Received: 11 April 2017 | Revised: 9 April 2018 | Accepted: 26 April 2018

DOI: 10.1002/sim.7822

RESEARCH ARTICLE

WILEY Statistics in Medicine

High-dimensional variable selection and prediction under competing risks with application to SEER-Medicare linked data

Jiayi Hou[1] | Anthony Paravati[2] | Jue Hou[3] | Ronghui Xu[3,4] | James Murphy[2]

[1]Altman Clinical and Translational Research Institute, University of California, San Diego, La Jolla, CA 92093, U.S.A.

[2]Department of Radiation Medicine and Applied Sciences, University of California, San Diego, La Jolla, CA 92093, U.S.A.

[3]Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, U.S.A.

[4]Department of Family Medicine and Public Health, University of California, San Diego, La Jolla, CA 92093, U.S.A.

Correspondence
Ronghui Xu, Family Medicine and Public Health, University of California, San Diego, 9500 Gilman Drive, Mail Code 0112, La Jolla, CA 92093, U.S.A.
Email: rxu@ucsd.edu

Funding information
American Society of Clinical Oncology (ASCO); National Institutes of Health Clinical and Translational Science Award (CTSA), Grant/Award Number: UL1TR001442

Competing risk analysis considers event times due to multiple causes or of more than one event types. Commonly used regression models for such data include (1) cause-specific hazards model, which focuses on modeling one type of event while acknowledging other event types simultaneously, and (2) subdistribution hazards model, which links the covariate effects directly to the cumulative incidence function. Their use in the presence of high-dimensional predictors are largely unexplored. Motivated by an analysis using the linked SEER-Medicare database for the purposes of predicting cancer versus noncancer mortality for patients with prostate cancer, we study the accuracy of prediction and variable selection of existing machine learning methods under both models using extensive simulation experiments, including different approaches to choosing penalty parameters in each method. We then apply the optimal approaches to the analysis of the SEER-Medicare data.

KEYWORDS
boosting, cumulative incidence function, electronic medical record, LASSO, machine learning, precision medicine

1 | INTRODUCTION

As an illustration project of how information contained in patients' electronic medical records can be harvested for the purposes of precision medicine, we consider the large data set linking the Surveillance, Epidemiology and End Results (SEER) Program database of the National Cancer Institute with the federal health insurance program Medicare database for prostate cancer patients of age 65 or older. Each year, 180 000 men are examined with prostate cancer in the United States, and the important clinical decision commonly encountered in this patient population is whether to pursue aggressive cancer-directed therapy in the presence of preexisting comorbidities. Prostate cancer can progress slowly, and a proportion of men will die of competing causes before their prostate cancer becomes symptomatic. Current clinical guidelines for the management of prostate cancer instruct clinicians to make treatment decisions based on 2 factors: (1) an estimation of the aggressiveness of a patient's tumor and (2) estimation of a patient's overall life expectancy.[1] Classical cancer-specific survival prediction relies on 3 main risk factors: tumor stage, Gleason score, and prostate-specific antigen. On the other hand, currently, no tool exists to predict noncancer survival for this patient population. As

noncancer and cancer survival are not independent, ie, so-called competing risks in the statistical literature, an accurate comprehensive survival prediction tool should consider both types of risks simultaneously.

We restrict our analysis to patients examined between 2004 and 2009 in the SEER-Medicare database. After excluding additional patients with missing clinical records, we have a total of 57 011 patients who have information available on 7 relevant clinical variables (age, prostate-specific antigen, Gleason score, American joint committee on cancer (AJCC) stage, and AJCC stage T, N, M, respectively), 5 demographical variables (race, marital status, metro, registry, and year of diagnosis), plus 8971 binary insurance claim codes. We assumed that the survival prediction would occur at the time of diagnosis; therefore, we used clinical and demographic information at the time of diagnosis and insurance claims data during the year prior to diagnosis. Insurance claims capture medical diagnoses and procedures through health care common procedure coding system (HCPCS) codes, international classification of diseases, 9th revision (ICD-9) diagnosis codes, and ICD-9 procedure codes. These claims indirectly describe events that occur in surgical procedures, hospitalization, and outpatient activities. We converted each unique insurance claim code into a binary variable denoted 1 if the claim appeared anytime in the year before diagnosis and 0 if the code was absent. Until December 2013 (end of follow-up for these data), there were a total of 1247 deaths due to cancer and 5221 deaths unrelated to cancer. It is well understood that for time-to-event data, the number of events dictates the effective sample size, so in this case, we have more predictors to consider than the effective sample size.

Classical statistical methods, such as stepwise regression, have been known to suffer from model inconsistency and are computationally infeasible when the number of covariates is equal to or greater than the (effective) sample size. A group of machine learning methods, in particular supervised learning, has shown good performance empirically when the data are of high dimensionality.[2] The goals of these methods are[3] (1) prediction, to find a set of covariates which results in minimal prediction error in independent test data, and (2) variable selection, estimate the true sparsity pattern with low false positive rate for each covariate. In theory, consistent variable selection requires stronger assumptions, known as neighborhood stability or irrepresentable condition, which roughly translates to not too strongly correlated design, and the beta-min condition, which requires all nonzero coefficients to be sufficiently large. Both these conditions might be difficult to achieve for high-dimensional data in practice.[3] Therefore, methods like least absolute shrinkage and selection operator (LASSO) can be consistent for estimating the underlying regression function, for a properly chosen penalty parameter (see below also), but can perform very poorly for variable selection with strongly correlated design. Fortunately in our application, the prediction of mortality rates is of interest, and in this paper, we will focus on the performance of machine learning methods in estimating the true cumulative incidence function (CIF) at any given time. Meanwhile, during the process of simulation experiments, we also obtain results on variable selection as a side product and an added aspect.

Researchers have studied different approaches to analyze survival data with high-dimensional covariates. Notably, Tibshirani[4] proposed the LASSO under the Cox proportional hazards model. Zhang and Lu[5] investigated the statistical properties of adaptive LASSO for the Cox proportional hazards model. Hothorn et al[6] introduced a random forest algorithm and a generic gradient boosting algorithm for right censoring data. When considering theoretical aspects, Bradic et al[7] studied a group of penalty functions and established strong oracle properties of nonconcave penalized methods for ultrahigh-dimensional covariates in the presence of right censoring. In comparison, very few high-dimensional methods have been developed in the presence of competing risks. Binder et al[8] first proposed a boosting approach for fitting the proportional subdistribution hazards (PSDH) model. Ha et al[9] considered variable selection for clustered competing risks data under the subdistribution hazard frailty model. Very recently, Fu et al[10] considered penalized approaches under the same model. Given the high-dimensional nature of our data, in this paper, we will investigate the accuracy of variable selection and prediction using existing computational software under 2 commonly used models: the proportional cause-specific hazards (PCSH) model and the PSDH model. This leads to the approach of Binder et al under the PSDH model, and LASSO and adaptive LASSO approaches under the PCSH model, both being readily implemented and applicable to our large data set. In addition, we have implemented a LASSO algorithm under the PSDH model. All these approaches rely critically on the selection of a "penalty" parameter, and there are different ways to select this parameter. We will empirically evaluate these different methods using Monte Carlo simulations. The ultimate goal is to assess the prediction accuracy of the CIF as a risk assessment tool.

The remainder of this paper is organized as follows: In Section 2, we review the PCSH and the PSDH models. In Section 3, we review the relevant machine learning methods that have been or can be feasibly implemented to analyze competing risks data under each model. In Section 4, we conduct comprehensive simulation studies on these methods with varying numbers of predictors (relative to the sample size) that are continuous or binary (and in case of binary, sparse, or not sparse). In Section 5, we apply the machine learning methods under either model to classify prostate patients from the SEER-Medicare linked data into different risk groups according to their predicted CIFs. Finally, Section 6 contains discussion and directions for future work.

2 | COMPETING RISK MODELS

Competing risks occur when multiple types of failures coexist and the occurrence of one type of failure may prevent the observation of the other types of failure. In addition, the failure times may be subject to right censoring. Let $\epsilon = 1,...,J$ be the cause or type (we use the 2 words interchangeably in the following) of failure. Let $T = \min_{j=1}^{J} \tilde{T}_j$, where $\tilde{T}_j$ is the (possibly) latent failure time due to cause j. Let $X_i = \min(T_i, C_i)$, $\delta_i = I(T_i \le C_i)$, where C_i is the potential censoring time and is assumed noninformative. Denote $S(t) = P(T > t)$ as the survival function of T. The CIF for failure type j is $F_j(t) = P(T \le t, \epsilon = j)$. Obviously, $S(t) = 1 - \sum_{j=1}^{J} F_j(t)$, and $\sum_{j=1}^{J} F_j(\infty) = 1$. Denote the cause-specific hazard function of type j as $\lambda_j(t) = \lim_{\Delta t \to 0+} P(t \le T < t + \Delta t, J = j | T \ge t)/\Delta t$. Then one can also show that

$$F_j(t) = \int_0^t \lambda_j(u) S(u) du, \tag{1}$$

leading to a nonparametric estimate of the CIF if we use[11]

$$\hat{\lambda}_j(t_i) = \frac{d_{ji}}{n_i}, \tag{2}$$

where d_{ji} denotes the number of failures from cause j at time t_i and n_i the number of subjects at risk at t_i, and

$$\hat{S}(t) = \prod_{j:t_i \le t} \left\{ 1 - \sum_{j=1}^{J} \hat{\lambda}_j(t_i) \right\}. \tag{3}$$

Then we have $\hat{F}_j(t) = \sum_{i:t_i \le t} \hat{p}_j(t_i)$, where $\hat{p}_j(t_i) = \hat{\lambda}_j(t_i) \hat{S}(t_i^-)$. While $\hat{F}_j(t)$ is a complex function of the $\hat{\lambda}_j(t_i)$'s, the $(1-\alpha)100\%$ pointwise (for each t) confidence intervals can calculated using the "Cuminc()" function in the R package "mstate."

2.1 | The PCSH model

In the regression settings, the Cox-type proportional hazards model can be used to model the above cause-specific hazards, and existing software for fitting the Cox model for classical survival data without competing risks can be used to fit the PCSH model.[12] Under this model, however, the dependence of the CIF of a particular failure type on the covariates involves also the effects of the covariates on the cause-specific hazards of all other types of failures.

Given a p-dimensional vector of covariates Z, under the proportional hazards assumption of the cause-specific hazard function, we have

$$\lambda_j(t|Z) = \lambda_{0j}(t) \exp(\beta_j' Z), \tag{4}$$

for $j = 1,...,J$. To estimate β_j, we can use any software for the regular Cox model for one cause at a time, by treating all other types of events as if censored. This is because the (partial) likelihood for all event types factors into a separate likelihood function for each event type, and the likelihood function for each event type treats all other types of events as if censored.

To estimate the CIF given $Z = z_0$, we have similar to the above nonparametric estimation:

$$\begin{aligned} \hat{F}_j(t; z_0) &= \int_0^t \hat{S}(u; z_0) d\hat{\Lambda}_j(u; z_0) \\ &= \sum_{i=1}^{n} \frac{\hat{S}(X_i; z_0) \delta_{ji} I(X_i \le t) \exp(\hat{\beta}_j' z_0)}{\sum_{i'=1}^{n} I(X_i \le X_{i'}) \exp(\hat{\beta}_j' Z_{i'})}, \end{aligned} \tag{5}$$

where $\hat{S}(u; z_0) = \exp\{-\sum_{j=1}^{J} \hat{\Lambda}_j(u; z_0)\}$, $\hat{\Lambda}_j(u; z_0) = \hat{\Lambda}_{0j}(u) \exp(\hat{\beta}_j' z_0)$, and $\hat{\Lambda}_{0j}(u)$ is a Breslow-type estimator of the baseline cumulative hazard.[13] Notice that in estimating the overall survival function $\hat{S}$, we need to fit the models for all event types, even if we are only interested in the CIF of type j.

A $(1-\alpha)100\%$ pointwise confidence interval can be computed following Cheng et al.[14] We implemented an R package "CompetingRisk"[15] to compute the above estimator of the CIF with its pointwise confidence intervals.

2.2 | The PSDH model

To link the covariates directly to the CIF, Fine and Gray[16] proposed to model the so-called subdistribution hazards. The proportional hazards modeling of the subdistribution hazards, also known as Fine-Gray model, has gained popularity in recent years.

Gray[17] introduced the subdistribution hazard function as $\tilde{\lambda}_j(t) = -\frac{d}{dt}\log\{1-F_j(t)\}$. Under the proportional hazards assumption of the subdistribution hazard function for cause 1, we have[16]

$$\tilde{\lambda}_1(t|Z) = \tilde{\lambda}_{01}(t)\exp(\beta' Z). \tag{6}$$

It is easy to see that model (6) provides a direct way to estimate the CIF of cause 1, so that there is no need to fit models for the other causes to estimate F_1.

Fine and Gray[16] proposed estimating equations for β. Geskus[18] further showed that these estimating equations can be solved using weighted Cox regression, ie, software for the regular Cox model incorporating weights. The baseline subdistribution hazard is again estimated using a modified version of the Breslow estimator. The $(1-\alpha)100\%$ pointwise confidence intervals can be constructed by sampling standard normal random variables and otherwise closed-form formulas.[16]

The PCSH and the PSDH models are typically not valid at the same time, and limited empirical experiences seem to indicate that in real data applications, the 2 models can lead to similar conclusions.[19] In addition, Lambert et al[20] considered flexible parametric modeling of the CIF given covariates using splines. Koller et al[21] argued favorably for using the PCSH for etiology or efficacy hypotheses and the PSDH for prognosis in clinical settings. Our data are more complex than the more classic clinical data in that they include many treatment procedures captured by the claims codes, as well as prognostic variables. From a modeling point of view, both models might be used to approximate the underlying but unknown data-generating mechanism. In the following, we consider both the PCSH and the PSDH models.

3 | REGULARIZATION

Unlike classical statistical methods, machine learning methods aimed at high-dimensional covariates often involve the selection of a tuning parameter, based on the minimal estimated prediction error. There are 2 ways to estimate this prediction error: cross-validation, which is computationally intensive, or approximation methods such as the C_p-type statistics. When a log-likelihood loss function is used, the latter leads to the well-known Akaike information criterion (AIC). Another commonly used information-based criterion is the Bayesian information criterion (BIC), which imposes a larger penalty than the AIC.

To choose the tuning parameter, denoted by λ below, we consider the following different methods:

- CV10: λ associated with the minimum 10-fold cross-validated (CV) prediction error (referred to as "error" in the following);
- CV + 1SE: λ associated with the minimum 10-fold CV error plus one standard error of the CV estimated errors;
- min AIC/BIC: λ associated with the minimum AIC or BIC criteria;
- elbow AIC / BIC: λ associated with the largest descent in AIC or BIC.

Under the Cox regression model, the AIC is defined as $-2\log(L) + 2s$, where L is the partial likelihood and $s = |S(\hat{\boldsymbol{\beta}})|$ is the number of nonzero estimated regression coefficients, ie, the size of the estimated active set $S(\hat{\boldsymbol{\beta}})$.[22,23] Bayesian information criterion under the Cox model is defined as $-2\log(L) + 2s\log(k)$, where k is the number of observed uncensored events.[24] We apply these definitions to the PCSH and the PSDH models below, where k is the number of observed events from the cause of interest, and the error is the negative log partial likelihood. The "elbow" criteria are described in Tibshirani et al[25] as a way to avoid overselection in practice.

3.1 | Least absolute shrinkage and selection operator

Least absolute shrinkage and selection operator is an L_1 penalization method proposed by Tibshirani[26] for building parsimonious models when the performance of classical methods such as stepwise regression or best subset selection is not satisfactory. For linear regression, LASSO solves a penalized least squares problem along the regularization path, where the regression coefficients associated with unimportant covariates shrink to exactly zero while granting nonzero coefficients for important covariates. The theoretical properties of LASSO have been extensively studied under the linear regression model. Meinshausen and Bühlmann[27] showed consistency of LASSO under the neighborhood stability condition, when the true nonzero coefficients are sufficiently large in absolute value. This condition is equivalent to the irrepresentable condition used by Zhao and Yu.[28] Although some of these theoretical conditions might be difficult to achieve in practice, LASSO has gained numerous attention as a technique to reduce dimensionality and construct predictive models. One of the main reasons for its popularity is its computational simplicity, involving convex optimization only.

Tibshirani[4] extended LASSO to the Cox regression model. Since the Cox regression software is typically used to fit the PCSH model, we may apply the same LASSO algorithm as proposed in Tibshirani.[4] Under the PCSH model, the partial likelihood for event type j is

$$L_j(\beta_j) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta_j' Z_i)}{\sum_{l \in R_i} \exp(\beta_j' Z_l)} \right\}^{\delta_i I(\epsilon_i = j)}, \tag{7}$$

where $R_i = \{l : X_l \geq X_i\}$ is the risk set at time X_i. The LASSO estimator of β_j minimizes the L_1-penalized log partial likelihood:

$$-\frac{1}{n} \log L_j(\beta_j) + \lambda \sum_{k=1}^{p} |\beta_{jk}|, \tag{8}$$

where the penalty parameter λ will be chosen by the methods described above in our simulations. We note that the R package "glmnet" implementation of LASSO has been widely used and is able to fix Cox regression model as well as the PCSH model with large data sets.

Under the PSDH model,[16] the pseudolikelihood used for inference is the same as the partial likelihood for complete (ie, no censoring) data, but otherwise with weights in the risk sets to account for censoring:

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta' Z_i)}{\sum_{l \in R_i} w_l(X_i) \exp(\beta' Z_l)} \right\}^{I(\delta_i \epsilon_i = 1)}, \tag{9}$$

where $R_i = \{l : X_l \geq X_i \text{ or } \delta_l \epsilon_l > 1\}$ is the risk set consisting of individuals who have not had any event or who have had an event of other causes and $w_l(t) = \hat{G}(t) I(t \geq X_l) \delta_l / \hat{G}(X_l) + I(t < X_l)$ where $\hat{G}(t)$ is the Kaplan-Meier estimate of $P(C > t)$. The LASSO estimator is similar to the above under the PSDH model, ie, it minimizes:

$$-\frac{1}{n} \log L(\beta) + \lambda \sum_{k=1}^{p} |\beta_k|, \tag{10}$$

where the penalty parameter λ will again be chosen by the methods described above. We note that while Fu et al[10] considered the LASSO and other penalized approaches under the PSDH model, we were not able to apply the associated R package "crrp" to the linked SEER-Medicare data as it ran out of memory. Instead, we have developed our own R program using gradient descent algorithm written in Fortran, which is in the process of being made into a package.

Tibshirani[26] discussed the standard errors of the LASSO estimator, which was approximated by a sandwich-type covariance estimator based on the penalized log-likelihood. However, recent works on inference under high dimensions by Zhang and Zhang[29] and van de Geer et al[30] have shown that such regularized estimators are generally biased (at $\sqrt{n}$-rate) and proposed bias correction procedures for inference purposes. A related group of authors of this paper is currently developing a similar inference procedure under the PSDH model in high dimensions (see https://arxiv.org/abs/1707.09561).

3.2 | Adaptive LASSO

Adaptive LASSO was initially developed by Zou.[31] As mentioned earlier, Zhang and Lu[5] proposed adaptive LASSO for the Cox proportional hazards model, where adaptively weighted L_1 penalties were used. Instead of the LASSO penalty above, the adaptive LASSO penalty has the form $\lambda\sum_{k=1}^{p}|\beta_k/\tilde{\beta}_k|$, where $\tilde{\beta} = (\tilde{\beta}_1, \ldots \tilde{\beta}_p)$ maximized the (partial) likelihood without any penalty. This way, small weights are given to large coefficients and large weights for small coefficients. The resulting estimator enjoys the oracle properties (not enjoyed by LASSO) and has a convex penalty form that ensures the existence of global optimizers and can be efficiently solved by standard algorithms just like LASSO.

The algorithms described above for LASSO can be readily applied to adaptive LASSO under both the PCSH and the PSDH models.

3.3 | Boosting

Freund and Schapire[32] introduced the AdaBoost algorithm to solve classification problems by combining rough and moderately accurate "rules of thumb" repeatedly. Later, Friedman[33] developed boosting methods for linear regression as a numerical optimization method to minimize the squared error loss function. Boosting can be viewed as a gradient descent optimization algorithm in function space and is essentially the same as the matching pursuit algorithm in signal processing.[34] Bühlmann [35] proved that boosting with the squared error loss is consistent in high-dimensional linear models, where the number of predictors is allowed to grow as fast as exponential to the sample size.

For the PSDH model with high-dimensional data, Binder et al[8] proposed a likelihood-based boosting approach. Binder et al[8] applied the componentwise boosting approach of Bühlmann and Yu[36] to the likelihood in (9): Starting from initial value zero for β, at step k, for component j of β, let $\gamma_{k,j} = \beta_{k,j} - \beta_{k-1,j}$ be the potential update of that component from the previous step $(k-1)$ to the current step k, and find $\gamma_{k,j}$ by minimizing

$$l_{pen}(\gamma_{k,j}) = -\log L_j(\beta_{k-1,1}, \ldots, \beta_{k-1,j-1}, \beta_{k,j}, \beta_{k-1,j+1}, \ldots, \beta_{k-1,p}) + \frac{\lambda}{2}\gamma_{k,j}^2; \tag{11}$$

then among all $j = 1,\ldots,p$, only update the component j that has the largest score statistic $U_{pen}'(0)I_{pen}^{-1}(0)U_{pen}(0)$ where $U_{pen}(\gamma) = \partial l_{pen}(\gamma)/\partial\gamma$ and $I_{pen}(\gamma) = \partial^2 l_{pen}(\gamma)/\partial\gamma^2$. For boosting, the number of steps is the main tuning parameter, while λ in (11) is chosen so that the number steps is greater than 50.[8] The R package "CoxBoost" can be used, and the number of boosting steps will be chosen by the methods described above in our simulations.

4 | SIMULATIONS

4.1 | Setup

To investigate the performance of LASSO, adaptive LASSO, and boosting under the PCSH and the PSDH models, respectively, we conducted comprehensive simulation studies with both continuous and dichotomized covariates in competing risks data. We assumed $J = 2$, and we considered sample size $n = 500$ and number of covariates $p = 20$, 500, and 1000. We repeated each simulation setting 100 times.

For continuous covariates, the covariate vector for each subject was generated for the following correlation structures:

1. Independent: Each covariate was independently generated from $N(0,1)$.
2. Exchangeable: The covariate vector was generated from a multivariate normal distribution with mean zero, marginal variance of one, and a block diagonal covariance matrix—each block of size 10 and within a block the pairwise correlation $\rho(i,i') = 0.5$.
3. AR(1): The covariate vector was generated from a multivariate normal distribution with mean zero, marginal variance of one, and a block diagonal covariance matrix—each block of size 10 and within a block the pairwise correlation $\rho(i,i') = 0.5^{|i-i'|}$.

For binary covariates, the covariate vector was first generated the same as in the above, then dichotomized at threshold a, with $<a$ coded as 1 and 0 otherwise. We considered $a = 0$, to give a balanced binary distribution, and

$a=-1$, to give a relatively sparse 16% of 1's. We set the number of nonzero regression coefficients, ie, the size of the active set, to be $s_1=5$ and $s_2=3$ for causes 1 and 2, respectively. We let $\beta_{1,1\cdots,5}=(1.96,-0.79,-0.5,-1.35,1.29)$, $\beta_{2,11\cdots,13}=(-1.16,-0.86,0.5)$, and the rest of the β_1 and β_2 values were zero. These β values were used under both the PCSH and the PSDH models.

To simulate survival outcomes under the PCSH model, we followed the approach described in Beyersmann et al[37]; that is, we simulated the event time T first, then we simulated the cause ϵ given T. We assumed the baseline hazard functions for types 1 and 2 failures to be $\lambda_{01}(t)=0.15$ and $\lambda_{02}(t)=0.10$, respectively. The overall (not cause-specific) cumulative hazard function for T was then $\Lambda(t|z)=t\{\lambda_{01}\exp(\beta_1{}'z)+\lambda_{02}\exp(\beta_2{}'z)\}$, and T was generated using the fact that $U=\exp(-\Lambda(T))\sim U(0,1)$ given z. The cause ϵ was generated proportional to the cause-specific hazard function, ie, $P(\epsilon=1|z)=\lambda_{01}\exp(\beta_1{}'z)/\{\lambda_{01}\exp(\beta_1{}'z)+\lambda_{02}\exp(\beta_2{}'z)\}$. Under this model, the true CIF for cause j was

$$F_j(t|z)=\int_0^t S(u|z)\lambda_{0j}\exp(\beta_j{}'z)du=\lambda_{0j}\exp(\beta_j{}'z)\frac{e^{tM}}{M}, \tag{12}$$

where $M=-\{\lambda_{01}\exp(\beta_1{}'z)+\lambda_{02}\exp(\beta_2{}'z)\}$. The censoring times were generated from $U(0,20)$, which resulted in an average event rate of 45.8% for cause 1 and 33.6% for cause 2 with continuous covariates, an average event rate of 51.8% for cause 1 and 27.2% for cause 2 with balanced binary covariates, and an average event rate of 59.8% for cause 1 and 17.8% for cause 2 with sparse binary covariates..

To simulate under the PSDH model, we followed the approach described in Fine and Gray.[16] The CIF for failure from cause 1 was given by

$$F_1(t|z)=P(T\le t,\epsilon=1|z)=1-\{1-p(1-e^{-t})\}^{\exp(\beta_1'z)}, \tag{13}$$

where we used $p=0.6$. As this was a subdistribution function, with a point mass $1-F_1(\infty|z)$ at infinity, the proper distribution function that was used to generate T was $F(t|z)=F_1(t|z)/F_1(\infty|z)$, so that $F(T)\sim U(0,1)$ given z. Note that $P(\epsilon=1|z)=F_1(\infty|z)$ and $P(\epsilon=2|z)=1-P(\epsilon=1|z)$. Finally, the event times for failure from cause 2 were generated according to an exponential distribution with rate $\exp(\beta_2'z)$. The censoring times were generated from $U(0,20)$, resulting in an average event rate of 53.5% for cause 1 and 35.1% for cause 2 with continuous covariates, an average event rate of 55.8% for cause 1 and 33.4% for cause 2 with balanced binary covariates, and an average event rate of 55.5% for cause 1 and 33.5% for cause 2 with sparse binary covariates.

We also considered higher cause 2 event rates at the suggestion of a reviewer, and more details are provided in the Supporting Information.

4.2 | Results

We evaluate the performance of prediction at a given covariate vector value z_0. We set $z_0=(0.5,\cdots,0.5)_{1\times p}$ for the continuous case; and for all the binary cases, each element of z_0 was independently drawn with a fixed seed from Bernoulli distribution with $p=0.5$.

Figures 1 to 3 show the empirical distributions of the estimated $F_1(2|z_0)$ over the 100 simulation runs, where the vertical line marks the true $F_1(2|z_0)$; the empirical distributions were plotted using the R function "density()." The results under the PCSH model using adaptive LASSO and under the PSDH model using adaptive LASSO and boosting are given in these figures for sparse binary covariates, while results for continuous and balanced binary covariate, as well as additional results under both the PCSH and the PSDH models using LASSO, are given in the Supporting Information due to limitation of space.

In the figures, the blue dashed lines are for the oracle estimator, which fits the exact true active set $S(\beta)$. The oracle estimator varied extremely slightly with the 3 correlation structures for Z, any variation appearing due to Monte Carlo, and the one under the AR(1) structure is plotted here. It is seen that the distribution of the oracle estimator is more concentrated for the continuous covariates than for the sparse binary covariates, which reflects the "effective sample size" that is reduced with the sparse binary covariates. The solid lines are the estimated $F_1(2|z_0)$ under each model after regularization, with different colors representing different correlation structures of Z.

Under the PCSH model using adaptive LASSO to regularize (Figure 1 and Supporting Information), the performances were generally not satisfactory as compared with the oracle estimator for $p=500$ and 1000. The worst performances were seen when using minimum AIC and BIC to choose the penalty parameter; some of these results

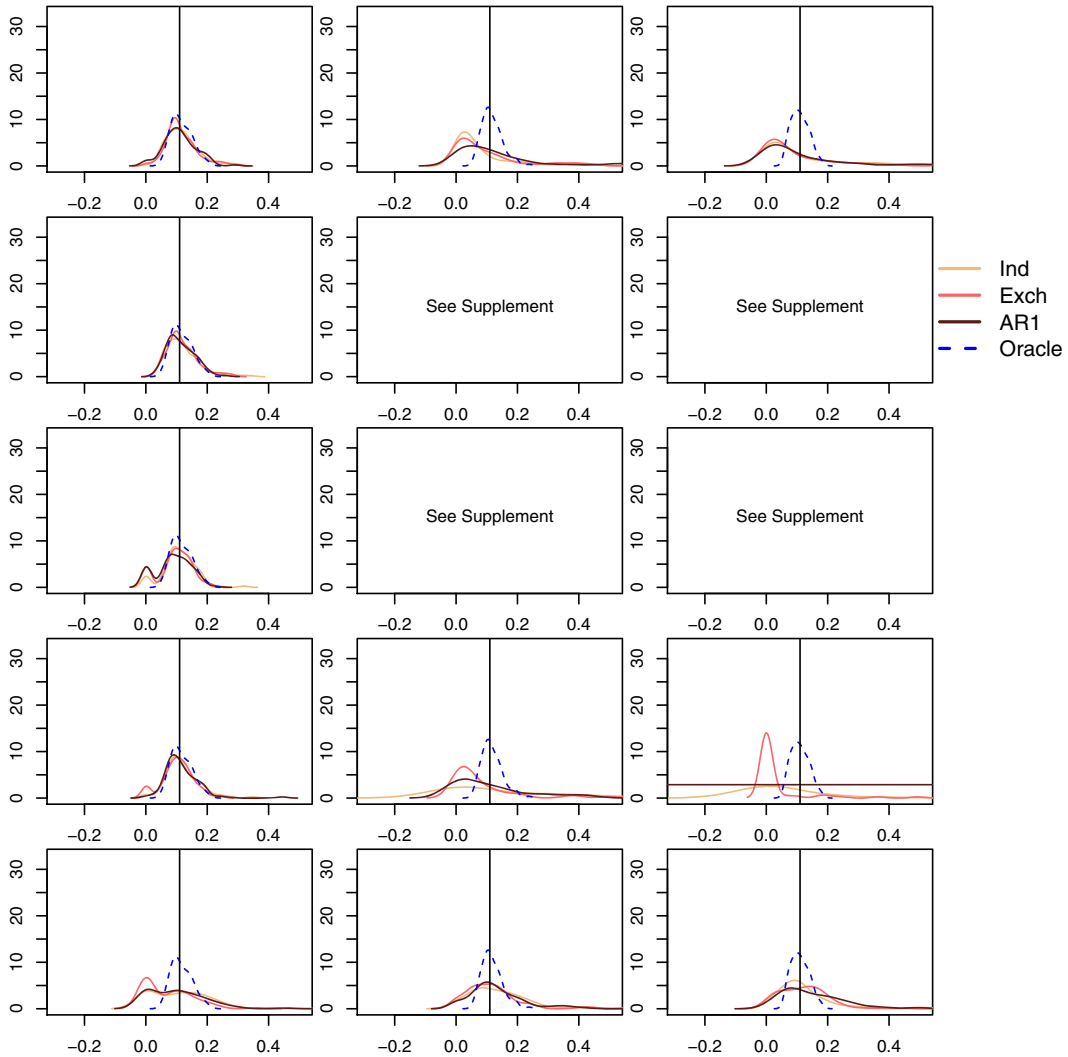

FIGURE 1 The (smoothed) empirical distribution of $\hat{F}_1(2|z_0)$ under the proportional cause-specific hazards model with adaptive least absolute shrinkage and selection operator, for sparse binary covariates. The 3 columns correspond to $p = 20$, 500, and 1000. The rows correspond to different ways of selecting penalty parameter λ, from top to bottom: (1) CV10, (2) minimum Akaike information criterion (AIC), (3) minimum Bayesian information criterion (BIC), (4) elbow AIC, and (5) elbow BIC. The true $CIF_1(t = 2|z_0) = 0.11$ [Colour figure can be viewed at wileyonlinelibrary.com]

were so extreme that "density()" failed to work and these were instead shown in the Supporting Information using boxplots. Elbow BIC appeared to perform the best for continuous covariates, but not so for binary covariates even when $p = 20$. CV10 had the best performance for binary covariates in general, but it too deteriorated for $p =$ 500 and 1000.

Under the PSDH model using adaptive LASSO (Figure 2 and Supporting Information), the results appeared to be similar to those under the PCSH model using adaptive LASSO for continuous covariates. For the sparse binary covariates, even the oracle estimator had a very wide spread, and the results were generally not satisfactory for $p = 500$ and 1000.

Under the PSDH model using boosting (Figure 3 and Supporting Information), we see that for continuous covariates, the estimators performed reasonably well when CV10 or min AIC/BIC was used to choose the number of boosting steps; with CV10, the estimation was perhaps the best. The performance deteriorated with binary covariates for $p = 500$ and 1000, with the performance under the independent structure the worst of all. We note that in Bühlmann[35] simulation studies (their table 1), the mean squared error for boosting with correlated design was also smaller than that with uncorrelated design, and their Figure 1 showed that boosting tended to overselect covariates in the uncorrelated design than the correlated design.

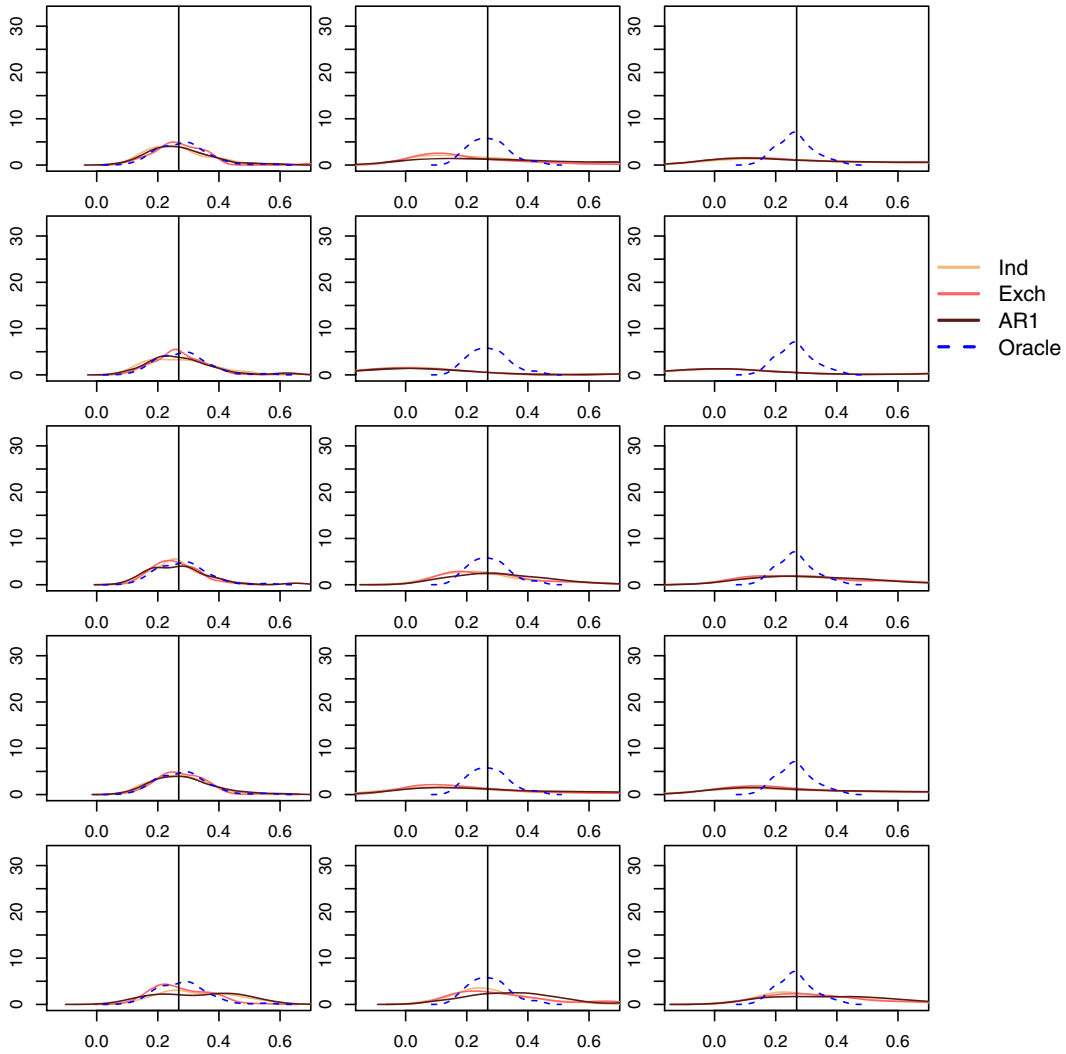

FIGURE 2 The (smoothed) empirical distribution of $\hat{F}_1(2|z_0)$ under the proportional subdistribution hazards model with adaptive least absolute shrinkage and selection operator, for sparse binary covariates. The 3 columns correspond to $p = 20$, 500, and 1000. The rows correspond to different ways of selecting penalty parameter λ, from top to bottom: (1) CV10, (2) minimum Akaike information criterion (AIC), (3) minimum Bayesian information criterion (BIC), (4) elbow AIC, and (5) elbow BIC. The true $CIF_1(t = 2|z_0) = 0.27$ [Colour figure can be viewed at wileyonlinelibrary.com]

Table 1 gives the mean integrated squared errors (MISEs) of $\hat{F}_1(t|z_0)$ from t= 0 to 20 (using numerical integration with increments of 0.1), for all 5 combinations of models and regularization methods when the penalty parameter was chosen using CV10. We note that $t = 20$ is the maximum follow-up time. In the Supporting Information, we also provide the MISEs for other methods of selecting the penalty parameter (min AIC, min BIC, elbow AIC, and elbow BIC), and in general, CV10 outperformed these other selection methods. It is seen from Table 1 that LASSO and adaptive LASSO were generally comparable under the PCSH model, with LASSO having slightly smaller MISE for sparse binary covariates and when p was large. Adaptive LASSO had smaller MISE than LASSO under the PSDH model for continuous covariates, but not always so for binary covariates, especially when p was large. Boosting under the PSDH model had much smaller MISE compared with LASSO or adaptive LASSO.

The Supporting Information provide additional simulation results including, among other things, variable selection results. Variable selection was not the main goal for our application, but it was a necessary step before prediction and the results helped to explain the prediction accuracy. Often when the prediction results were poor (as discussed above), it was because variable selection was poor; for example, when there were a couple of hundred false positives, the corresponding prediction results were extremely poor. Boosting had no more than 5 false positives in all cases. We also note that elbow BIC had a tendency to underselect, ie, with relatively more false negatives.

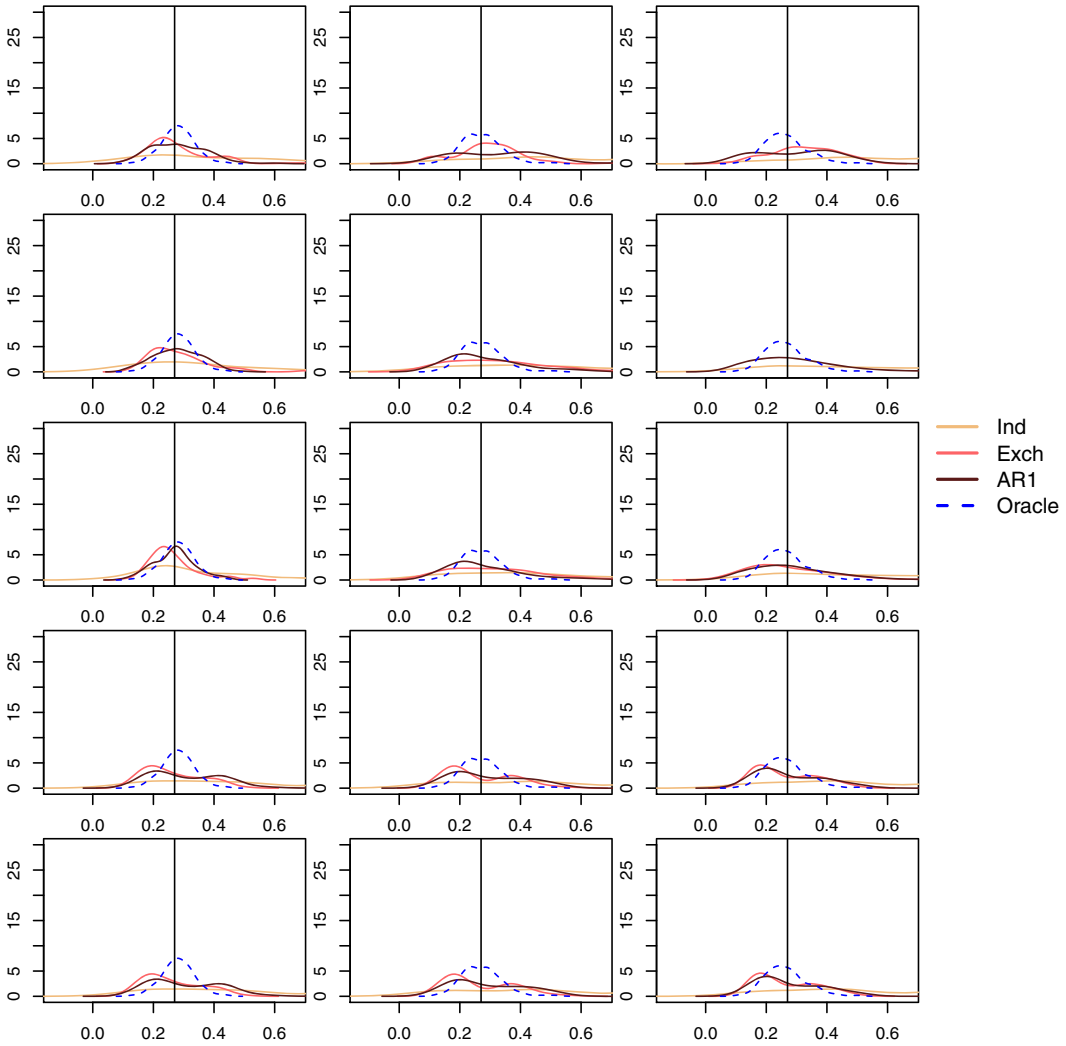

FIGURE 3 The (smoothed) empirical distribution of $\hat{F}_1(2|z_0)$ under the proportional subdistribution hazards model with boosting, for sparse binary covariates. The 3 columns correspond to $p = 20$, 500, and 1000. The rows correspond to different ways of selecting γ, from top to bottom: (1) CV10, (2) minimum Akaike information criterion (AIC), (3) minimum Bayesian information criterion (BIC), (4) elbow AIC, and (5) elbow BIC. The true $F_1(2|z_0) = 0.27$ [Colour figure can be viewed at wileyonlinelibrary.com]

5 | SEER-MEDICARE LINKED DATA

We randomly split the SEER-Medicare data set into approximate equal-sized training ($n = 28{,}505$) and test ($n = 28{,}506$) data sets. As a first step, we excluded binary claim codes having less than 10 ones, with rest all being zero. We then used univariate screening to further reduce dimensionality, eliminate noise, and increase the performance of subsequent variable selection methods.[38]

5.1 | PCSH model with LASSO and adaptive LASSO

Univariate screening under the PCSH model with P value cutoff of .05 gave $p_1 = 2188$ and $p_2 = 1079$ claim codes for noncancer and cancer mortalities, respectively. For each type of mortality, we first applied LASSO under the PCSH model on the training data with the above prescreened claim codes plus the clinical and demographic variables. Based on the simulation results, CV10 was used to choose the penalty parameter. The final model contained 143 predictors for noncancer mortality and 9 predictors for cancer mortality. Since the regression coefficients from LASSO are biased, we refit the PCSH model with the selected predictors.

TABLE 1 Mean integrated squared errors of $\hat{F}_1(t|z_0)$; the penalty parameter was chosen using CV10

Covariates	PCSH		PSDH		
	LASSO	Ad. LASSO	LASSO	Ad. LASSO	Boosting
Continuous					
$p = 20$					
Independence	0.07	0.05	0.08	0.06	0.03
Exchangeable	0.03	0.03	0.03	0.03	0.02
AR1	0.04	0.04	0.04	0.03	0.02
$p = 500$					
Independence	0.11	0.21	0.50	0.19	0.03
Exchangeable	0.27	0.23	0.43	0.28	0.02
AR1	0.30	0.20	0.40	0.31	0.02
$p = 1000$					
Independence	0.12	0.28	0.65	0.24	0.02
Exchangeable	0.25	0.24	0.50	0.39	0.03
AR1	0.29	0.36	0.48	0.34	0.02
Balanced binary					
$p = 20$					
Independence	0.13	0.16	0.11	0.08	0.09
Exchangeable	0.17	0.17	0.11	0.10	0.02
AR1	0.20	0.21	0.10	0.07	0.04
$p = 500$					
Independence	0.62	0.72	0.47	0.52	0.09
Exchangeable	0.71	0.97	0.47	0.50	0.07
AR1	0.74	0.82	0.59	0.66	0.05
$p = 1000$					
Independence	0.71	1.13	0.79	1.12	0.18
Exchangeable	0.68	1.20	0.71	0.90	0.04
AR1	0.81	1.27	1.14	1.28	0.05
Sparse binary					
$p = 20$					
Independence	0.36	0.34	0.30	0.21	0.03
Exchangeable	0.35	0.34	0.23	0.20	0.01
AR1	0.31	0.38	0.30	0.26	0.01
$p = 500$					
Independence	0.95	1.87	1.10	1.40	0.08
Exchangeable	1.28	1.50	1.23	1.70	0.02
AR1	1.32	1.50	1.19	1.98	0.03
$p = 1000$					
Independence	1.29	1.84	2.00	2.70	0.09
Exchangeable	1.30	1.99	2.21	2.15	0.03
AR1	1.39	2.06	1.27	2.30	0.04

Abbreviations: LASSO, least absolute shrinkage and selection operator; PCSH, proportional cause-specific hazards; PSDH, proportional subdistribution hazards.

To evaluate the resulting prediction model on the test data, we first calculated the risk score $\hat{\beta}_j'Z$ for each patient in the test data, $j = 1,2$. For each mortality type j, we then divided the test set into 4 risk strata: low (L), median low (ML), median high (MH), and high (H) according to the quartiles. Combining the 2 types of mortalities, we formed a total of 16 strata for their predicted CIF. Using the average Z values in each of the 16 strata, we plotted their predicted CIFs for both cancer and noncancer mortalities, and the results are shown in the Supporting Information. It was clear that instead of 16 groups, 5 distinct risk groups emerge for both cancer and noncancer mortalities. We note that these are not the same 5 groups for the 2 types of mortality, and the Supporting Information provides the definition for each of them. It is perhaps not surprising to see that each mortality risk was most influenced by the corresponding cause-specific risk and also secondarily by its competing risk.

In Figures 4 and 5, we plot the nonparametric CIF (solid lines) and its 95% confidence intervals (shaded) for each of the above 5 risk groups for noncancer and cancer mortalities, respectively. The clear separation of the 5 groups shows the usefulness of the final PCSH model in classifying patients according to their different prognosis for both cancer and noncancer mortalities. For comparison purposes, we also plot the predicted CIF for each of the 5 risk groups. While the prediction is more accurate for the noncancer CIF, the predicted cancer CIF seems less accurate especially for the high (H) risk group.

We then applied adaptive LASSO under the PCSH model in a similar fashion, and the results are also given in Figures 4 to 5 and the Supporting Information. The final model contained 113 predictors for noncancer mortality and 11 predictors for cancer mortality. Overall, the results were somewhat similar to those from LASSO, with also 5 visual groups from the original 16. The prediction results also appeared generally comparable between these 2 regularization methods, with prediction more accurate using one method sometimes and less so the other times.

5.2 | PSDH model with boosting, LASSO, and adaptive LASSO

Univariate screening with P value cutoff of .05 under the PSDH model initially gave $p_1 = 4634$ and $p_2 = 6088$ claim codes for noncancer and cancer mortalities, respectively. We further reduced the dimension by retaining only the top 2000 claim codes (ranked by P value) for each type of mortality, to be comparable with the fitting of the PCSH model above, as well as for the boosting algorithm to be able to run on our Dell R630 computer (2 Intel Xeon E5-2660 v3 2.6 GHz, each processor with 10 cores [20 threads] for a total of 20 cores [40 threads], and 128 GB of DDR3). We applied boosting under the PSDH model to the training data with these claim codes plus the clinical and demographic variables. Although CV10 performed best in our simulation results, AIC was a close second especially for binary covariates and was less computationally intensive for this procedure where boosting itself was computationally intensive already. Therefore, AIC was used to choose the optimal step. The final model contains 53 predictors for noncancer mortality and 13 predictors for cancer mortality. We refit the PSDH model with the selected predictors to obtain the unbiased estimator.

Similar to the above, we calculated the risk score $\hat{\beta}_j'Z$ for each patient in the test data, $j = 1,2$. For each mortality type j, we divided the test set into 5 risk strata: low (L), median low (ML), median (M), median high (MH), and high (H) according to the quintiles. Again, the classification was not the same 5 groups for the 2 types of morality. In Figures 4 and 5, we plot the nonparametric CIF (solid lines) and its 95% confidence intervals (shaded) for the above 5 risk groups for both noncancer and cancer mortalities. We also plot the predicted CIF for each of the 5 risk groups.

We then applied LASSO and adaptive LASSO under the PSDH model. While LASSO selected 143 predictors for noncancer mortality and 22 predictors for cancer mortality in the final models, adaptive LASSO selected 1045 predictors for noncancer mortality and 153 predictors for cancer mortality. The predicted and nonparametric estimates of the CIFs are also plotted in Figures 4 and 5. It is seen that the prediction results are generally similar to those from boosting. The prediction was still more accurate for the noncancer CIF, and less so for cancer CIF especially for the high (H) risk group.

6 | DISCUSSION

The rapid accumulation of data across many fields, medicine in particular, has created unique challenges in statistics. The distinct issues with high-dimensional data have come to be recognized recently, including, for example, the rapid noise accumulation, the unrealistic independence assumption, and the necessity for novel robust data analysis methods.[39] While researchers work to meet these challenges, some of the methods proposed in the literature do not necessarily scale well to large data sets.

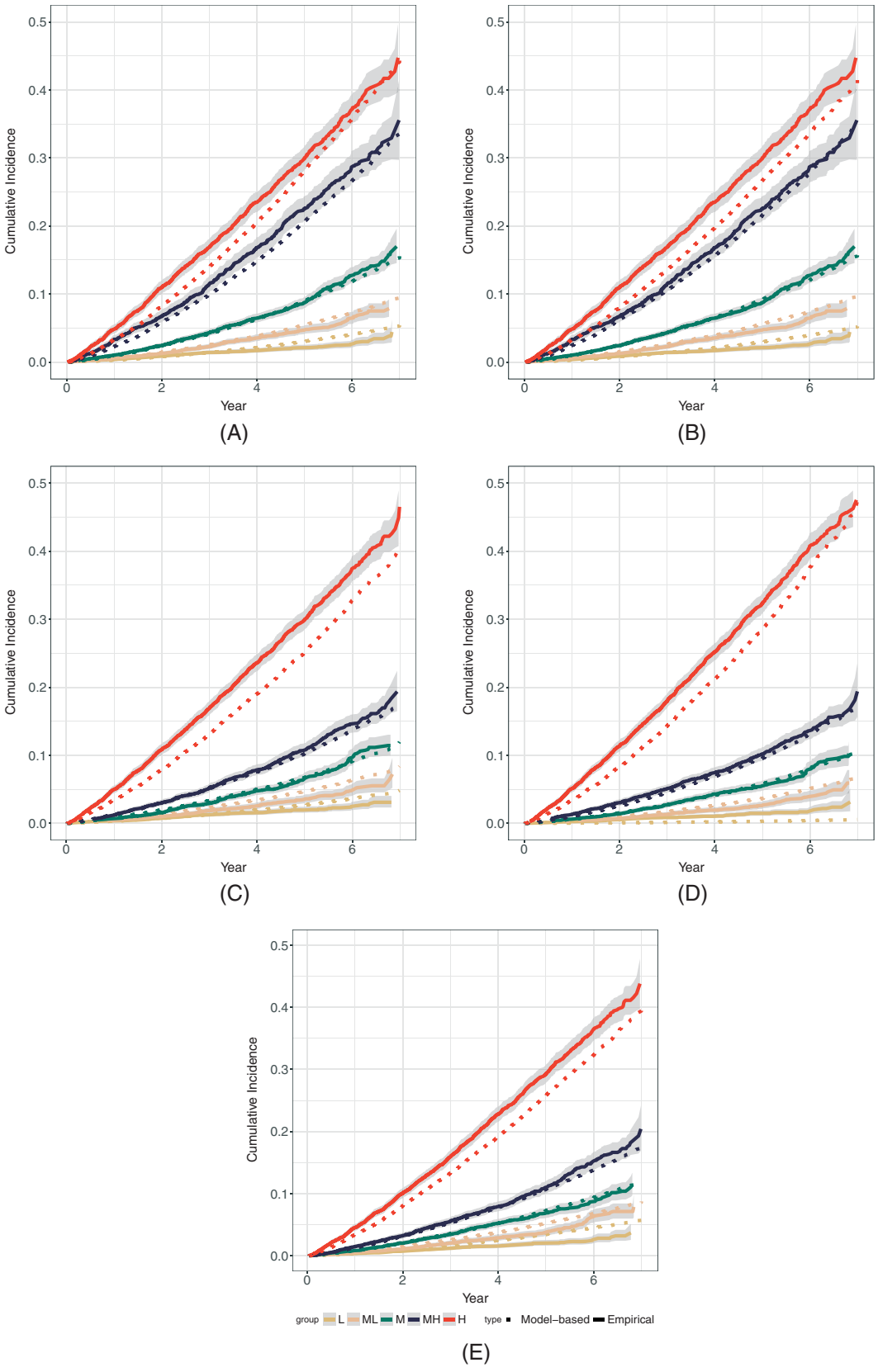

FIGURE 4 Cumulative incidence functions for noncancer mortalities, with classification and prediction using A, LASSO and B, adaptive LASSO under the PCSH model and C, LASSO, D, adaptive LASSO, and E, boosting under the PSDH model. The shaded area is the 95% pointwise confidence intervals based on the nonparametric estimate. LASSO, least absolute shrinkage and selection operator; PCSH, proportional cause-specific hazards; PSDH, proportional subdistribution hazards [Colour figure can be viewed at wileyonlinelibrary.com]

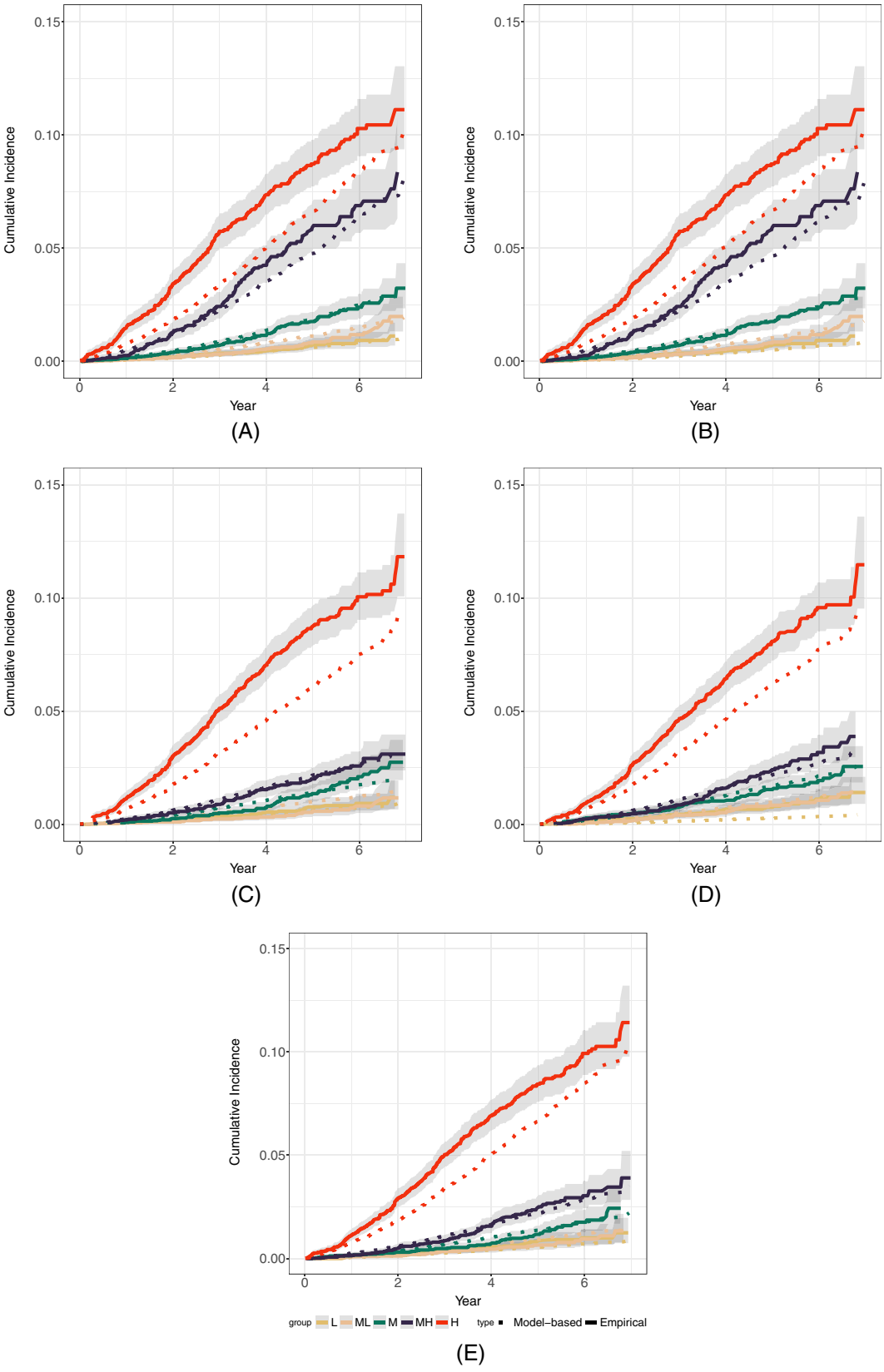

FIGURE 5 Cumulative incidence functions for cancer mortalities, with classification and prediction using A, LASSO and B, adaptive LASSO under the PCSH model and C, LASSO, D, adaptive LASSO, and E, boosting under the PSDH model. The shaded area is the 95% pointwise confidence intervals based on the nonparametric estimate. LASSO, least absolute shrinkage and selection operator; PCSH, proportional cause-specific hazards; PSDH, proportional subdistribution hazards [Colour figure can be viewed at wileyonlinelibrary.com]

In this paper, in addition to existing implementations of machine learning methods, ie, LASSO under the PCSH model and boosting under the PSDH model, we have also developed efficient algorithms for LASSO under the PSDH model and a corresponding R package is expected to be completed in the near future. We empirically studied the performance of these methods in variable selection and prediction through comprehensive simulations in both low- and high-dimensional settings with different covariate structures. From the simulation results, we see that the performance is generally better for continuous covariates than for binary ones, and further worse if the binary covariates are sparse, as in the case of claims data. This finding seems to echo a recent paper by Mukherjee et al[40] who showed that when a binary design matrix is sufficiently sparse, no signal can be detected irrespective of its strength. More work, both methodological and theoretical, appears to be needed to study binary data in high dimension, especially for sparse binary data.

In comparing these machine learning methods, adaptive LASSO had advantage over LASSO in some scenarios, but this was not universally the case especially when the covariates were binary. While it is known that good prediction does not necessarily require exact selection of the true predictors, poor performance in prediction was generally due to overselection of variables that were not true predictors in our simulation. Boosting performed well for prediction, largely due to the fact that there was not severe overselection of the variables.

For each of the machine learning methods, we compared different approaches to choose the penalty parameters, which was an important step in applying these methods. We found the 10-fold cross-validation to be generally as good as the other approaches, if not better. For LASSO-type methods, including adaptive LASSO, AIC including elbow AIC selected overly large numbers of false positive variables, resulting in poor performance for prediction purposes. For boosting however, this did not appear to be the case, and the computational ease of AIC over cross-validation seems a worthwhile advantage.

By applying the above methods to analyze a rich data set with claim codes describing disease diagnoses, surgical procedures, hospitalization, and outpatient activities, we created an individualized patient prediction tool aimed at helping prostate cancer patients and their physicians to better understand the prognosis for both cancer and other morbidities, which can in turn aid in clinical decision making. For the linked SEER-Medicare data that we have considered, the sample size is much larger than what we can afford in simulations. The main differences in prediction results for these data seem to be between the different models and less so between the regularization methods under the same model. As we have described in details, under the PCSH model, we initially had 16 strata that were then combined into 5 strata based on visual inspection, and the 5 strata were of different sizes (see tables in the Supporting Information). On the other hand, under the PSDH model, we directly divided the data into 5 equal-sized strata. How to best divide the strata including deciding the number of strata may worth future investigation. However, the purpose of dividing into strata here was mainly to illustrate that the fitted models have predictive capability. For the purposes of precision medicine, it is the individual prediction given a vector of covariates that is of most importance. We see that for noncancer mortality, both models did a reasonable job in prediction, and it shows the usefulness of claims database such as Medicare in predicting noncancer mortality. For cancer mortality, on the other hand, the clinical information captured in the SEER data might be relatively limited, or the models we considered might be partly misspecified; additional investigation including the use of individual patient electronic medical records might be able to provide more clinical information for better prediction.

Finally, our work here assumed the proportional hazards under both models. We note that there has been recent work considering other modeling approaches such as the additive hazards in the presence of competing risks.[41] Machine learning methods are still yet to be developed and studied under these models.

ACKNOWLEDGMENTS

This research was supported by a grant from the American Society of Clinical Oncology (ASCO) and was partially supported by the National Institutes of Health Clinical and Translational Science Award (CTSA)UL1TR001442.

ORCID


Ronghui Xu http://orcid.org/0000-0002-2822-0561

REFERENCES

1. Thompson I, Thrasher JB, Aus G, et al. Guideline for the management of clinically localized prostate cancer: 2007 update. *J Urol*. 2007;177:2106-2131.

2. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001;96:1348-1360.
3. Bühlmann P, van de Geer S. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin Heidelberg: Springer; 2011.
4. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med*. 1997;16:385-395.
5. Zhang HH, Lu W. Adaptive lasso for Cox's proportional hazards model. *Biometrika*. 2007;94:691-703.
6. Hothorn T, Bühlmann P, Dudoit S, Molinaro A, Van Der Laan MJ. Survival ensembles. *Biostatistics*. 2006;7:355-373.
7. Bradic J, Fan J, Jiang J. Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann Stat*. 2011;39:3092.
8. Binder H, Allignol A, Schumacher M, Beyersmann J. Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics*. 2009;25:890-896.
9. Ha ID, Lee M, Oh S, Jeong J, Sylvester R, Lee Y. Variable selection in subdistribution hazard frailty models with competing risks data. *Stat Med*. 2014;33:4590-4604.
10. Fu Z, Parikh CR, Zhou B. Penalized variable selection in competing risks regression. *Lifetime Data Anal*. 2017;23:353-376.
11. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis*, Vol. 169. New York: John Wiley & Sons; 2011.
12. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*, Vol. 360. New York: John Wiley & Sons; 2011.
13. Breslow N. Covariance analysis of censored survival data. *Biometrics*. 1974;30(1):89-99.
14. Cheng S, Fine JP, Wei L. Prediction of cumulative incidence function under the proportional hazards model. *Biometrics*. 1998;54(1):219-228.
15. Hou J, Xu R. CompetingRisk: the semi-parametric cumulative incidence function. R package version 1.0. https://CRAN.R-project.org/package=CompetingRisk; 2017.
16. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc*. 1999;94:496-509.
17. Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Ann Stat*. 1988;16:1141-1154.
18. Geskus RB. Cause-specific cumulative incidence estimation and the Fine-Gray model under both left truncation and right censoring. *Biometrics*. 2011;67:39-49.
19. Geskus RB. *Data Analysis with Competing Risks and Intermediate States*. Boca Raton, FL: Taylor & Francis Group, LLC; 2016.
20. Lambert PC, Wilkes SR, Crowther MJ. Flexible parametric modelling of the cause-specific cumulative incidence function. *Stat Med*. 2017;36:1429-1446.
21. Koller MT, Raatz H, Steyerberg EW, Wolbers M. Competing risks and the clinical community: irrelevance or ignorance. *Stat Med*. 2012;31:1089-1097.
22. Verweij PJ, Van Houwelingen HC. Cross-validation in survival analysis. *Stat Med*. 1993;12:2305-2314.
23. Xu R, Vaida F, Harrington DP. Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models. *Stat Sin*. 2009;19:819-842.
24. Volinsky CT, Raftery AE. Bayesian information criterion for censored survival models. *Biometrics*. 2000;56:256-262.
25. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a dataset via the gap statistic. *J R Stat Soc, Ser B*. 2001;63:411-423.
26. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc. Ser B (Methodological)*. 1996;58(1):267-288.
27. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. *Ann Stat*. 2006;34(3):1436-1462.
28. Zhao Peng, Yu Bin. On model selection consistency of Lasso. *J Mach Learn Res*. 2006;7:2541-2563.
29. Zhang CH, Zhang SS. Confidence intervals for low dimensional parameters in high dimensional linear models. *J R Stat Soc, Ser B*. 2014;76:217-242.
30. van de Geer S, Bühlmann P, Ritov Y, Dezeure R. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann Stat*. 2014;42:1166-1202.
31. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2006;101:1418-1429.
32. Freund Y, Schapire RE. A desicion-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci*. 1997;55:119-139.
33. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29:1189-1232.
34. Mallat SG, Zhang Z. Matching pursuits with time-frequency dictionaries. *IEEE Transa Signal Process*. 1993;41:3397-3415.
35. Bühlmann P. Boosting for high-dimensional linear models. *Ann Stat*. 2006;34:559-583.
36. Bühlmann P, Yu B. Boosting with the L_2 loss: regression and classification. *J Am Stat Assoc*. 2003;98:324-339.
37. Beyersmann J, Latouche A, Buchholz A, Schumacher M. Simulating competing risks data in survival analysis. *Stat Med*. 2009;28:956-971.
38. Wasserman L, Roeder K. High dimensional variable selection. *Ann Stat*. 2009;37:2178-2201.
39. Fan J, Han F, Liu H. Challenges of big data analysis. *Nat Sci Rev*. 2014;1:293-314.

40. Mukherjee R, Pillai NS, Lin X. Hypothesis testing for high-dimensional sparse binary regression. *Ann Stat*. 2015;43(1):352.

41. Zheng C, Dai R, Hari PN, Zhang M. Instrumental variable with competing risk model. *Stat Med*. 2017;36:1240-1255.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: 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;37:3486–3502. https://doi.org/10.1002/sim.7822