Penalized Casebase in Survival Analysis

Trevor Kwan

August 29th, 2022

University of British Columbia

Table of contents

THE UNIVERSITY OF BRITISH COLUMBIA

Introduction

Motivating Example

- Consider Karl.
    - Age: 58
    - Karl has prostate cancer.
    - Doctors want to know the 5 or 10 year risk of death given Karl's covariate profile.

Cox Proportional Hazards Method

$$h(t|\mathbf{x}_i) = h_0(t) * exp(\mathbf{x}_i^T \boldsymbol{\beta})$$

- Advantages:
  - A flexible way of measuring the risk of one patient relative to the risk of another patient with a hazard ratio.
  - Proportional Hazards Assumption: The ratio of the hazards does not depend on time.
  - $\frac{h(t|\mathbf{x}_1)}{h(t|\mathbf{x}_2)} = \frac{h_0(t) * exp(\mathbf{x}_1^T \boldsymbol{\beta})}{h_0(t) * exp(\mathbf{x}_2^T \boldsymbol{\beta})} = \frac{exp(\mathbf{x}_1^T \boldsymbol{\beta})}{exp(\mathbf{x}_2^T \boldsymbol{\beta})}$
- Disadvantages:
  - To recover the full hazard function, we need to separately estimate the baseline hazard $h_0(t)$.

Popular Methods in Survival Analysis

THE UNIVERSITY OF BRITISH COLUMBIA

We want the absolute risk:

- Parametric models (e.g. Weibull, exponential)
  - Requires us to make distributional assumptions on the survival time.
- Breslow estimator on the Cox regression model
  - Leads to stepwise estimates in absolute risk curves that are difficult to interpret.

Popular Methods in Survival Analysis

- Survival analysis rarely produces prognostic functions, even though the software is widely available in Cox regression packages.
- This could be because the stepwise nature of the cumulative incidence curves reduces the interpretability of absolute risk for clinicians.

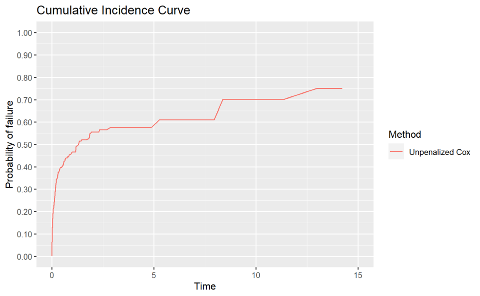

Figure 1: Cox Proportional Hazards Cumulative Incidence Curve

The Casebase Framework

THE UNIVERSITY OF BRITISH COLUMBIA

- Casebase is a method for estimating fully parametric hazard models through logistic regression.
- Compares person-moments when the failure event occurred "cases", with person-moments when patients were at risk "bases".
- Casebase allows us to create smooth cumulative incidence curves and consider time and censoring.

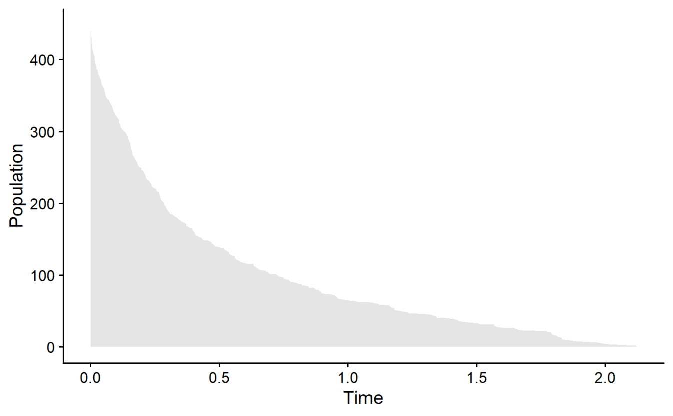

Figure 2: Casebase Population Time Plot

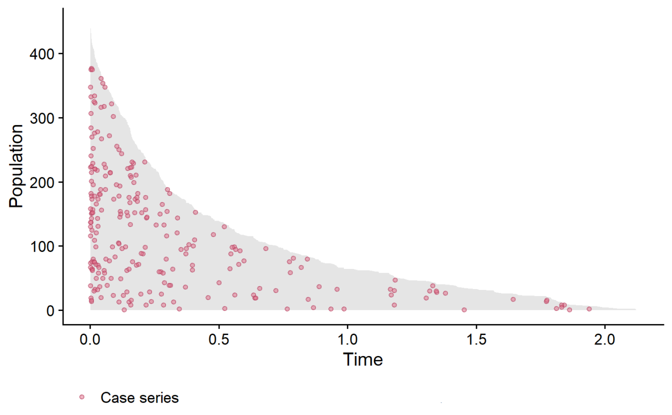

Figure 3: Casebase Population Time Plot

The Casebase Framework

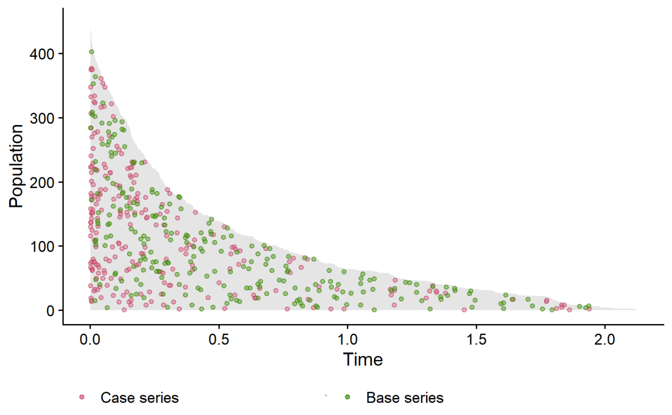

Figure 4: Casebase Population Time Plot

My Project Goal

The goal of my project is to compare Casebase methods with more classical survival analysis methods in terms of prediction performance and variable selection.

Methods

Generating the Data: The Settings

Simulation settings that were explored:

- n = number of observations
- p = number of predictors
- z = number of β coefficients that were set to 0
- snr = signal-to-noise ratio
- ρ = fixed correlation between any 2 predictors X_i and X_j

Generating the Data: The Process

THE UNIVERSITY OF BRITISH COLUMBIA

- Standard Gaussian predictor data: $X_{ij} \sim N(0,1)$
- True $\boldsymbol{\beta}$ coefficients: $\beta_j = (-1)^j * e^{-2(j-1)/(p-z)}$
- Failure times: $Y_i = e^{\sum_{j=1}^{p} X_{ij}^T \beta_j + k*E_i}$
    - $E_i \sim N(0,1)$
- Censored times: $C_i = e^{k*E_i}$
- Event times: $T_i = min\{Y_i, C_i\}$
    - If $C_i < Y_i$, the observation was censored.

Generating the Data: The Dataset

status	*time*(T)	x_1	x_2	x_3	x_4	x_5	...
0	1.420	-0.019	0.373	1.498	-1.532	1.605	...
1	0.091	-0.270	-0.633	0.283	0.706	-0.032	...
1	0.274	1.224	-0.466	0.681	1.075	0.554	...
0	2.731	-2.946	-3.041	-3.178	-4.0398	3.273	...
1	0.198	-0.080	1.402	-0.512	-0.561	-0.721	...
...	...	...	...	...	...	...	

Table 1: Simulated Dataset

Fitting the Models: Cox Proportional Hazards

Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ be the $p \times 1$ vector of covariates.
Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ be the $p \times 1$ vector of regression coefficients.

Unpenalized Cox:

- $\arg\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta})$
- $L(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{e^{\boldsymbol{x}_{j(i)}^T \boldsymbol{\beta}}}{\sum_{j \in R_i} e^{\boldsymbol{x}_j^T \boldsymbol{\beta}}}$
    - $\prod_{i=1}^{m}$ multiplies through observations by order of event times.
    - $j(i)$ is the index of the observation failing at time t_i.
    - $j \in R_i$ is the set of indices of all observations still at risk at time t_i.

Penalized Cox:

- $\arg\max_{\boldsymbol{\beta}} \left(log(L(\boldsymbol{\beta})) - \lambda(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2) \right)$
    - λ controls the level of penalization.
    - The penalty function turns into a lasso penalty when $\alpha = 1$ or a ridge penalty when $\alpha = 0$.
    - λ is selected by 10-fold cross validation.

Transform the data into a Casebase Series:

- b = ratio*c
  - b is the number of bases
  - c is the number of cases
- Sample b bases from n observations, each observation i with probability $\frac{t_i}{B}$ of being selected.
  - t_i is the event time of observation i.
  - B is the sum of all event times in the data.
- Transform the event time T column into one of the predictors $log(T)$
  - Weibull hazard: $log(h(t;X)) = \beta_0 + \beta_1 log(t) + X^T\boldsymbol{\beta}$

Fitting the Models: Casebase

y (death)	$log(time)$	x_2	x_3	x_4	x_5	...
0	-5.773	-0.019	0.373	1.498	-1.532	...
1	-3.396	-0.270	-0.633	0.283	0.706	...
1	-0.961	1.224	-0.466	0.681	1.075	...
0	-1.526	-2.946	-3.041	-3.178	-4.0398	...
1	-1.789	-0.080	1.402	-0.512	-0.561	...
...	...	...	...	...	...	...

Table 2: Casebase Series

Fitting the Models: Casebase

Fitting the Casebase Series:

- Logistic regression uses the logit link function $log\left(\frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)}\right) = \beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}$ to model it's log likelihood.
  - $p(\mathbf{x}_i)$ is the probability of failure for observation $\mathbf{x}_i$.

$$\textbf{arg min}_{\boldsymbol{\beta}} \left(-[\sum_{i=1}^{N} y_i * (\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}) - log(1 + e^{\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}})] + \lambda(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2)\right)$$

- y_i denotes censored or failure.
- λ controls the level of penalization.
- The penalty function turns into a lasso penalty when $\alpha = 1$ or a ridge penalty when $\alpha = 0$.
- λ is selected by 10-fold cross validation.

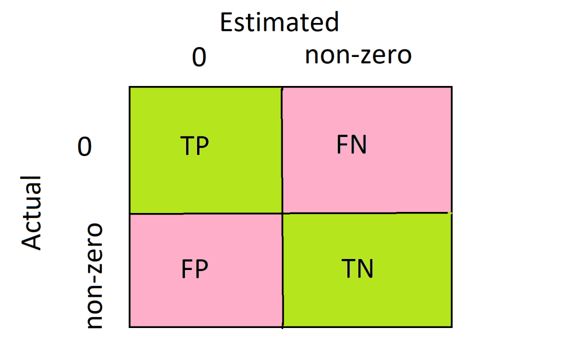

Figure 5: Confusion Matrix

β_j	Actual $\boldsymbol{\beta}$'s	Model Estimated $\boldsymbol{\beta}$'s	Status
β_1	-0.368	0.078	TP
β_2	0.333	-0.087	TP
β_3	0.000	0.000	TN
β_4	0.000	0.000	TN
β_5	0.000	0.054	FP
β_6	0.000	-0.676	FP
β_7	-0.202	0.000	FN
β_8	0.183	0.000	FN
β_9	-0.165	0.034	TP
β_{10}	0.000	0.000	TN

Table 3: $\boldsymbol{\beta}$ Comparison

Evaluating Metrics: Variable Selection

- True Positive Rate (TPR): $\frac{TP}{(TP+FN)}$
- True Negative Rate (TNR): $\frac{TN}{(TN+FP)}$
- False Positive Rate (FPR): $1 - TNR$
- False Negative Rate (FNR): $1 - TPR$
- Matthew's Correlation Coefficient (MCC)
  $MCC = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

Evaluating Metrics: Concordance

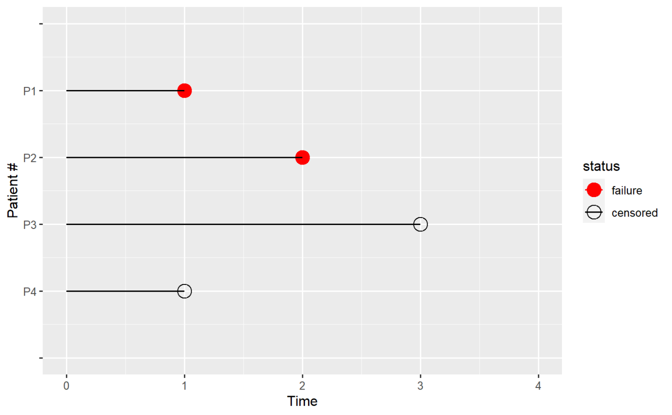

Figure 6: The Problem of Censoring in Concordance

Evaluating Metrics: Concordance

Concordant: correct prediction of which observation in the pair is going to ”fail” first.

Discordant: incorrect prediction of which observation in the pair is going to ”fail” first.

The observation with the higher risk score is predicted to fail first.

- Risk score for $\boldsymbol{x}_i$ in Cox models: $\boldsymbol{x}_i^T\boldsymbol{\beta}$
- Risk score for $\boldsymbol{x}_i$ in Casebase models: $p(y_i = 1) = \frac{1}{1+exp(-\boldsymbol{x}_i^T\boldsymbol{\beta})}$

Evaluating Metrics: Concordance

Let R_1 and R_2 represent the risk scores for a given pair, and T_1 and T_2 the true event times for a given pair.

For each evaluable pair, there are only 4 possible cases:

1. $R_1 > R_2$ and $T_1 < T_2$ OR $R_1 < R_2$ and $T_1 > T_2$: The pair is concordant (C).
2. $R_1 > R_2$ and $T_1 > T_2$ OR $R_1 < R_2$ and $T_1 < T_2$: The pair is discordant (D).
3. $R_1 == R_2$: The risk scores are equal (R).
4. $T_1 == T_2$: The times are equal (T).

$$Concordance = \frac{C + \frac{R}{2}}{C + D + R}$$

If all our observations were uncensored...

$BS(t) = \sum_i \{I(Y_i \leq t) - \hat{F}(t|\mathbf{x}_i)\}^2$

...then the Brier Score at each time t is the squared difference between the failure status and probability of failure for observation $\mathbf{x}_i$.

Evaluating Metrics: Time-Dependent Brier Scores

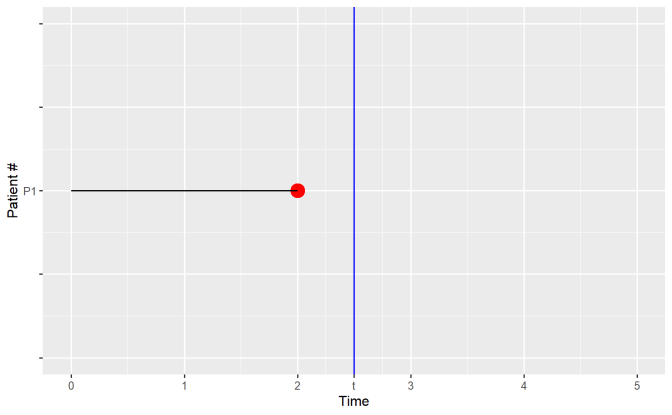

Figure 7: The Problem of Censoring in Brier Scores (Evaluable Case)

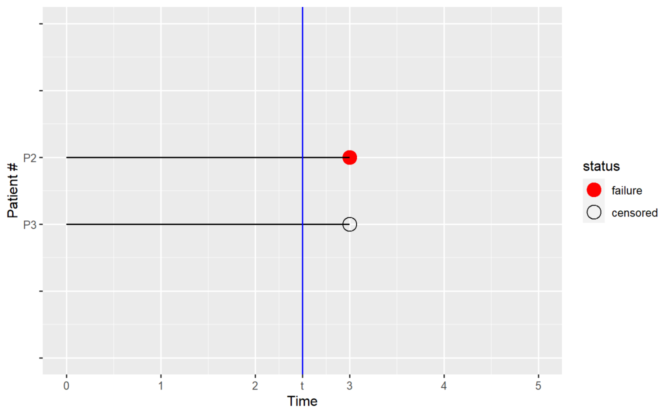

Figure 8: The Problem of Censoring in Brier Scores (Evaluable Case)

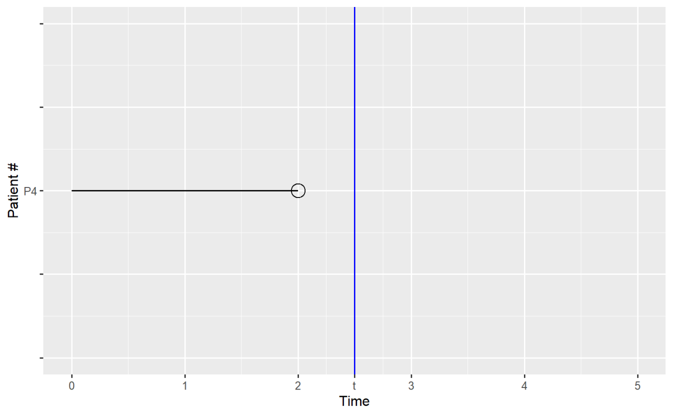

Figure 9: The Problem of Censoring in Brier Scores (Non-evaluable Case)

Evaluating Metrics: Time-Dependent Brier Scores

$$BS(t) = \sum_i \{\frac{I(C_i \geq min(Y_i, t))}{\hat{G}(min(Y_i, t))}\}\{I(Y_i \leq t) - \hat{F}(t|\mathbf{x}_i)\}^2$$

- Y_i is the failure time.
- C_i is the censoring time.
- t is the time the observation is evaluated at.
- $\hat{F}(t|\mathbf{x}_i)$ is the probability of failure.
- $\{I(Y_i \leq t) - \hat{F}(t|\mathbf{x}_i)\}^2$ is the squared difference between the actual event status and prediction probability.
- $\frac{I(C_i \geq min(Y_i, t))}{\hat{G}(min(Y_i, t))}\}$ gives more weight to observations near the end because there will be more non-evaluable cases.

Results

Simulation Setting 1: n>p

$n = 500,\ p = 120,\ z = 100,\ snr = 3,\ \rho = 0.5$

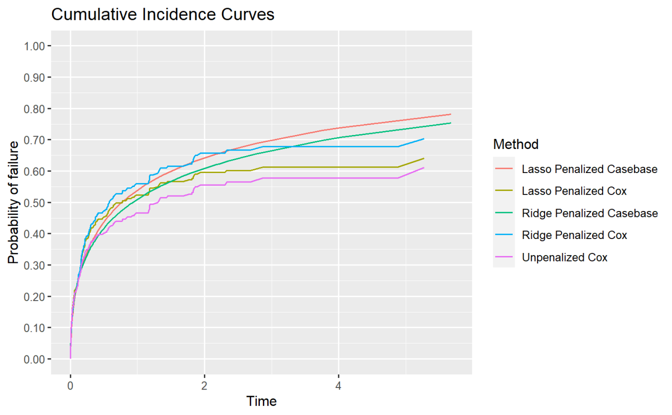

Figure 10: Cumulative Incidence Curves (Setting 1)

Simulation Setting 1: n>p

$n = 500$, $p = 120$, $z = 100$, $snr = 3$, $\rho = 0.5$

Measure	Lasso Cox	Lasso Casebase
MCC	0.525	0.369
TPR	0.850	0.700
TNR	0.800	0.760
FPR	0.200	0.240
FNR	0.150	0.300

Table 4: Variable Selection (Setting 1)

Simulation Setting 1: n>p

$n = 500$, $p = 120$, $z = 100$, $snr = 3$, $\rho = 0.5$

Method	Concordance
Unpenalized Cox	0.760
Lasso Penalized Cox	0.830
Ridge Penalized Cox	0.804
Lasso Penalized Casebase	0.873
Ridge Penalized Casebase	0.888

Table 5: Concordance (Setting 1)

Simulation Setting 1: n>p

$n = 500, p = 120, z = 100, snr = 3, \rho = 0.5$

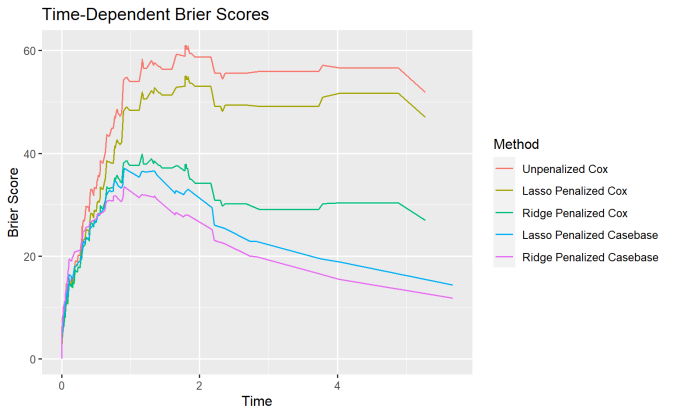

Figure 11: Time-Dependent Brier Scores (Setting 1)

Simulation Setting 2: p>n

$n = 100, p = 200, z = 150, snr = 3, \rho = 0.5$

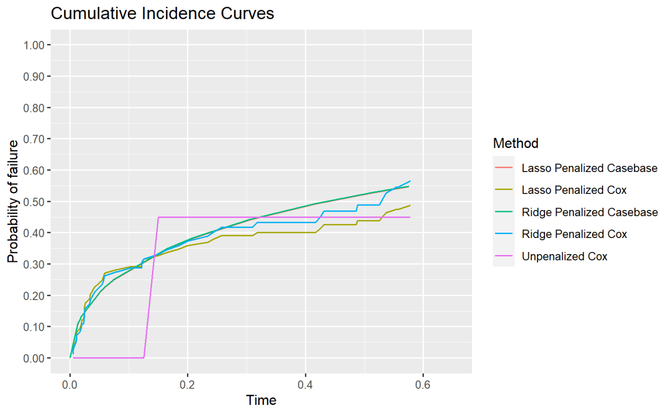

Figure 12: Cumulative Incidence Curves (Setting 2)

Simulation Setting 2: p>n

$n = 100$, $p = 200$, $z = 150$, $snr = 3$, $\rho = 0.5$

Measure	Lasso Cox	Lasso Casebase
MCC	0.243	0.176
TPR	0.160	0.140
TNR	0.973	0.960
FPR	0.027	0.040
FNR	0.840	0.860

Table 6: Variable Selection (Setting 2)

Simulation Setting 2: p>n

$n = 100, p = 200, z = 150, snr = 3, \rho = 0.5$

β_j	Actual $\boldsymbol{\beta}$'s	Lasso Cox $\boldsymbol{\beta}$'s	Lasso Casebase $\boldsymbol{\beta}$'s
β_1	-1.000	0.000	-0.350
β_2	0.905	-0.167	-0.399
β_3	-0.819	0.079	0.000
β_4	0.741	-0.176	-0.161
β_5	-0.670	0.068	0.000
β_6	0.606	0.000	-0.232
β_7	-0.549	0.028	0.044
β_8	0.497	0.000	0.000
β_9	-0.449	0.000	0.000
β_{10}	0.407	0.000	0.000

Table 7: $\boldsymbol{\beta}$ Comparison (Setting 2)

Simulation Setting 2: p>n

$n = 100$, $p = 200$, $z = 150$, $snr = 3$, $\rho = 0.5$

Method	Concordance
Unpenalized Cox	0.342
Lasso Penalized Cox	0.623
Ridge Penalized Cox	0.342
Lasso Penalized Casebase	0.894
Ridge Penalized Casebase	1.000

Table 8: Concordance (Setting 2)

Simulation Setting 2: p>n

$n = 100, p = 200, z = 150, snr = 3, \rho = 0.5$

β_j	Actual β's	Unpen Cox β's	Ridge Cox β's	Ridge Casebase β's
β_1	-1.000	20.837	0.006	-0.688
β_2	0.905	198.217	-0.017	-0.526
β_3	-0.819	406.150	0.012	5.160e-39
β_4	0.741	613.539	-0.022	-1.227e-38
β_5	-0.670	1000.745	0.015	7.100e-39
β_6	0.606	-1408.957	-0.009	-1.563e-38
β_7	-0.549	123.732	0.012	9.142e-39
β_8	0.497	214.724	-0.008	-1.225e-38
β_9	-0.449	999.303	0.011	1.025e-38
β_{10}	0.407	-652.457	-0.009	-6.516e-39

Table 9: β Comparison (Setting 2)

Simulation Setting 2: p>n

$n = 100, p = 200, z = 150, snr = 3, \rho = 0.5$

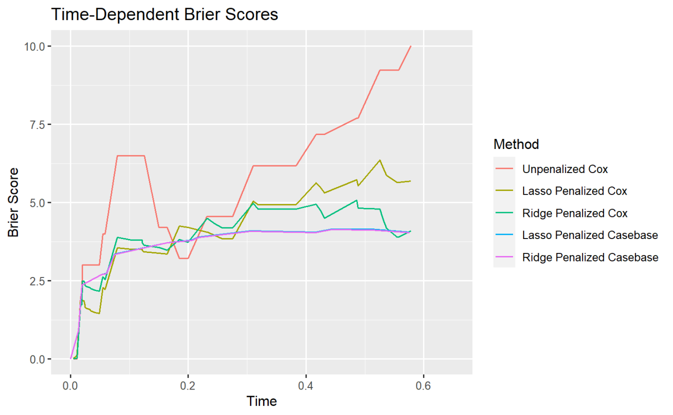

Figure 13: Time-Dependent Brier Scores (Setting 2)

Simulation Setting 3: Correlated Predictors

THE UNIVERSITY OF BRITISH COLUMBIA

$n = 500, p = 120, z = 100, snr = 3, \rho = 0.9$

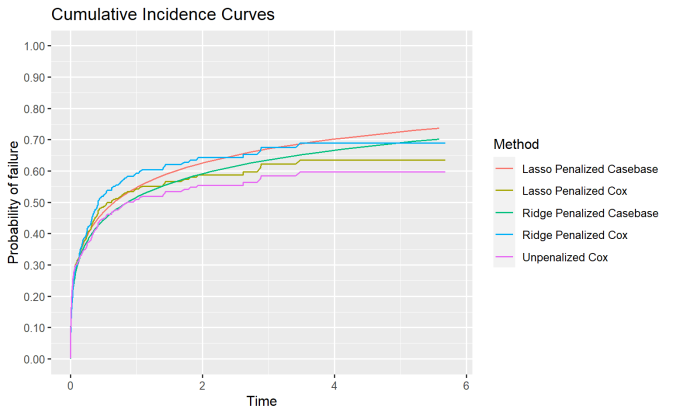

Figure 14: Cumulative Incidence Curves (Setting 3)

Simulation Setting 3: Correlated Predictors

$n = 500$, $p = 120$, $z = 100$, $snr = 3$, $\rho = 0.9$

Measure	Lasso Cox	Lasso Casebase
MCC	0.454	0.376
TPR	0.800	0.850
TNR	0.770	0.650
FPR	0.230	0.350
FNR	0.200	0.150

Table 10: Variable Selection (Setting 3)

Simulation Setting 3: Correlated Predictors

THE UNIVERSITY OF BRITISH COLUMBIA

$n = 500$, $p = 120$, $z = 100$, $snr = 3$, $\rho = 0.9$

Method	Concordance
Unpenalized Cox	0.755
Lasso Penalized Cox	0.808
Ridge Penalized Cox	0.794
Lasso Penalized Casebase	0.887
Ridge Penalized Casebase	0.935

Table 11: Concordance (Setting 3)

Simulation Setting 3: Correlated Predictors

$n = 500, p = 120, z = 100, snr = 3, \rho = 0.9$

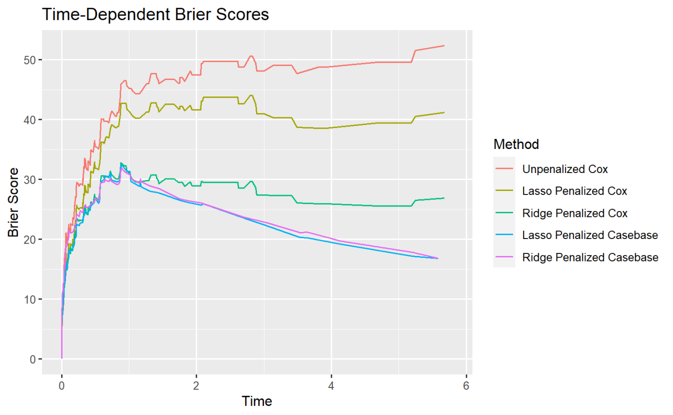

Figure 15: Time-Dependent Brier Scores (Setting 3)

Simulation Setting 4: Uncorrelated Predictors

$n = 500$, $p = 120$, $z = 100$, $snr = 3$, $\rho = 0.1$

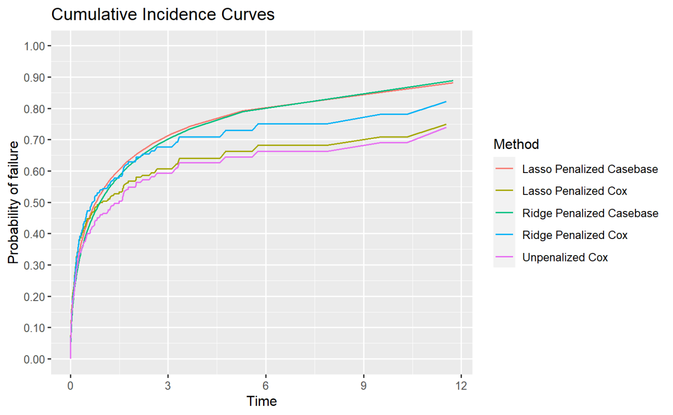

Figure 16: Cumulative Incidence Curves (Setting 4)

Simulation Setting 4: Uncorrelated Predictors

$n = 500$, $p = 120$, $z = 100$, $snr = 3$, $\rho = 0.1$

Measure	Lasso Cox	Lasso Casebase
MCC	0.538	0.307
TPR	0.900	0.800
TNR	0.780	0.610
FPR	0.220	0.390
FNR	0.100	0.200

Table 12: Variable Selection (Setting 4)

Simulation Setting 4: Uncorrelated Predictors

$n = 500$, $p = 120$, $z = 100$, $snr = 3$, $\rho = 0.1$

Method	Concordance
Unpenalized Cox	0.762
Lasso Penalized Cox	0.829
Ridge Penalized Cox	0.807
Lasso Penalized Casebase	0.878
Ridge Penalized Casebase	0.866

Table 13: Concordance (Setting 4)

Simulation Setting 4: Uncorrelated Predictors

$n = 500$, $p = 120$, $z = 100$, $snr = 3$, $\rho = 0.1$

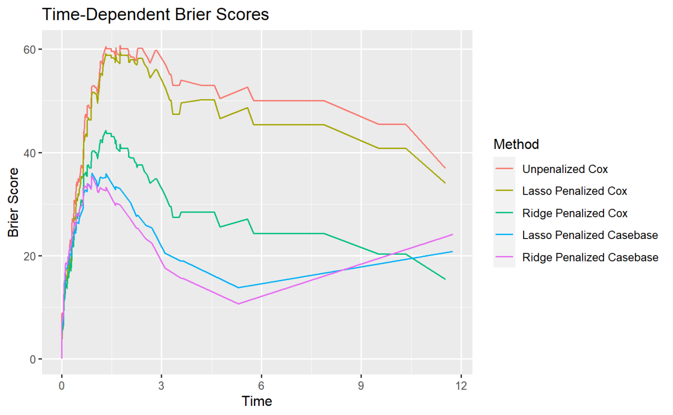

Figure 17: Time-Dependent Brier Scores (Setting 4)

Simulation Setting 5: High Signal-to-noise Ratio

$n = 500, p = 120, z = 100, snr = 7, \rho = 0.5$

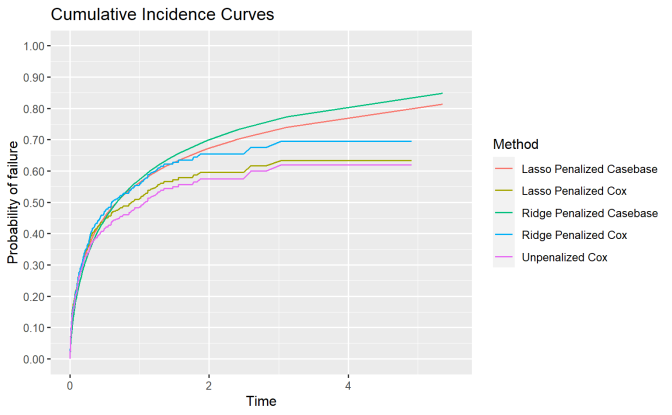

Figure 18: Cumulative Incidence Curves (Setting 5)

Simulation Setting 5: High Signal-to-noise Ratio

$n = 500$, $p = 120$, $z = 100$, $snr = 7$, $\rho = 0.5$

Measure	Lasso Cox	Lasso Casebase
MCC	0.502	0.313
TPR	0.950	0.950
TNR	0.710	0.460
FPR	0.290	0.540
FNR	0.050	0.050

Table 14: Variable Selection (Setting 5)

Simulation Setting 5: High Signal-to-noise Ratio

$n = 500$, $p = 120$, $z = 100$, $snr = 7$, $\rho = 0.5$

Method	Concordance
Unpenalized Cox	0.832
Lasso Penalized Cox	0.881
Ridge Penalized Cox	0.865
Lasso Penalized Casebase	0.842
Ridge Penalized Casebase	0.870

Table 15: Concordance (Setting 5)

Simulation Setting 5: High Signal-to-noise Ratio

$n = 500$, $p = 120$, $z = 100$, $snr = 7$, $\rho = 0.5$

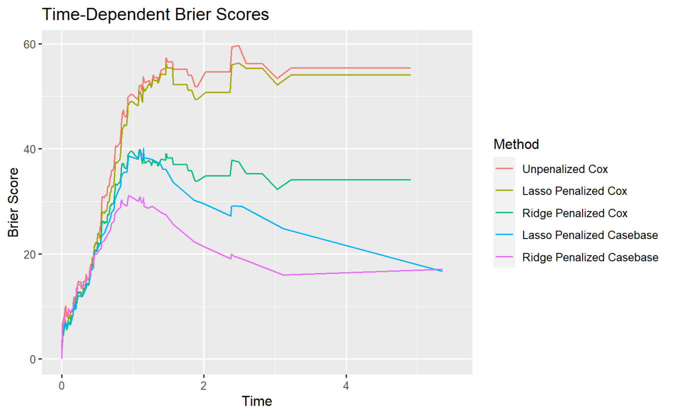

Figure 19: Time-Dependent Brier Scores (Setting 5)

Simulation Setting 6: Low Signal-to-noise Ratio

$n = 500, p = 120, z = 100, snr = \frac{1}{7}, \rho = 0.5$

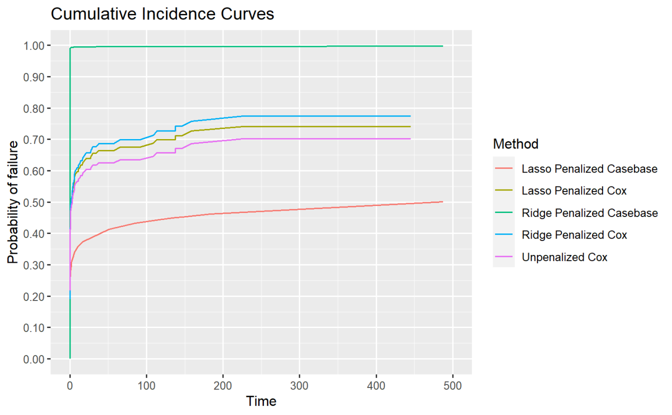

Figure 20: Cumulative Incidence Curves (Setting 6)

Simulation Setting 6: Low Signal-to-noise Ratio

$n = 500$, $p = 120$, $z = 100$, $snr = \frac{1}{7}$, $\rho = 0.5$

Measure	Lasso Cox	Lasso Casebase
MCC	0.297	0.212
TPR	0.250	0.200
TNR	0.960	0.950
FPR	0.040	0.050
FNR	0.750	0.800

Table 16: Variable Selection (Setting 6)

Simulation Setting 6: Low Signal-to-noise Ratio

$n = 500$, $p = 120$, $z = 100$, $snr = \frac{1}{7}$, $\rho = 0.5$

Method	Concordance
Unpenalized Cox	0.558
Lasso Penalized Cox	0.592
Ridge Penalized Cox	0.558
Lasso Penalized Casebase	0.943
Ridge Penalized Casebase	0.998

Table 17: Concordance (Setting 6)

Simulation Setting 6: Low Signal-to-noise Ratio

$n = 500, p = 120, z = 100, snr = \frac{1}{7}, \rho = 0.5$

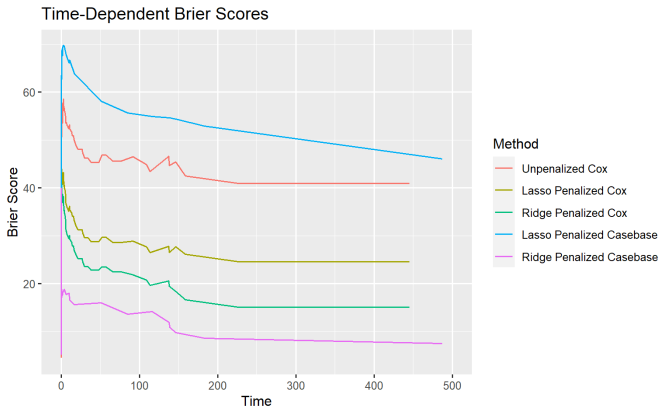

Figure 21: Time-Dependent Brier Scores (Setting 6)

Results: Summary

- Penalized Casebase models performed better on average in any setting compared to unpenalized and penalized Cox models in terms of prediction performance.
- Lasso Cox was slightly better at variable selection compared to lasso Casebase in all settings.
- In $p > n$ settings, penalized models showed better prediction performance, especially penalized Casebase models.
- Correlation between predictors had little effect on prediction performance between models.
- All models suffered in the low signal-to-noise ratio setting except in the Concordance metric for Casebase models.
- In settings where the signal-to-noise ratio was low or when $p > n$, lasso models yielded low TPR and high TNR.

Acknowledgements

Main Advisor:

Prof. Gabriela V. Cohen Freue (UBC)

Co-Advisors:

Prof. David Soave (Wilfrid Laurier University)

Prof. Sahir Bhatnagar (McGill University)

Funding:

CANSSI Collaborative Research Team: Improving robust high-dimensional causal inference and prediction modelling.

References

- Bhatnagar, S.R., Turgeon, M., Islam, J., Hanley, J.A., Saarela, O. (2020). casebase: An Alternative Framework For Survival Analysis and Comparison of Event Rates. *Journal of Statistical Software*. VV(II), doi:10.18637/jss.v000.i00
- Breslow N (1972). "Discussion of the paper by DR Cox cited below." Journal of the Royal Statistical Society, Series B, 34, 187–220.
- Cox DR (1972). "Regression models and life-tables." Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187–202.
- Simon, N., Friedman, J., Hastie, T., Tibshirani, R. (2011) Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent, Journal of Statistical Software, Vol. 39(5) 1-13 https://www.jstatsoft.org/v39/i05/

References

- Hanley JA, Miettinen OS (2009). “Fitting smooth-in-time prognostic risk functions via logistic regression.” The International Journal of Biostatistics, 5(1).
- Bhatnagar S, Turgeon M, Islam J, Saarela O, Hanley J (2020). casebase: Fitting Flexible Smooth-in-Time Hazards and Risk Functions via Logistic and Multinomial Regression. R package version 0.9.0, URL https://CRAN.R-project.org/package=casebase.
- Friedman, J., Hastie, T. and Tibshirani, R. (2008) Regularization Paths for Generalized Linear Models via Coordinate Descent, https://web.stanford.edu/ hastie/Papers/glmnet.pdf Journal of Statistical Software, Vol. 33(1), 1-22 Feb 2010 https://www.jstatsoft.org/v33/i01/
- Soave, D., Lawless, J.F. (2021). Regularized Regression for Two Phase Failure Time Studies. DOI: 10.1002