat{G}(t)$ is the Kaplan-Meier estimate for $G(t)$. At a given time t, if the individual is right censored or failed due to an event of interest, $\omega_i(t)Y_i(t) = 0$; if failed due to competing risks, then $\omega_i(t)Y_i(t)$ is between 0 and 1 and decreasing over time; otherwise, $\omega_i(t)Y_i(t) = 1$.

The pseudo log-partial likelihood can also be written as[20]

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} I(\delta_i \varepsilon_i = 1) \left[\boldsymbol{\beta}^T X_i - \log \left\{ \sum_{j \in R_i} \omega_j(t_i) \exp(\boldsymbol{\beta}^T X_j) \right\} \right], \tag{3}$$

where the risk set $R_i = \{j : (T_j \ge T_i) \cup ((T_j \le T_i) \cap (\delta_j = 1) \cap (\varepsilon_i \ne 1))\}$, including subjects still at risk and those who have already failed from competing cause prior to time t.

Recently, penalization methodology is extended to the PSH model by proposing a generalized objective function and rigorously established the asymptotic properties of the proposed penalized estimators:[21]

$$Q(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{4}$$

where $l(\boldsymbol{\beta})$ is defined in Equation (3), $p_\lambda(|\beta_j|)$ is the penalty function, and λ is a tuning parameter that controls the complexity of selected models.

As can see from the literature,[17,19,21,22] Lasso-type penalties are the most widely used ones among all regularized models. However, when the dimension of the feature space and/or the number of samples are extremely large, solving common optimization algorithms for the lasso-type problem remains challenging.

In this research, using SAFE feature screening rules, we can quickly remove a significant number of features without solving the L_1 optimization problems and these discarded features are guaranteed not to appear in an optimal solution. Consequently, the computational burden associated with computationally intensive optimization problems can be substantially reduced.

2.2 | SAFE rules for Lasso-PSH and adaptive Lasso-PSH models

In this subsection, we will derive SAFE rules for PSH model with Lasso and adaptive Lasso penalties:

- Lasso penalty: $p_\lambda(|\beta_j|) = \lambda|\beta_j|$.
- Adaptive Lasso: $p_\lambda(|\beta_j|) = \lambda w_j|\beta_j|$, where w_j is a data-adaptive weight assigned to each regression parameter. Generally speaking, a default of $w_j = 1/|\tilde{\beta}_j|$ can yield the oracle properties, and the penalized estimator $\tilde{\boldsymbol{\beta}} = [\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_p]^T$ is the maximizer of the log partial likelihood $l(\boldsymbol{\beta})$ in Equation (3).

First, we derive the SAFE screening rule for the Lasso-PSH model. The primal optimal problem for the Lasso PSH model is as follows:

$$\max_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) - \lambda||\boldsymbol{\beta}||_1 = \max_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} I(\delta_i \varepsilon_i = 1) \left[\boldsymbol{\beta}^T X_i - \log \left\{ \sum_{j \in R_i} \omega_j(t_i) \exp(\boldsymbol{\beta}^T X_j) \right\} \right] - \lambda||\boldsymbol{\beta}||_1 \right\}. \tag{5}$$

Let $\boldsymbol{\beta}^*$ be the optimum of the primal problem. To get the dual form of the optimization problem (5), we introduce the following notations. An event time of interest matrix can be defined as an indicator matrix $I := (I_{ij}) \in \{0, 1\}^{f \times n}$ where f is the number of unique observed events of interest failure times ($\sum_{i=1}^{n} I(\delta_i \varepsilon_i = 1)$) and $I_{ij} = 1$ if $j \in R_i$. Assuming $Z := (z_{ij}) = \mathbf{1}\boldsymbol{\beta}^T X^T \in \mathbb{R}^{f \times n}$, $\mathbf{1} = (1, \dots, 1)^T \in \mathbb{R}^f$.

Then the optimization problem (5) for PSH model can be written as

$$\begin{aligned}&\max_{\boldsymbol{\beta},Z}\left\{\sum_{\{i|\delta_i\varepsilon_i=1\}}\left[\boldsymbol{x_i\beta}-\log\left\{\sum_{j\in R_i}\omega_j(t_i)\exp(z_{ij})\right\}\right]-\lambda||\boldsymbol{\beta}||_1\right\}\\&=\max_{\boldsymbol{\beta},Z}\left\{c^T\boldsymbol{\beta}-\sum_{i=1}^{f}\log\left\{\sum_{j=1}^{n}I_{ij}\omega_j(t_i)\exp(z_{ij})\right\}-\lambda||\boldsymbol{\beta}||_1\right\},\end{aligned}\tag{6}$$

where $\boldsymbol{c}=\sum_{\{i|\delta_i\varepsilon_i=1\}}\boldsymbol{x_i}\in R^p$.

For (6), introducing a dual variable $U:=(u_{ij})\in R^{f\times n}$, the dual form can be written as

$$\min_{U}\max_{\boldsymbol{\beta},Z}\boldsymbol{c}^T\boldsymbol{\beta}-\sum_{i=1}^{f}\log\left(\sum_{j=1}^{n}I_{ij}\omega_j(t_i)\exp(z_{ij})\right)-\lambda||\boldsymbol{\beta}||_1+tr(U(Z^T-X\boldsymbol{\beta}\mathbf{1}^T)),\tag{7}$$

where $tr(A)$ refers to the trace of the matrix A.

Following the derivation detailed in Appendix A, the dual problem given in Equation (7) can be rewritten as[4]:

$$\begin{aligned}&\min_{U}\sum_{i=1}^{f}\sum_{j=1}^{n}u_{ij}(\log u_{ij}-\log\omega_j(t_i)),\\&\text{s.t. }||\boldsymbol{X}^TU\mathbf{1}-c||_\infty\le\lambda,\ U^T\mathbf{1}=\mathbf{1},U\ge\mathbf{0},U\circ(1-I)=0,\end{aligned}\tag{8}$$

where $\circ$ denotes the element-wise multiplication.

Theorem 1 below gives the SAFE rule for the Lasso PSH model. The proof of this theorem is available in Appendix B.

Theorem 1 (SAFE rule for Lasso-PSH). *Consider the optimization problem Lasso-PSH in (5). Denote by $\boldsymbol{x_k}$ the kth feature (column) of the matrix X. We can obtain the index set for all inactive (excluded) features:*

$$\zeta=\left\{k|\lambda>\max\left(c_k-\sum_{i=1}^{f}\min_{j:I_{ij}=1}x_{jk},\ \sum_{i=1}^{f}\max_{j:I_{ij}=1}x_{jk}-c_k\right)\right\},\tag{9}$$

where $c=\sum_{\{i|\delta_i\varepsilon_i=1\}}\boldsymbol{x_i}\in\mathbb{R}^p$, f is the number of unique events of interest failure times and $I_{ij}=I_{i\in R_j}$, $R_j=\{k:(T_k\ge T_j)\cup((T_k\le T_j)\cap(\delta_k=1)\cap(\varepsilon_k\ne 1))\}$.

According to the above SAFE screening rule (9), for every index $k\in\zeta$, the kth entry of $\boldsymbol{\beta}^*$ (the optimum of the primal problem) is zero, that is, $(\beta^*)_k=0$, and feature $\boldsymbol{x_k}$ can be safely eliminated from X, a priori to solving the optimization problem (6).

With the adaptive Lasso penalty, we can obtain the SAFE rule for the adaptive Lasso PSH model as stated in following Theorem 2.

Theorem 2 (SAFE rule for adaptive Lasso-PSH). *Similar to the Lasso PSH model, the primal optimal problem for the adaptive Lasso PSH model can be written as:*

$$\begin{aligned}&\max_{\boldsymbol{\beta}}l(\boldsymbol{\beta})-\lambda\sum_{j=1}^{p}|\beta_j|/|\tilde{\beta}_j|\\&=\max_{\boldsymbol{\beta}}\left\{\sum_{i=1}^{n}I(\delta_i\varepsilon_i=1)\left[\boldsymbol{\beta}^Tx_i-\log\left\{\sum_{j\in R_i}\omega_j(t_i)\exp(\boldsymbol{\beta}^Tx_j)\right\}\right]-\lambda\sum_{j=1}^{p}|\beta_j|/|\tilde{\beta}_j|\right\}.\end{aligned}\tag{10}$$

Following the procedure in Theorem 1, we can obtain the index set for all inactive (excluded) features:

$$\zeta=\left\{k|\frac{\lambda}{|\tilde{\beta}_k|}>\max(c_k-\sum_{i=1}^{f}\min_{j:I_{ij}=1}x_{jk},\ \sum_{i=1}^{f}\max_{j:I_{ij}=1}x_{jk}-c_k)\right\}.\tag{11}$$

Based on the SAFE screening rule in (11), for every index $k \in \zeta$, the kth entry of $\boldsymbol{\beta}^*$ is zero, that is, $(\beta^*)_k = 0$, and feature $\boldsymbol{x_k}$ can be safely eliminated from X, a priori to solving the optimization problem (10).

In penalized Lasso-type problems, the tuning parameter λ plays an important role. Typically, the optimal value of λ is chosen via cross-validation, Akaike information criterion (AIC), Bayesian information criterion (BIC), generally cross-validation (GCV), and/or other criteria. Since optimization problems over a sequence of tuning parameter values are involved, such procedures are usually time consuming.

With the help of the proposed screening method, the computational burden on solving Lasso or adaptive Lasso PSH models can be greatly alleviated. Hence, for a given λ, some inactive features of (5) or (10) can be identified and discarded. In other words, for a specified λ, only a partial data matrix is involved to solve the optimization problem in (5) or (10), and consequently the algorithm efficiency is substantially improved.

3 | NUMERICAL STUDIES

Since the Lasso penalty leads to biased estimates for true nonzero coefficients[23] and tends to select too many noninformative variables[24] while the adaptive Lasso ensures the existence of global optimizers, produces less biased estimators and reduces the number of false positives, here we only provide empirical results of the adaptive Lasso penalty.

In the following, we will systematically evaluate the screening and predictive performance of the proposed PSH-SAFE method with adaptive Lasso-type penalty (shorten as "PSH_SAFE_aLasso") on simulation and real-world datasets.

3.1 | Evaluation criteria

To evaluate the performance of our proposed algorithm, the following evaluation criteria are specified before presenting the experimental results.

3.1.1 | Efficiency of screening

To measure the efficiency of SAFE screening rules, we choose two popular criteria, that is, the rejection ratio[25,26] and screen ratio:[27,28]

$$\text{Rejection ratio} = \frac{\text{Number of eliminated features by screening}}{\text{Number of inactive features in original Lasso solution } \beta^*},$$
$$\text{Screen ratio} = \frac{\text{Number of retained features}}{\text{Original feature dimension}(p)}.$$

Here, inactive features are those features whose coefficients will be set to zero in the original Lasso optimization problem. The rejection ratio less than or equal to 1 implies the screening is SAFE (coefficients of discarded features are guaranteed to be zero in the targeted optimal solution), and a lower screening ratio means feature dimensionality is dramatically decreased.

In this article, we adopt the approach in a previous study[27] to build the solution path: initialize $\lambda_{\max}$ to a sufficiently large value, which force all $\boldsymbol{\beta}$ to a zero vector, and then gradually decrease λ in each iteration. Hence, $\lambda_{\max}$ is obtained by setting all $\hat{\beta}_j$ to 0. As for $\lambda_{\min}$, if $n \geq p$, we set $\lambda_{\min} = 0.001\lambda_{\max}$, else we set $\lambda_{\min} = 0.05\lambda_{\max}$. In our experiments, we search m different λ values in total. For the kth step, $\lambda_k = \lambda_{\max}(\lambda_{\min}/\lambda_{\max})^{k/m}$.

3.1.2 | Prediction performance

After variable screening and/or variable selection procedures, it is usually necessary to evaluate the predictive performance power of the model in question. In this study, a concordance index (C-index)[29] is adopted to evaluate the accuracy of survival models in competing risks. The C-index metric is calculated by comparing a risk score $\tilde{M}(X)$ at time t with the survival time of each subject. Here, a higher value of $\tilde{M}(X)$ implies a higher risk of the event of interest. For two subjects

X_i, X_j, the concordance value can be obtained by

$$C_1(t) := P(M(t, X_i) > M(t, X_j) | \varepsilon_i = 1 \text{ and } T_i \leq t \text{ and } (T_i < T_j \text{ or } \varepsilon_j = 2)). \quad (12)$$

In this study, the above risk scores are based on estimates of the cumulative incidence function obtained by the Fine-Gray's model. Here, the reported C-index values are evaluated over the generated 100 λs on the test sets for each of the three methods (PSH_Lasso, PSH_aLasso, PSH_SAFE_aLasso).

3.2 | Simulated study

We first investigate the screening performance on data with different censoring rates, dimensions and different correlations between covariates. Then, we explore the computational efficiency under different combinations of dimensions and samples.

3.2.1 | Simulation settings

Similar to Reference 20, we consider the following two settings: $(n, p) = (100, 500)$ and $(n, p) = (200, 800)$. And covariates $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ are marginally standard normal with pairwise correlations $\text{corr}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \rho^{|i-j|}$. In the experiments, we set $\rho = 0.6, 0.9$ to reflect moderate and high correlated cases among the covariates. Censoring times are generated from a uniform distribution $U(0, c_0)$, where c_0 is chosen to obtain the low (about 25%), moderate (about 50%) an high (about 70%) censoring rates. Here, we consider two events, one primary event and one competing event. In the following, denote the primary event by 1 and the competing event by 2. Also denote the regression parameter of cause 1 by $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \ldots, \beta_{1p})^T$ and set $\boldsymbol{\beta}_1 = (0.5, 0.5, -0.5, 0.5, 0, \ldots, 0)^T$; and for cause 2, $\boldsymbol{\beta}_2 = -\boldsymbol{\beta}_1$. The CIF of cause 1 is:

$$F_i(t|X) = P(T \leq t, \varepsilon = 1|X) = 1 - [1 - pr\{1 - \exp(-t)\}]^{\exp(\boldsymbol{\beta}_1^T X)},$$

which is a unit exponential mixture with mass $1 - pr$ at ∞ when $X = 0$. The value of pr is set to 0.3. The CIF for cause 2 is obtained by taking $P(\varepsilon = 2|X) = 1 - P(\varepsilon = 1|X)$ and then using an exponential distribution with rate $\exp(\boldsymbol{\beta}_2^T X)$ for the conditional CIF, $P(T \leq t, |\varepsilon = 2, X)$.

Hence, altogether 12 simulated scenarios are investigated. Simulated datasets with moderate correlations (Case 1: $\rho = 0.6$) and high correlations (Case 2: $\rho = 0.9$) are summarized in Tables 1 and 2, respectively.

3.2.2 | Comparison results on simulated data

First, we present the simulated results for Case 1 ($\rho = 0.6$) when moderated correlations between covariates are present. The screening related results, namely, screen ratio and rejection ratio with different λs are shown in Figures 1 and 2, respectively. While the predictive results in terms of C-index with different λs is presented in Figure 3.

TABLE 1 Simulated datasets with moderated correlations (Case 1: $\rho = 0.6$)

Dataset	#instance	#feature	Censoring rate	Event 1 proportion
S1	100	500	31%	35%
S2	100	500	62%	19%
S3	100	500	80%	7%
S4	200	800	25%	44%
S5	200	800	54%	23%
S6	200	800	79%	9.5%

TABLE 2 Simulated datasets with high correlations (Case 2: $\rho = 0.9$)

Dataset	#instance	#feature	Censoring rate	Event 1 proportion
S7	100	500	21%	39%
S8	100	500	49%	22%
S9	100	500	70%	13%
S10	200	800	27.5%	34%
S11	200	800	56%	15.5%
S12	200	800	74.5%	8%

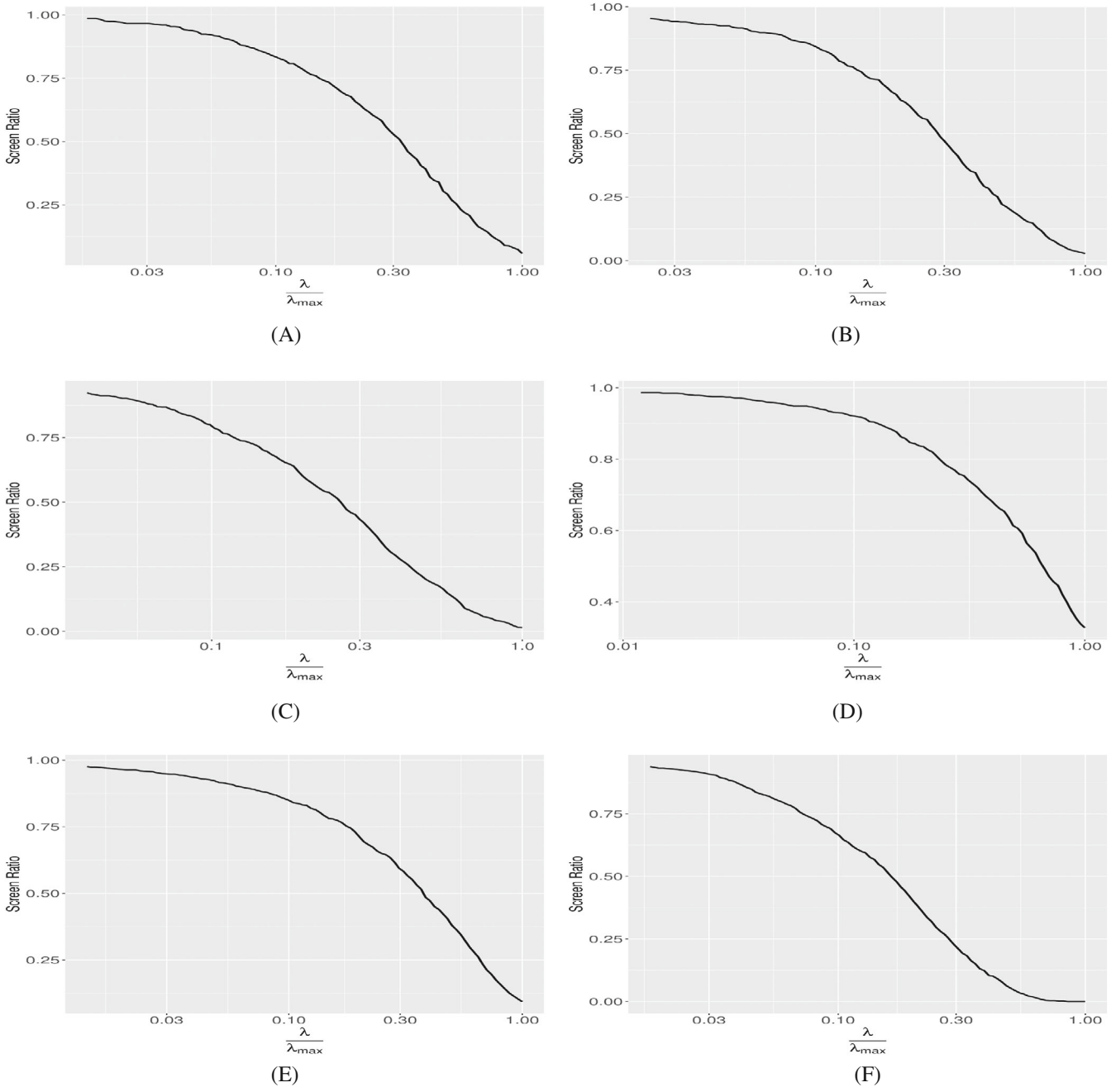

FIGURE 1 Screen ratio for Case 1 given 100 λs parameters. (A) S1, (B) S2, (C) S3, (D) S4, (E) S5, and (F) S6

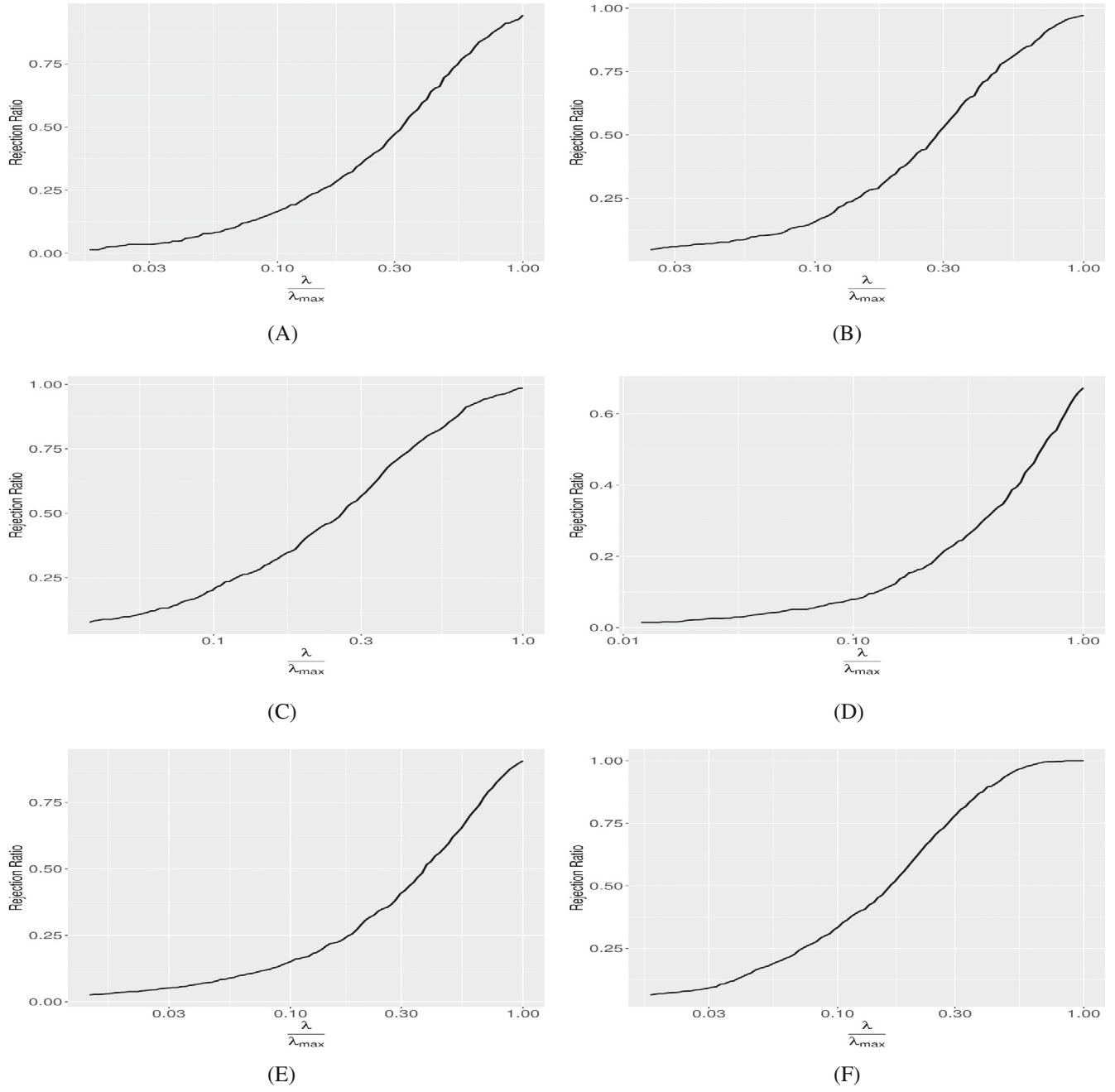

FIGURE 2 Rejection ratio for Case 1 given 100 λs parameters. (A) S1, (B) S2, (C) S3, (D) S4, (E) S5, and (F) S6

From Figure 1, one can observe that the screen ratio decreases very rapidly with the increase of $\lambda/\lambda_{\max}$, indicating that it is very effective in reducing the dimensionality of data. According to Figure 2, we can see that under all these six simulated scenarios, the rejection ratio is always less than or equal to 1, which satisfies the SAFE property, that is, the rejection ratio will not be greater than one.[3]

According to Figure 3, our algorithm (PSH_SAFE_aLasso, the rightmost one of each subfigure) outperforms the original PSH_aLasso in all these cases and gain better results than PSH_Lasso in most cases (moderate censoring rate with/without a larger sample size). We also notice that, when highly censored data (S3 and S6) are present, all compared models produces unsatisfactory or incorrect predictions as most C-index predictions are below 0.5. This not something unexpected since highly censoring rates (80% and 79%) imply only a small portion of data (7% and 9.5%) have events of interest. With such few data, parameter estimation for PSH may not be reliable and these inaccurate parameters will lead to low predictive performance.

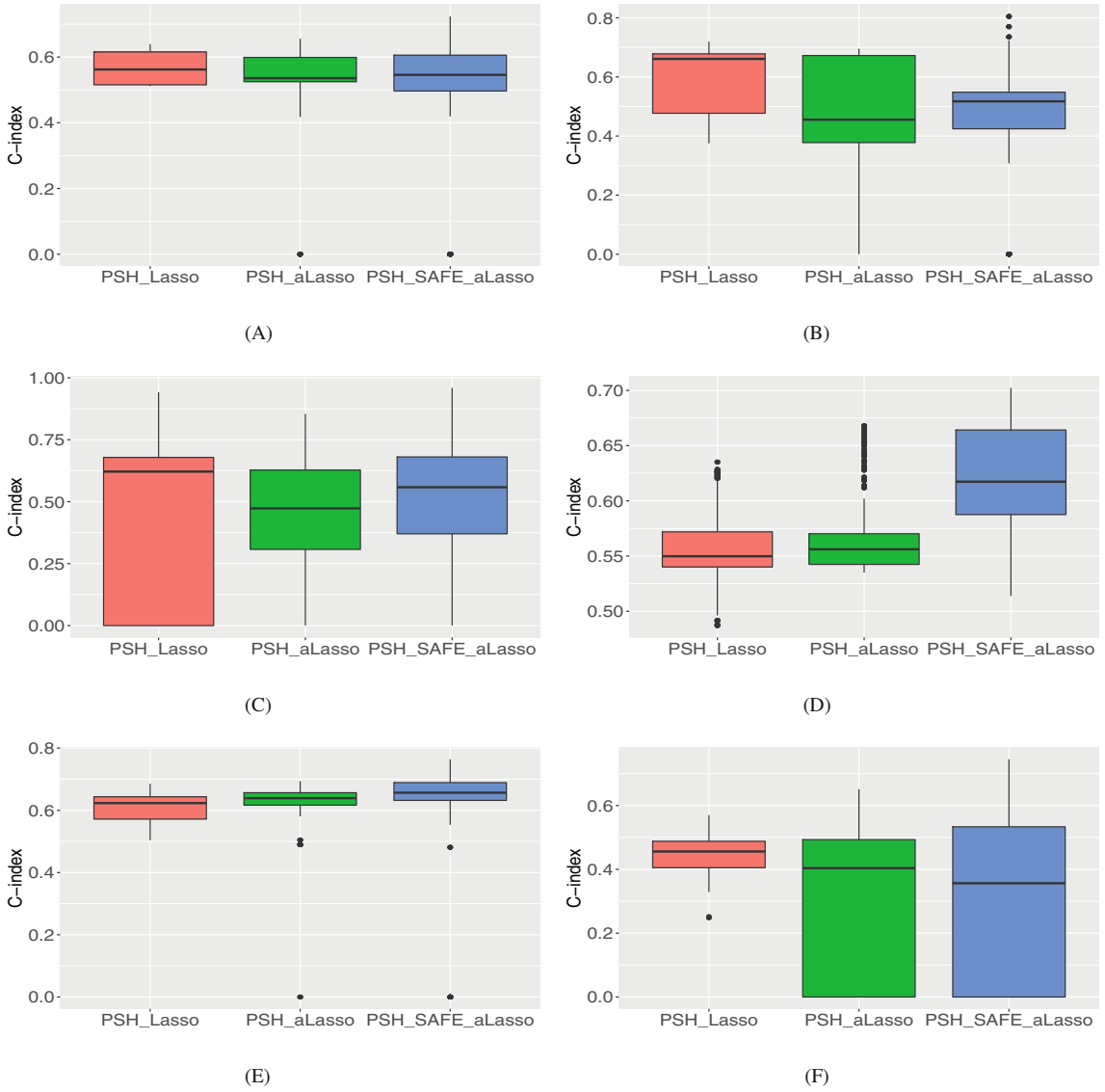

FIGURE 3 C-index boxplots for Case 1 over different λs. (A) S1, (B) S2, (C) S3, (D) S4, (E) S5, and (F) S6

Next, we present the simulated results for Case 2 ($\rho = 0.9$) when high correlations between covariates are present. This is often the case when Omics data are involved. The screening related results, namely, screen ratio and rejection ratio with different λs are shown in Figure 4 and Figure 5, respectively. And the predictive results in terms of C-index with different λs is presented in Figure 6.

According to Figures 4 and 5, the results on screen ratio and rejection ratio is very similar to the results in Case 1. These results again demonstrate that the proposed method is effective in screening efficiency and safe in eliminating inactive features.

From Figure 6, one can find that, in terms of predictive capability, the proposed method (PSH_SAFE_aLasso, the rightmost one in each subfigure) again show similar performance: PSH_SAFE_aLasso beats PSH_aLasso in almost all cases and outperforms the lasso method in most scenarios, except the heavily censoring case S9 where all three methods achieve comparable results.

From both Figures 5 and 6, the proposed algorithm generally performs the best in terms of C-index and the superiority stands out when moderate censoring and/or high correlated covariates are present. However, with a relative small sample size ($n = 100,\ 200$), if highly censored rates are encountered, only a small proportion (about 10% in four scenarios) event of interest data will be used for the screening procedure and estimating the weight $\tilde{\boldsymbol{\beta}}$. Consequently, the noise or instability incurred during both procedures make PSH_SAFE_aLasso may not work as expected.

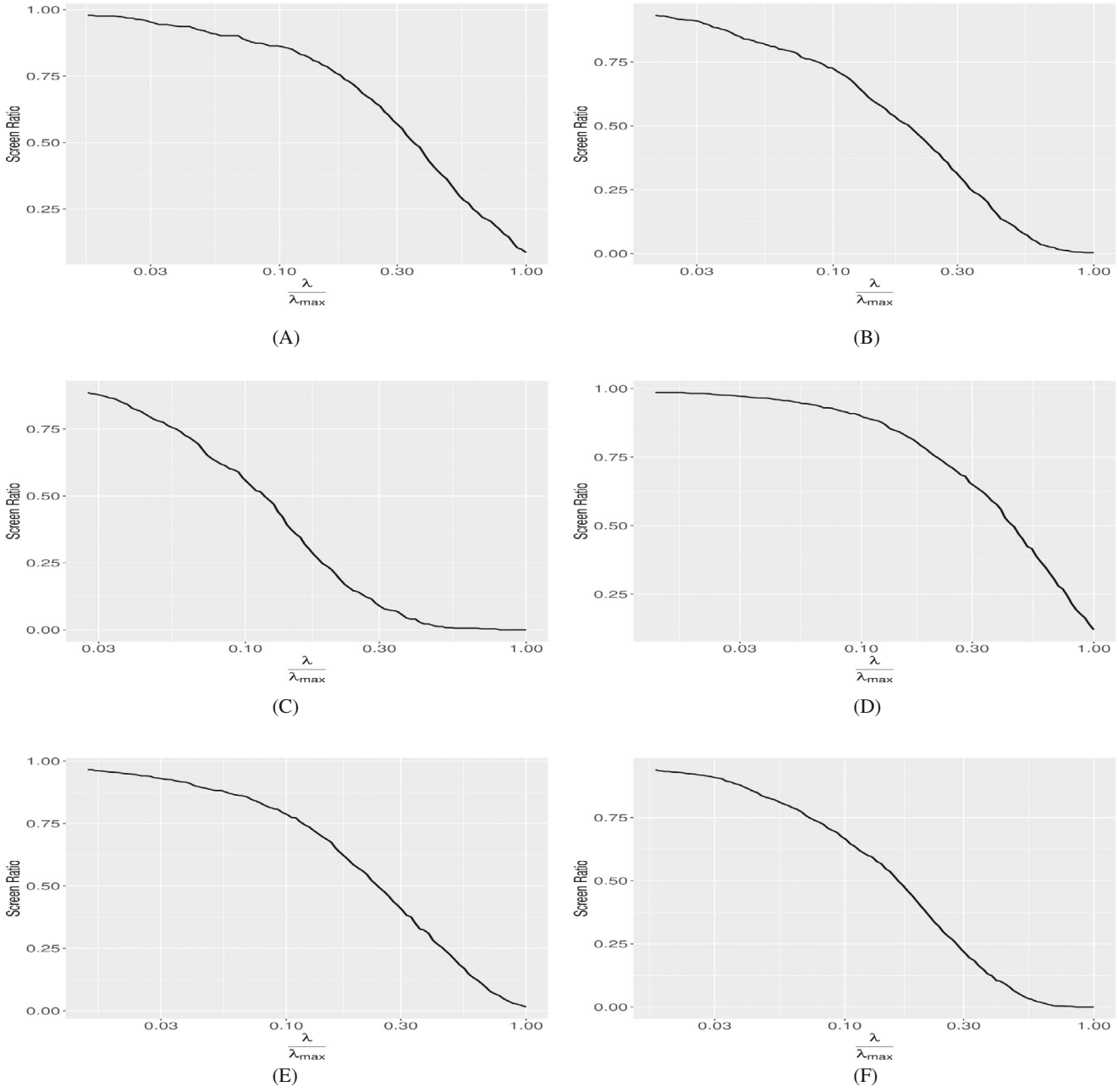

FIGURE 4 Screen ratio for Case 2 given 100 λs parameters. (A) S7, (B) S8, (C) S9, (D) S10, (E) S11, and (F) S12

In the above experiments, the proportion of the primary event (interest of event 1) varies from 8% to 44% and all the rejection and screen ratios results from 12 scenarios suggest that this does not have any effect on the screening efficiency or the SAFE property of the algorithm. As to the effect on predictive capability, we find that the proposed method (PSH_SAFE_aLasso) works best when the primary event data occupy a large proportion (20%) (S4, S5, S7, S8, S10). But, PSH_SAFE_aLasso also beats the other two models on some low proportion cases (S3, S11). Therefore, we may conclude that the proposed method is not too much sensitive to the proportion of primary event.

3.2.3 | Results on computation efficiency

In survival data with competing risks, we may come across different kinds of big data such as large sample sized and/or ultra-high dimensional data. The computation efficiency of different kinds of method under such settings is our

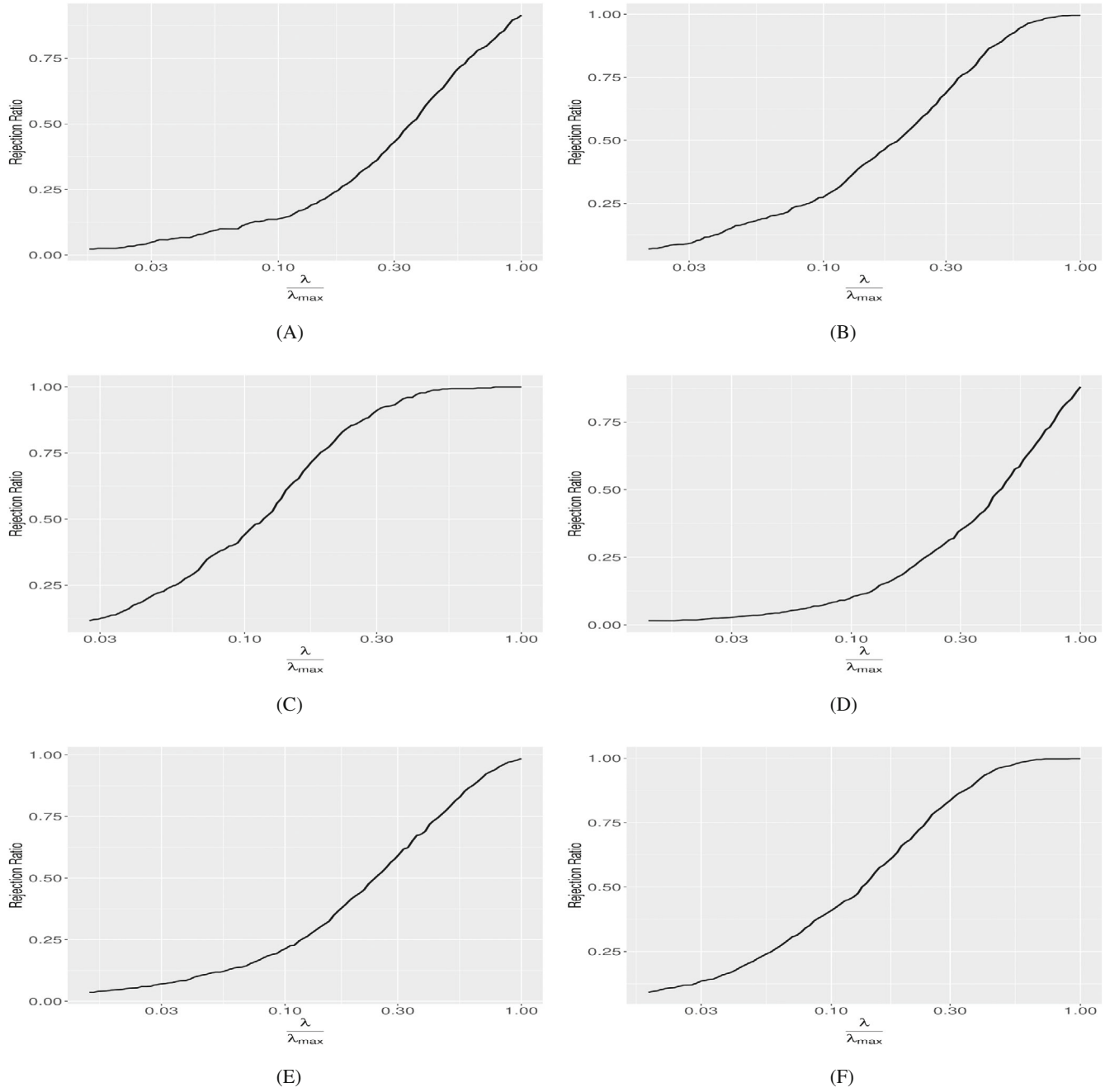

FIGURE 5 Rejection ratio for Case 2 given 100 λs parameters. (A) S7, (B) S8, (C) S9, (D) S10, (E) S11, and (F) S12

primary concern. To validate the effectiveness of the proposed method in reducing computational burdens, we evaluate the running times of our algorithm and other competitive algorithms under different settings.

In this experiment, simulation settings are almost the same with the above comparison study in the case of $\rho = 0.9$ and moderate censoring. However, we vary the values of n and p to denote the high dimensional case, ultra-high dimensional case, large sample sized case and the case of both large sample size and ultra-high dimensional data. Here, cases C1, C2, and C3 have the same dimensionality ($p = 500$) but with different sample sizes ($n = 100,\ 200, 2000$). Cases C2, C4, and C5 have the fixed sample size ($n = 200$) but different dimensionality ($p = 500,\ 800, 5000$). The most challenging case is C6, where one can find both a large sample size ($n = 2000$) and a high dimensionality ($p = 5000$). In all these cases, λs are chosen to screening out about half of the dimensionality and all models are trained 100 times. Detailed information for these simulated datasets can be found in Table 3.

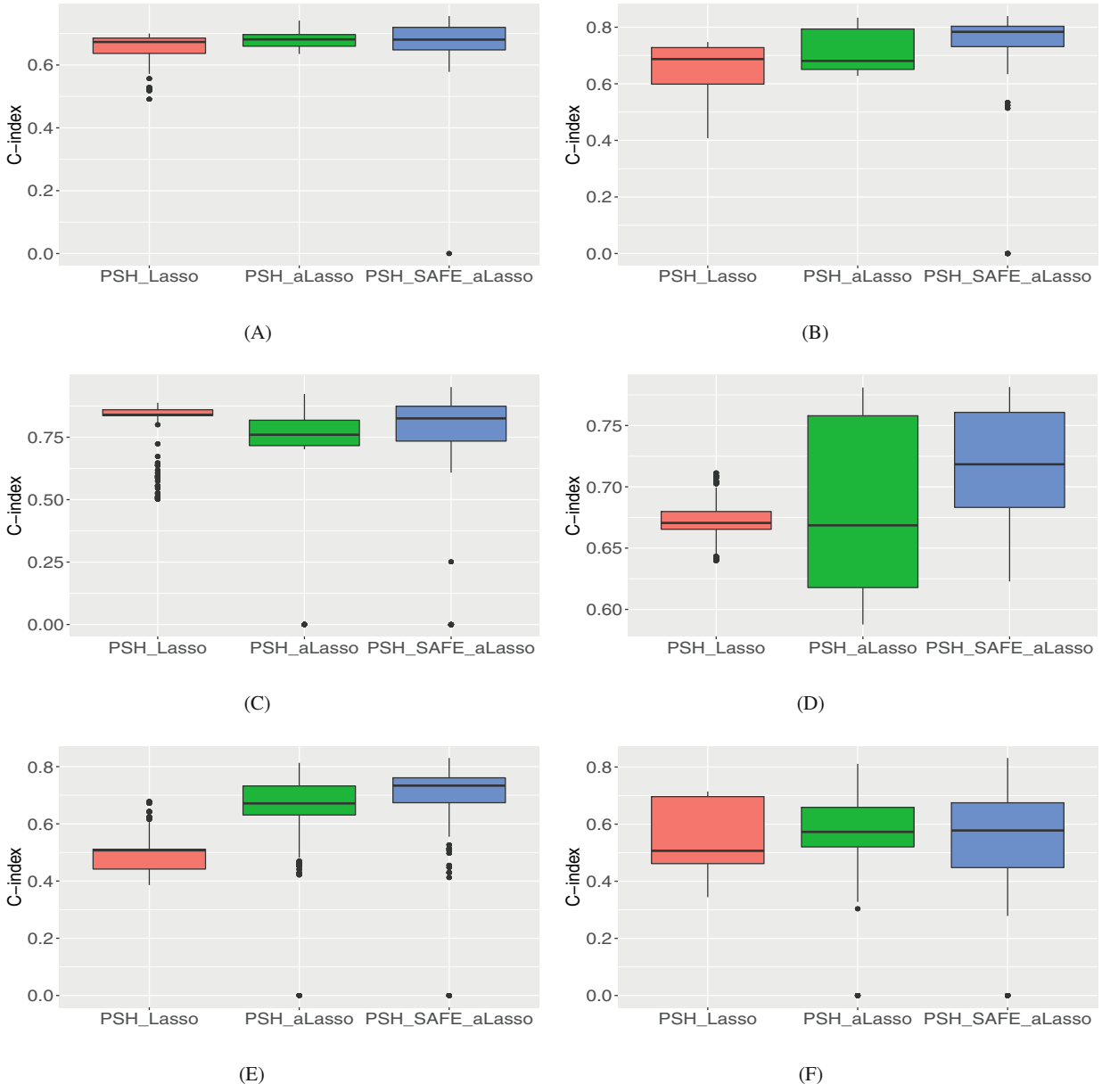

FIGURE 6 C-index boxplots for Case 2 over different λs. (A) S7, (B) S8, (C) S9, (D) S10, (E) S11, and (F) S12

TABLE 3 Simulated cases for computational efficiency comparison

Dataset	#instance	#feature	Censoring rate	Event 1 proportion
C1	100	500	58%	21%
C2	200	500	62%	19%
C3	2000	500	49%	25%
C4	200	800	49%	19%
C5	200	5000	51%	24%
C6	2000	5000	51%	25%

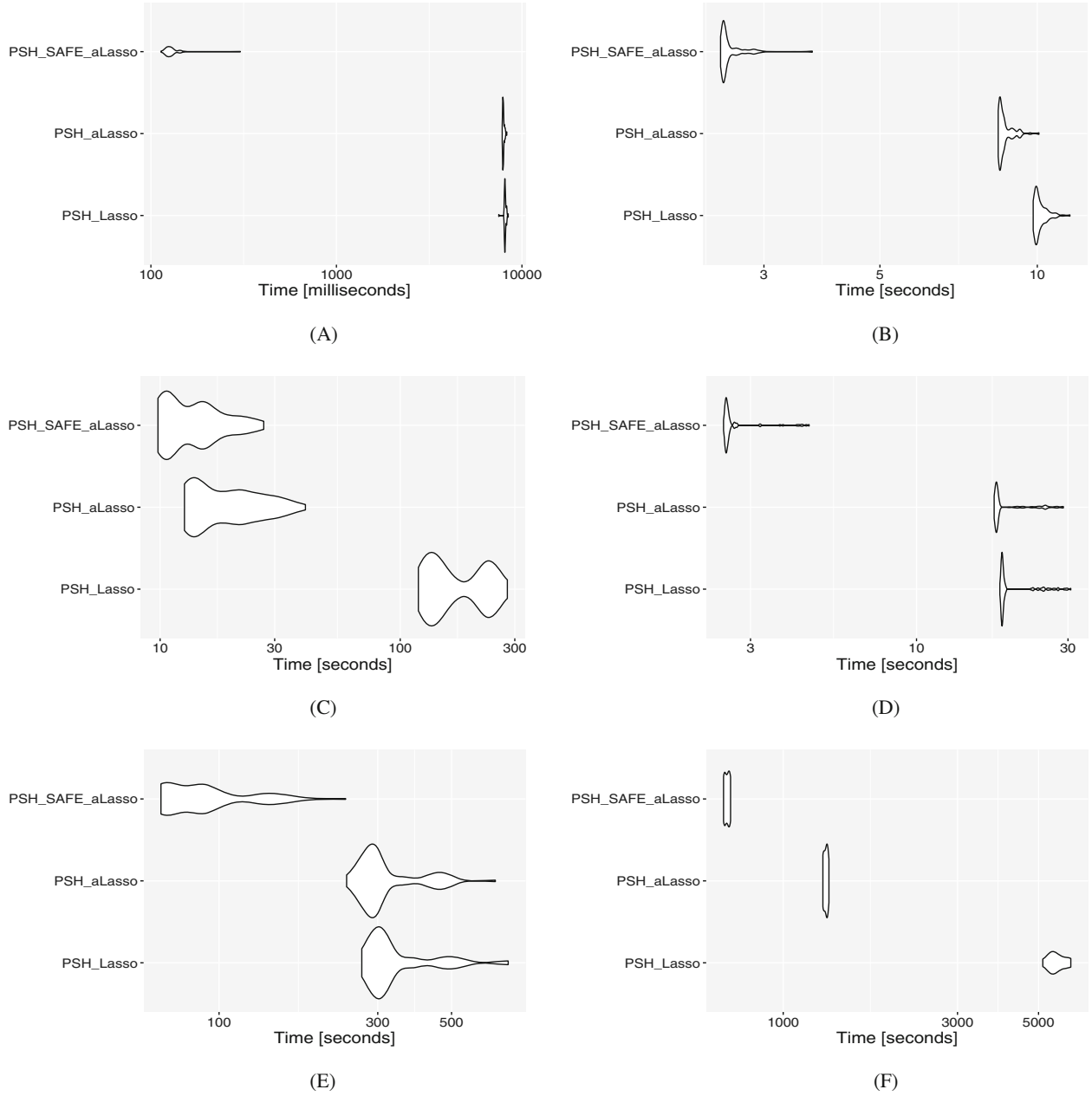

FIGURE 7 Runtime comparison for simulation datasets. (A) C1, (B) C2, (C) C3, (D) C4, (E) C5, and (F) C6

Figure 7 shows the running times of all three compared algorithms over 100 runs. Table 4 gives the average running times of PSH_SAFE_aLasso (with screening) and PSH_aLasso algorithms (without screening). We also provide the speed-ups of the proposed PSH_SAFE_aLasso method in all simulated cases.

As can be seen from Figure 7, in terms of time efficiency, the proposed algorithm always takes the lead in all six simulated scenarios. PSH_aLasso takes the second while PSH_Lasso always takes the most amount of running times. We also observe that there is a sharp difference in running times between the proposed PSH_SAFE_aLasso and methods without safe screening for small sample-sized high dimensional and ultra-high dimensional data. However, with the increase of sample size, PSH_SAFE_aLasso begins to deteriorate but is still faster than PSH_aLasso and much faster than PSH_Lasso.

From Table 4, one may find that adaptive lasso method with safe screening generally outperforms its no-screening counterpart by a noticeable margin in terms of computational efficiency. With safe screening, the speedups can be several times or even several dozens of times in our simulations.

TABLE 4 Average runtime comparison for the PSH-aLasso with and without SAFE screening rule over 100 runs

Dataset	With screening	Without screening	Speed up
C1 ($n = 100, p = 500$)	131.29 ms	7947.51 ms	60.53
C2 ($n = 200, p = 500$)	2.59 s	8.66 s	3.35
C3 ($n = 2000, p = 500$)	14.19 s	19.73 s	1.39
C4 ($n = 200, p = 800$)	2.69 s	19.47 s	7.25
C5 ($n = 200, p = 5000$)	94.06 s	324.56 s	3.45
C6 ($n = 2000, p = 5000$)	702.00 s	1307.87 s	1.86

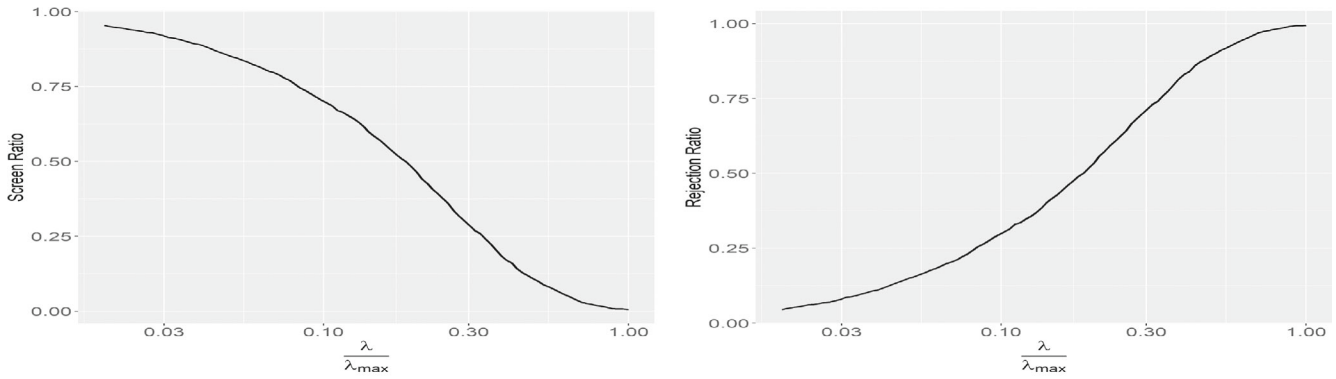

FIGURE 8 Empirical analysis of screening efficiency and safety: Screen ratio given 100 λs parameters. (A) Screen ratio and (B) rejection ratio

3.3 | Real application

In this part, we use a publicly available bladder cancer dataset to perform an empirical analysis of the proposed method.[30] This dataset is further preprocessed by eliminating all columns have the same values or data with missing values. The resulting dataset includes 329 samples, each corresponding to 1381 publicly available preprocessed custom platform microarray features.

In this dataset, the response of interest is the time to progression or death from bladder cancer. And death from other or unknown causes is the competing event. For the former event, 57 patients were observed while for the latter, 49 were observed. The remaining 223 patients are censored samples.

3.3.1 | Results on screening efficiency and safety

The screening efficiency and safety of the proposed screening procedure is shown in Figure 8. According to Figure 8(a), the screen ratio decreases fast with the increase of dimensionality, indicating that the method can dramatically decrease the feature dimensionality. From Figure 8(b), we know that the higher the dimension, the higher the rejection ratio. The rejection ratio ≤ 1, which means the PSH_SAFE_aLasso can successfully identify a majority of the inactive features and they only eliminate features that are guaranteed to be absent after solving the optimization problem.

3.3.2 | Results on prediction accuracy

In the experiments, results obtained are based on the 5×2 cross-validation procedure.[31] In 5×2 folds cross-validation, the dataset is randomly divide into two equal-sized blocks. The model is trained on the first block and evaluated

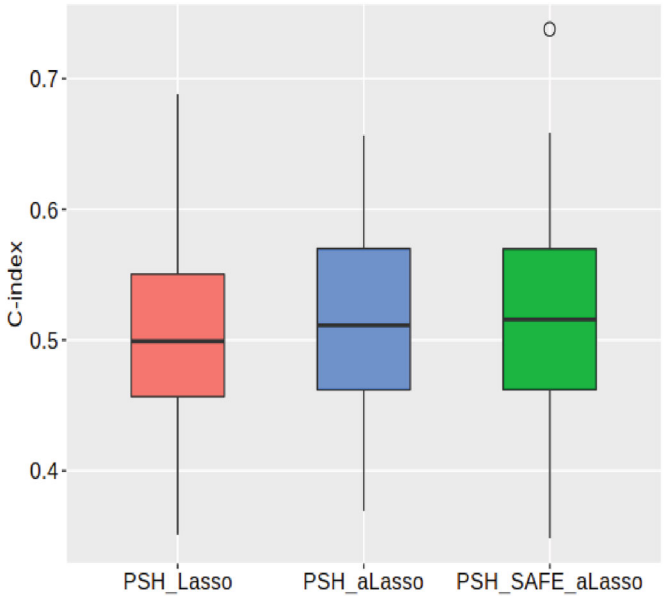

FIGURE 9 Prediction performance comparison: C-index on the real datasets with best chosen λs

on the second block and vice versa. This process is repeated five times. Here, following the reviewers' suggestions, instead of comparing all the predictive power over the 100 generated λs values, we compare the predictive power of all three models by choosing their best parameter λs, respectively. Here, the best tuning parameter is chosen via cross-validation.

From Figure 9, we can clearly see that in terms of C-index, adaptive methods (with screening or not) give almost identical results and both methods outperform the competing (PSH_Lasso) method. This again demonstrates the effectiveness of the proposed method.

3.3.3 | Results on time efficiency

In simulated study, we have seen that with a fixed λ value (screening out about half of the dimensionality), the proposed method achieves the best performance in terms of time efficiency in all simulated cases. Here, we want to explore the running times at different values of tuning parameter λs. Here, we select four representative quantiles (ie, the upper 5th, 25th, 75th, and 95th) from the generated 100 λs and compare the runtimes of all these three methods with these four λs. Again, for each λ, the experiment is run a hundred times. Figure 10 shows the runtimes of all three compared models on the bladder cancer dataset.

From Figure 10, we can clearly see that the PSH_SAFE_aLasso takes the least time compared with PSH_aLasso and PSH_Lasso. In order to show the advantages of our algorithm's fast speed more clearly, we display average running time values in the form of a Table 5.

It is seen from Table 5 that a larger λ value (such as λ_5 or λ_{25}) is associated with a larger speed-up. This is reasonable, since with a large λ value, a large proportion of inactive features will be removed prior to training the adaptive Lasso PSH model. Hence, the model is trained on a much reduced data matrix, which consequently leads to greater savings in the computational time.

4 | DISCUSSION

We have proposed fast SAFE screening algorithms for the PSH model for competing risks data with ultrahigh dimensional covariates. Our empirical studies demonstrate that our algorithm is able to efficiently and safely eliminate some features whose corresponding coefficients of optimization problem are guaranteed to be zero. Moreover, it can significantly reduce

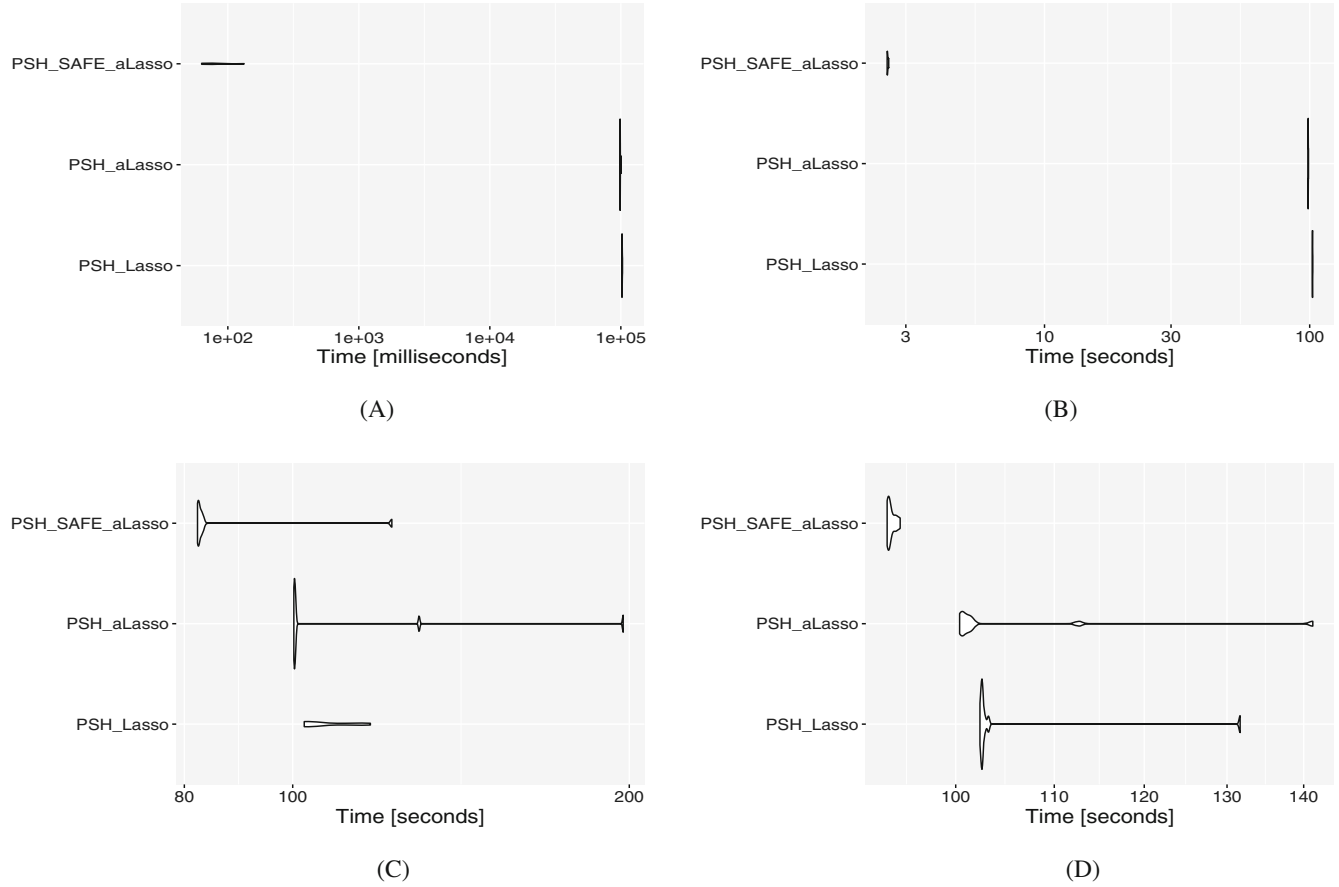

FIGURE 10 Runtime on the real dataset with four specific λs over 100 runs. (A) λ_5, (B) λ_{25}, (C) λ_{75}, and (D) λ_{95}

TABLE 5 Average runtime comparison for the PSH-aLasso with and without SAFE screening rule over 100 runs

λ	With screening	Without screening	Speed up
λ_5	82.68 ms	98886.73 ms	1196.02
λ_{25}	2.56 s	98.74 s	38.57
λ_{75}	82.4 s	100.45 s	1.31
λ_{95}	93.94 s	106.10 s	1.13

the running time for large high dimensional competing risks data while maintaining competitive prediction performance. In particular, the computational superiority stands out for large penalty parameter values.

We point out that the proposed PSH-SAFE screening strategy works with any PSH Lasso solvers. Inspired by the fact that the computational complexity for the log-pseudo likelihood and its derivatives for the PSH model can be reduced from $O(n^2)$ to $O(n)$,[17] our team is currently working on a more efficient PSH-SAFE implementation to handle competing risks survival data with both ultrahigh dimensionality and massive sample size.

An R package "SFEcmprsk" has been developed for the proposed screen procedure and the code is available at https://github.com/whcsu/safecomp.

ACKNOWLEDGEMENTS

Hong Wang is supported by National Social Science Foundation of China(No.17BTJ019), National Statistical Scientific Research Project of China(2022LZ28), Changsha Municipal Natural Science Foundation (No.kq2202080), Zhenyuan Shen

is supported by Fundamental Research Funds for the Central Universities of Central South University (2020zzts361), and Gang Li is supported by National Institutes of Health (P30 CA-16042, UL1TR000124-02, and P01AT003960).

DATA AVAILABILITY STATEMENT

The bladder cancer dataset is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5479.

ORCID

Hong Wang https://orcid.org/0000-0002-6938-9507
Gang Li https://orcid.org/0000-0002-4753-9420

REFERENCES

1. Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Stat Soc Ser B (Methodol)*. 1996;58(1):267-288.
2. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2006;101(476):1418-1429. doi:10.1198/016214506000000735
3. Ghaoui LE, Viallon V, Rabbani T. Safe feature elimination in sparse supervised learning. *Pacif J Optim*. 2010;8(4):667-698. doi:10.1007/s10255-012-0191-1
4. Ko J. *Solving the Cox Proportional Hazards Model and Its Applications*. Master's thesis. EECS Department, University of California, Berkeley; 2017.
5. Wang Y, Xiang ZJ, Ramadge PJ. Lasso screening with a small regularization parameter. Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing; May 26, 2013:3342-3346; IEEE.
6. Fercoq O, Gramfort A, Salmon J. Mind the duality gap: safer rules for the Lasso; 2015:abs/1505.03410.
7. Ndiaye E, Fercoq O, Gramfort A, Salmon J. GAP safe screening rules for sparse multi-task and multi-class models. *Advances in Neural Information Processing Systems*. New York: Curran Associates, Inc: 2015:811-819.
8. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc*. 1999;94(446):496-509.
9. Hu XS, Tsai WY. Linear rank tests for competing risks model. *Stat Sin*. 1999;9:971-983.
10. Andersen PK, Abildstrom SZ, Rosthøj S. Competing risks as a multi-state model. *Stat Methods Med Res*. 2002;11(2):203-215.
11. Pintilie M. Analysing and interpreting competing risk data. *Stat Med*. 2007;26(6):1360-1367.
12. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001;96(456):1348-1360. doi:10.1198/016214501753382273
13. Binder H, Allignol A, Schumacher M, Beyersmann J. Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics*. 2009;25(7):890-896. doi:10.1093/bioinformatics/btp088
14. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat*. 2010;38(2):894-942. doi:10.1214/09-aos729
15. Kuk D, Varadhan R. Model selection in competing risks regression. *Stat Med*. 2013;32(18):3077-3088. doi:10.1002/sim.5762
16. Tapak L, Saidijam M, Sadeghifar M, Poorolajal J, Mahjub H. Competing risks data analysis with high-dimensional covariates: an application in bladder cancer. *Genom Proteom Bioinform*. 2015;13(3):169-176. doi:10.1016/j.gpb.2015.04.001
17. Kawaguchi ES, Shen JI, Suchard MA, Li G. Scalable algorithms for large competing risks data. *J Comput Graph Stat*. 2021;30(3):685-693. doi: 10.1080/10618600.2020.1841650.
18. Chen X, Li C, Zhang T, Gao Z. On correlation rank screening for ultra-high dimensional competing risks data. *J Appl Stat*. 2022;49(7):1848-1864. doi:10.1080/02664763.2021.1884209
19. Li E, Tian M, Tang ML. Variable selection in competing risks models based on quantile regression. *Stat Med*. 2019;38(23):4670-4685. doi:10.1002/sim.8326
20. Erqian L, Bo M, Maozai T. Feature screening based on ultrahigh dimensional competing risks models. *Sci Sin Math*. 2018;48(8):1061.
21. Fu Z, Parikh CR, Zhou B. Penalized variable selection in competing risks regression. *Lifetime Data Anal*. 2017;23(3):353-376. doi:10.1007/s10985-016-9362-3
22. Ren X, Li S, Shen C, Yu Z. Linear and nonlinear variable selection in competing risks data. *Stat Med*. 2018;37(13):2134-2147. doi:10.1002/sim.7637
23. Wu Y. Elastic net for Cox's proportional hazards model with a solution path algorithm. *Stat Sin*. 2012;22(1):271-294. doi:10.5705/ss.2010.107
24. Zou H, Hastie T. Regression shrinkage and selection via the elastic net, with applications to microarrays. *J R Stat Soc Ser B*. 2003;67:301-320.
25. Wang J, Zhou J, Liu J, Wonka P, Ye J. A safe screening rule for sparse logistic regression. *Adv Neural Inf Process Syst*. 2014;27:1053-1061.
26. Ndiaye E, Fercoq O, Gramfort A, Salmon J. Gap safe screening rules for sparsity enforcing penalties. *J Mach Learn Res*. 2017;18(1):4671-4703.
27. Li Y, Wang L, Wang J, Ye J, Reddy CK. Transfer learning for survival analysis via efficient L2, 1-norm regularized Cox regression. Proceedings of the IEEE 16th International Conference on Data Mining (ICDM); 2016.
28. Bao R, Gu B, Huang H. Fast oscar and owl regression via safe screening rules. PMLR; 2020:653-663.
29. Wolbers M, Blanche P, Koller MT, Witteman JC, Gerds TA. Concordance for prognostic models with competing risks. *Biostatistics*. 2014;15(3):526-539.
30. Dyrskjøt L, Zieger K, Real FX, et al. Gene expression signatures predict outcome in non–Muscle-invasive bladder carcinoma: a multicenter validation study. *Clin Cancer Res*. 2007;13(12):3545-3551. doi:10.1158/1078-0432.ccr-06-2940

31. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput*. 1998;10(7):1895-1923.

How to cite this article: Wang H, Shen Z, Tan Z, Zhang Z, Li G. Fast Lasso-type safe screening for Fine-Gray competing risks model with ultrahigh dimensional covariates. *Statistics in Medicine*. 2022;41(24):4941-4960. doi: 10.1002/sim.9545

APPENDIX A. DERIVATION OF THE DUAL FORM (8)

In this appendix, we provide the detailed derivation of the dual form (8) of Lasso PSH.

For the dual form,

$$\min_{U}\max_{\beta,Z} \boldsymbol{c}^T\boldsymbol{\beta} - \sum_{i=1}^{f}\log\left(\sum_{j=1}^{n} I_{ij}\omega_j(t_i)\exp(z_{ij})\right) - \lambda||\boldsymbol{\beta}||_1 + tr(U(Z^T - X\boldsymbol{\beta}\mathbf{1}^T))$$

using $U = [\boldsymbol{u_1}, \boldsymbol{u_2}, \ldots, \boldsymbol{u_f}], Z = [\boldsymbol{z_1}, \boldsymbol{z_2}, \ldots, \boldsymbol{z_f}]$, where $\boldsymbol{u_i}, \boldsymbol{z_i} \in R^n$, the above formula can be rewritten as[4]

$$P^* = \min_{U}\sum_{i=1}^{f}\max_{\boldsymbol{z_i}} \boldsymbol{u_i}^T\boldsymbol{z_i} - \log\left(\sum_{j=1}^{n} I_{ij}\omega_j(t_i)\exp(z_{ij})\right) + \max_{\boldsymbol{\beta}} \boldsymbol{c}^T\boldsymbol{\beta} - \lambda||\boldsymbol{\beta}||_1 - tr(\mathbf{1}^T UX\boldsymbol{\beta}). \tag{A1}$$

In order to get P^*, first consider the situation containing only $\boldsymbol{\beta}$

$$\begin{aligned} f(\beta) &= \max_{\boldsymbol{\beta}} \boldsymbol{c}^T\boldsymbol{\beta} - \lambda||\boldsymbol{\beta}||_1 - tr(\mathbf{1}^T U\, X\, \boldsymbol{\beta}) \\ &= \max_{\boldsymbol{\beta}} \boldsymbol{\beta}^T(\boldsymbol{c} - X^T U^T\mathbf{1}) - \lambda||\boldsymbol{\beta}||_1. \end{aligned} \tag{A2}$$

For (A2), $f(\boldsymbol{\beta})$ is convex but not smooth. Next, we need to consider its subgradient

$$\frac{\partial f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{c} - X^T U^T\mathbf{1} - \lambda\boldsymbol{v} = 0. \tag{A3}$$

In which $\boldsymbol{v}$ is the subgradient of $||\boldsymbol{\beta}||_1$, its definition by element-wise is: $\forall k,\ k = 1, 2, \ldots, p$

$$v_k = \begin{cases} +1, & \beta_k > 0, \\ -1, & \beta_k < 0, \\ [-1, +1], & \beta_k = 0. \end{cases}$$

Plugging into Equation (A2), we can obtain that if $||\boldsymbol{X}^T U\mathbf{1} - c||_\infty \le \lambda$, then $f^*(\boldsymbol{\beta}) = 0$. So

$$P^* = \min_{U}\sum_{i=1}^{f}\max_{\boldsymbol{z_i}} \boldsymbol{u_i}^T\boldsymbol{z_i} - \log\left(\sum_{j=1}^{n} I_{ij}\omega_j(t_i)\exp(z_{ij})\right) :\ ||\boldsymbol{X}^T U\mathbf{1} - c||_\infty \le \lambda. \tag{A4}$$

For every i, consider every optimization problem

$$\max_{\boldsymbol{z_i}} \boldsymbol{u_i}^T\boldsymbol{z_i} - \log\left(\sum_{j=1}^{n} I_{ij}\omega_j(t_i)\exp(z_{ij})\right) :\ ||\boldsymbol{X}^T U\mathbf{1} - c||_\infty \le \lambda. \tag{A5}$$

The solution is

$$= \begin{cases} \sum_{j=1}^{n} u_{ij}(\log u_{ij} - \log \omega_j(t_i)), & u_i \geq 0, \mathbf{1}^T u_i = 1, \forall j : u_{ij}(1 - I_{ij}) = 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

So,

$$P^* = \min_U \sum_{i=1}^{f} \sum_{j=1}^{n} u_{ij}(\log u_{ij} - \log \omega_j(t_i)),$$
$$\text{s.t. } ||\boldsymbol{X}^T U\mathbf{1} - c||_\infty \leq \lambda,\ U^T\mathbf{1} = \mathbf{1}, U \geq \mathbf{0}, U \circ (1 - I) = 0, \tag{A6}$$

where $\circ$ is the multiplication between the element-wise.

APPENDIX B. PROOF OF THEOREM 1

After solving the dual problem of optimization, we can obtain the SAFE rule. In this appendix, we present the proof of Theorem 1.

We restate the optimization problem for PSH model here:

$$\max_{\boldsymbol{\beta}, Z} \left\{ c^T \boldsymbol{\beta} - \sum_{i=1}^{f} \log \left\{ \sum_{j=1}^{n} I_{ij} \omega_j(t_i) \exp(z_{ij}) \right\} - \lambda ||\boldsymbol{\beta}||_1 \right\},$$

where $\boldsymbol{c} = \sum_{\{i|\delta_i \varepsilon_i = 1\}} \boldsymbol{x_i} \in R^p$. For each feature $k = 1, 2, \ldots, p$, based on the convex optimization theory, if the following holds, then $\beta_k = \beta_k^*$, this is that β_k is at optimum.

$$\lambda > \max_U |\boldsymbol{x_k}^T U^T \mathbf{1} - c_k| \ : \ U\mathbf{1} = \mathbf{1}, U \geq 0, U \circ (1 - I) = 0. \tag{B1}$$

At the same time, from the dual form equation

$$\min_U \max_{\boldsymbol{\beta}, Z} \boldsymbol{c}^T \boldsymbol{\beta} - \sum_{i=1}^{f} \log \left(\sum_{j=1}^{n} I_{ij} \omega_j(t_i) \exp(z_{ij}) \right) - \lambda ||\boldsymbol{\beta}||_1 + tr(U(Z^T - X\boldsymbol{\beta}\mathbf{1}^T)),$$

we have if $\beta_k^* = 0$, the following must be true at optimum β_k^* .

$$-\lambda |\beta_k^*| + (\boldsymbol{x_k}^T U\mathbf{1} - c_k)\beta_k^* < 0. \tag{B2}$$

If all U in the feasible set satisfies Equation (B2), then the feature can be safely eliminated. To obtain the feature screening rule, the maximization problem (B1) must be solved. First, consider the below expression for each $\boldsymbol{x_k}$.

$$S_+(\boldsymbol{x_k}) = \max_U \boldsymbol{x_k}^T U\mathbf{1} \ : \ U^T\mathbf{1} = \mathbf{1},\ U \geq 0, U \circ (1 - I) = 0.$$

Based on duality, we can get

$$\begin{aligned} S_+(x_k) &= \min_Z \max_{U \geq 0, U^T\mathbf{1} = \mathbf{1}} \boldsymbol{x_k}^T U\mathbf{1} + trZ^T((I - \mathbf{1}\mathbf{1}^T) \circ U) \\ &= \min_Z \max_{U \geq 0, U^T\mathbf{1} = \mathbf{1}} trU((I - \mathbf{1}\mathbf{1}^T) \circ Z + \mathbf{1}\boldsymbol{x_k}^T) \\ &= \min_Z \sum_{i=1}^{f} \max_{1 \leq j \leq n} ((I_{ij} - 1)z_{ij} + x_{jk}) \\ &= \sum_{i=1}^{f} \min_z \max_{1 \leq j \leq n} (x_{jk} + (I_{ij} - 1)z_{ij}) \end{aligned}$$

$$= \sum_{i=1}^{f} \min_{z} \max \left(\max_{j:I_{ij}=1} x_{jk}, \ \max_{j:I_{ij}=0} x_{jk} - z_{ij} \right)$$
$$\geq \sum_{i=1}^{f} \max_{j:I_{ij}=1} x_{jk}. \tag{B3}$$

It can also be shown that $S_+(\boldsymbol{x_k}) \leq \sum_{i=1}^{f} \max_{j:I_{ij}=1} x_{jk}$ by choosing $z_{ij} = 0$ for $I_{ij} = 1, z_{ij} = \max_{h,I_{hj}=0} x_{hj} - \max_{h,I_{hj}=1} x_{hj}$ otherwise. As a result, the expression can be written as

$$S_+(\boldsymbol{x_k}) = \sum_{i=1}^{f} \max_{j:I_{ij}=1} x_{jk}. \tag{B4}$$

Similarly, we can get

$$\begin{aligned} S_-(\boldsymbol{x_k}) &= \min_{U} \boldsymbol{x_k}^T U^T \mathbf{1} : \quad U\mathbf{1} = \mathbf{1}, \ U \geq 0, U \circ (1 - I) = 0 \\ &= -S_+(-\boldsymbol{x_k}) \\ &= \sum_{i=1}^{f} \min_{j:I_{ij}=1} x_{jk}. \end{aligned} \tag{B5}$$

Based on these two expressions, Equation (B1) can be written as

$$\lambda > \max \left(|S_+(\boldsymbol{x_k}) - c_k|, |S_-(\boldsymbol{x_k}) - c_k| \right), \tag{B6}$$

$$\lambda > \max \left(c_k - \sum_{i=1}^{f} \min_{j:I_{ij}=1} x_{jk}, \ \sum_{i=1}^{f} \max_{j:I_{ij}=1} x_{jk} - c_k \right). \tag{B7}$$

If the kth feature satisfies the SAFE condition (B7), then $\beta_k = 0$ at optimum.