JBE Journal of Biostatistics and Epidemiology
J Biostat Epidemiol. 2025;11(1): 63-81

ORIGINAL ARTICLE
DOI: *https://doi.org/10.18502/jbe.v11i1.19318*
*Received **November** 24, 2024; Accepted **January** 06, 2025;*
*Published online **March** 15, 2025.*

Sparse Variable Selection in Competing Risks Additive Hazards Regression: An Application for Identifying Biomarkers Related to Prognosis of Bladder Cancer

Leili **Tapak**[1,2*], Michael R. **Kosorok**[3], Omid **Hamidi**[4*], Mahya **Arayeshgari**[2]

[1]Modeling of Noncommunicable Diseases Research Center, Hamadan University of Medical Sciences, Hamadan, Iran.
[2]Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran.
[3]Department of Biostatistics, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, USA.
[4]Department of Science, Hamedan University of Technology, Hamedan, Iran.

ABSTRACT

Introduction: Variable selection is increasingly becoming a key step in biomedical research, particularly in high-throughput genomic data analysis. One major focus is selecting relevant gene expression profiles associated with time-to-event outcomes, such as death. A significant challenge in this context is competing risks, where identifying a small subset of gene expression profiles related to the cumulative incidence function (CIF) is essential.

Methods: Several methods have been proposed for directly modeling CIF, primarily by modeling the subdistribution hazard function for the event of interest. We proposed a regularized method for variable selection in the additive subdistribution hazards model by integrating five penalized likelihood approaches—Least Absolute Shrinkage and Selection Operator (LASSO), Adaptive LASSO (ALASSO), Elastic Net (ENET), Adaptive Elastic Net (AENET), and Smoothly Clipped Absolute Deviation (SCAD)—with the pseudoscore method. We conducted Monte Carlo simulations to evaluate the performance of our proposed method. Furthermore, the method was applied to a publicly available dataset of 301 patients diagnosed with non-muscle-invasive bladder carcinoma from five countries between 1987 and 2000.

Results: Our proposed method was evaluated through simulation studies and applied to genomic data, focusing on progression-free survival as the endpoint and identifying the genes associated with the CIF of bladder cancer in the presence of competing events. Five genes, namely DCTD, IGF1R, NCF2, PLEK, and CDC20, were consistently identified across different penalties. Notably, the overexpression of DCTD and IGF1R was associated with a decreased cumulative incidence of bladder cancer progression or death. In contrast, the overexpression of NCF2, PLEK, and CDC20 correlated with an increased cumulative incidence of these events.

Conclusion: According to the findings of the simulation studies, all penalties yielded comparable results in terms of sensitivity and specificity. However, the AENET and ALASSO penalties demonstrated superior estimation accuracy.

Key words: Competing risks; Subdistributions; Microarray; Additive hazards model; Variable selection; LASSO

*Corresponding Authors:
l.tapak@umsha.ac.ir &
Omid_hamidi@hut.ac.ir

INTRODUCTION

Over the past decade, significant advancements in molecular biology experimental technologies, such as next-generation sequencing and microarray gene expression, have led to the accumulation of vast amounts of biomedical data. This progress has facilitated the discovery and understanding of molecular mechanisms, biomarker identification, and the development of personalized medicine.[1,2] In particular, there is a growing interest in analyzing high-throughput data to correlate gene expression profiles with the timing of survival outcomes, such as death.[3] However, efficient analysis of such data presents challenges due to its high dimensionality—where the number of covariates significantly exceeds the number of observations—and the complications arising from survival outcomes, such as censoring and truncation.[3,4] The need for appropriate statistical methods to analyze this type of data, particularly high-dimensional right-censored data, where many classical inference techniques may not be applicable, has spurred numerous theoretical and computational advancements. Among these, variable selection has emerged as a crucial technique for identifying a small subset of features that help mitigate overfitting in high-dimensional settings. This approach enhances both the predictive power and the interpretability of the model.[4] A key challenge is the simultaneous selection of variables and estimation, which is effectively addressed using regularized regression models. Regularization works by adding a penalty term to the model's loss function, which not only shrinks the coefficients toward zero but also sets some coefficients exactly to zero.[5] Penalization methods such as Least Absolute Shrinkage and Selection Operator (LASSO),[5] Smoothly Clipped Absolute Deviation (SCAD),[6] and Elastic Net (ENET) (7) are particularly well-suited for handling high-dimensional data, where traditional variable selection techniques encounter substantial challenges.[8]

Regularization techniques for variable selection in high-dimensional time-to-event data have been developed beyond the Cox model,[9,10] including the Lin and Ying additive hazards model[11] as a beneficial alternative. For instance, Lin and Lv introduced a class of regularization methods for simultaneous variable selection and estimation in the additive hazards model by combining the non-concave penalized likelihood approach with the pseudo-score method.[4] Other studies have also employed the pseudo-score estimating function for regularized estimation in the high-dimensional additive hazards model.[12-14] For example, Liu et al. integrated the composite penalty and the pseudoscore in the additive hazards regression model under the high-dimensional framework.[14] The objective function of an additive hazards model offers computationally simpler least-squares estimations than proportional hazards models, which is particularly advantageous in high-dimensional studies where computational cost is a significant concern.[15] Additionally, additive models possess notable characteristics that make them especially relevant in epidemiological and clinical research; they pertain to the risk difference or excess risk measure, providing insightful information for such studies.[4]

This model has been utilized only for single survival endpoints. However, competing risks are a fundamental aspect of medical research, where treatment responses can be classified based on failures due to disease processes or non-disease-related causes. In a competing risk scenario, the occurrence of one type of failure precludes the occurrence of others. A typical approach for analyzing such data

involves cause-specific hazard regression models. While this method is valuable for investigating disease dynamics and gaining insights into disease mechanisms and biological processes, it is less suitable for clinical decision support, where considering cumulative incidence probabilities—reflecting the marginal probability of failure for specific causes—is preferable.[1] Several methods have been proposed for directly modeling cumulative incidence functions (CIF), involving modeling the subdistribution hazard function of the event or cause of interest in a low-dimensional context.[16-21] The analysis of high-dimensional data becomes increasingly complex in the presence of competing risks, as the relevant genes associated with different causes of failure can vary significantly. Despite the importance of this issue, only a limited number of studies have addressed the challenges in analyzing high-dimensional competing risks data. Binder et al. (2009) developed a component-wise likelihood-based boosting algorithm designed for variable selection in high-dimensional competing risks scenarios, directly modeling the proportional subdistribution hazards (PSH) model.[2] Tapak et al. (2015) employed the penalized cause-specific hazards method for analyzing high-dimensional competing risks data.[22] Moreover, Ambrogi and Scheike (2016) proposed a penalized method for competing events using a direct binomial regression model.[1] In addition, Fu et al. (2017)[23] and Kawaguchi et al. (2021)[24] developed penalized variable selection methods for competing risks in the presence of high-dimensional data based on the PSH model. However, a significant gap persists in the literature concerning the modeling and variable selection for high-dimensional competing risks data in the additive subdistribution hazards model. The only relevant attempt to address this issue was a 2016 study conducted by Tapak et al., where a cause-specific penalized additive hazards model was applied.[25] To address this gap, the current study aimed to develop a penalized additive subdistribution hazards model for variable selection capable of handling competing risks in high-dimensional time-to-event data. Specifically, we sought to compare the performance of five widely used penalized variable selection methods: LASSO, Adaptive LASSO (ALASSO), SCAD, ENET, and Adaptive Elastic Net (AENET), and to identify genes associated with the progression or death from bladder cancer, which may serve as potential therapeutic targets.

We detailed the proposed methodology in Sections 2.1 and 2.2. Then, in Section 2.3, we presented simulation studies to evaluate the performance of the approach. Finally, we demonstrated its practical applicability using a publicly available bladder cancer dataset.

MATERIALS AND METHODS

The Regularized Additive Subdistribution Hazards Model

For a sample with k competing risk types, let T_k be the time to the kth type of failure, $T = \min(T_1, ...,T_k)$ be the failure time, and C be the censoring time. Denote the censored failure time by $T^* = (T \wedge C)$ and the failure indicator by $\Delta = I\ (T \leq C)$, where $I\ (\cdot)$ is the indicator function. Let Z be a p-dimensional vector of predictable covariate processes and assume that T and C are conditionally independent given Z. So, the observed data consists of $(T_i^*, \Delta_i \varepsilon_i, Z_i)$, where $\varepsilon_i \in \{1,...,K\}$ indicates the (potentially unobserved) type of the event.[2, 4] Here, the interest is modeling the cumulative incidence

function for failure from cause 1 conditional on the covariates, $F_1(t;Z) = P(T \le t, \varepsilon = 1)$, which is the expected proportion of patients suffering event 1 over time.

The subdistribution hazard function is defined as:

$$\lambda_1(t;Z) = \lim_{\Delta t \to 0} \frac{P\left(t \le T \le t + \Delta t, \varepsilon = 1 \mid T \ge t \cup \left(T \le t \cap \varepsilon \ne 1\right), Z\right)}{\Delta t} = \frac{dF_1(t,Z)/dt}{1 - F_1(t,Z)} \tag{1}$$

In terms of counting process notation, let $N_{i1}(t) = I(T \le t, \Delta_i \varepsilon_i = 1)$ be the number of observed events due to cause 1 and $Y_{i1}(t) = I\left(T_i \ge t \cup \left(T_i \le t \cap \varepsilon_i \ne 1\right)\right)$ be the risk indicator for the *i*th individual specific to cause j = 1. Furthermore, let $r_i(t)$ be the vital status of the *i*th individual, where $r_i(t) = I[C_i \ge (T_i \wedge t)]$ indicates that individual *i* has not been censored by the minimum time between T_i and *t*. In counting process notation, $N_{i1}(t)$ and $Y_{i1}(t)$ can only be computed when $r_i(t)$=1, leading to $r_i(t)N_{i1}(t) = N_{i1}(t)$ and $r_i(t)Y_{i1}(t) = Y_{i1}(t)$. When an individual is censored, ri(t) =0, and the functions Ni1(t) and Yi1(t) cannot be calculated. Nevertheless, it can be shown that $r_i(t)N_{i1}(t)$= 0 and $r_i(t)Y_{i1}(t)$ =0. The risk set for the subdistribution hazards model includes full contributions of individuals who have neither failed nor been censored by time t as well as weighted contributions from individuals who failed prior to *t* with $\varepsilon \notin \{0,1\}$. A time-dependent weight, $w_i(t)$, is defined as the inverse probability of the censoring distribution for right-censored data, $w_i(t) = r_i(t) \times \left[\hat{G}(t)/\hat{G}(T_i^*, t)\right]$, where $\hat{G}(.)$ is the Kaplan-Meier estimate of the survival function of the censoring distribution using $1 - \Delta_i$.[26] Therefore, the contribution to the risk set at time t for individual i is given by

$$r_i(t)Y_{i1}(t) = I\left\{C_i \ge (T_i \wedge t) \cap \left[\left(T_i^* \ge t\right) \cup \left(\varepsilon_i \notin \{0,1\}\right)\right]\right\} \times \frac{\hat{G}(t)}{\hat{G}(T_i^*, t)}. \tag{2}$$

We adapt the Lin and Ying additive hazards model for subdistribution hazards. Then, the hazard function of a failure time *T* conditional on a p-vector of possibly time-dependent covariates *Z* is specified as:

$$\lambda_1(t;Z) = \lambda_{01}(t) + \beta_0^T Z \tag{3}$$

where $\lambda_0(.)$ is an unspecified baseline hazard function that is shared among all subjects and β_0 is a p-vector of regression coefficients (4).

So, the counting process martingale is defined as:

$$M_i(t) = N_i(t) - \int_0 \mathrm{w_i(s)Y_i\ (s)}\left\{\lambda_{01}(s) + \beta_0^T Z\right\} ds. \tag{4}$$

The pseudo score linear in the β function of the model can be defined as:

$$\mathrm{U}(\beta) = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \int_0^{\tau} \{\mathrm{Z_i} - \bar{Z}\} \left\{\mathrm{d}N_i(t) - w_i(\mathrm{t})\mathrm{Y_i\ (t)Z_i}dt\right\} \tag{5}$$

where $\bar{Z}=\sum_{j=1}^{n} w_i(t)Y_j(t)Z_j / \sum_{j=1}^{n} w_i(t)Y_j(t)$ and τ is the maximum follow-up time. After some algebraic manipulation, it can be written as follows:

$$U(\beta)=b-V\beta \tag{6}$$

where $b=\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{Z_i-\bar{Z}\}dN_i(t)$ and $V=\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} w_i(t)Y_i(t)\left\{(Z_i-\bar{Z})(Z_i-\bar{Z})^T\right\}dt$ which is positive semi-definite. Integrating - $U(\beta)$ with respect to β leads to the least squares type loss function $L(\beta)=\frac{1}{2}\beta^T V\beta-b^T\beta$.[4]

Then, the penalized estimator $\hat{\beta}$ is a solution to the regularization problem:

$$\hat{\beta}=\arg\min_{\beta\in\mathbb{R}^p}\left\{Q(\beta)\equiv L(\beta)+\sum_{j=1}^{p} p_\lambda(|\beta_j|)\right\} \tag{7}$$

where $p_\lambda(\theta),\theta\geq 0,$ is a penalty function that depends on the regularization parameter $\lambda\geq 0$ and often rewrites as $p_\lambda(\cdot)=\lambda\rho(\cdot)$.[4]

In this study, we considered five commonly used sparse penalty functions, as listed below.

- The LASSO uses the L1-penalty, i.e. $\rho(\theta)=\theta;\theta\geq 0$.[5]
- The ALASSO is obtained by applying the LASSO with the assumption of $\theta=\hat{w}_j|\beta_j|$, where $\hat{w}_j$ represents the weight of the jth variable, which can be computed by $\hat{w}_j=|\hat{\beta}_j^{ini}|^{-\gamma}$. γ denotes a positive constant, and $\hat{\beta}^{ini}$ comprises a set of initial parameters that can be estimated using ordinary least squares (OLS) or ridge regression.[27]
- The ENET combines the L1-penalty $\rho(\theta)=\theta$ and the L2-penalty $\rho(\theta)=\theta^2$, yielding a penalty in the form of $\rho(\theta)=(1-a)\theta+a\theta^2;0<a<1$.[7]
- The AENET is a combination of the ENET and the ALASSO.[28]
- The SCAD is defined by the derivative $\rho'_\lambda(\theta)=I(\theta\leq\lambda)+\frac{(a\lambda-\theta)_+}{(a-1)\lambda}I(\theta>\lambda),\ \theta\geq 0$, with some a>2 as a shape parameter.[6]

Estimation of $\hat{\beta}$ is accomplished through the coordinate descent algorithm (4).

Choosing the Tuning Parameter

After generating a solution path, the optimal regularization parameter λ is selected using a cross-validation score obtained through *M*-fold cross-validation. The cross-validation score is defined as follows:

$$CV(\lambda)=\frac{1}{M}\sum_{m=1}^{M} L^{(m)}\left(\hat{\beta}^{(-m)}(\lambda)\right), \tag{8}$$

where $L^{(m)}(.)$ represents the least squares type loss function computed from the *m*th subset of the data, and $\hat{\beta}^{(-m)}(\lambda)$ denotes the estimate derived from the data with the *m*th subset removed.[4] To

determine the optimal value of the regularization parameter, λ, this study employed 10-fold cross-validation. SCAD and ENET penalties include an additional parameter a that needs to be tuned. In this study, the method described in (4) was used to determine the value of a for ENET.

Simulation Studies

In this subsection, we conducted Monte Carlo simulations to evaluate the performance of our proposed method. Competing risk data with two possible events were simulated: the event of interest (I) and the competing event (C). Specifically, we considered two different sample sizes n{200, 400}.

The performance of the method was assessed through the following simulation scenarios with p{20, 2000} :

Case 1: Only predictors 1 and 10 among the p covariates were informative;

Case 2: Predictors 1, 5, 10, and 15 among the p covariates were informative.

For the informative covariates, we assigned values of 2 and -2 to represent increasing and decreasing effects, respectively. Covariates with no direct effect on the hazards were assigned a value of 0. In all scenarios, covariates were generated from N(0,1), both independently and in correlation (in the correlated case, the correlation between x_i and x_j was defined as $0.5^{|i-j|}$). Additionally, we considered coefficient values of 0.5 and -0.5 for the informative variables to assess the performance of the proposed method for smaller signals.

Following the methodology outlined by Beyersmann et al.,[29] we generated event times based on proportional subdistribution hazards. After simulating the covariates, we defined f as the ratio of the event of interest (I) to the competing risk (C). The subdistribution for the event I was generated as follows:

$$\Pr(T_i \le t, \varepsilon_i = 1 \mid \boldsymbol{x}_i) = 1 - \left[1 - f\left\{1 - \exp(-t)\right\}\right]^{\exp\left(\sum_{i=1}^{p} \beta_{i1} \boldsymbol{x}_i\right)} \tag{9}$$

which is a unit exponential mixture with mass 1-f at ∞ when all covariates are zero. Here, we considered f as {0.2, 0.5, 0.8}. The subdistribution for the second event type was generated using an exponential distribution with a rate of $\exp(\sum_{i=1}^{p} \beta_i x_i)$ by taking $\Pr(\varepsilon_i = 2 \mid x_i) = 1 - \Pr(\varepsilon_i = 1 \mid x_i)$.

Censoring times were generated using a uniform distribution $U(0, a)$, with the value of a selected to achieve censoring for approximately 35% of the observations. Given that the estimated coefficients were biased toward zero, we focused on the probability of selecting relevant covariates.

To evaluate model selection consistency, we assessed the performance of different variable selection methods by calculating the rate of correctly selected non-zero (informative) coefficients (sensitivity) and the rate of correctly non-selected zero (non-informative) coefficients (specificity). These metrics indicate the effectiveness of the methods in identifying important variables and shrinking unimportant variables to zero, thus serving as criteria for model selection consistency. A total of one hundred

replications were conducted for these simulations.

To compare the estimation accuracy of the different methods, we used the L_2-loss ($\left\|\hat{\beta}-\beta\right\|_2$) and L_1-loss ($\left\|\hat{\beta}-\beta\right\|_1$) as described by Lin and LV (4).

Bladder Cancer Data

In this study, a publicly available dataset (GEO with series accession no. GSE5479) was used to illustrate the applicability of the proposed model. The dataset comprised complete information on 301 patients diagnosed with non-muscle-invasive bladder carcinoma who underwent surgery at hospitals across Denmark, Sweden, Spain, France, and England between 1987 and 2000. This dataset includes 1381 measurements of gene expression along with five clinical covariates: age, sex, BCG/MMC treatment, grade, and the pathological stage of the disease.[30]

RESULTS

Simulation Study Results

We reported the results of our simulation studies for low-dimensional settings with s = 2 and s = 4 for n = 200 and n = 400 under independent and correlated structures for covariates in Tables 1 and 2, respectively. The standard deviations of the number of selected variables ranged from a minimum of 0.62 to a maximum of 3.51 in both tables. Overall, all methods demonstrated high sensitivity and acceptable specificity, with no significant performance differences noted between the two sample sizes. The results indicated that the variable selection methods performed similarly among all scenarios presented in Tables 1 and 2.

In a further analysis, we examined the same settings with the absolute values of the true coefficients set to 0.5 for non-zero coefficients. Table 3 summarizes the results for p = 20 under this condition. Compared to the previous settings, both sensitivities and specificities decreased; however, they remained in acceptable ranges. Additionally, AENET achieved higher specificity and sensitivity in most cases. Another notable observation is that LASSO and ENET tended to select more variables than the other methods, leading to lower specificity in the majority of scenarios, particularly for LASSO. Furthermore, in terms of sensitivity, SCAD exhibited lower values in half of the cases, especially when f = 0.8 and the sample size was smaller (n = 200).

We also investigated the performance of different penalization techniques in a high-dimensional setting, with findings summarized in Table 4. In this setting, we reported sensitivity and specificity metrics for the variable selection methods. Notably, SCAD exhibited superior performance in both sensitivity and specificity in high-dimensional simulations in approximately half of the scenarios, which may be attributed to the concavity of its penalty function. Nevertheless, all methods performed well in specificity across all scenarios. Moreover, LASSO performed the poorest in sensitivity in about half of the cases. An additional noteworthy aspect is that, for f = 0.2, LASSO, ALASSO, ENET, and AENET tended to select more variables than SCAD. However, for f = 0.5 and 0.8, the

number of selected variables was comparable between the methods.

Table 1. Results of various penalized methods in simulation studies under the independent covariates scenario (p = 20), with sample sizes of n = 200 and n = 400, number of informative variables (s) equal to 2 and 4, an effect size of 2, and a censoring rate of approximately 35%, across 100 replicates, in terms of true positive (TP; Sensitivity) and true negative (TN; Specificity).

		n = 200						n = 400					
		s = 2			s = 4			s = 2			s = 4		
f = I/C	Method	v	Sensitivity (TP)	Specificity (TN)	v	Sensitivity (TP)	Specificity (TN)	v	Sensitivity (TP)	Specificity (TN)	v	Sensitivity (TP)	Specificity (TN)
0.2	LASSO	2.30	1.00	0.983	4.45	1.00	0.972	2.37	1.00	0.979	4.23	1.00	0.986
	ALASSO	2.25	1.00	0.986	4.27	1.00	0.983	2.07	1.00	0.996	4.48	1.00	0.970
	SCAD	2.03	1.00	0.998	4.00	1.00	1.00	2.02	1.00	0.999	4.05	1.00	0.997
	ENET	2.22	1.00	0.988	4.22	1.00	0.986	2.13	1.00	0.993	4.14	1.00	0.991
	AENET	2.10	1.00	0.996	4.20	1.00	0.988	2.12	1.00	0.993	4.10	1.00	0.994
0.5	LASSO	2.21	1.00	0.988	4.45	1.00	0.972	2.37	1.00	0.979	4.23	1.00	0.986
	ALASSO	2.19	1.00	0.989	4.26	1.00	0.984	2.11	1.00	0.993	4.23	1.00	0.986
	SCAD	2.00	1.00	1.00	4.02	1.00	0.999	2.00	1.00	1.00	4.00	1.00	1.00
	ENET	2.04	1.00	0.998	4.11	1.00	0.993	2.15	1.00	0.992	4.24	1.00	0.985
	AENET	2.04	1.00	0.998	4.08	1.00	0.995	2.13	1.00	0.993	4.12	1.00	0.992
0.8	LASSO	2.21	1.00	0.988	4.26	1.00	0.984	2.10	1.00	0.994	4.19	1.00	0.988
	ALASSO	2.18	1.00	0.990	4.19	1.00	0.988	2.10	1.00	0.994	4.23	1.00	0.986
	SCAD	2.07	1.00	0.996	4.02	1.00	0.999	2.01	1.00	0.999	4.02	1.00	0.999
	ENET	2.05	1.00	0.997	4.11	1.00	0.993	2.02	1.00	0.999	4.19	1.00	0.988
	AENET	2.05	1.00	0.997	4.09	1.00	0.994	2.02	1.00	0.999	4.07	1.00	0.995

v, The average number of selected variables; ALASSO, Adaptive least absolute shrinkage and selection operator; SCAD, Smoothly clipped absolute deviation; AENET, Adaptive elastic net

Table 2. Results of various penalized methods in simulation studies under the correlated covariates scenario (p = 20), with sample sizes of n = 200 and n = 400, number of informative variables (s) equal to 2 and 4, an effect size of 2, and a censoring rate of approximately 35%, across 100 replicates, in terms of true positive (TP; Sensitivity) and true negative (TN; Specificity).

		n = 200						n = 400					
		s = 2			s = 4			s = 2			s = 4		
f = I/C	Method	v	Sensitivity (TP)	Specificity (TN)	v	Sensitivity (TP)	Specificity (TN)	v	Sensitivity (TP)	Specificity (TN)	v	Sensitivity (TP)	Specificity (TN)
0.2	LASSO	2.53	1.00	0.970	4.60	1.00	0.962	2.49	1.00	0.973	4.92	1.00	0.942
	ALASSO	2.29	1.00	0.984	4.48	1.00	0.970	2.24	1.00	0.987	4.40	1.00	0.975
	SCAD	2.03	1.00	0.998	4.02	1.00	1.00	2.05	1.00	0.997	4.06	1.00	0.996
	ENET	2.10	1.00	0.994	4.40	1.00	0.975	2.17	1.00	0.990	4.46	1.00	0.971
	AENET	2.07	1.00	0.996	4.12	1.00	0.992	2.13	1.00	0.993	4.12	1.00	0.992
0.5	LASSO	2.30	1.00	0.983	4.58	1.00	0.964	2.32	1.00	0.982	4.64	1.00	0.960
	ALASSO	2.22	1.00	0.988	4.32	1.00	0.980	2.11	1.00	0.994	4.40	1.00	0.975
	SCAD	2.03	1.00	0.988	4.06	1.00	0.996	2.01	1.00	0.999	4.02	1.00	0.999
	ENET	2.03	1.00	0.998	4.32	1.00	0.980	2.02	1.00	0.999	4.56	1.00	0.965
	AENET	2.04	1.00	0.998	4.30	1.00	0.981	2.07	1.00	0.996	4.06	1.00	0.996
0.8	LASSO	2.23	1.00	0.987	4.54	1.00	0.966	2.12	1.00	0.993	4.40	1.00	0.975
	ALASSO	2.15	1.00	0.992	4.42	1.00	0.974	2.20	1.00	0.989	4.32	1.00	0.980
	SCAD	2.01	1.00	0.999	4.02	1.00	1.00	2.01	1.00	0.999	4.00	1.00	1.00
	ENET	2.03	1.00	0.998	4.62	1.00	0.961	2.06	1.00	0.997	4.34	1.00	0.979
	AENET	2.02	1.00	0.999	4.10	1.00	0.994	2.04	1.00	0.998	4.18	1.00	0.989

v, The average number of selected variables; ALASSO, Adaptive Least absolute shrinkage and selection operator; SCAD, Smoothly clipped absolute deviation; AENET, Adaptive elastic net

Table 3. Results of various penalized methods in simulation studies under the independent and correlated covariates scenario (p = 20), with sample sizes of n = 200 and n = 400, number of informative variables (s) equal to 4, an effect size of 0.5, and a censoring rate of approximately 35%, across 100 replicates, in terms of true positive (TP; Sensitivity) and true negative (TN; Specificity).

f = I/C	Method	n = 200						n = 400					
		iid covariates			correlated covariates			iid covariates			correlated covariates		
		v	Sensitivity (TP)	Specificity (TN)	v	Sensitivity (TP)	Specificity (TN)	v	Sensitivity (TP)	Specificity (TN)	v	Sensitivity (TP)	Specificity (TN)
0.2	LASSO	6.78	0.965	0.818	7.46	0.910	0.761	6.64	0.990	0.833	8.40	1.000	0.725
	ALASSO	6.30	0.960	0.846	5.94	0.930	0.861	7.18	1.000	0.801	6.72	1.000	0.830
	SCAD	5.80	0.895	0.861	6.66	0.895	0.808	5.18	1.000	0.926	7.20	0.990	0.798
	ENET	6.34	0.965	0.845	7.42	0.915	0.765	7.02	1.000	0.811	7.96	0.990	0.750
	AENET	5.78	0.965	0.880	5.82	0.855	0.850	5.34	1.000	0.916	6.14	0.995	0.865
0.5	LASSO	5.78	0.970	0.881	6.94	0.965	0.808	5.68	0.995	0.894	6.98	0.995	0.813
	ALASSO	5.44	0.945	0.904	5.86	0.925	0.878	5.94	0.995	0.879	5.38	0.990	0.914
	SCAD	5.04	0.975	0.921	5.72	0.965	0.874	4.38	1.000	0.975	4.88	1.000	0.943
	ENET	6.10	0.970	0.863	6.66	0.920	0.825	5.36	1.000	0.915	7.56	0.995	0.778
	AENET	4.90	0.965	0.936	5.00	0.925	0.918	4.68	1.000	0.959	5.26	0.995	0.920
0.8	LASSO	4.74	0.935	0.945	6.46	0.960	0.828	5.50	1.000	0.906	6.80	1.000	0.824
	ALASSO	4.54	0.935	0.950	4.74	0.960	0.944	5.06	1.000	0.934	5.24	1.000	0.923
	SCAD	4.54	0.900	0.941	4.74	0.815	0.908	4.28	0.985	0.979	5.38	0.985	0.910
	ENET	5.10	0.965	0.923	6.32	0.925	0.836	4.86	1.000	0.946	6.78	0.995	0.825
	AENET	4.54	0.980	0.961	4.86	0.875	0.915	4.50	1.000	0.969	5.44	0.985	0.906

v, The average number of selected variables; ALASSO, Adaptive Least absolute shrinkage and selection operator; SCAD, Smoothly clipped absolute deviation; AENET, Adaptive elastic net

Table 4. Results of various penalized methods in simulation studies under the independent and correlated covariates scenario (p = 2000), with sample sizes of n = 200 and n = 400, number of informative variables (s) equal to 4, an effect size of 0.5, and a censoring rate of approximately 35%, across 100 replicates, in terms of true positive (TP; Sensitivity) and true negative (TN; Specificity).

f = I/C	Method	n = 200						n = 400					
		iid covariates			correlated covariates			iid covariates			correlated covariates		
		v	Sensitivity (TP)	Specificity (TN)	v	Sensitivity (TP)	Specificity (TN)	v	Sensitivity (TP)	Specificity (TN)	v	Sensitivity (TP)	Specificity (TN)
0.2	LASSO	7.42	0.745	0.998	9.62	0.763	0.997	8.76	0.945	0.998	12.86	0.898	0.995
	ALASSO	7.95	0.850	0.998	8.32	0.795	0.997	7.40	0.965	0.998	10.24	0.985	0.997
	SCAD	3.95	0.988	1.000	4.80	0.575	0.999	6.32	0.920	0.999	5.22	0.970	0.999
	ENET	8.28	0.715	0.997	8.44	0.790	0.997	7.07	0.981	0.998	9.30	0.988	0.997
	AENET	7.90	0.790	0.998	7.44	0.783	0.998	6.92	0.971	0.998	9.45	0.975	0.997
0.5	LASSO	4.44	0.710	0.999	6.98	0.747	0.998	4.90	0.910	0.999	4.58	0.968	0.999
	ALASSO	3.62	0.680	1.000	6.61	0.975	0.998	4.22	0.950	1.00	5.28	0.955	0.999
	SCAD	4.00	1.000	1.000	5.10	0.830	0.999	4.00	0.975	1.00	4.14	1.000	1.000
	ENET	4.57	0.858	0.999	5.64	0.898	0.999	4.18	0.972	1.00	6.00	0.988	0.999
	AENET	5.67	0.875	0.999	5.35	0.775	0.999	4.10	0.958	1.00	5.25	0.913	0.999
0.8	LASSO	5.73	0.868	0.999	5.43	0.750	0.998	3.52	0.784	1.00	4.85	0.985	0.995
	ALASSO	4.93	0.905	0.999	5.29	0.813	0.999	3.58	0.855	1.00	4.85	0.930	0.999
	SCAD	4.00	1.000	1.000	4.33	0.958	1.000	4.22	1.00	0.999	4.10	0.960	1.000
	ENET	3.00	0.612	1.000	5.50	0.813	0.999	4.35	1.00	0.999	4.38	0.990	1.000
	AENET	5.14	0.785	0.999	5.00	0.918	0.999	4.42	1.00	0.999	5.37	1.000	0.999

v, The average number of selected variables; ALASSO, Adaptive Least absolute shrinkage and selection operator; SCAD, Smoothly clipped absolute deviation; AENET, Adaptive elastic net

Table 5 presents the estimation accuracy of various methods in our simulation studies, focusing on three specific cases under the assumption of s=4, p=20, and an effect size of 0.5. These cases include independent covariates with two sample sizes (n=200 and n=400) and correlated covariates with n=400. ALASSO and AENET produced coefficient estimates closer to the Oracle estimator than those obtained from LASSO, ENET, and SCAD in all scenarios.

Table 5. Estimation accuracy results for various penalized methods in simulation studies under the independent and correlated covariates scenario (p = 20), with sample sizes of n = 200 and n = 400, number of informative variables (s) equal to 4, an effect size of 0.5, and a censoring rate of approximately 35%, across 100 replicates, in terms of L_2-loss ($\|\hat{\beta}-\beta\|_2$) and L_1-loss ($\|\hat{\beta}-\beta\|_1$).

f = I/C	Method	n = 200 Independent covariates				n = 400 Independent covariates				n = 400 Correlated covariates			
		L_2-loss	sd	L_1-loss	sd	L_2-loss	sd	L_1-loss	sd	L_2-loss	sd	L_1-loss	sd
0.2	LASSO	0.955	0.016	1.927	0.026	0.950	0.008	1.916	0.021	0.953	0.010	1.918	0.019
	ALASSO	0.934	0.014	1.905	0.041	0.937	0.013	1.904	0.029	0.936	0.009	1.904	0.031
	SCAD	0.956	0.018	1.946	0.050	0.953	0.010	1.913	0.030	0.951	0.014	1.917	0.034
	ENET	0.960	0.012	1.936	0.028	0.954	0.013	1.920	0.020	0.945	0.009	1.919	0.019
	AENET	0.933	0.015	1.895	0.048	0.936	0.013	1.895	0.026	0.939	0.012	1.989	0.027
	Oracle	0.926	0.020	1.851	0.039	0.933	0.011	1.869	0.023	0.934	0.012	1.869	0.023
0.5	LASSO	0.942	0.019	1.895	0.040	0.936	0.015	1.890	0.024	0.937	0.013	1.889	0.024
	ALASSO	0.906	0.025	1.849	0.043	0.918	0.012	1.858	0.029	0.919	0.013	1.856	0.031
	SCAD	0.952	0.023	1.912	0.049	0.943	0.016	1.885	0.034	0.941	0.018	1.887	0.037
	ENET	0.953	0.016	1.905	0.037	0.946	0.013	1.894	0.026	0.944	0.014	1.889	0.028
	AENET	0.916	0.018	1.844	0.047	0.922	0.014	1.853	0.034	0.924	0.017	1.854	0.031
	Oracle	0.901	0.021	1.802	0.042	0.914	0.014	1.828	0.028	0.912	0.014	1.825	0.028
0.8	LASSO	0.950	0.024	1.912	0.045	0.944	0.020	1.895	0.035	0.945	0.013	1.896	0.028
	ALASSO	0.913	0.030	1.842	0.055	0.919	0.016	1.849	0.032	0.914	0.019	1.857	0.032
	SCAD	0.951	0.033	1.923	0.055	0.948	0.020	1.891	0.043	0.946	0.023	1.897	0.043
	ENET	0.956	0.025	1.917	0.043	0.945	0.018	1.899	0.043	0.946	0.016	1.899	0.034
	AENET	0.911	0.032	1.839	0.059	0.923	0.014	1.842	0.039	0.918	0.017	1.846	0.035
	Oracle	0.896	0.030	1.792	0.061	0.912	0.020	1.823	0.040	0.914	0.018	1.827	0.036

ALASSO, Adaptive least absolute shrinkage and selection operator; SCAD, Smoothly clipped absolute deviation; AENET, Adaptive elastic net

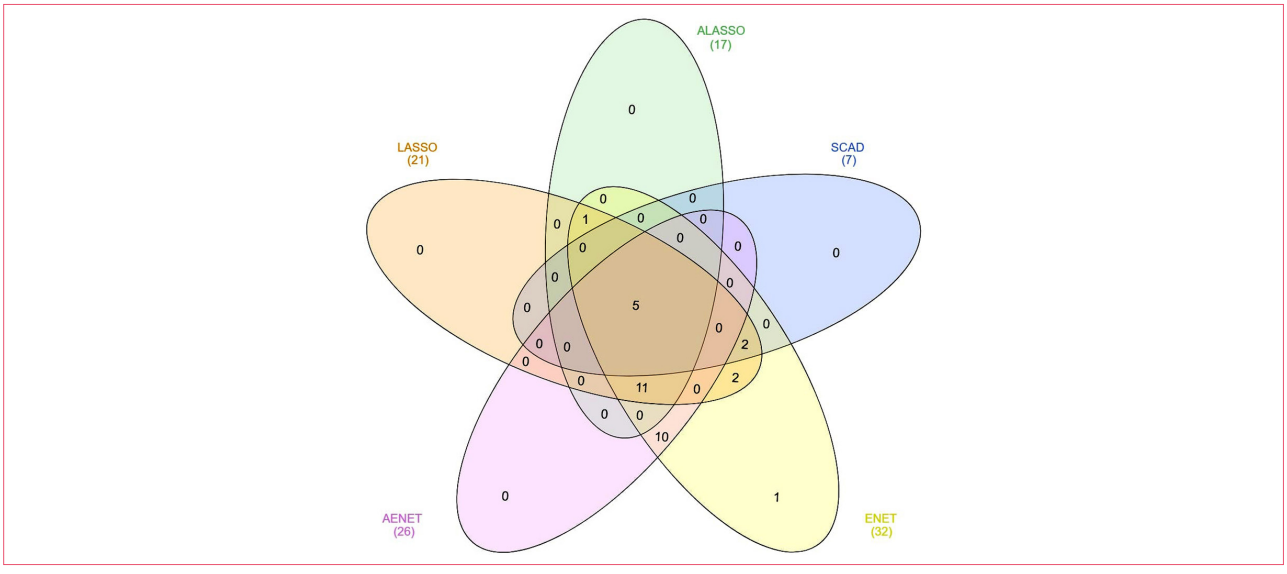

Figure 1. Number of shared genes identified by different penalized methods for predicting bladder cancer progression or death
ALASSO, Adaptive least absolute shrinkage and selection operator; SCAD, Smoothly clipped absolute deviation; AENET, Adaptive elastic net

Bladder Cancer Data Results

In our analysis of bladder cancer data, we identified two competing events: time to progression or death from bladder cancer (the event of interest observed for 74 patients) and death from other or unknown causes (observed for 33 patients). The survival times of an additional 194 patients were censored. A 10-fold cross-validation was utilized to choose the optimal tuning parameters. We repeated the variable selection process 100 times. Table 6 lists the probes consistently selected in all 100 iterations. Interestingly, four probes overlapped with those identified by Dyrskjøt et al.30 Figure 1 illustrates the overlap among probes selected by different penalization methods. Five genes—CDC20, PLEK, NCF2, IGF1R, and DCTD—were repeatedly detected through various methods. Table 7 presents the results of fitting the Lin and Ying additive subdistribution hazards model for these five shared genes using two different modeling approaches (univariate and multivariate).

Table 6. Selected genes identified by different penalized methods for predicting bladder cancer progression or death

Gene ID	GenBank accession No.	Symbol	LASSO	ALASSO	SCAD	ENET	AENET
SEQ1014	NM_002447.1	MST1R	✓		✓	✓	
SEQ1036	NM_012164.2	FBXW2				✓	✓
SEQ1037	NM_005127.2	CLEC2B	✓	✓		✓	✓
SEQ1082	NM_207521.1	RTN4	✓	✓		✓	✓
SEQ1126	-	-	✓			✓	
SEQ1197	NM_003103.5	SON	✓	✓		✓	
SEQ1226	NM_001921.1	DCTD	✓	✓	✓	✓	✓
SEQ1259	NM_014216.3	ITPK1				✓	✓
SEQ1262	NM_000875.2	IGF1R	✓	✓	✓	✓	✓
SEQ1298	NM_000961.2	PTGIS	✓	✓		✓	✓
SEQ1330	NM_003094.1	SNRPE	✓	✓		✓	✓
SEQ1337	NM_170744.2	UNC5B	✓	✓		✓	✓
SEQ139	NM_007002	ADRM1				✓	✓
SEQ162	XM_088569	PTGR1	✓			✓	
SEQ188	AL117536	NA				✓	✓
SEQ227	NM_007008	RTN4	✓	✓		✓	✓
SEQ260	NM_016442	ERAP1				✓	✓
SEQ288	NM_022126	LHPP				✓	✓
SEQ312	NM_058242	KRT6C				✓	✓
SEQ34	NM_000433	NCF2	✓	✓	✓	✓	✓
SEQ347	NM_001129	AEBP1	✓	✓		✓	✓
SEQ377	NM_002664	PLEK	✓	✓	✓	✓	✓
SEQ399	NM_004663	RAB11A	✓	✓		✓	✓
SEQ494	M87507	IL1BCE				✓	✓
SEQ567	NM_004046	ATP5F1A				✓	
SEQ634	NM_004453	ETFDH	✓	✓		✓	✓
SEQ696	NM_000067	CA2				✓	✓
SEQ813	NM_002206.1	ITGA7	✓	✓		✓	✓
SEQ820	NM_005916	MCM7	✓		✓	✓	
SEQ833	NM_001255.1	CDC20	✓	✓	✓	✓	✓
SEQ919	NM_024665.2	IRA1				✓	✓
SEQ940	NM_020159.1	SMARCAD1	✓	✓		✓	✓
			21	17	7	32	26

ALASSO, Adaptive least absolute shrinkage and selection operator; SCAD, Smoothly clipped absolute deviation; AENET, Adaptive elastic net

Table 7. Results of fitting the Lin and Ying additive subdistribution hazards model using five shared genes identified by different penalized methods for predicting bladder cancer progression or death

Gene	Univariate approach			Multivariate approach		
	Coefficient	Standard Error	P-value	Coefficient	Standard Error	P-value
DCTD	-0.0044	0.0007	<0.001	-0.002	0.0008	0.007
IGF1R	-0.0044	0.0008	<0.001	-0.003	0.0008	<0.001
NCF2	0.0037	0.0007	<0.001	0.002	0.0008	0.014
PLEK	0.0033	0.0007	<0.001	0.0008	0.0007	0.275
CDC20	0.0024	0.0003	<0.001	0.0017	0.0003	<0.001

DISCUSSION

In this paper, we evaluated the performance of five commonly used penalized variable selection methods in identifying important variables associated with the CIF in the additive subdistribution hazards model for competing risks.

Several studies have investigated the efficiency of penalized methods in the context of additive and proportional hazards models. In a study by Martinusse and Scheike, different methods were used for variable selection in the additive hazards model when analyzing survival data in both high- and low-dimensional settings. Although the Dantzig selector was the most effective in selecting the correct models, LASSO and ALASSO achieved lower MSE. Furthermore, as the sample size increased from 100 to 200 and 400, ALASSO demonstrated superior performance compared to LASSO.[12] However, based on our results, the advantage of ALASSO over LASSO in estimation accuracy was more pronounced with a smaller sample size (n = 200) than a larger sample size (n = 400). Lin and Lv conducted simulation studies to evaluate the performance of various penalized methods such as LASSO, ENET, and SCAD in a high-dimensional, low-sample-size setting using the additive hazards model. The results indicated that SCAD outperformed LASSO and ENET.[4] However, in our study, conducted in the presence of competing risks, these methods achieved comparable estimation accuracy. Wang et al. investigated the use of AENET for variable selection in the proportional odds model and compared its performance against LASSO, ALASSO, and ENET. The simulation findings suggested that, in most cases, AENET delivered better performance than the others in terms of variable selection accuracy and mean squared error (MSE).[31] In the current study, the estimation accuracy of AENET and ALASSO was similar and superior to that of the other methods. Bradic et al. evaluated the performance of penalized methods in the Cox proportional hazards model through simulation studies. According to the results, LASSO underperformed to SCAD in high-dimensional settings with 100 observations and either 1000 or 5000 predictors. Specifically, LASSO detected fewer true positives, produced more false positives, and demonstrated a higher median prediction error (MPE) than SCAD.[32] In the present study, in a high-dimensional setting with 2000 predictors and 200 or 400 observations, SCAD achieved the highest sensitivity in half of the scenarios, whereas LASSO showed the weakest in a similar proportion.

We also compared our findings with studies that focused on competing risks. For example, Fu

et al. introduced a penalized approach in the proportional subdistribution hazards (PSH) model to identify key predictors for the CIF. Among the penalty techniques explored were LASSO, ALASSO, and SCAD. The simulation findings revealed that, for sample size n = 200, ALASSO provided superior performance in terms of the Median of Mean Squared Error (MMSE). However, for n = 400, both ALASSO and SCAD exhibited comparable results, outperforming LASSO.[23] Hou et al. applied variable selection methods, including LASSO, ALASSO, and boosting, in the proportional cause-specific hazards model and the PSH model for high-dimensional data. A set of comprehensive simulation studies was designed and conducted to evaluate the performance of these models. Although ALASSO outperformed LASSO in some scenarios, it demonstrated weaker estimation accuracy compared to the boosting method.[33]

In this study, all the selected genes significantly affected the subdistribution hazard, thereby influencing the CIF of death due to bladder cancer in an unadjusted setting. Moreover, all genes except for PLEK were significant in an adjusted setting. The overexpression of DCTD and IGF1R genes was significantly associated with a decreased cumulative incidence of progression or death from non-muscle-invasive bladder cancer. In contrast, the overexpression of NCF2, PLEK, and CDC20 genes showed a significant association with an increased risk of progression or death from this cancer.

A few studies have shown a link between elevated expression of the cell division cycle 20 homolog (CDC20) and poor prognosis in bladder cancer. For instance, Shen et al. found that the high expression of CDC20 was significantly associated with poor overall survival, suggesting its role in bladder cancer mortality.[34] Likewise, Liu et al. supported the idea that excessive CDC20 expression accelerates tumor progression in bladder cancer.[35] Other investigations showed that CDC20 expression was notably higher in bladder cancer tissues compared to normal bladder tissues.[36, 37] Similarly, our results suggested that the overexpression of CDC20 increased the risk of progression or death in patients with bladder cancer. The impact of pleckstrin (PLEK) on bladder cancer has been reported in only a limited number of studies. A study by Zhu et al. indicated that the PLEK gene was upregulated in muscle-invasive bladder cancer tissues compared to non-muscle-invasive counterparts.[38] Similar findings have been observed in other cancers as well. For instance, Yan et al. identified differentially expressed genes (DEGs) between gastric cancer and normal gastric samples by analyzing three expression profiles. They identified 85 upregulated genes, including PLEK.[39] Furthermore, Vuong et al. highlighted four potential cancer genes, including PLEK, whose expression levels were associated with poorer overall survival rates in patients with melanoma, lung cancer, or colorectal cancer.[40] Our findings indicated that heightened PLEK expression was associated with a greater risk of disease progression or death in patients with bladder cancer. Neutrophil cytosolic factor 2 (NCF2) has been rarely reported to have prognostic value in bladder cancer. Recently, Ke et al. proposed a gene screening method for PSH regression and applied it to non-muscle-invasive bladder carcinoma datasets. In their sensitivity analysis, both LASSO and PSH-CSIS+LASSO models selected the same five genes as the CoxBoost model. Among these, NCF2 was identified as a risk gene.[41] Xie et al. also reported that NCF2 expression was higher in advanced bladder cancer tissues compared to those in early-stage bladder cancer.[42] Our findings also demonstrated that NCF2 overexpression was

associated with an increased risk of progression or death in bladder cancer patients. Insulin and insulin-like growth factors, including IGF1R, are key regulators of energy metabolism and growth.[43] Previous studies have reported the overexpression of IGF1R in muscle-invasive bladder cancer, highlighting its association with tumor outcomes.[44, 45] However, our findings revealed that in patients with non-muscle-invasive bladder cancer, the overexpression of IGF1R was related to a decreased cumulative incidence of progression or death from non-muscle-invasive bladder cancer. This finding aligns with a study by Faraj et al., which demonstrated that higher levels of IGF1R expression were associated with a favorable tumor recurrence-free survival [OR: 0.58, p=0.021] in patients diagnosed with non-muscle-invasive bladder cancer.[46] Therefore, further research is required to clarify the role of IGF1R in this population. Similarly, the overexpression of the DCTD gene (dCMP deaminase) has been associated with shorter survival rates in various cancers, including malignant glioma.[47] In a study investigating the prognostic role of metabolic genes in cancer immunotherapy, Ou et al. identified an upregulation of DCTD in bladder cancer cells.48 In contrast, our study found that DCTD overexpression in patients with non-muscle-invasive bladder cancer was associated with a decreased cumulative incidence of progression or death from bladder cancer. This discrepancy highlights the need for further research to understand better the role of DCTD in bladder cancer.

Our study has some limitations. First, to our knowledge, no prior research has compared all five penalty functions examined in our study within the additive hazards model or in the context of competing risks. Consequently, our findings could not be comprehensively compared to previous studies. Second, we included only continuous variables in the simulation and did not account for discrete variables, which may limit the generalizability of our findings to real-world scenarios where both continuous and discrete variables exist. Third, we focused solely on moderate effect sizes; therefore, the performance of our methods in scenarios with smaller effect sizes remained uncertain.

CONCLUSION

The primary objective of this study was to investigate variable selection methods for low- and high-dimensional competing risk data based on the additive subdistribution hazards model. We evaluated five popular penalized variable selection methods: LASSO, ALASSO, SCAD, ENET, and AENET. Our Monte Carlo simulation results indicated that while all penalty functions exhibited comparable sensitivity and specificity, those based on AENET and ALASSO penalties outperformed the others in estimation accuracy. These findings suggested that AENET and ALASSO were promising methods for variable selection in competing risk analysis in the additive subdistribution hazards model. However, further studies should explore alternative variable selection methods, such as genetic algorithms and Particle Swarm Optimization, and compare their performance with penalized approaches. Moreover, developing survival trees and random forests based on additive hazards models presents an intriguing avenue for future research.

Abbreviations

CIF, Cumulative incidence function;
LASSO, Least absolute shrinkage and selection operator;
ALASSO, Adaptive least absolute shrinkage and selection operator;
ENET, Elastic net;
AENET, Adaptive elastic net;
SCAD, Smoothly clipped absolute deviation;

Funding

This study was supported by Hamadan University of Medical Sciences (Grant No. 14040202670).

Conflict of interest

None declared.

ACKNOWLEDGMENTS

The authors would like to thank the Vice-Chancellor for Research and Technology at Hamadan University of Medical Sciences. The first author is deeply grateful to Professor Dr. Jelle Goeman and Professor Dr. Hein Putter (Leiden University Medical Center) for their invaluable assistance in completing this work.

REFERENCES

1. Ambrogi F, Scheike THJB. Penalized estimation for competing risks regression with applications to high-dimensional covariates. Biostatistics. 2016;17(4):708-21.

2. Binder H, Allignol A, Schumacher M, Beyersmann J. Boosting for high-dimensional time-to-event data with competing risks. Bioinformatics. 2009;25(7):890-6.

3. Gaïffas S, Guilloux A. High-dimensional additive hazards models and the lasso. Electronic Journal of Statistics. 2012;6:522-46.

4. Lin W, Lv J. High-dimensional sparse additive hazards regression. Journal of the American Statistical Association. 2013;108(501):247-64.

5. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1996;58(1):267-88.

6. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties.

Journal of the American Statistical Association. 2001;96(456):1348-60.

7. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2005;67(2):301-20.

8. Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. Statistica Sinica. 2010;20(1):101.

9. Tibshirani R. The lasso method for variable selection in the Cox model. Statistics in Medicine. 1997;16(4):385-95.

10. Zhang HH, Lu W. Adaptive Lasso for Cox's proportional hazards model. Biometrika. 2007;94(3):691-703.

11. Lin D, Ying Z. Semiparametric analysis of the additive risk model. Biometrika. 1994;81(1):61-71.

12. Martinussen T, Scheike TH. Covariate selection for the semiparametric additive risk model. Scandinavian Journal of Statistics. 2009;36(4):602-19.

13. Zhang H, Sun L, Zhou Y, Huang J. Oracle inequalities and selection consistency for weighted Lasso in high-dimensional additive hazards model. Statistica Sinica. 2017;27(4):1903-20.

14. Liu L, Su W, Zhao X. Bi-selection in the high-dimensional additive hazards regression model. Electronic Journal of Statistics. 2021;15(1):748-72.

15. Ma S, Huang J. Additive risk survival model with microarray data. BMC Bioinformatics. 2007;8(1):1-10.

16. Eriksson F, Li J, Scheike T, Zhang MJJB. The proportional odds cumulative incidence model for competing risks. Biometrics. 2015;71(3):687-95.

17. Fine JP, Gray RJJJotAsa. A proportional hazards model for the subdistribution of a competing risk. Journal of the American Statistical Association. 1999;94(446):496-509.

18. Sun L, Liu J, Sun J, Zhang MJSS. Modeling the subdistribution of a competing risk. Statistica Sinica. 2006;16(4):1367.

19. Zheng C, Dai R, Hari PN, Zhang MJJSim. Instrumental variable with competing risk model. Statistics in Medicine. 2017;36(8):1240-55.

20. Scheike TH, Zhang M-JJLda. Flexible competing risks regression modeling and goodness-of-

fit. Lifetime Data Analysis. 2008;14(4):464.

21. Scheike TH, Zhang M-J, Gerds TAJB. Predicting cumulative incidence probability by direct binomial regression. Biometrika. 2008;95(1):205-20.

22. Tapak L, Saidijam M, Sadeghifar M, Poorolajal J, Mahjub HJG, proteomics, bioinformatics. Competing risks data analysis with high-dimensional covariates: an application in bladder cancer. Genomics, proteomics & Bioinformatics. 2015;13(3):169-76.

23. Fu Z, Parikh CR, Zhou B. Penalized variable selection in competing risks regression. Lifetime Data Analysis. 2017;23:353-76.

24. Kawaguchi ES, Shen JI, Suchard MA, Li G. Scalable algorithms for large competing risks data. Journal of Computational and Graphical Statistics. 2021;30(3):685-93.

25. Tapak L, Mahjub H, Sadeghifar M, Saidijam M, Poorolajal JJIjoph. Predicting the survival time for bladder cancer using an additive hazards model in microarray data. Iranian Journal of Public Health. 2016;45(2):239.

26. Dixon SN, Darlington GA, Desmond AF. A competing risks model for correlated data based on the subdistribution hazard. Lifetime Data Analysis. 2011;17(4):473-95.

27. Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 2006;101(476):1418-29.

28. Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. Annals of Statistics. 2009;37(4):1733.

29. Beyersmann J, Latouche A, Buchholz A, Schumacher M. Simulating competing risks data in survival analysis. Statistics in Medicine. 2009;28(6):956-71.

30. Dyrskjøt L, Zieger K, Real FX, Malats N, Carrato A, Hurst C, et al. Gene expression signatures predict outcome in non–muscle-invasive bladder carcinoma: a multicenter validation study. Clinical Cancer Research. 2007;13(12):3545-51.

31. Wang C, Li N, Diao H, Lu L. Variable selection through adaptive elastic net for proportional odds model. Japanese Journal of Statistics and Data Science. 2024;7(1):203-21.

32. Bradic J, Fan J, Jiang J. Regularization for Cox's proportional hazards model with NP-dimensionality. Annals of Statistics. 2011;39(6):3092.

33. Hou J, Paravati A, Hou J, Xu R, Murphy J. High-dimensional variable selection and prediction

under competing risks with application to SEER-Medicare linked data. Statistics in Medicine. 2018;37(24):3486-502.

34. Shen P, He X, Lan L, Hong Y, Lin M. Identification of cell division cycle 20 as a candidate biomarker and potential therapeutic target in bladder cancer using bioinformatics analysis. Bioscience Reports. 2020;40(7):BSR20194429.

35. Liu Y, Zou S-h, Gao X. Bioinformatics analysis and experimental validation reveal that CDC20 overexpression promotes bladder cancer progression and potential underlying mechanisms. Genes & Genomics. 2024;46(4):437-49.

36. Verma S, Shankar E, Lin S, Singh V, Chan ER, Cao S, et al. Identification of key genes associated with progression and prognosis of bladder cancer through integrated bioinformatics analysis. Cancers. 2021;13(23):5931.

37. Duan H, Yu S, Xia W, Wang C, Zhang S, Shen Y, et al. Prognostic implications of a four-gene signature in non-muscle invasive bladder cancer. 2023.

38. Zhu H, Chen H, Wang J, Zhou L, Liu SJO, therapy. Collagen stiffness promoted non-muscle-invasive bladder cancer progression to muscle-invasive bladder cancer. OncoTargets and Therapy. 2019;12:3441.

39. Yan P, He Y, Xie K, Kong S, Zhao W. In silico analyses for potential key genes associated with gastric cancer. PeerJ. 2018;6:e6092.

40. Vuong H, Cheng F, Lin C-C, Zhao Z. Functional consequences of somatic mutations in cancer using protein pocket-based prioritization approach. Genome Medicine. 2014;6:1-14.

41. Ke C, Bandyopadhyay D, Sarkar D. Gene Screening for Prognosis of Non-Muscle-Invasive Bladder Carcinoma under Competing Risks Endpoints. Cancers. 2023;15(2):379.

42. Xie J, Zhang H, Wang K, Ni J, Ma X, Khoury CJ, et al. M6A-mediated-upregulation of lncRNA BLACAT3 promotes bladder cancer angiogenesis and hematogenous metastasis through YBX3 nuclear shuttling and enhancing NCF2 transcription. Oncogene. 2023;42(40):2956-70.

43. Neuzillet Y, Chapeaublanc E, Krucker C, De Koning L, Lebret T, Radvanyi F, Bernard-Pierrot IJBc. IGF1R activation and the in vitro antiproliferative efficacy of IGF1R inhibitor are inversely correlated with IGFBP5 expression in bladder cancer. BMC Cancer. 2017;17(1):636.

44. Gonzalez-Roibon N, Kim JJ, Faraj SF, Chaux A, Bezerra SM, Munari E, et al. Insulin-like growth factor-1 receptor overexpression is associated with outcome in invasive urothelial carcinoma of urinary bladder: a retrospective study of patients treated using radical cystectomy.

Urology. 2014;83(6):1444. e1-. e6.

45. Rochester MA, Patel N, Turney BW, Davies DR, Roberts IS, Crew J, et al. The type 1 insulin-like growth factor receptor is over-expressed in bladder cancer. BJU International. 2007;100(6):1396-401.

46. Faraj S, Gonzalez-Roibon N, Bezerra S, Munari E, Sharma R, Rezaei K, et al. MP28-10 IGF1R IMMUNOEXPRESSION IN SUPERFICIAL NON-MUSCLE INVASIVE UROTHELIAL CARCINOMA OF URINARY BLADDER. The Journal of Urology. 2014;191(4S):e300.

47. Hu H, Wang Z, Li M, Zeng F, Wang K, Huang R, et al. Gene expression and methylation analyses suggest DCTD as a prognostic factor in malignant glioma. Scientific reports. 2017;7(1):11568.

48. Ou Q, Lu Z, Cai G, Lai Z, Lin R, Huang H, et al. Unraveling the influence of metabolic signatures on immune dynamics for predicting immunotherapy response and survival in cancer. MedComm–Future Medicine. 2024;3(2):e89.