Penalised Cox regression models for survival data

Article *in* International Journal of Medical Engineering and Informatics · January 2017

DOI: 10.1504/IJMEI.2017.10000836

Duo Zhou
Rutgers, The State University of New Jersey

Penalized Cox Regression Models for Survival Data

1. Abstract

In the recent years, the advances in genetic technology have enabled collection of large amount of genetic information; as a result, analysis tools for microarray data have been in high demand. Among all type of data, survival outcome is particularly more interest to study. Such data has become very typical in biomedical research or genetic lab. However, typical statistical models do not work for such scenarios; in this paper, several Cox model based Penalized regression approaches will be evaluated.

2. Keywords

Cox proportional hazard model, nonlinear, survival, prediction errors, AUCs, time-dependent, lasso, ridge, elastic-net, penalized regression

3. Introduction

In the recent years, the advances in genetic technology have enabled collection of large amount of genetic information; as a result, analysis tools for microarray data have been in high demand. Among all type of data, survival outcome is particularly more interest to study. However, if the total number of predictors is more than the total number of observations or the number of events, typical statistical models may not work; if predictors are highly correlated with each other, typical statistical models may not work efficiently. Nevertheless, highly correlated or high dimensional survival data has become very typical in biomedical research or genetic lab. A recent example was the study published by Beer et al. in 2002, which was performed on lung adenocarcinoma microarray expression data, the expression data were collected from 86 subjects with 7129 probe sets[1]. Another similar study published by Bhattacharjee et al. in 2001, was designed to use mRNA expressions to reveal distinct adenocarcinoma subclasses. A third study published by Garber et al. in 2001, was to study the diversity of lung adenocarcinoma with gene expressions. Many other similar studies can be found in literature. Thus efficient analysis tools for highly correlated or high dimensional survival data have become of great interest.

In this paper, several penalized regression approaches were built on top of Cox proportional hazard models, such as lasso[[2, 3, 4] regression as proposed by Tibshirani et a. in 1996, ridge[5] regression as proposed by Marquardt in 1970 and elastic-net[6] regression as proposed by Zou et al in 2005, were studied. The prediction performance of these models was evaluated.

4. Method

The Cox model based penalized regression approaches, were assessed via a simulation study and real world case study. For both studies, the original data were randomly partitioned into training and testing sets with 3:1 ratio; the training set was used for cross validations; the testing set was used for evaluating the prediction performance. The prediction performance was measured using time-dependent prediction errors and time-dependent AUCs.

4.1 Simulation Studies with Time-Varying Treatment Effect

The simulation study consisted of a total of 2000 subjects; 7 factors included age, sex, race, systolic blood pressure (SBP), diastolic blood pressure (DBP), body mass index (BMI) and treatment were simulated from computer programs using R[7] Software. SBP and DBP were simulated from two correlated normal distributions with coefficients of 0.9; and they were correlated with BMI with coefficient of 0.3 (SBP and BMI) and 0.2 (DBP and BMI), respectively. These factors were then used as the covariates to simulate time-to-events outcomes following a pre-specified exponential model.

4.2 Real World Case Study

The study was based on a breast cancer data downloaded from the Netherlands Cancer Institute, which was referred to as NKI70[8] data onward; the outcome is the metastasis-free survival. The data included 5 clinical factors and 70 gene signatures; a total of 144 independent lymph-node-positive breast cancer subjects were followed for 17 months; of these 144 subjects, 48 subjects experienced a metastasis event.

For this study, the number of factors were more than the number of event ($p \gg N$); typical Cox models did not work, penalized Cox regression models were further studied and evaluated.

4.3 Analysis Method

4.3.1 Cox Proportional Hazard (PH) Model

Typical Cox PH model can be expressed as,

$$\lambda(t|X) = \lambda(t)\exp(X\beta) \quad \text{.................. Eq. 1}$$

4.3.2 Penalized regression

For penalized regression, such as lasso, ridge and elastic-net, the penalization terms can be obtained through minimization of the loss function:

$$\bar{\beta} = \operatorname{argmin}\left\{\sum_{j=1}^{N}\left(y_j - \hat{y}_j\right)^2 + \lambda\alpha\sum_{l=1}^{\nu}|\beta_l| + \frac{1}{2}\lambda(1-\alpha)\sum_{l=1}^{\nu}\beta_l^2\right\} \quad \text{................ Eq. 2}$$

Lasso and ridge regression are just special case of this formula; if α=1, it becomes lasso regression; if α=0, it becomes ridge regression; it becomes elastic-net regression, if $0<\alpha<1$.

4.3.3 Interactions and Nonlinearity

For typical Cox regression models, interactions and nonlinearity should be addressed before implementing the analysis model, however for high dimensional data, it is usually impractical to adjust for interactions and nonlinearity due to too many predictors. For simulation study, all factors were adjusted using 5-degree polynomial transformations; pair-wise interactions were constructed between the nonlinear forms of all factors, thus a total of 342 covariate terms were considered. For real world case study, the models with adjustment of interactions and nonlinearity were not assessed, but can be provided upon request.

4.3.4 Evaluation of Model Performance

The model performance are measured using predictions errors and time-dependent AUCs. The prediction errors is calculated using Brier scores[9] as proposed by Gerds et al. (2006). The Brier Score at time, t, is defined as:

$$L(\mathrm{t}) = \frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}}\left\{\frac{\hat{\mathrm{S}}(\mathrm{t}|\mathrm{x_i})^2 \mathrm{I}(\mathrm{t_i} \leq \mathrm{t} \wedge \delta_\mathrm{i} = 1)}{\widehat{\mathrm{G}}(\mathrm{t_i})} + \frac{\left[1 - \hat{\mathrm{S}}(\mathrm{t}|\mathrm{x_i})\right]^2 \mathrm{I}(\mathrm{t_i} > \mathrm{t} \wedge \delta_\mathrm{i} = 0)}{\widehat{\mathrm{G}}(\mathrm{t})}\right\} \quad \text{................Eq. 3}$$

where $Y(t)$ is the observed survival status at time t, which is $I(T_j \geq t)$, $\hat{Y}(t)$ is the predicted survival probability at time t, N is the total number of subjects. The prediction error, L, is used to estimate the prediction performance at time t.

The time-dependent AUC measurement is were very much related to the c-index, as recommended by Harrell (2012), which is the probability of concordant pairs of predicted and observed survival status among all pairs of responses. Specifically, if the predicted survival probability is larger for the subject who (actually) lived longer, the predictions for that pair are said to be concordant with the (actual) outcomes. The c-index is an overall measurement of the prediction performance. But similar to the time-dependent prediction errors, the probability of concordance at each time point can also be evaluated by assigning each subject a survival status at the given time point: if the time point of interest occurs prior to the event time for a specific subject, the subject should be alive (status of 0) at that particular time point; if the time point of interest occurs after the subject had an event, the subject should have already been censoring (status of 1) at the particular time point; otherwise, the subject should have status of 0 at the time point. Consequently, at each time point, the predicted survival probability of each subject from the test set could be checked against the survival status of the same subject at the corresponding time point; averaging all subjects in the test set at the same time point, the probability of all concordant pairs could be obtained for the given time point; then the probabilities of concordance at different time points are referred to as time-dependent AUCs, it can be estimated as

$$AUC(t) = Pr\{\hat{S}_i < \hat{S}_j | T_i < T_j,\ T_i \leq t\} = \frac{\sum_{T_i \leq t} I(\hat{S}_i < \hat{S}_j) \times I(T_i < T_j)}{\sum_{T_i \leq t} I(T_i < T_j)} \quad \text{.................. Eq. 4}$$

where $\hat{\mathrm{S}}_\mathrm{i}$ and $\hat{\mathrm{S}}_\mathrm{j}$ are the predicted survival probability for subject i and j at time t, $i, j = 1, 2, \dots, n$ and $i \neq j$.

For prediction errors and time-dependent AUC measurements, the robust estimates and the corresponding 95 percentile credible intervals (PCI) were obtained based on the mean and the 95 percentiles of the measurements from 1000 bootstrapped samples of the test set, unless noted otherwise.

5. Analysis Results

5.1 Results of the Simulation Study

5.1.1 Summary Statistics of the Simulation Study

The demographics of the study population are summarized in Table 1; as can be seen from the table that all factors were equally distributed between the two randomization groups. A summary of subject survival status at 1-year post baseline and treatment switching after 1-year post baseline is presented in **Error! Reference source not found.**. For this simulation study, a total of 106 (10.6%) subjects from active treatment arm had events prior to 1 year and 163 (16.3%) subjects from the placebo (SOC) arm had events prior to 1 year. Of the 891 subjects randomized to active treatment who survived 1 year, none of the subjects switched to placebo; of the 840 subjects randomized to placebo and survived 1 year, 201 (23.9%) switched to the active treatment, and the rest of the 639 subjects stayed in the placebo (SOC) arm until they censored or failed. In this study, the treatment switching from placebo to active was assumed to be independent of the failure event, which however may not be reasonable in practice; in a clinical study, patients may switch treatment due to adverse events or lack of efficacy.

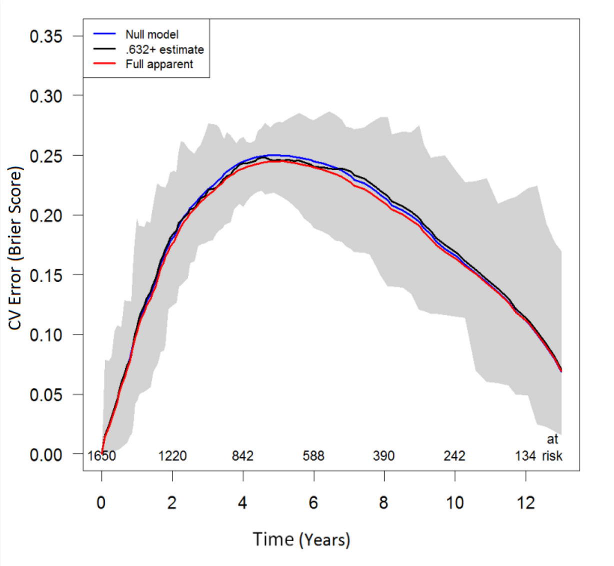
Table 1. Demographics of Simulation Study

Factors	Statistics	Enrollment Randomization: Active Treatment (N=997)	Enrollment Randomization: Placebo SOC (N=1003)	All Enrollment (N=2000)
Age	Mean ± SD	49.8 ± 11.73	50.5 ± 11.73	50.2 ± 11.73
	Median (25%, 75%)	49.4 (42.2, 57.9)	50.3 (42.7, 58.7)	49.7 (42.5, 58.2)
	Min – Max	16.8 – 85.11	15.8 – 94.5	15.8 – 94.5
Sex	Male	596 (60%)	623 (62%)	1219 (61%)
	Female	401 (40%)	380 (38%)	781 (39%)
Race	White	534 (54%)	546 (54%)	1080 (54%)
	Black	215 (22%)	242 (24%)	457 (23%)
	Hispanic	172 (17%)	181 (18%)	353 (18%)
	Asian	76 (8%)	24 (3%)	110 (6%)
SBP	Mean ± SD	110.6 ± 8.64	110.3 ± 8.64	110.4 ± 8.64
	Median (25%, 75%)	110.2 (104.8, 116.5)	110.5 (104.6, 115.7)	110.3 (104.6, 116.1)
	Min – Max	82.9 – 138.6	83.9 – 137.2	82.9 – 138.6
DBP	Mean ± SD	75.2 ± 4.87	75.2 ± 4.87	75.2 ± 4.87
	Median (25%, 75%)	75.2 (71.9, 78.6)	75.3 (72.0, 78.4)	75.3 (72.0, 78.5)
	Min – Max	58.4 – 89.66	58.0 – 91.5	58.0 – 91.5
BMI	Mean ± SD	28.8 ± 3.98	28.7 ± 3.98	28.7 ± 3.98
	Median (25%, 75%)	28.6 (26.0, 31.6)	28.7 (26.1, 31.3)	28.7 (26.1, 31.4)
	Min – Max	13.3 – 41.4	15.1 – 40.5	13.3 – 41.4

5.1.2 Lasso Cox Regression

The lasso Cox regression model was cross validated (CV) to achieve the minimum CV error as defined in Eq. 3. The model reached the best performance with $\lambda = 0.2855$. The CV errors (Brier scores) are displayed in Figure 1. In the figure, the blue solid line is cross validation error of the null model in which the survival probability was estimated with the Cox model without any covariates over the entire training set and the red solid line is the CV error of the full model with penalization parameters obtained from the entire training set; the black solid line is the CV error of the selected model from the CV, in which the survival probability was estimated from lasso Cox regression over the 10^{th} (left-out) CV samples and adjusted with .632 rule as suggested by Efron et al. (1997)[10] based on the selected model with the penalization parameters obtained from 9 CV samples; the gray shaded area is covered by the resampling data.

With λ = 0.2855, a total of 10 covariate terms were kept by the lasso Cox model and the corresponding coefficients for the reduced lasso Cox regression model are presented below. The regression coefficients were kept 5 decimal places, in order to retain the interaction term $\text{Age}^5\text{: Race} = \text{White}$, which is presented as $\text{Age}^5\text{: \{White\}}$.



$$\text{Prob}\{T \geq t\} = S_0(t)^{e^{X\beta}}, \text{where}$$

$$\begin{aligned} X\hat{\beta} = {} & 0.00327\ \text{MAP}^4 + 0.00650\ \text{MAP}^5 + \\ & 0.00298\ \text{Age}^4\text{: \{Placebo\}} - \\ & 0.00018\ \text{Age}^5\text{: \{Male\}} - \\ & 0.00004\ \text{Age}^5\text{: \{White\}} + \\ & 0.00667\ \text{MAP}^5\text{: \{Placebo\}} + \\ & 0.00178\ \text{MAP}^5\text{: \{Male\}} - \\ & 0.00378\ \text{BMI}^4\text{: \{Male\}} - \\ & 0.00104\ \text{BMI}^5\text{: \{Placebo\}} - \\ & 0.00094\ \text{BMI}^5\text{: \{Hispanic\}} \end{aligned}$$

Figure 1. CV Error for Lasso Cox Regression with λ = 0.2855 – Simulation Study

5.1.3 Ridge Cox Regression

Similarly ridge Cox regression model was cross validated to achieve the minimum CV error. The model reached its best performance with λ = 1.8153; the cross validation Errors are displayed in Figure 2. The ridge Cox regression achieved the optimum CV Brier Scores with almost all covariate terms; the coefficients are not presented in this paper.

5.1.4 Elastic-Net Cox Regression

For elastic-net Cox regression model, the cross validation was performed via interval search algorithm; in this algorithm, both penalization terms (α and λ) were searched simultaneously. The interval search paths for penalization parameters are displayed in Figure 3; the partial log likelihood deviances are presented on the top border of the figure; the penalization parameters are presented on the x- and y-axis and the number of covariate terms is labelled next to each point. The coordinates to the intersection of the red solid lines reflect the penalization terms when the elastic-net Cox regression model achieved the optimum CV performance, where α =0.2321 and λ = 0.2234. With these penalization terms, the model retained a total of 16 covariates; the coefficients of the 16 covariates for the elastic-net Cox regression model are summarized in Table 2; the corresponding elastic-net Cox regression model can be formulated with the coefficients from the table.

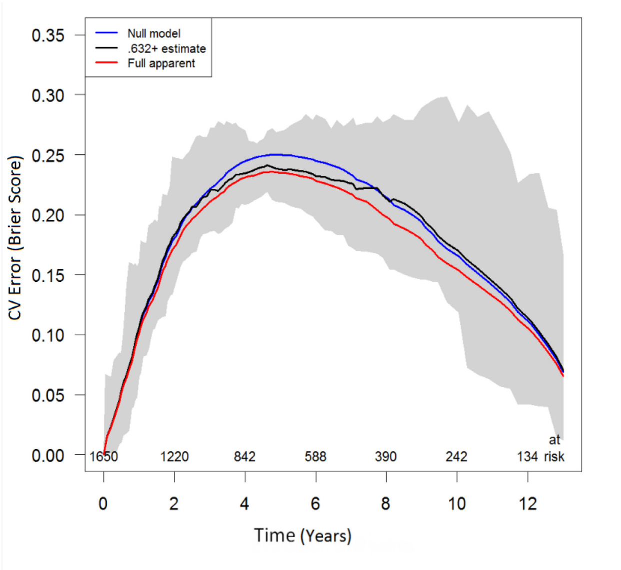


Figure 2. CV Error for Ridge Cox Regression with λ = 1.8153 – Simulation Study

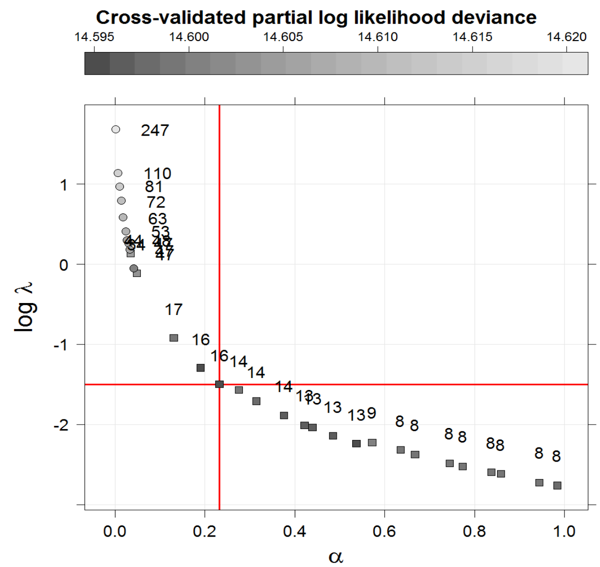


Figure 3. Interval Search Paths for Elastic Net Cox Regression – Simulation Study

Table 2. Coefficients of the Best Elastic-Net Cox Model with Penalization Parameters (α =0.2321 and λ = 0.2234) from Interval Search – Simulation Study

Terms	Coeffs
MAP	0.0482
MAP^3	0.0078
MAP^5	0.0008
Age:MAP:BMI^2	0.0039
Treatment=Placebo	0.2457
Sex=Male	-0.0794
Age^2:Treatment=Placebo	0.0089
MAP:Treatment=Placebo	0.0629
MAP^3:Treatment=Placebo	0.0063
MAP^5:Sex=Male	0.0001
Age:MAP:BMI:Treatment=Placebo	0.0112
Age^3:MAP:BMI:Race=Hispanic	0.0018
BMI^2:Sex=Male	-0.0200
Age^2:BMI^2:Race=Asian	-0.0046
Age:MAP:BMI^2:Treatment=Placebo	0.0165
Sex=Male:Race=White	-0.0387

$\text{Prob}\{T \geq t\} = S_0(t)^{e^{X\beta}}$, where

$$
\begin{aligned}
X\hat{\beta} = {} & 0.0482\ \text{MAP} + 0.0078\ \text{MAP}^3 + \\
& 8 \times 10^{-4}\ \text{MAP}^5 + \\
& 0.0039\ \text{Age: MAP: BMI}^2 + \\
& 0.2457\ \{\text{Placebo}\} - \\
& 0.0794\ \{\text{Male}\} + \\
& 0.0089\ \text{Age}^2\text{: }\{\text{Placebo}\} + \\
& 0.0629\ \text{MAP: }\{\text{Placebo}\} + \\
& 0.0063\ \text{MAP}^3\text{: }\{\text{Placebo}\} + \\
& 10^{-04}\ \text{MAP}^5\text{: }\{\text{Male}\} + \\
& 0.0112\ \text{Age: MAP: BMI: }\{\text{Placebo} \\
& 0.0018\ \text{Age}^3\text{: MAP: BMI: }\{\text{Hispar} \\
& 0.0200\ \text{BMI}^2\text{: }\{\text{Male}\} - \\
& 0.0046\ \text{Age}^2\text{: BMI}^2\text{: }\{\text{Asian}\} + \\
& 0.0165\ \text{Age: MAP: BMI}^2\text{: Placebc} \\
& -0.0387\ \{\text{Male}\}\text{: }\{\text{White"}\}
\end{aligned}
$$

5.1.5 Prediction Performance

Prediction performance measurements of the three selected penalized Cox models (lasso, ridge and elastic-net Cox regression models via interval search) were assessed based on the test set. The prediction errors and time-dependent AUCs of the selected penalized Cox regression models are summarized in

Table 3; the corresponding 95% PCIs for prediction errors and time-dependent AUCs obtained from 1000 bootstrap samples are also presented.

Table 3. Prediction Errors and Time Dependent AUCs for Lasso, Ridge and Elastic-Net Cox Models – Simulation Study Test Set

	Years	Lasso (95% PCI)	Ridge (95% PCI)	Elastic-Net IS (95% PCI)
Prediction Errors	1	0.113 (0.093, 0.137)	0.115 (0.093, 0.141)	0.115 (0.096, 0.138)
	2	0.175 (0.157, 0.193)	0.175 (0.155, 0.196)	0.173 (0.157, 0.190)
	3	0.213 (0.200, 0.226)	0.213 (0.196, 0.229)	0.209 (0.197, 0.222)
	4	0.230 (0.222, 0.237)	0.227 (0.215, 0.238)	0.225 (0.218, 0.233)
	5	0.243 (0.238, 0.247)	0.241 (0.233, 0.251)	0.238 (0.232, 0.243)
	6	0.244 (0.236, 0.251)	0.242 (0.237, 0.251)	0.239 (0.230, 0.24)
	7	0.242 (0.230, 0.252)	0.239 (0.243, 0.251)	0.237 (0.224, 0.249)
	8	0.234 (0.219, 0.248)	0.233 (0.216, 0.248)	0.233 (0.216, 0.247)
	9	0.227 (0.206, 0.246)	0.227 (0.205, 0.246)	0.227 (0.206, 0.247)
	10	0.213 (0.188, 0.236)	0.218 (0.192, 0.242)	0.214 (0.189, 0.240)
	11	0.206 (0.176, 0.236)	0.209 (0.180, 0.238)	0.208 (0.178, 0.239)
	12	0.199 (0.171, 0.236)	0.203 (0.175, 0.237)	0.201 (0.172, 0.239)
	13	0.193 (0.159, 0.231)	0.196 (0.165, 0.236)	0.196 (0.163, 0.235)
	14	0.180 (0.144, 0.221)	0.182 (0.150, 0.226)	0.184 (0.149, 0.229)
Time Dependent AUCs	1	0.644 (0.584, 0.715)	0.666 (0.604, 0.728)	0.690 (0.631, 0.750)
	2	0.599 (0.553, 0.649)	0.634 (0.583, 0.681)	0.649 (0.603, 0.697)
	3	0.590 (0.548, 0.632)	0.627 (0.584, 0.670)	0.636 (0.591, 0.677)
	4	0.590 (0.550, 0.628)	0.625 (0.584, 0.662)	0.634 (0.593, 0.672)
	5	0.586 (0.549, 0.623)	0.616 (0.578, 0.652)	0.626 (0.590, 0.662)
	6	0.582 (0.549, 0.620)	0.614 (0.579, 0.649)	0.625 (0.591, 0.661)
	7	0.579 (0.545, 0.616)	0.614 (0.579, 0.647)	0.625 (0.591, 0.657)
	8	0.576 (0.542, 0.611)	0.611 (0.576, 0.643)	0.621 (0.586, 0.651)
	9	0.575 (0.542, 0.609)	0.609 (0.575, 0.640)	0.619 (0.585, 0.650)
	10	0.574 (0.541, 0.607)	0.606 (0.573, 0.636)	0.618 (0.584, 0.648)
	11	0.574 (0.542, 0.608)	0.607 (0.574, 0.636)	0.612 (0.584, 0.647)
	12	0.574 (0.542, 0.607)	0.606 (0.573, 0.636)	0.618 (0.585, 0.647)
	13	0.573 (0.542, 0.607)	0.606 (0.573, 0.635)	0.617 (0.584, 0.647)
	14	0.573 (0.542, 0.606)	0.605 (0.573, 0.635)	0.617 (0.584, 0.647)

The graphic display of the prediction errors, the time-dependent AUCs and the corresponding 95% PCIs are presented in Figure 4. The solid black line and the grey shaded area indicate the prediction errors (or AUCs) and the corresponding the 95% PCIs for the lasso Cox model; the dotted blue line and the light blue shaded area are the prediction errors and the corresponding 95% PCIs for ridge Cox regression model; the prediction errors for elastic-net was the orange line covered by the orange shaded area for the corresponding 95% PCIs.

In terms of the prediction errors, the three penalized Cox regression models were very similar to each other, as can be seen from the figure; however, in terms of the time-dependent AUCs, the elastic-net had the best performance and the lasso Cox model had the worst performance. However in terms of the variable selection, the lasso Cox model selected a total of 10 covariates, the elastic-net selected a total of 16 covariates and the lasso selected almost all covariate terms; thus the lasso selected the minimum number of covariates with some sacrifice of prediction powers; elastic-net Cox model selected moderate number of covariates with excellent prediction accuracy; and the ridge Cox regression model did not select covariates.

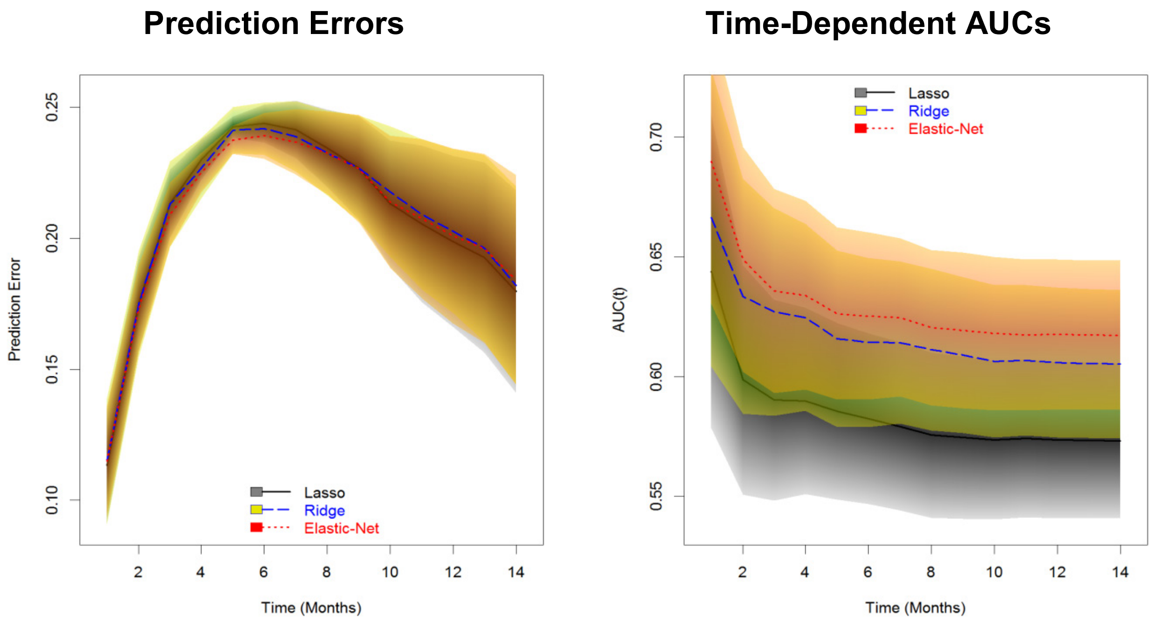


Figure 4. Prediction Errors and Time-Dependent AUCs for Lasso, Ridge and Elastic-Net Cox Models – Simulation Study Test Set

5.2 Results of the Real World Case Study

5.2.1 Summary Statistics of the Simulation Study

Considering it was impossible to present all factors in one page and it was probably overwhelmed to look at the descriptive summary for all 75 factors; only the 5 clinical factors as well as 2 of the 70 gene signatures are summarized in Table 9. Contingency tables for categorical factors are presented in the top half of the table. Continuous factors, including Age and 2 of the 70 gene signatures are summarized in the bottom half. One of the clinical factors, Grade of tumor, is a normal factor, with 3 category levels; thus two dummy variable were created for reference to the relative difference between the 3 category levels.

For this study, only the linear form of the original 76 factors (74 + 2 dummy factors) were considered to construct the initial model.

5.2.2 Lasso Cox Regression

For lasso Cox regression model, the coefficient solution paths for all factors included in the lasso Cox regression linear model are displayed in Figure 5; an increasing λ could shrink all coefficients towards 0; i.e., if λ is sufficiently large, all coefficients could be shrunk to 0, in which case the model would have no covariates left; smaller penalization term (λ) could keep more covariates. When λ was close to 0, the performance of the lasso model would be dominated by ridge regression; most (if not all) of the covariates should be kept in the model. For the lasso Cox regression model, the cross validation error reached the minimum with the penalization term, $\lambda = 0.0441$; the cross validation error curves are displayed in Figure 6.

Table 4. Brief Summary of the NKI70 Data

Factors	Statistics (N=144)
Survival: Metastasis	48 (33%)
Diameter of Tumor	
≤ 2 cm	73 (51%)
> 2 cm	71 (49%)
Number of affected lymph nodes	
1-3	106 (74%)
≥ 4	38 (26%)
Estrogen receptor status	
Negative	27 (19%)
Positive	115 (80%)
Missing	2 (1%)
Grade of tumor	
Poorly Differentiated	48 (33%)
Intermediate	52 (38%)
Well Differentiated	41 (28%)
Missing	3 (2%)

Factors (71)	n/nmiss	Mean ± SD	Median	Quartiles	Ranges
Age	142/2	44.31 ± 5.34	45	41, 49	16 – 53
TSPYL5	144/0	-0.109 ± 0.33	-0.089	-0.331, 0.117	-1.08 – 0.6018
DIAPH3	144/0	-0.033 ± 0.24	-0.022	-0.179, 0.241	-0.679, 0.618
⋮	⋮	⋮	⋮	⋮	⋮
C20OR46	144/0	-0.086 ± 0.25	-0.133	-0.256, 0.020	-0.451 – 0.992

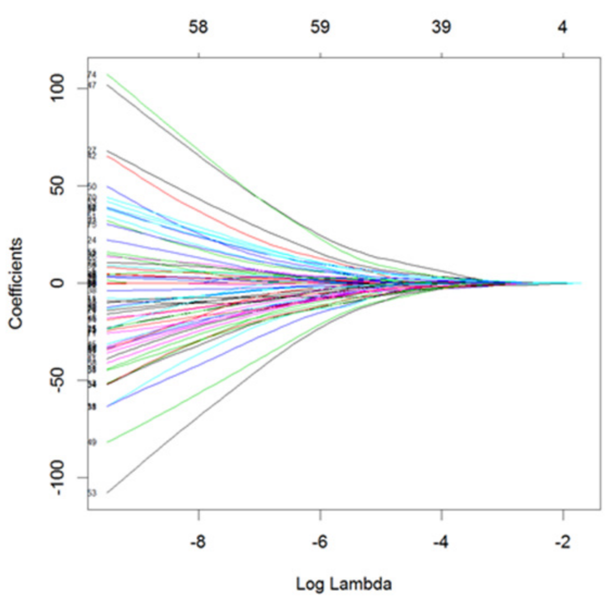


Figure 5. Coefficients Solution Path for all Factors – NKI70 Data

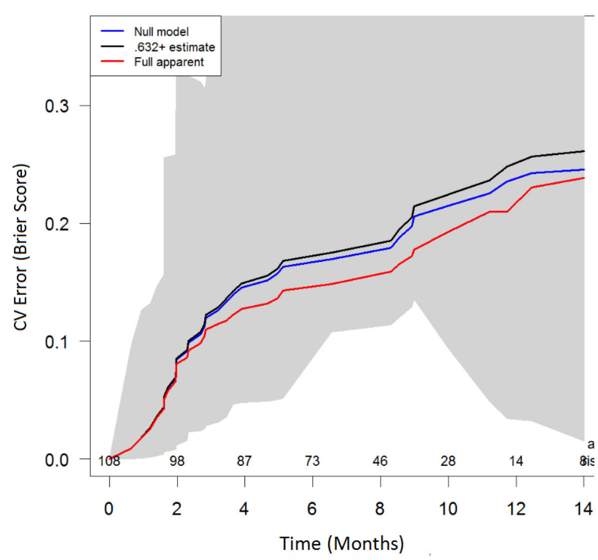


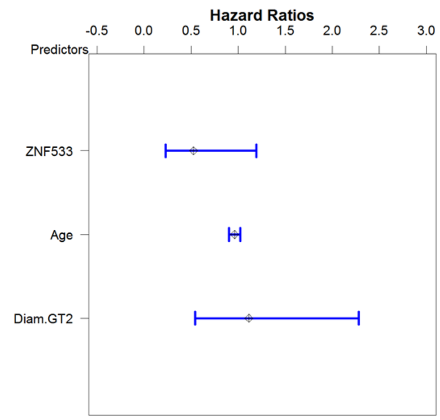
Figure 6. CV Errors for Lasso Cox Linear Model – NKI70 Data

The coefficients obtained from the cross validated lasso Cox linear model are presented in Table 5, as a comparison, the unbiased coefficients estimates for the corresponding factors obtained from a typical Cox regression model are also presented. From the table, it can be seen that smaller coefficients (in absolute values), should have less bias, the bigger the absolute values of the coefficients, the bigger the bias. For lasso Cox model, 3 factors were selected, Diam.GT2, Age and ZNF533; of which, Age had the smallest coefficient (0.04) in absolute values, the unbiased coefficient estimate for Age was 0.05, the observed bias was very small; for ZNF533, the coefficient estimate was -0.6451, the absolute value of the coefficient was the biggest; the unbiased coefficient estimate was -1.6468; the bias was quite substantial. The forest plot of the estimated hazard ratios from Lasso Cox linear model is presented in Figure 8.

Table 5. Biased of Coefficients from Lasso (λ=0.0441) Cox Linear Model and Unbiased Estimates from Typical Cox Regression Model – NKI70 Data

	Lasso Cox Regression (Biased)			
	Coef	**HR**	**SE (Coef)**	**P-val**
Diam.GT2	0.1075	1.1135	0.3666	0.7693
Age	-0.0407	0.9601	0.0316	0.1987
ZNF533	-0.6451	0.5246	0.4186	0.1233
	Cox Regression (Unbiased)			
Diam.GT2	**Coef**	**HR**	**SE (Coef)**	**P-val**
Age	0.8182	2.2664	0.3894	0.0356
ZNF533	-0.0585	0.9432	0.0312	0.0607

Figure 7. Forest Plot of Hazard Ratios for Lasso Linear Cox model (λ=0.0441) – NKI70 Data



5.2.3 Ridge Cox Regression

Similarly, ridge Cox linear model started with the original 76 factors; the model was cross validated to obtain the optimal penalization term, λ. The solution paths of coefficients and penalization terms for ridge Cox linear model are not presented. The CV errors are presented in Figure 7. The λ corresponding to the selected model was 0.04111.

For this model, the CV errors were much worse than the full apparent model, which was possibly due to overfitting; the ridge Cox linear model had probably picked up too many noise covariates with the penalization term. The CV errors from ridge regression ($.632^+$ estimates) were estimated via the predicted survival probability from the 10th left-out sample based on the model with the penalization parameters obtained from the 9-fold CV samples. And the CV error of the "full apparent" model was obtained by fitting the full model with the penalization term estimated from the entire training set. Therefore, the CV errors for the full apparent ridge Cox linear model were different from the ones for full apparent Lasso Cox linear model, since the penalization terms were different. The penalization parameter λ=0.04111 achieved the minimum cross validation errors for the ridge Cox linear model by keeping all covariates.

With the penalization parameter, λ = 0.04111, the model kept all 76 factors; the coefficient estimates for a sample of 11 factors (randomly selected) are presented in Table 6; the complete summary of the coefficient estimates from the ridge Cox linear regression is not presented.

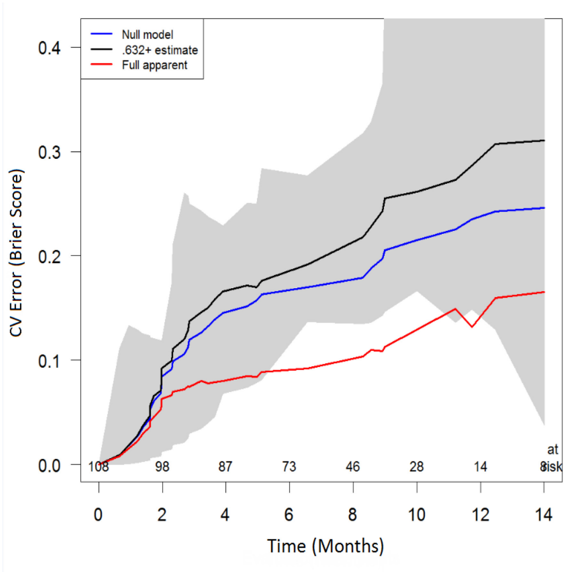


Figure 8. CV Errors for Ridge Cox Linear Regression Model with λ = 0.0411 – NKI70 Data

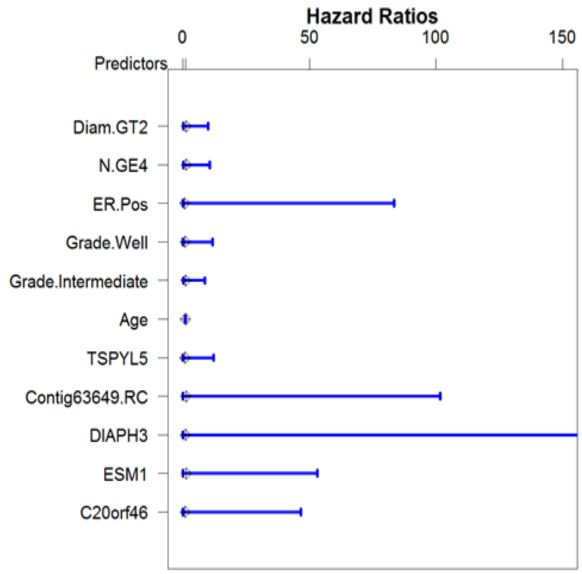


Figure 9. Forest Plot of the Subset of Hazard Ratios from Ridge Cox Linear Regression – NKI70 Data

Table 6. A Subset of Coefficients from the Ridge Cox Linear Regression with λ =0.0411 – NKI70 Data

	Coef	HR	SE (coef)	z	P-val
Diam.GT2	0.3059	1.358	1.0199	0.2999	0.7642
N.GE4	0.2979	1.347	1.0546	0.2825	0.7776
ER.Pos	-0.3960	0.673	2.4597	-0.1610	0.8721
Grade.Well	-0.1101	0.896	1.3102	-0.0840	0.9331
Grade.Intermediate	0.0664	1.069	1.0708	0.0620	0.9505
Age	-0.0472	0.954	0.0904	-0.5219	0.6017
TSPYL5	-0.1953	0.823	1.3696	-0.1426	0.8866
Contig63649.RC	0.1990	1.220	2.2566	0.0882	0.9297
DIAPH3	0.0101	1.010	3.8526	0.0026	0.9979
⋮	⋮	⋮	⋮	⋮	⋮
ESM1	0.2628	1.301	1.8927	0.1389	0.8896
C20orf46	-0.1501	0.861	2.0366	-0.0737	0.9412

Forest plot of the hazard ratios is displayed in Figure 8; the hazard ratio for gene signature, DIAPH3, had a very wide 95% CI of [0.0005, 1921.8], which was almost 20 times wider than the second widest, therefore the confidence interval was truncated at the maximum of 150 in order to see the CIs for the rest of factors. Again, these coefficient estimates were biased. The unbiased coefficient estimates could not be obtained from the typical Cox regression model due to nonestimability; the typical Cox model could not solve for the coefficients due to singular design matrix.

5.2.4 Elastic-Net Cox Regression

Similar to the simulation study, Elastic-net Cox regression model was cross validated using interval search algorithm; Figure 9 displays the last iteration of the search paths for the penalization parameters (α and λ); the model reached the minimum cross validation error with the penalization parameters of α = 0.00686 and λ = 1.127488; the model retained a total of 68 covariate terms.

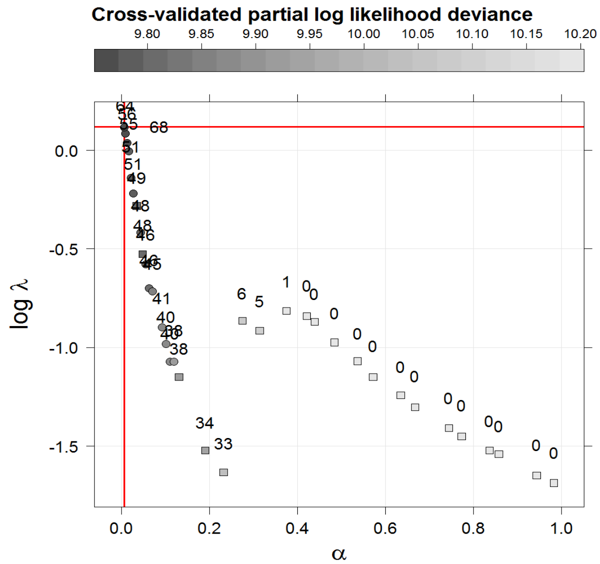


Figure 10. Interval Search Paths for Elastic Net Cox Linear Model – NKI70 Data

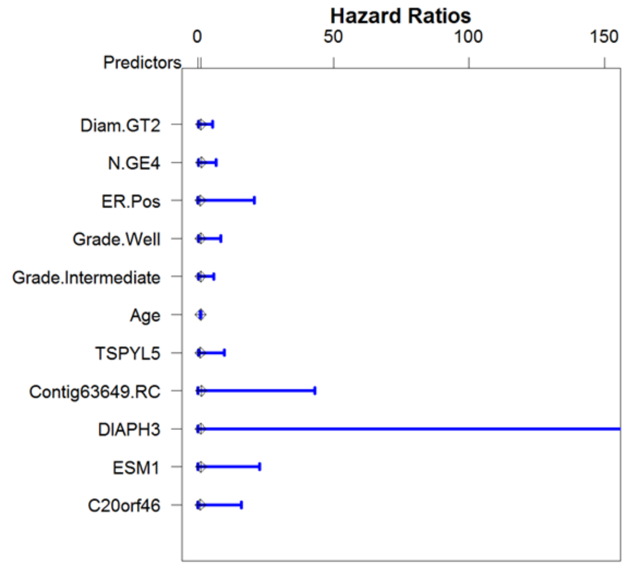


Figure 11. HR for a Subset of 11 Factors from Elastic-Net Cox Linear Model with Penalization Parameters (α = 0.0069 and λ = 1.1275) – NKI70 Data

Table 7. Subset of Coefficients for the Elastic-Net Cox Linear Model with Penalization Parameters (α = 0.00686 and λ = 1.127488) – NKI70 Data

	Coef	HR	SE (Coef)	z	P-val
Diam.GT2	0.0840	1.088	0.8061	0.10	0.9170
N.GE4	0.1531	1.165	0.8841	0.17	0.8625
ER.Pos	-0.1239	0.883	1.6094	-0.08	0.9386
Grade.Well	-0.0468	0.954	1.1061	-0.04	0.9663
Grade.Intermediate	0.0060	1.006	0.8958	0.01	0.9946
Age	-0.0090	0.991	0.0707	-0.13	0.8983
TSPYL5	-0.0665	0.936	1.1965	-0.06	0.9557
Contig63649.RC	0.1670	1.182	1.8354	0.09	0.9275
DIAPH3	0.0144	1.015	3.0319	0.00	0.9962
⋮	⋮	⋮	⋮	⋮	⋮
ESM1	0.1222	1.130	1.5323	0.08	0.9427
C20orf46	-0.1032	0.902	1.4690	-0.07	0.9440

Of all 68 covariate terms, the coefficients of 11 random selected covariate terms are presented in Table 7. It can be seen from Table 7 that the p-values were extremely big and none were significant; since the elastic-net model selected covariates not by p-values, but by the regularization of the regression coefficients. On the other hand, the elastic-net linear Cox model selected all covariates that were highly correlated, and the number of covariates well exceeded the maximum number of parameters that could be estimated with typical Cox regressions (due to singularity); thus unbiased estimates of the coefficients could not be produced. The forest plot of hazard ratios corresponding to 11 covariate terms is displayed in Figure 10. The factor, DIAPH3, again had extremely wide confidence intervals, which was cutoff at 150.

5.2.5 Prediction Performance

The selected penalized Cox regression models were evaluated against the test set to assess the prediction errors and the time-dependent AUCs; results are summarized in Table 8.

Table 8. Prediction Errors and Time Dependent AUCs for Lasso, Ridge and Elastic-Net Cox Models – NKI70 Test Set

	Years	Lasso (95% PCI)	Ridge (95% PCI)	Elastic-Net IS (95% PCI)
Prediction Errors	1	0.053 (0.000, 0.130)	0.050 (0.000, 0.127)	0.045 (0.000, 0.127)
	2	0.090 (0.027, 0.175)	0.082 (0.021, 0.161)	0.082 (0.021, 0.161)
	3	0.162 (0.075, 0.255)	0.131 (0.055, 0.221)	0.131 (0.055, 0.221)
	4	0.187 (0.105, 0.282)	0.152 (0.074, 0.235)	0.152 (0.074, 0.235)
	5	0.234 (0.148, 0.327)	0.189 (0.110, 0.280)	0.189 (0.110, 0.280)
	6	0.230 (0.147, 0.319)	0.184 (0.108, 0.273)	0.184 (0.108, 0.273)
	7	0.238 (0.154, 0.323)	0.195 (0.117, 0.284)	0.195 (0.117, 0.284)
	8	0.238 (0.154, 0.323)	0.195 (0.117, 0.284)	0.195 (0.117, 0.284)
	9	0.247 (0.182, 0.316)	0.200 (0.134, 0.268)	0.200 (0.134, 0.268)
	10	0.240 (0.181, 0.302)	0.193 (0.131, 0.254)	0.193 (0.131, 0.254)
	11	0.241 (0.181, 0.304)	0.193 (0.131, 0.255)	0.193 (0.131, 0.255)
	12	0.234 (0.184, 0.374)	0.192 (0.131, 0.386)	0.192 (0.131, 0.386)
	13	0.261 (0.194, 0.355)	0.236 (0.141, 0.394)	0.236 (0.141, 0.394)
	14	0.261 (0.194, 0.355)	0.236 (0.141, 0.394)	0.236 (0.141, 0.394)
Time Dependent AUCs	1	0.738 (0.592, 1.000)	0.805 (0.693, 1.000)	0.776 (0.633, 1.000)
	2	0.725 (0.587, 0.855)	0.798 (0.692, 0.893)	0.768 (0.633, 0.900)
	3	0.722 (0.562, 0.852)	0.794 (0.673, 0.893)	0.765 (0.625, 0.894)
	4	0.714 (0.500, 0.851)	0.787 (0.500, 0.892)	0.757 (0.500, 0.894)
	5	0.694 (0.500, 0.848)	0.762 (0.500, 0.892)	0.737 (0.500, 0.888)
	6	0.680 (0.500, 0.845)	0.753 (0.500, 0.888)	0.728 (0.500, 0.894)
	7	0.702 (0.500, 0.847)	0.776 (0.500, 0.888)	0.743 (0.500, 0.889)
	8	0.706 (0.500, 0.845)	0.781 (0.500, 0.888)	0.751 (0.500, 0.894)
	9	0.716 (0.500, 0.849)	0.790 (0.627, 0.888)	0.763 (0.613, 0.894)
	10	0.730 (0.554, 0.855)	0.793 (0.673, 0.888)	0.763 (0.611, 0.894)
	11	0.722 (0.575, 0.852)	0.797 (0.691, 0.892)	0.765 (0.625, 0.889)
	12	0.724 (0.585, 0.852)	0.798 (0.692, 0.892)	0.767 (0.633, 0.894)
	13	0.724 (0.587, 0.852)	0.798 (0.692, 0.892)	0.767 (0.633, 0.894)
	14	0.724 (0.587, 0.852)	0.798 (0.692, 0.892)	0.767 (0.633, 0.894)

The graphical display of the prediction errors and time-dependent AUCs of the selected penalized Cox regression models are presented in Figure 11. In terms of prediction errors, the ridge and elastic-net Cox regression models were similar to each other and both were better than the lasso Cox regression model; in terms of the AUCs, the ridge cox regression model retain all covariates to achieve the best performance; elastic net Cox selected a total of 68 factors with moderate prediction accuracy and the lasso Cox regression model kept a total of 3 predictors, but it also had the worst performance.

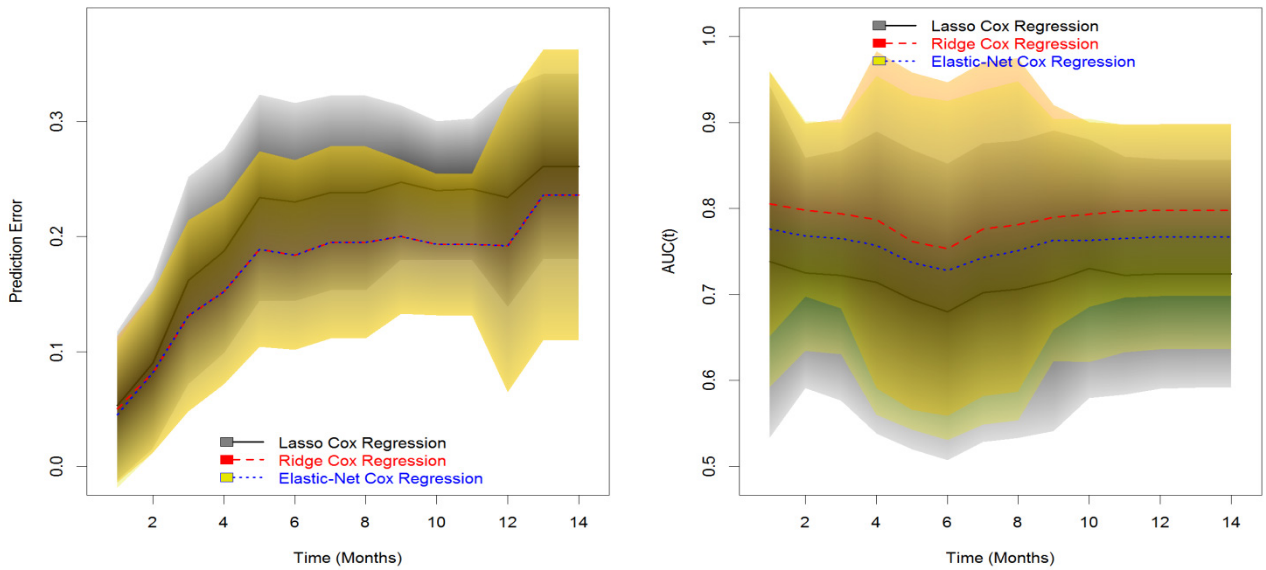


Figure 12. Prediction Errors and Time-Dependent AUCs for Lasso, Ridge and Elastic-Net Cox Models – NKI70 Test Set

6. Conclusions and Discussions

In this research paper, three penalized Cox regression models including lasso Cox regression, ridge Cox regression and elastic-net Cox regression models were studied.

In terms of prediction errors, the three penalized Cox models were very similar to each other for the simulation study; but for the real world case study, ridge and elastic-net Cox regression models were similar to each other; and lasso had the worst performance.

In terms of the time-dependent AUCs, the elastic-net Cox model via interval search had the best time-dependent AUCs for the simulation study and second best for the real world case study. The ridge Cox model had the second best AUCs for the simulation study and the best for the real world case study by keeping all prognostic factors or covariate terms; the lasso Cox regression model had the worst prediction performance.

As found in the both studies, the lasso Cox regression had some nice features; it provided regularization (variable selection) and shrinkage simultaneously, however it also had some problems, in cases when the number of factors were bigger than the number of observations (n), the lasso Cox regression selects at most n factors; if high correlations existed among factors, the lasso regression only randomly selected one from the correlated group of factors and the prediction performance of the lasso was dominated by ridge regression. Another disadvantage was persistent to all penalized Cox models; the parameter estimates were biased, which would have been even worse if the true unknown parameter was large. As found from the real world case study, polynomial transformations did not improve the model performance for lasso Cox models; inclusion of pair-wise interactions was able to achieve slight improvement to the prediction AUCs for lasso Cox regression.

In general, ridge Cox regression model had very good prediction performance, but it did not select predictors; instead, it shrank most of the coefficients towards zero, but never reached zero, the exceptional prediction performance was achieved by keeping most if not all covariates. Additionally, it was found in the case study that, nonlinear (polynomial) transformation and inclusion of pair-wise interactions did not improve the prediction performance, instead it cost substantial deterioration to the prediction performance of ridge Cox regression.

On the other hand, Elastic-net Cox regression served the purpose for variable selection with relatively good prediction accuracy comparing to lasso and ridge Cox regression. In the real world case study, elastic-net Cox regression models should have relative excellent prediction performance compared to most of the other survival models; however, it was found that the elastic-net interaction model did not perform as good due to the lasso-like penalization terms ($\alpha \approx 1$), which should be considered as an exception. Otherwise, the elastic-net Cox model was the most effective of all intended survival models.

7. Future Work to be Done

In the simulation study, some placebo treated subjects had switched treatment during the study; the placebo treated subjects who had switched treatment were considered as two different observations, since the same subject had different treatment over different study period. The time-varying treatment effect should be handled using Anderson-Gill[11] extension of the Cox proportional hazard model, however, the penalized Cox regression model cannot handle multiple observations per subject, thus different observation from the same subject were considered as independent subjects, which could be a potential problem if the observations from the same subjects were correlated with each other. Thus, an extension of penalized Cox model with adjustment of multiple observations per subject should be more considered for further research.

.

8. References

[1] Beer, D. G. et al. (2002) Gene-expression profiles predict survival of subjects with lung adenocarcinoma. Nature Medicine, 8, 816-824.

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

[3] Jelle J. Goeman (2012) Penalized R package, version 0.9-42.

[4] Jelle J. Goeman (2010) L1 penalized estimation in the Cox proportional hazards model, Biometrical Journal, 52 (1) 70-84.

[5] Marquardt, D. W. (1970) Generalized inverses, ridge regression, biased linear estimation and non-linear estimation, Technometrics, 12, 591–612.

[6] Zou H., Hastie T. (2005) Regularization and variable selection via the elastic net. J.R. Statist. Soc. B 67, Part 2, PP. 201-320.

[7] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/

[8] M.J. van de Vijver, Y.D. He, L.J. van 't Veer, H. Dai, A.A.M. Hart, D.W. Voskuil, G.J. Schreiber, J.L. Peterse, C. Roberts, M.J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E.T. Rutgers, S.H. Friend, and R. Bernards (2002). A gene-expression signature as a predictor of survival in breast cancer. New England Journal of Medicine 347 (25), 1999–2009.

[9] Gerds & Schumacher (2006), Consistent estimation of the expected Brier score in general survival models with right-censored event times, Biometrical Journal 48, 1029-1040.

[10] Efron B., Tibshirani R. (1997) Improvements on Cross-Validation: The .632+ Bootstrap Method. Journal of the American Statistical Association, 92 (438), pp. 548-560. URL: http://links.jstor.org/sici?sici=0162-1459%28199706%2992%3A438%3C548%3AIOCT.B%3E2.0.CO%3B2-I

[11] Anderson, T. W.; Darling, D. A. (1952). "Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes". Annals of Mathematical Statistics 23: 193–212.