Penalized regression for left-truncated and right-censored survival data

Author list

December 23, 2020

Abstract

High-dimensional data are common in the medical field as high-throughput screening, detailed electronic health record (EHR) information, and comprehensive genomics testing are now used to capture the patient profile. Statistical models that attempt to study the effects of many predictors on survival typically implement feature selection or penalized methods to mitigate the undesirable consequences of overfitting. In some cases survival data is also left truncated which can create an immortal time bias, but penalized survival methods that adjust for left truncation are unavailable in common software. To address these challenges, we implement a penalized Cox proportional hazards model for left-truncated and right-censored survival data and assess implications of left truncation adjustment on bias and inference. We use simulation studies and examples from high-dimensional real-world EHR and genomics data to highlight the pitfalls of failing to account for left truncation in survival modeling. The extension of survival modeling to left truncated data is available in the latest release of **glmnet**.

Keywords— left truncation, penalized regression, Lasso, high-dimensional data, survival analysis, Cox model

1 Introduction

The recent development and accessibility of genomic (DNA sequencing, microarrays), proteomic (immunohistology, tissue arrays), digital (wearables), and imaging (PET, fMRI) technologies have led to an increasing complexity and dimensionality of patient health information. The level of person-level and disease-specific detail captured by these technologies offers an exciting new opportunity to understand patient history and prognosis by combining these molecular data with macro-scale clinical data on patient characteristics, treatment, and outcomes. Specifically, these data may offer insights into risk factors affecting patient-related outcomes leading to interest in developing prognostic tools to improve clinical decision-making and patient care (Gui and Li 2005; Wishart et al. 2010; Ow and Kuznetsov 2016; Yousefi et al. 2017). Given the hundreds or even thousands of molecular, patho-physiological, and clinical parameters that may be readily available for each patient, developing an appropriate statistical model that is robust to overfitting is critical to ensure

good model performance and generalizability. Survival analysis is of particular interest given that overall survival is often a primary outcome of interest.

In a high-dimensional setting, computationally efficient regularization and shrinkage techniques have been developed (Tibshirani 1997; Friedman et al. 2010; Simon et al. 2011) for the Cox model. Similar to ordinary least squares regression (OLS), these penalized regression methods estimate regression coefficients by minimizing the sum of squared errors, but place a constraint in the equation to minimize the sum of coefficients. This constraint (controlled by a parameter, λ) penalizes a high number of variables in the model and shrinks the coefficients of less influential features towards zero - acting, in cases when some coefficients are shrunk fully to zero, as automatic feature selection. Common penalized approaches include lasso (Tibshirani 1996), ridge (Hoerl and Kennard 1970a,b), and elastic net (Zou and Hastie 2005), with extensions such as the adaptive lasso (Zou 2006).

A common challenge in survival data is that patients are often included in the data only after having been at risk (e.g. diseased) for some time so that patients with shorter survival times are unobserved. In other words, the data are left truncated because patients are only included if they survive beyond some initial milestone that marks entry into the observed data. An analysis of the observed patients is subject to an immortal time bias because observed patients cannot die prior to entering the study cohort (Lévesque et al. 2010; Giobbie-Hurder et al. 2013). An example in medicine is a dataset describing patients who receive a specific diagnostic or genetic test; patients who die before having the chance to be tested are excluded from the data so their survival times do not contribute to overall survival estimates. Traditional survival models are inappropriate in this case and predicted survival probabilities are overestimated because of the exclusion of short survival times.

Left truncation is a well known feature of medical data and there are methods to adjust for it (Kalbfleisch and Prentice 2011). The Kaplan-Meier estimator of the survival function requires almost no modifications as it is only necessary to change the risk set so that individuals are only at risk of an event after study entry (Tsai et al. 1987). Similarly, with left truncation, the likelihood function in a parametric model can be adjusted so that it is conditional on survival to the entry time.

Here, we implement a penalized Cox proportional hazards for left-truncated and right-censored survival data and critically assess scenarios in which bias arises from and is exacerbated by left truncation. We apply the method to simulated data and a real-world use case from the Flatiron Health and Foundation Medicine Clinico-Genomic Database (CGDB). Our results highlight the importance of adjusting for left truncation for both inference and prediction. Furthermore, we show that use of inappropriate metrics for model evaluation can result in overly optimistic conclusions about model performance. For implementation of the proposed method, we introduce the left truncation extension to survival modeling in the R package **glmnet** v.X.X (CITE LATEST PACKAGE).

The remainder of this paper is organized as follows: In Section 2, we introduce the penalized regression and left truncation framework and present the efficient method that adjusts for both. Section 3 presents the simulation results and in Section 4 the method is applied to a real-world dataset combining electronic health records and genomic testing. Finally, we discuss general takeaways and survival modeling recommendations in Section 5.

2 Methods

2.1 Data and notation

Consider a typical survival analysis framework in the absence of truncated survival times. For each patient i, we have data of the form y_i, e_i, x_i where y_i is the observed survival time, defined as the time between baseline and event, e_i is the event or failure of interest (e.g. death), and $x_i = (x_{i,1}, x_{i,2} \ldots, x_{i,p})$ is a vector of covariates of length p used in the regression model. The event or failure of interest e_i is 1 if the event is observed or 0 if the event is right-censored. Suppose there are m unique event times, and let $t_1 < t_2 < \ldots < t_m$ be the ordered, unique event times. In this framework, all subjects are assumed to be at risk of the event starting at a commonly-defined "time zero" (a milestone that serves as a reference point for survival such as date of diagnosis) and observed over the interval $[0, T]$ where T is the maximum follow-up time. Let $j(i)$ be the index of the observation with an event at time t_i. A Cox proportional hazards model is commonly used to model the relationship between survival times y_i and predictors x_i assuming a semi-parametric hazard of the form

$$h_i(t) = h_0(t)e^{x_i^\mathsf{T}\beta} \tag{1}$$

where $h_i(t)$ is the hazard for patient i at time t, $h_0(t)$ is the baseline hazard shared across subjects, and β is a length p vector of predictor coefficients.

Cox (1972) proposed a partial likelihood for β without involving the baseline hazard $h_0(t)$

$$\prod_{i=1}^{m} \frac{e^{x_{j(i)}^\mathsf{T}\beta}}{\sum_{j(i)\in R_i} e^{x_{j(i)}^\mathsf{T}\beta}} \tag{2}$$

where $R_i = \{j : y_j \geq t_i\}$ denotes the set of indices $j(i)$ who are "at risk" for failure at time t_i, called the risk set. Maximizing the partial likelihood solves for β and allows for inferences on the relationship between survival time and the set of predictors.

2.2 Left truncation

In the absence of left truncation, the risk set $R_i = \{j : y_j \geq t_i\}$ assumes that every individual is at risk of an event at time 0 and continues to be at risk until the event occurs or the individual is censored. In this sense, all individuals enter the risk set at the same time and leave only at death or censoring. However, when survival times are left truncated, individuals are truly at risk prior to entering the study (e.g. prior to being observed for follow-up). Take Figure 1 as an example.

Survival is defined as the time from diagnosis until death, but the study does not begin until 2011 (dashed blue vertical line). Patients 2 and 5 (in red) are excluded from the analysis because of death and a right censoring event, respectively, prior to study eligibility. A naive analysis on this data would therefore be

Figure 1: Patient follow-up. Left-truncated, right-censored patient follow-up in a hypothetical study cohort. Patients become eligible to enter the study after diagnosis (closed circle) and are followed until death (open circle) or censoring (cross). However, patients who die before the start of the study (dashed blue line) are left truncated (in red), and only those who have survived until the study start (in black) are observed.
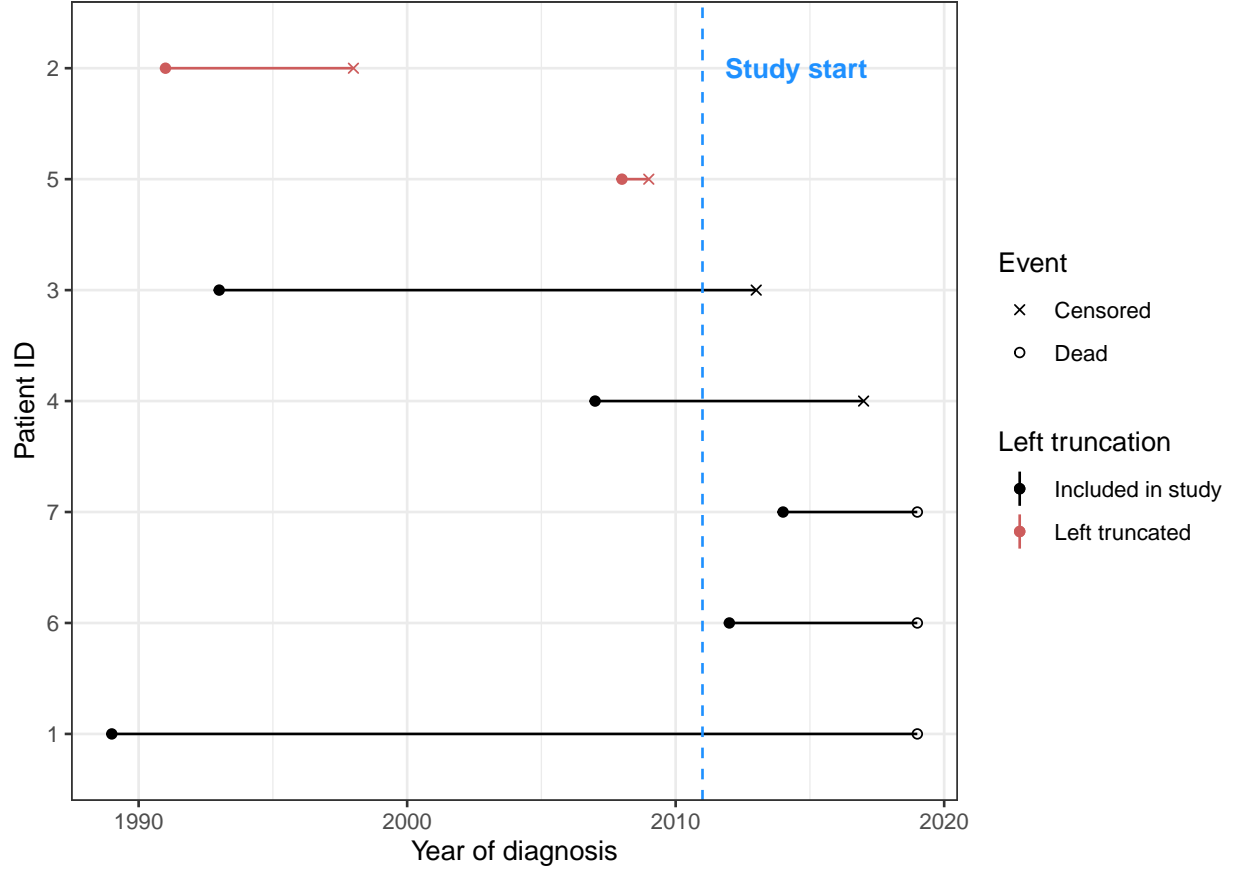
biased because it only contains survival time for "super-survivors" (those who have survived a long time) and the recently-diagnosed.

Fortunately, the partial likelihood can be easily modified to make R_i the set of individuals who are at risk (have entered the study and are not censored or failed) at that time. Let individual i have survival data of the form (v_i, y_i, e_i, x_i), where individual i enters the study at time v_i. The left-truncated risk set $R_{i*} = \{j : y_j \geq t_i > v_j\}$. In other words, subjects are only considered at risk in the analysis after they have been observed to be at risk (e.g. exceeded their left truncation time). Here, event times $t_1 < t_2 < ... < t_m$ are ordered relative to the earliest chronological event time observed in the patient cohort, rather than assuming that all patients are at risk starting at the same referent time 0.

The partial likelihood in Equation 2 can then be evaluated by substituting R_{i*} to make inferences on β in the case of left truncation.

2.3 Penalized regression with left-truncated and right-censored data

Penalized regression methods such as l_1 (lasso) (Tibshirani 1996) and l_2 (ridge) (Tikhonov 1963) regression, elastic net regression (Zou and Hastie 2005), and SCAD (Smoothly Clipped Absolute Deviation) (Xie and Huang 2009) have been developed to overcome the limitation of fitting a model to a large number of covariates. In a high dimensional feature space, where the number of predictors is larger than the sample size (i.e., $p > n$), a model may exhibit low bias but suffer from poor prediction accuracy and limited generalizability, a consequence of overfitting the data. Penalized approaches, which add a regularization penalty to constrain the size of the β, reduce the complexity of the model and mitigate the negative effects of overfitting.

The extension of penalized regression methods to survival data have been extensively described and applied, for example in microarray gene expression data (Gui and Li 2005), transcriptomes (Wu et al. 2011), and injury scale codes (Mittal et al. 2013). Like linear regression, the performance and fit of a Cox proportional hazards model is known to deteriorate with large p, providing an advantage to penalized approaches in high-dimensional, time-to-event feature sets.

Simon et al. (2011) introduced a fast, pathwise algorithm to fit a regularized Cox proportional hazards model solving for

$$\hat{\beta} = argmax_\beta \left[\frac{2}{n} \left(\sum_{i=1}^{m} x_{j(i)}^\intercal \beta - log \left(\sum_{j(i) \in R_i} e^{x_{j(i)}^\intercal \beta} \right) \right) - \lambda P_\alpha \beta \right] \tag{3}$$

where β is the vector of regression coefficients and is maximized over Equation 3 subject to a constraint defined by the penalty $\lambda \cdot P_\alpha \beta$. As described elsewhere (Simon et al. 2011), the penalty can take the form of the l_1 (lasso), l_2 (ridge), or elastic net penalty, and the algorithm fits over a path of λ values via cyclical coordinate descent. Instead of solving a least squares problem at each iteration, the algorithm solves a penalized reweighted least squares problem, in which observation weights w are defined in the coordinate descent least squares minimization step. The optimal λ can be identified through cross-validation. Tied survival times are handled via the Breslow approximation (Breslow 1972) and details provided in (Simon et al. 2011).

Applying the standard penalized Cox implementation in the software **glmnet** to left-truncated survival data will result in biased estimates. However, the extension to adjust for left truncation is straightforward. Given survival data with left truncation and right censoring, we solve Equation 3 with a modification to the risk set R_i:

$$\hat{\beta} = argmax_\beta \left[\frac{2}{n} \left(\sum_{i=1}^{m} x_{j(i)}^\intercal \beta - log \left(\sum_{j(i) \in R_{i*}} e^{x_{j(i)}^\intercal \beta)} \right) \right) - \lambda P_\alpha \beta \right] \tag{4}$$

where the left-truncated risk set $R_{i*} = \{j : y_j \geq t_i > v_j\}$; that is, it considers subjects at risk only after their left truncation times. The method can take lasso, ridge, and elastic net penalties for $\lambda P_\alpha \beta$ defined

elsewhere ([Friedman et al. 2010](#)) and is a simple extension of default methods in **glmnet**. The efficient implementation of this method is non-trivial, as the usual updating of risk set components cannot be used when left truncation is present. This implementation is available in the release of **glmnet** v. X.X (CITE LATEST PACKAGE).

2.4   Real-world use case

We consider the Clinico-Genomic Database (CGDB) offered jointly by Flatiron Health and Foundation Medicine. The CGDB is an oncology research platform that combines real-world, patient-level clinical data and outcomes with patient-level genomic data to provide a comprehensive patient and tumor profile. Patients in the CGDB have received at least one Foundation Medicine tumor genomic test as well as have had their electronic health records captured by Flatiron Health. Together, these data sources represent thousands of potential predictors per patient, ranging from demographic, lab, and treatment information to the specific mutations detected on each biomarker included in a Foundation Medicine baitset.

We take a subset of $4,429$ stage IV, non-small cell lung cancer (NSCLC) patients in the CGDB as a use case for this study. The outcome of interest is survival time predicted from the date of Stage 4 diagnosis. Survival times are left-truncated because patients who die before receiving a Foundation Medicine test are logically not included in the cohort of Foundation Medicine test recipients.

As shown in [Figure 2](#) the distribution of left truncation times (time from diagnosis to genomic testing) is right-tailed with approximately 21.5% of patients diagnosed more than one year prior to receiving the Foundation Medicine genomic test. Moreover, the correlation between diagnosis year and left truncation time is strongly negative ($\rho = -0.63$), suggesting the presence of left truncation survivor bias as patients with records in the CGDB diagnosed longer ago needed to survive more years before being tested.

Next, we use these data as the basis for a simulation study and a real-world application.

Figure 2: Distribution of left truncation time (days) in non-small cell lung cancer patients in the CGDB
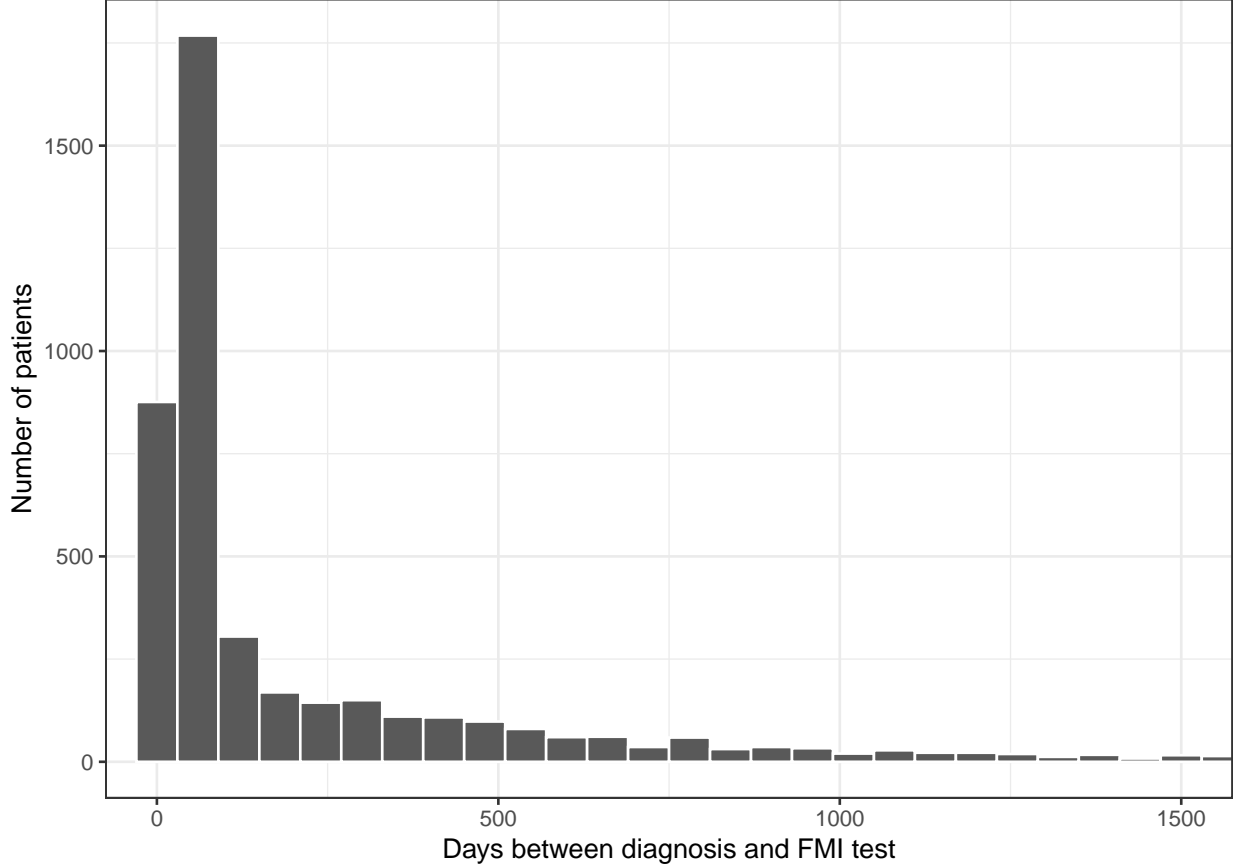
3 Simulation study

Simulations were conducted to assess the performance of the method with left truncated survival data. To set notation, let T_i, U_i, and V_i be latent survival, right censoring, and study entry times for patient i, respectively. The simulation proceeds by simulating T_i, U_i, and V_i from suitable survival distributions and assumes that T_i, U_i, and V_i are uncorrelated. The observed survival time is given by $Y_i = \min(T_i, U_i)$ and a patient is right-censored if $T_i > U_i$. Patients are left-truncated (i.e., hidden from the sample) if $V_i > Y_i$.

3.1 Data generating process

We generated n patients and p predictors to construct a $n \times p$ matrix, X. The first 11 predictors were from the CGDB and the remaining $\tilde{p}$ variables were simulated binary variables. The CGDB variables were the same 11 variables used in the small p model in Section 4 and generated by randomly sampling rows of the original data with replacement. The matrix of binary variables, $\tilde{X}$ was simulated using a (latent) multivariate probit approach

$$\begin{pmatrix} \tilde{x}_{i,1}^* \\ \vdots \\ \tilde{x}_{i,\tilde{p}}^* \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_{\tilde{p}} \end{pmatrix}, \Sigma \right), \tag{5}$$

with the kth binary variable $\tilde{x}_{i,k} = 1$ if $\tilde{x}_{i,k}^* > 0$ and 0 otherwise. Each diagonal element of Σ equals 1 and non-diagonal elements determine the correlation between variables. The probability, π_k that variable k was equal to 1 was randomly sampled from a uniform distribution ranging from .2 to .8; that is, we sampled $\pi_k \sim U(0.2, 0.8)$ and set $\mu_k = \Phi^{-1}(\pi_k)$ where $\Phi(\cdot)$ is the cumulative density function of the normal distribution. Σ was generated by filling a $\tilde{p} \times \tilde{p}$ matrix, $\tilde{\Sigma}$, with $N(0,1)$ random variables and setting $\Sigma = S/D^T D$ with $S = \tilde{\Sigma}^T \tilde{\Sigma}$ and D a vector containing the diagonal elements of S.

Latent survival time was simulated from a proportional hazards Weibull distribution with shape parameter a and scaled parameter m_i based on evidence from the CGDB as detailed in Section 3.2. Covariates predict latent survival time through their effect on the scale parameter,

$$m_i = \exp(\alpha + x_i^T \beta), \tag{6}$$
$$T_i \sim \text{Weibull}(a, m_i), \tag{7}$$

where α is an intercept and β is a vector of coefficients. α and values of β for the CGDB variables were estimated from the CGDB data. The probability that the coefficient of a binary variables was non-zero was 0.5 and drawn from a Bernoulli distribution. Non-zero coefficients were drawn from a uniform distribution with lower bound of -0.5 and upper bound of 0.5 so that the minimum and maximum hazard ratios were 0.61 and 1.65, respectively.

Time to right censoring was also sampled from a Weibull distribution but did not depend on covariates,

$$U_i \sim \text{Weibull}(\tilde{a}, \tilde{m}). \tag{8}$$

Lastly, a two-part model was used to simulate time to study entry so that a proportion $\tilde{\pi}$ of patients could enter the study at $t = 0$ while a proportion $1 - \tilde{\pi}$ would enter at $t > 0$. The probability of an entry time greater than 0 was drawn from a Bernoulli distribution and positive entry dates were modeled using a lognormal distribution.

$$d_i \sim \text{Bernoulli}(\tilde{\pi}) \tag{9}$$
$$V_i^* \sim \text{Lognormal}(\mu, \sigma^2) \tag{10}$$
$$V_i = \begin{cases} 0 \text{ if } d_i = 0 \\ V_i^* \text{ if } d_i = 1 \end{cases} \tag{11}$$

3.2 Calibration

The simulation was calibrated to be consistent with the CGDB NSCLC data. The Nelson-Aalen estimator of the cumulative hazard of overall survival was generally monotonically decreasing, which suggested that a Weibull distribution could adequately capture the baseline hazard. This was further supported by the similarity in fit between a spline-based survival model with one internal knot at the median of uncensored survival time and the Weibull model. As such, an intercept only Weibull model was used to estimate a and α in the simulation. Similarly, a Weibull distribution was used to model time to right censoring (by reversing the censoring indicator in the model for overall survival. Both the model for overall survival and right censoring adjusted for left truncation.

Coefficients for the CGDB variables were estimated using a Cox proportional hazards model. The full coefficient vector combined the coefficients from the CGDFB variables with the coefficients from the simulated binary variables. The covariates were centered, $m = \alpha + x^T\beta - E(x^T\beta)$, so that predicted survival was consistent with the intercept only Weibull model.

The latent time to study entry is inherently unobserved, but we believe a two-part model can characterize the large fraction of tests in the observed data that occur near the diagnosis date (Figure 2). The probability of a positive entry time was set equal to 0.2 and the lognormal distribution for positive entry times was derived using a right skewed distribution with a median value of 1 and a mean value of 1.6, both measured in years. We found that this parameterization generated median survival times in the simulation that were similar to median survival in Kaplan-Meier estimates that did not adjust for left truncation in the CGDB.

3.3 Implementation

Datasets were simulated where each row contained information on a patient's observed survival time (y_i), a censoring indicator (e_i), the time of study entry (v_i). 75% of patients were assigned to a training set and the remaining 25% were assigned to a test set. Within the test set, Cox models with lasso penalties were fit among the observed (i.e., non-truncated) patients. Cross-validation with 10 folds was used to tune the model and the value of λ that minimized the partial likelihood deviance was selected. Models were fit with and without accounting for left truncation.

Predictions and model evaluations were performed on the test set. Testing was performed on both the "observed" sample, (non-truncated patients) and the "complete" sample (both truncated and non-truncated sample). The complete sample provides a true test of the method because there is non bias due to truncation.

This process was repeated 20 times. For each iteration, models were evaluated based on a dataset consisting of $n = 5,000$ patients. Simulations were peformed in both big and small p scenarios. In the small p scenario, 10 binary variables were simulated in addition to the 11 CGDB variables so that $p = 21$; in the big p scenario, $1,000$ binary variables were simulated so that $p = 1,011$.

3.4 Results

We used calibration curves (Harrell Jr 2015) to compare models with and without a left truncation adjustment. Each point in the plot consists of simulated patients in a given decile of predicted survival at

a given time point. The x-axis is the averaged predicted survival probability and the y-axis is the pseudo observed survival probability computed using the Kaplan-Meier estimator. Perfectly calibrated predictions are those that lie along the 45 degree line so that predicted and observed survival probabilities are equal. The plots consist of patients in the complete sample and are therefore a good test of the effectiveness of the left truncation adjustment.

Calibration curves are displayed in Figure 3 for both the big and small p scenarios. Predicted survival probabilities from the model that does not adjust for left truncation are too high because the model is fit on the observed sample consisting only of patients that survived long enough to enter the sample. In other words, failing to adjust for left truncation results in an overestimation of survival probabilities. In contrast, the left-truncated adjusted model corrects for this bias. The small p model is almost perfectly calibrated with points along the 45 degree line.

The results of the big p scenario are similar, although calibration is more difficult than in the lower dimensional setting with points lying farther from the 45 degree line on both extremes. Predictions tend to be more accurate for patients with predicted survival close to 0.5, but underestimated for patients with higher survival probabilities and overestimated for patients with lower survival probabilities. Probability rescaling has been suggested in such cases (e.g. characteristic sigmoid-shaped calibration curves produced by different learning algorithms), for example with the use of Platt scaling (Platt 1999) or isotonic regression (Niculescu-Mizil and Caruana 2005).

A common metric used to evaluate survival models is the C-index (Harrell Jr et al. 1996), which provides a measure of how well a model will discriminate prognosis between patients. We computed the C-index using predictions from the Cox model with lasso penalty on high dimensional data. The results are presented in Table 1 and suggest that care should be taken when using the C-index to evaluate survival models in the presence of left-truncated data. When using the observed sample to evaluate the model, the C-index is higher in model that does not adjust for left truncation, even though it is poorly calibrated. Furthermore, even when using the complete sample, the C-indices are very similar.

Table 1: Comparison of model discrimination in the simulation

Left truncation adjustment	Test sample	C-index
No	Observed	0.70
Yes	Observed	0.65
No	Complete	0.65
Yes	Complete	0.65

Cox models with lasso penalty were fit using the high-dimensional simulated data.

These results are perhaps not surprising. For example, when computing the C-index in the observed sample the model is evaluated in a test set that suffers from the same bias as the training set. Moreover, the similarity in the values of the C-index in the complete sample may be caused by the assumption of no direct correlation between the predictors and latent time to study entry. In other words, the bias in the coefficients is based only on the indirect correlation between the predictors and left truncation induced by (i) the correlation between the predictors and survival and (ii) the correlation between survival and the probability a patient is

10

Figure 3: Calibration of survival predictions for lasso model in simulation
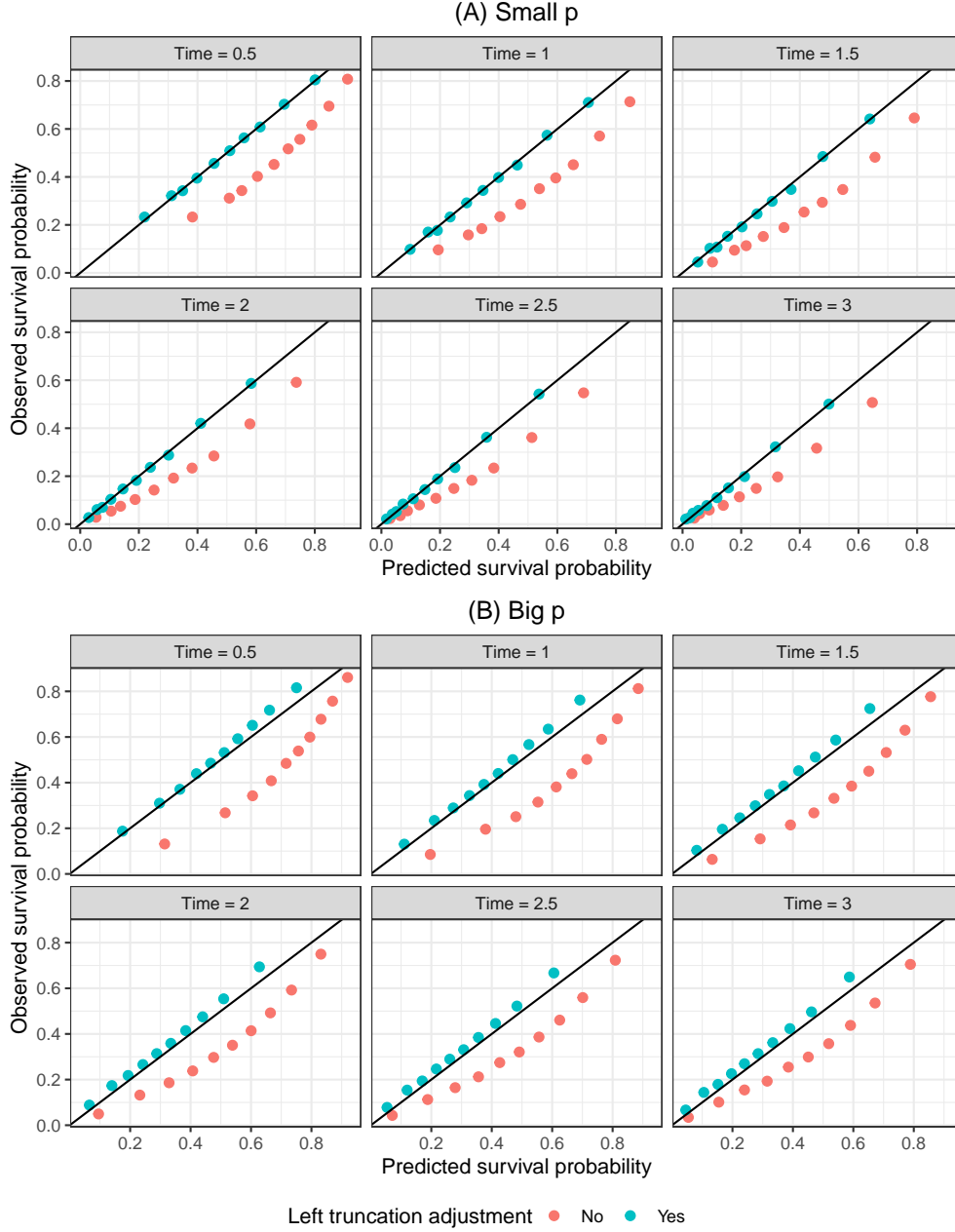
Notes: The Cox model with lasso penalty using the training data and was subsequently used to predict the survival function for each patient in the test set. The small and big p models contained 21 and 1,011 predictors, respectively. Patients were divided into deciles at each time point based on their predicted survival probabilities. Each point in the plot represents patients within a decile. "Predicted survival probability" is the average of the predicted survival functions from the Cox model across patients within each decile and the "Observed survival probability" is the Kaplan-Meier estimate of the survival function within each decile. A perfect prediction lies on the black 45 degree line.

left-truncated. As we will see in Section 4, the impact of left truncation adjustment on the estimated values of the coefficients can be considerably larger if there is a more direct correlation.

4 Real-world data application: Non-small cell lung cancer patients in the Clinico-Genomic Database (CGDB)

We fit models using 4,429 stage IV, NSCLC patients. 75% of patients were assigned to a training set and the remaining 25% were assigned to a test set. The training set was then used to train and tune both small p and large p models with 13 and 639 predictor variables, respectively. Both un-penalized and lasso penalized Cox models were fit, with and without a left truncation adjustment. In the penalized models, the optimal value of the shrinkage parameter, λ, was selected using the value that minimized the partial likelihood deviance in 10-fold cross-validation. The models were evaluated by assessing predictions on the test set.

In this real-world application, both small p and large p models were designed to highlight left truncation in a real-world dataset and its impact on inference, as opposed to optimizing predictive performance and the prognostic value of predictors. For this reason, models were kept non-complex and relatively uninformative, e.g. modeling linear and additive effects, applying no transformations to predictors (for instance, modeling mutations as binary), and including a variable, year of diagnosis, that is highly correlated with left truncation time but not necessarily with survival (Figure 2). Survival was predicted from diagnosis. For an evaluation of meaningful prognostic clinical and genomic factors in NSCLC, see for example Yu-Heng et al. (2020).

Specifically, the smaller model consisted of the following variables: year of diagnosis, age at diagnosis, race, practice type (academic or community center), and the mutational status (mutated/not mutated) of 3 known prognostic biomarkers for NSCLC (TP53, KRAS, and EGFR). For each biomarker, 3 alteration types were considered: copy number (CN), short variant (SV), and rearrangement (RE). 3,022 patients remained in the training set after dropping patients with missing values on at least one covariate.

The larger model included all non-biomarker variables from the smaller model and added all mutations tested by Foundation Medicine. As the data come from both historic and current clinico-genomic testing products, the set of tested genes has changed over time. To resolve the resultant missing data problem, we constrained the mutation data only to columns with less than 50% missing indicators and imputed the remaining mutation data using k nearest neighbors (Troyanskaya et al. 2001), with k = 20.

Figure 4 displays hazard ratios for each model. The plot for the large p only contains coefficients ranked in the top 10 by absolute value of the hazard ratio in either the left truncation adjusted or non-adjusted model. The arrows denote the impact of adjusting for left truncation on the hazard ratio. In some cases the adjustment can have a large impact on the hazard ratios, often moving a large hazard ratio all the way to 1 (i.e., to a coefficient of 0). This is most notable with diagnosis year, which has a hazard ratio above 2 in both the small and large p models but is set to 1 after the left truncation adjustment. This effect is likely due to the strong negative correlation between calendar time and the probability that a patient is left truncated.

Table 2 and Figure 5 evaluate the models based on predictions on the test set. Table 2 compares models using the C-index. The C-index is relatively low in all of these illustrative models, though is artificially higher in the unadjusted models that do not account for left truncation - sometimes considerably so - and decreases after adjustment as the predictor hazard ratios are pushed toward 1 and there are fewer differences in predicted relative risks across patients. As expected, the lasso model has a higher C-index in the larger p

setting as the Cox model is considerably overfitted.

Table 2: Comparison of model discrimination in the CGDB

		C-index	
Model	Covariate size	No adjustment	Adjustment
Cox	Small	0.647	0.579
Cox (lasso)	Small	0.647	0.576
Cox	Big	0.603	0.573
Cox (lasso)	Big	0.665	0.609

Note: "Adjusted" models incorporate left truncation times.

Figure 5 displays the same survival calibration curves presented in the simulation section. As in the simulation study, the unadjusted models overpredict survival probabilities. In contrast, the adjusted model tends to underpredict survival, which contrasts from the well calibrated adjusted models from the simulations. Here, the adjusted models tend to be adequately calibrated for patients with lower survival probabilities but less well calibrated for patients with higher survival probabilities. One possible reason for this result is that the adjusted models pushed many of the hazard ratios to 1 and do not adequately separate low and high risk patients.

Figure 4: Hazard ratios from Cox lasso model

(A) Small p

(B) Large p

Left truncation adjustment ● No ● Yes
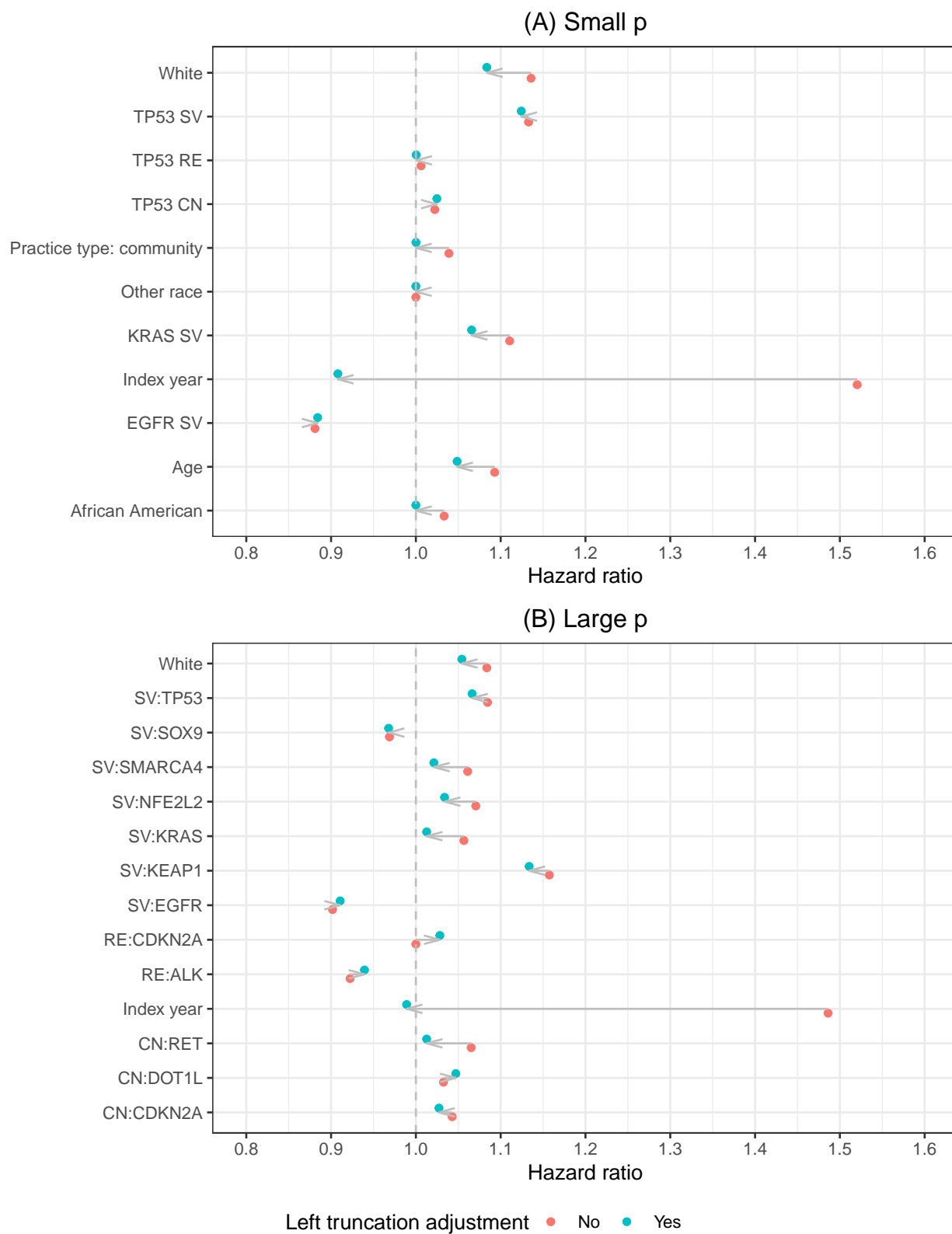
Notes: The figure for the "Large" p model only includes variables ranked in the top 10 by the absolute value of the hazard ratio in either the left truncation adjusted or non-adjusted model.

Figure 5: Calibration of survival predictions in the CGDB from the Cox lasso model

(A) Small p

(B) Large p

Left truncation adjustment ● No ● Yes
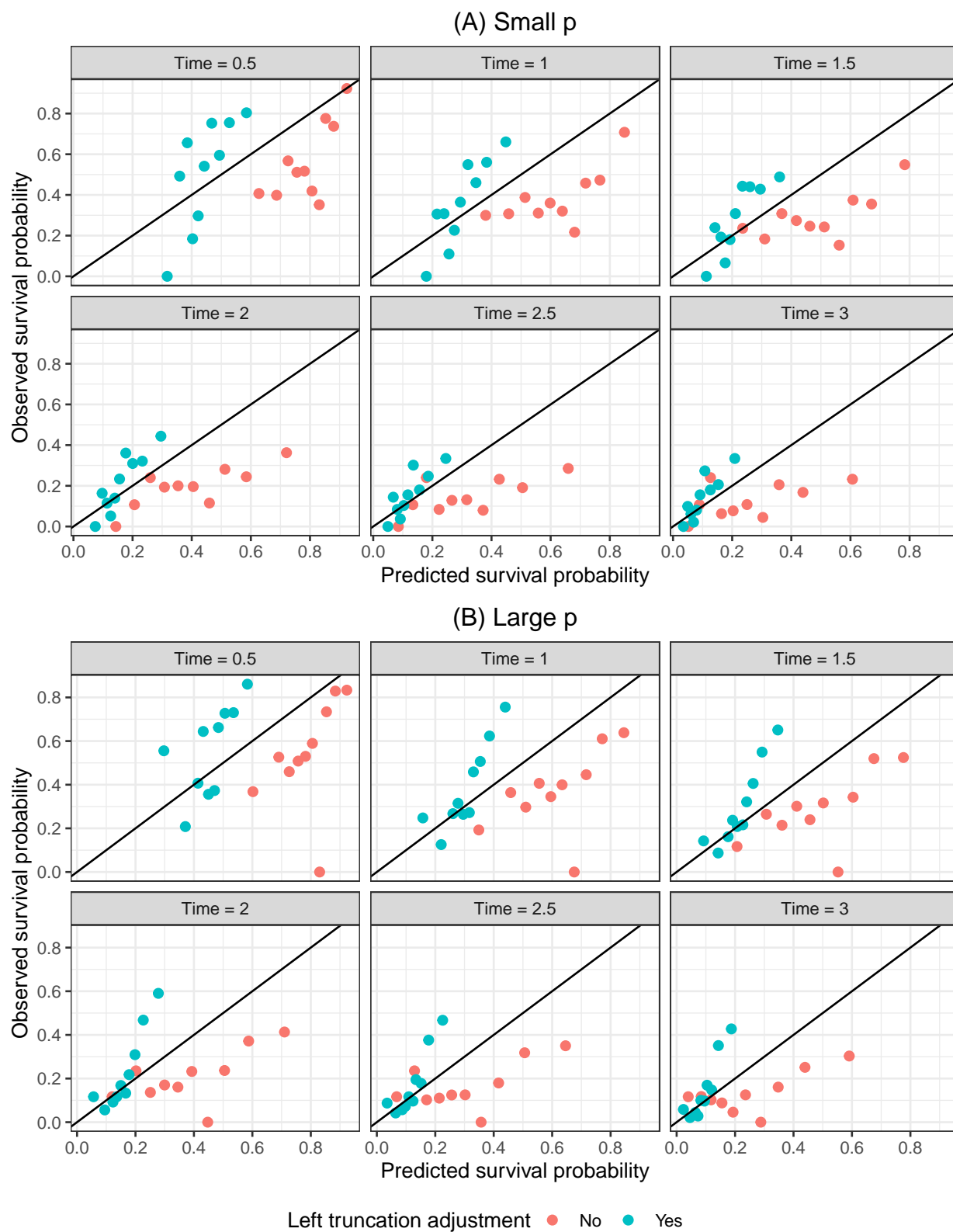
5   Discussion

In this paper, we implement an efficient approach for a left truncation-adjusted, penalized Cox proportional hazards model and critically assess its impact on bias and inference in left truncated data. Using simulation studies and examples from real-world EHR and genomics data, we showed that there is a need for approaches that can both adjust for left truncation and model high dimensional data. In particular, our simulation shows that predictions from models that fail to adjust for left truncation will overestimate true survival probabilities whereas models that properly adjust can yield well calibrated survival predictions, even with high-dimensional data. Moreover, in our real-world use case, there were significant differences in coefficient estimates between models that did and did not adjust for left truncation.

Our results have important implications for specification, interpretation, and evaluation of prognostic survival models. Much of the difficulty arises because training and test sets are both subject to the same selection bias. In other words, a careful understanding of the data generating process is crucial and naive approaches can lead analysts astray. For example, our simulations suggest that commonly used C-index cannot differentiate between models that mimic the true data generating process and those that do not and may therefore lead to misleading conclusions in the presence of left truncated data. On the other hand, metrics that focus on predicted survival probabilities such as calibration curves are better able to determine whether a model can accurately predict the risk of an event.

In our real-world use case, an analyst could have mistakenly concluded that the high C-index in the non-adjusted model was indicative of good model performance when it was in fact driven primarily by a high correlation between predictors and left truncation. For instance, in our large p model the hazard ratio for the predictor "year of diagnosis" moved from being the largest in absolute value to 1 after adjusting for left truncation. The high C-index in the non-adjusted model was artificial because patients diagnosed earlier appeared to survive longer due to immortal time bias.

Although not the primary focus of this paper, our approach generated findings that were both consistent with prior literature in genomics and clinically plausible. One of the mutation features that had the highest importance was the presence of short variants in the gene KEAP1. This finding is supported by similar reports of risk-conferring status in an independent lung cancer cohort (Shen et al. 2019). Conversely, variants in EGFR and ALK were found to be protective in the CGDB data, putatively because mutations in these genes qualify patients to receive targeted therapies and therefore benefit from improved response and longer survival.

We conclude with several recommendations. When making predictions with survival data is of interest, it is critical to define the risk set appropriately. To do this requires recognizing the underlying process by which individuals are captured in the data in the first place; i.e., when and how individuals came to be included in the database of interest. One of the motivations for the current work derives from the advent of biobank and genomic test databases that store information on patient illness - presumably after the patient has been diagnosed and "at risk" for some time. In these data, overlooking this "delayed entry" of at-risk patients would lead to adverse consequences including misleading conclusions, as we demonstrate in this work. In addition, we recommend that models should be evaluated using both discrimination and calibration based metrics.

The implementation of this approach is available in **glmnet** v.X.X.X on CRAN.

References

Breslow, N. E. (1972). Contribution to the discussion of the paper by d.r. cox. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):216–217.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Giobbie-Hurder, A., Gelber, R. D., and Regan, M. M. (2013). Challenges of guarantee-time bias. *Journal of clinical oncology*, 31(23):2963.

Gui, J. and Li, H. (2005). Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21(13):3001–3008.

Harrell Jr, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis.* Springer.

Harrell Jr, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387.

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*, volume 360. John Wiley & Sons.

Lévesque, L. E., Hanley, J. A., Kezouh, A., and Suissa, S. (2010). Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *Bmj*, 340:b5087.

Mittal, S., Madigan, D., Cheng, J., and Burdc, R. (2013). Large-scale parametric survival analysis. *Statistics in Medicine*, 32(23):3955–3971.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. *ICML '05: Proceedings of the 22nd International Conference on Machine Learning*.

Ow, G. S. and Kuznetsov, V. A. (2016). Big genomics and clinical data analytics strategies for precision cancer prognosis. *Scientific reports*, 6:36493.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. *Advances in Large Margin Classifiers*, pages 61–74.

Shen, R., Martin, A., Ni, A., Hellmann, M., Arbour, K. C., Jordan, E., Arora, A., Ptashkin, R., Zehir, A., Kris, M. G., Rudin, C. M., Berger, M. F., Solit, D. B., Seshan, V. E., Arcila, M., Ladanyi, M., and Riely, G. J. (2019). Harnessing Clinical Sequencing Data for Survival Stratification of Patients with Metastatic Lung Adenocarcinomas. *JCO precision oncology*, 3.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39(5):1.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395.

Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics*, 4:1035–1038.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics (Oxford, England)*, 17(6):520–525.

Tsai, W.-Y., Jewell, N. P., and Wang, M.-C. (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, 74(4):883–886.

Wishart, G. C., Azzato, E. M., Greenberg, D. C., Rashbass, J., Kearins, O., Lawrence, G., Caldas, C., and Pharoah, P. D. (2010). Predict: a new uk prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Research*, 12(1):R1.

Wu, T. T., Gong, H., and Clarke, E. M. (2011). A transcriptome analysis by lasso penalized cox regression for pancreatic cancer survival. *Journal of Bioinformatics and Computational Biology*, 9(1):63–73.

Xie, H. and Huang, J. (2009). Scad-penalized regression in high-dimensional partially linear models. *Annals of Statistics*, 37(2):673–696.

Yousefi, S., Amrollahi, F., Amgad, M., Dong, C., Lewis, J. E., Song, C., Gutman, D. A., Halani, S. H., Vega, J. E. V., Brat, D. J., et al. (2017). Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific reports*, 7(1):1–11.

Yu-Heng, L., Chen, W.-N., Hsu, T.-C., Lin, C., Tsao, Y., and Wu, S. (2020). Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. *Scientific reports*, 10(1):4679.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.