Title
Survival Analysis on United Network for Organ Sharing(UNOS) Kidney Transplant Program

Author
Lee, Seungyeon

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Survival Analysis on United Network for Organ Sharing(UNOS) Kidney Transplant Program

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Applied Statistics

by

Seungyeon Lee

2023

ABSTRACT OF THE THESIS

Survival Analysis on
United Network for Organ Sharing(UNOS)
Kidney Transplant Program

by

Seungyeon Lee
Master of Applied Statistics
University of California, Los Angeles, 2023
Professor Xiaowu Dai, Chair

This paper conducts survival analysis on the kidney transplant-related data collected by United Network for Organ Sharing (UNOS) since 1987. We investigate which kidney transplant-related variables from UNOS have significant effect on recipients' survival time and the extent of the effects. Standard Cox Proportional-Hazards model, Cox-Proportional Hazards model with ridge, LASSO and elastic net penalty and lastly random survival forest model are used to compute the survival function after kidney transplantations, which we use to estimate patients' survival probability at time t of a given transplant. Length of stay(LOS) at the hospital after transplant comes out to be the most significant variable in determining patient's survival probability. The longer the patient stayed post transplant, the higher the survival probability. Following the modeling, we compare these models based on two different scores: concordance index and the brier score. Concordance index checks whether the models generate reliable ranking of survival times(i.e. discrimination), while the brier score calculates the average squared distances(i.e. calibration). Random Survival Forest model

provides the best result based on the brier score, while Kaplan-Meier model produces the best outcome based on the concordance index.

The thesis of Seungyeon Lee is approved.

Nicolas Christou

Hongquan Xu

Xiaowu Dai, Committee Chair

University of California, Los Angeles

2023

To my family and friends . . .

who have supported, motivated and inspired me

through graduate school

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Chronic kidney disease is a serious health problem that poses a risk to the lives of countless individuals around the globe. Kidney damage resulting from this disease can lead to permanent loss of organ function, ultimately resulting in kidney failure. To sustain life in such cases, individuals need to undergo ongoing dialysis treatment or receive a kidney transplant.

Transplantation is the favored method of treatment for kidney failure. However, there is a significant shortage of available donor kidneys compared to the demand. Patients can receive a transplanted kidney from either a deceased individual or a living person. Approximately two-thirds of transplanted kidneys are obtained from deceased donors, while the remaining one-third comes from healthy living donors who willingly offer their kidneys [8]. Given the scarcity of kidneys available for transplant, it is crucial to understand what factors are related to a higher graft survival rate post transplant. Information about which factors are related to higher graft survival rate can be used during pre-transplant phase to more efficiently distribute kidney and during post-transplant phase to better care for the patients. In this paper, we analyze the kidney transplant data provided by United Network for Organ Sharing (UNOS) organization compiled since 1987 and build three survival models to unveil which tranplant-related variables significantly affect patients' survival probability and quantify their extent.

Our three models consist of the standard Cox Proportional-Hazards model, Cox Proportional-Hazards with different penalty such as ridge, LASSO and elastic net, and lastly the random survival forest model. Upon building the models, we compare the performance of the them

on two different scales: the concordance index and the brier score. With concordance index, we estimate the discrimination power of the model, while with the brier score we quantifies the calibration power of the model.

CHAPTER 2

Data

2.1 Data Source

Data for this paper comes from the United Network for Organ Sharing (UNOS) organization. UNOS has compiled transplant data for different organs since 1987 . In this paper, we are specifically interested in the kidney transplant program and how different variables affect patients's survival after the transplant.

To further describe the data, it includes both deceased and living donor transplants. Each record is per waiting list registration/transplant event, and each record includes the most recent follow-up information such as graft survival reported to Organ Procurement and Transplantation Network (OPTN) on the date of file creation. If a patient registered for a transplant, but either was removed before the transplant occurred or still waiting, all transplant-related data is set to null. In other words, we have registration-related data but not transplant-related data. On the other hand, if a patient received a transplant but was never on the waiting list, we will have all transplant-related data but not the registration-related data. Each registration and transplant has a unique code assigned in the data for identification.

The original data has 982,456 entries and 491 variables. Through data cleaning and preparation, the size of the data we conduct survival analysis is reduced to 7,257 entries and 47 variables. I will explain in more details about the data cleaning and preparation process in the following sections.

2.2 Data Cleaning

As mentioned in the previous section, the original data contains 982,456 entries and 491 variables. We filter out certain variables and entries based on the following logic:

- The dataset contains information on both kidney and pancreas transplant data. We filter out kidney transplant entries when WL_ORG variable is KI.

- Some variables were overlapping in information. For instance, HLAMIS variable is based on A1/DA1, A2/DA2, B1/DB1, B2/DB2, DR1/DDR1, DR2/DDR2 variables. Therefore in situations like these, we proceed with the most inclusive variable and drop the variables used for calculation.

- Some variables convey the same information but in a slightly different way. In these situations, we proceed with the more informative variable and drop the other variables[1]. For example, PREV_TX_ANY shows whether patient has received any prior transplants and NUM_PREV_TX is the count of the total number of previous transplants for a given patient. Since NUM_PREV_TX variable includes information portrayed in PREV_TX_ANY, we drop PREV_TX_ANY and proceed with NUM_PREV_TX.

- We drop variables with more than 60 percent of entries missing. We are assuming entries are missing at random. I considered imputation however when more than handful of entries are missing, imputation is not very feasible. Therefore we proceed with dropping variables with more than 60 percent of entries missing [12].

- We also drop variables that we don't have a clear understanding of due to lack of documentation. For instance, DGN_TCR is a categorical variable with 78 unique categories; however it is unclear what each category represents.

Then we drop all entries with any missing information from the remaining variables.

2.3 Data Preparation

After we clean the data as described in the previous section, AGE and AGE_DON variables are included in the dataset. However to utilize Cox Proportional Hazard model, which I further explain in chapter four, we need all variables to be time-independent. Therefore instead of these two variables, we calculate the difference in age between the recipient and donor and take the absolute value and create a variable called AGE_DIFF and add it to our dataset. This allows to retain some degree of information regarding recipients' and donors' age while keeping all variables time-independent.

Based on the selected entries and variables for our final dataset, we plot a correlation matrix(see figure 2.1) to see the relationships between different numeric variables. We confirm that all numeric variables have low correlation, with correlation below 0.5 except correlation between KDPI and AGE_DON at correlation 0.70. We proceed with both variables since we would like to estimate the effect on hazard ratio for both covariates.

2.4 Exploratory Data Analysis

In this section, I will provide visualization of our final dataset to provide better understanding of it. First in figure 2.2, we see how many people are alive after kidney transplant at the time of data collection. We are happy to see that our final dataset has a balanced count between alive and dead patients.

Figure 2.3 provides a visual comparison of the survival functions between male and female. We see a general trend where female has a higher survival function compared to male. However, we do have to keep in mind visual comparisons are highly subjective and for a more conclusive result on difference in survival functions, log-rank test is more appropriate.

In figure 2.4, we see a few interesting trends. Before I describe the plots, I'll state what each variable in figure 2.4 means: LOS stands for recipient length of stay post transplant,

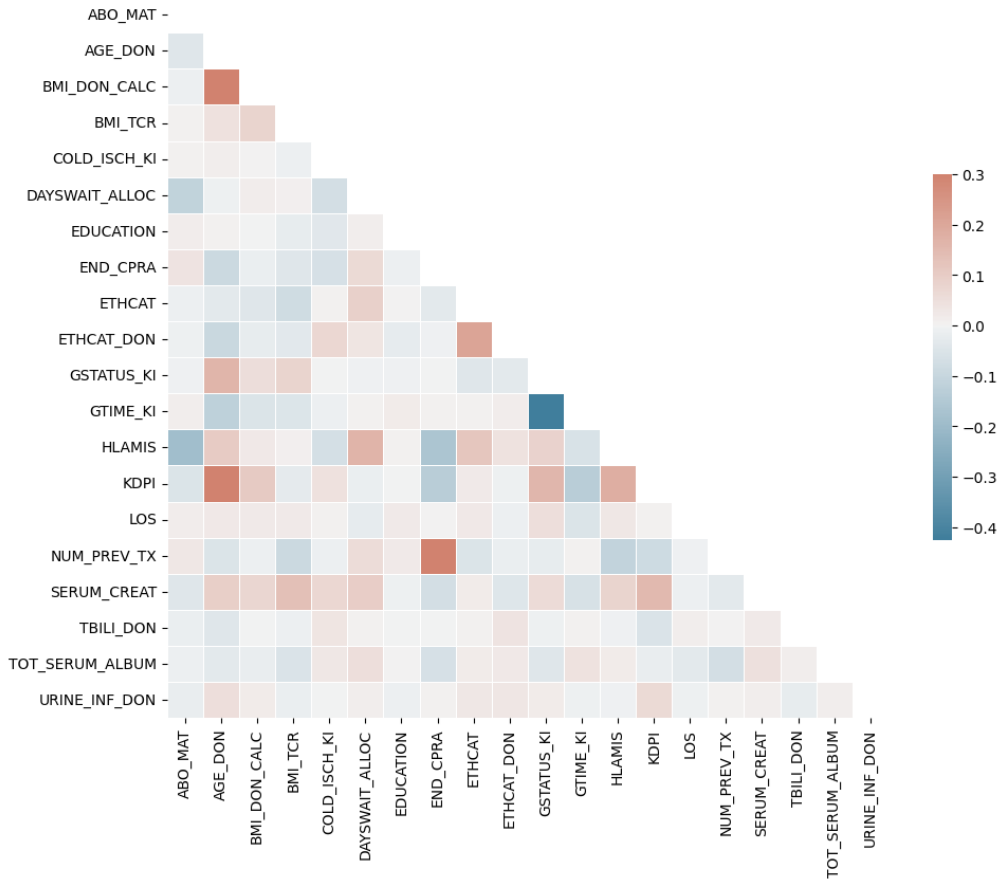

Figure 2.1: Correlation Matrix

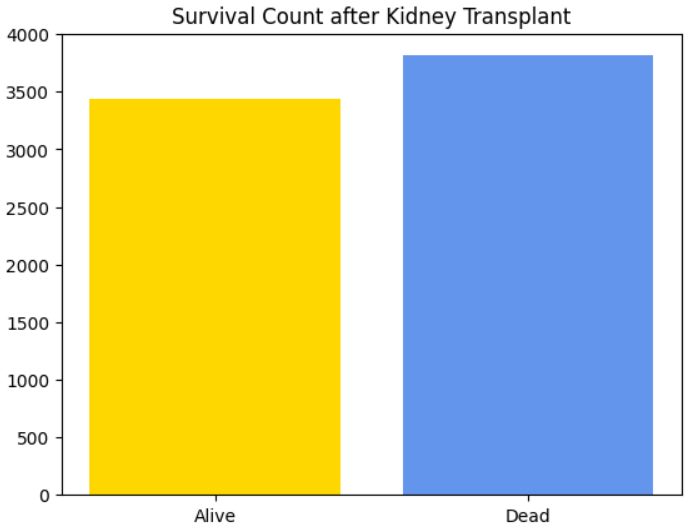

Figure 2.2: Survival count after kidney transplant at the time of data collection

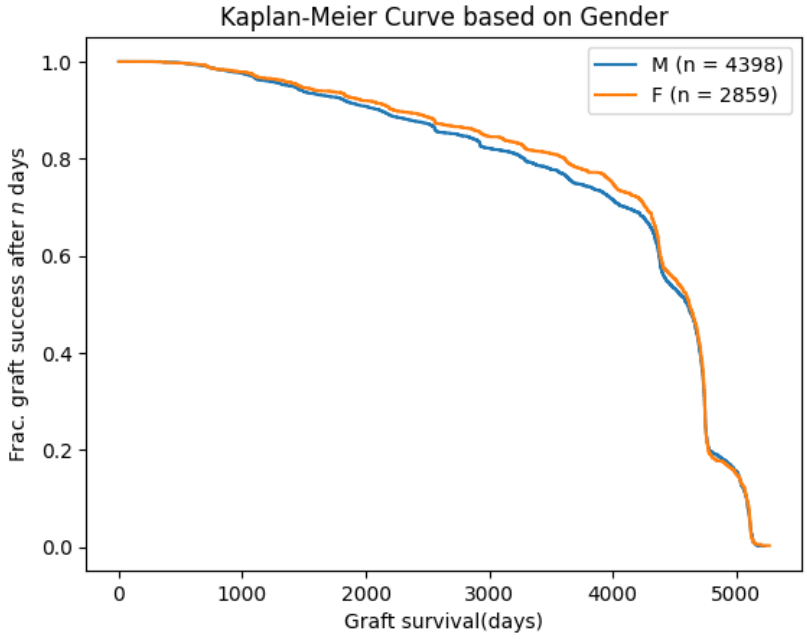

Figure 2.3: Kaplan-Meier Curve based on Gender

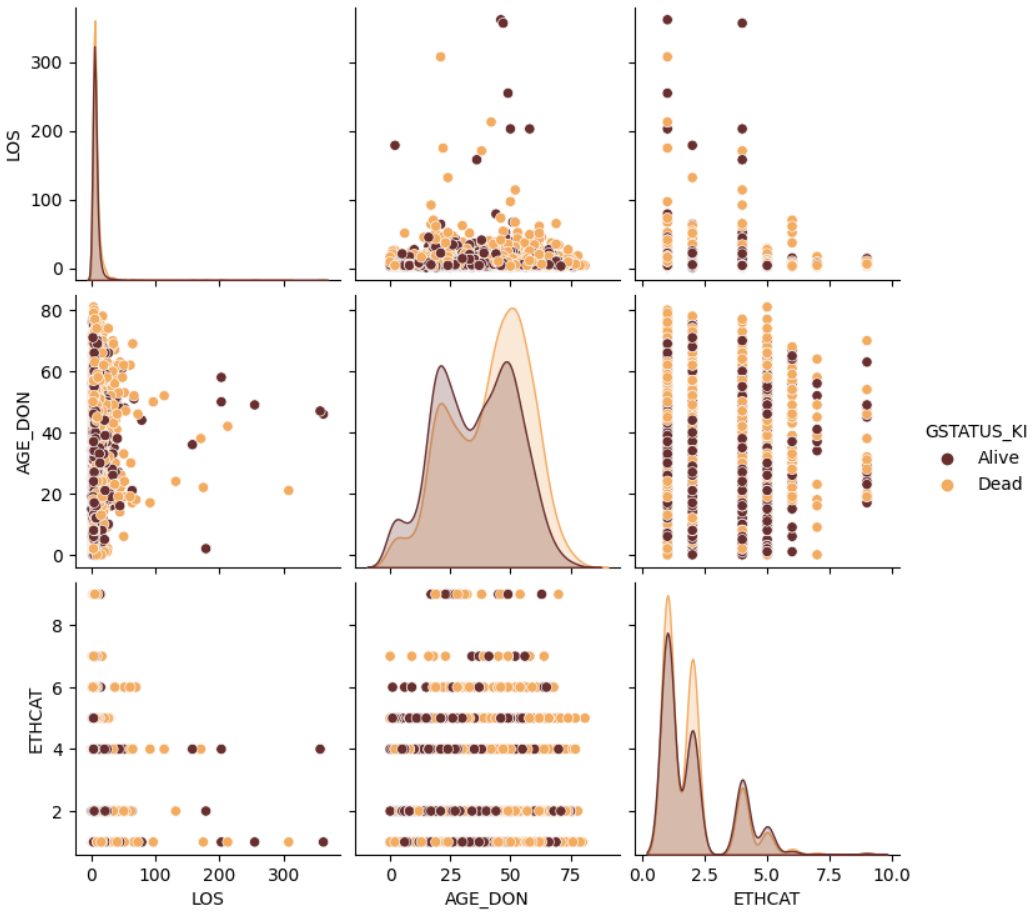

Figure 2.4: LOS, AGE_DON, ETHCAT of Living and Deceased Recipients

AGE_DON is the age of the donor, and ETHCAT is the ethnicity category of the recipient. First for deceased transplant recipients, the age of donors seem to be left skewed; whereas for living recipients we see a relatively normally distributed curve. In other words, recipients that passed away tend to receive kidneys from older donors. Moreover, we see recipients with certain ethnic category with higher death rate. For instance, we see ethnic category 1 and 2 has a higher death rate than ethnic category 4 and 5. Lastly recipients that had a longer length of stay(LOS) seem to have a lower death rate, which corresponds to the cox proportional-hazards model results described later in the paper.

CHAPTER 3

Methodology

3.1 Survival Analysis

Survival analysis is used to analyze time until an event of interest occurs. It is often used in medical research but also in other areas such as engineering to calculate hardware failure time, customer analytics for customer churn rate and even inventory management to track time-to-sale. Survival analysis is particularly useful when dealing with censored data, where the event of interest has not occurred for all subjects by the end of the study or when they are lost to follow-up since it takes this into account this information to calculate time to event and the probability of experiencing the event.

With survival analysis we are primarily interested in survival time, which represents the time until the event of interest occurs. One thing to note is event doesn't necessarily have to be a negative outcome like death. It can also be a positive event, such as the time to recovery or the time to achieving a particular milestone.

Currently, Kaplan-Meier estimator and Cox proportional hazard regression are most commonly used survival analysis methods. Kaplan-Meier estimator is a non-parametric estimator to measure the fraction of surviving units for a certain time after treatment taking censored observations into account. However if we want to consider multiple covariates, Kaplan-Meier quickly becomes infeasible because the size of subgroups will become very small. Unlike Kaplan-Meier estimator, Cox proportional hazard regression allows for multiple covariates and estimate the impact of each variable has on survival time. In this section,

I will be explaining both Kaplan-Meier and Cox proportional Hazard methods as well as Random survival forest, which are used in my modeling to establish connection between covariates and survival time of kidney transplant patients from the time of transplant.

3.1.1 Censored Data

In survival analysis, we often find censored data where we do not know the exact time of the event of interest. Depending on the industry, the event of interest can be referred to as "failures", "deaths" or simply "events". For simplicity, I will either use "event of interest" or "deaths" among the aforementioned options in this paper moving forward.

Generally there are two types of censored data: right censored and left censored. Right censored data refers to data where survival time becomes incomplete on the right side of the study end date or when the patient is lost to follow-up or is withdrawn. In other words, the patients or subjects survived pass the study end date and death is expected to occur some time in the future but not observed. In our analysis, kidney transplant patients that were alive at the end of the data collection date would be considered "right censored" as well as patients that where lost to follow-up and have withdrawn. The plot below demonstrates the different types of right censored scenarios that I have mentioned previously. Survival time of subject A is unknown since he or she is lost. As for subject C, we do not know the survival time since the subject dropped out of the study. Lastly, survival time of subject E surpasses the study end date therefore, we do not have information on exact survival time of subject E.

Moreover, data can be left censored. Left censored data can occur when true survival time is shorter than or equal to observed survival time. For example, let's say we are following subjects until they test positive for a certain virus and our event of interest is the first time they test positive for the given virus. Oftentimes, we may not know the exact time of subjects' first exposure to the virus, and therefore do not know exactly when the patients first started testing positive to the virus. In other words when data is left censored, true

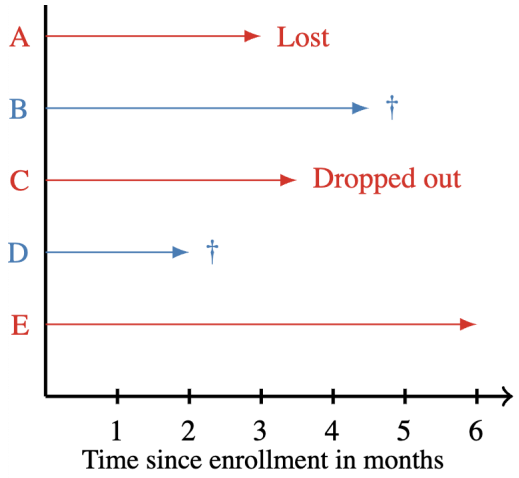

Figure 3.1: Right Censored Data [9]

survival time which ends at event of interest(i.e. first exposure in the virus case), is less than the observed length for subjects to test positive. Differently put if a subject is left-censored, we know event of interest occurred between time 0 and t, but do not know the exact time of event.

On another note, right censored data are much more common than left censored data. In our UNOS data, we indeed do see a handful of right censored observations; on the other hand, left censored data does not apply to our UNOS dataset since each observation is based on either a registration or transplant date.

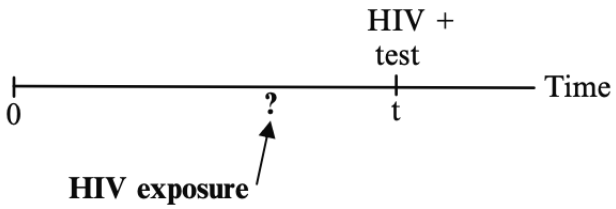

Figure 3.2: Left Censored Data [4]

3.1.2 Kaplan-Meier estimator

As mentioned previously, Kaplan-Meier is a non-parametric estimator which calculates the survival function at a given time. In this section, I will describe how to estimate and graph survival curves using Kaplan-Meier. Moreover, I will also explain how log-rank test can be used to compare two or more survival curves.

$$\begin{aligned} S(t_f) &= \prod_{i=1}^{f} Pr[T > t_i | T \geq t_i] \\ &= S(t_{f-1}) \cdot Pr[T > t_f | T \geq t_f] \end{aligned} \tag{3.1}$$

Kaplan-Meier survival probability at event time of interest is as the formula above. It is the probability of surviving past t_{f-1}, multiplied by the conditional probability of surviving past time t_f given surviving past t_{f-1}. This formula can be also expressed as a product limit when survival probability $S(\hat{t_{f-1}})$ is substituted with the product of all conditional probabilities for t_{f-1} or before.

We can understand Kaplan-Meier estimator in a pretty simple manner using the probability of joint event. As described below, probability of a joint event, say A and B, is equal to the probability of event A times the conditional probability of the event B, given A[4].

$$Pr(A \cap B) = Pr(A) \cdot Pr(B|A)$$

Furthermore, we often are curious in survival analysis if two different subgroups behave similarly in their survival curves. Although comparing two different Kaplan-Meier curves can provide visual comparison, log-rank test offers a statistical metric on whether the two subgroups' survival curves are equivalent. However, it is important to keep in mind that the result of log-rank test does not indicate we have proof that the true survival curves are equivalent.

Log-rank test is essentially a large-sample chi-square test where we compare the expected and observed values of deaths at a given f where f signifies each ordered failure time. Let's say we are interested in whether two subgroups' Kaplan-Meier curves are equivalent. Assuming there are two groups for comparison, we calculate below for each subgroup for each time period t to get the expected count of deaths. For the first group, $\frac{n_{1f}}{n_{1f}+n_{2f}}$ is the proportion of first group in risk set given time t and $[m_{1f} + m_{2f}]$ is the total number of failures over both groups at time t. We can apply the same logic to calculate the expected count of death for the second group.

$$e_{1f} = \frac{n_{1f}}{n_{1f} + n_{2f}} \cdot [m_{1f} + m_{2f}] \tag{3.2}$$

$$e_{2f} = \frac{n_{2f}}{n_{1f} + n_{2t}} \cdot [m_{1f} + m_{2f}] \tag{3.3}$$

Then, we calculate the difference between observed and expected for each subgroup for each time period f and sum up the difference. Equation below describes this process. i signifies which group the value is for and F means total number of failure times.

$$O_i - E_i = \sum_{i=1}^{F} (m_{if} - e_{if}) \text{ for i = 1,2} \tag{3.4}$$

Finally, we calculate thee log-rank statistics like below.

$$\text{Log-rank statistics} = \frac{(O_i - E_i)^2}{\text{Var}(O_i - E_i)} \tag{3.5}$$

where

$$Var(O_i - E_i) = \sum_f \frac{n_{1f} n_{2f} (m_{1f} + m_{2f})(n_{1f} + n_{2f} - m_{1f} - m_{2f})}{(n_{1f} + n_{2f})^2 (n_{1f} + n_{2f} - 1)} \text{ for i = 1,2} \tag{3.6}$$

Note that the variance is the same for two subgroups we are comparing. Finally, we find the p-value of the log-rank statistic using chi-square distribution with one degree of freedom,

which will determine whether to reject the null hypothesis that states there is no difference between survival curves.

We can also use the log-rank test to compare more than two Kaplan-Meier curves. We simply update the null hypothesis to state all curves are the same, and use both variances and covariances of difference between summation of observed minus expected for each subgroup.

Although Kaplan-Meier is a very intuitive way of presenting survival functions, it is not very feasible if have more than one or two covariates since the size of subgroup can be very small. When conducting survival analysis for data with multiple covariates, we can use Cox's proportional hazard's model, which I will cover in the next section.

3.1.3 Cox's Proportional Hazard's Model

For survival analysis with multiple covariates, it is useful to use Cox's proportional hazard's model as it provides information on the impact of different covariates on survival outcomes, as well as it mitigates the issue of subgroups becoming too small which often occur in Kaplan-Meier with multiple covariates[4].

With Cox's proportional hazard's model, our main metric of interest is the hazard ratio. The hazard ratio is the ratio of hazard rates between two groups, which can be expressed like below.

$$\frac{\lambda(t, \mathbf{Z})}{\lambda_0(t)} = e^{\sum_{i=1}^{p} \beta_i Z_i} \tag{3.7}$$

The hazard ratio will tell us how much each covariate contributes to the survival outcome. As with the logistic model, this hazard ratio is expressed in terms of an exponential of one or more regression coefficients in the model.To obtain the hazard ratio, we use the product over the likelihood contribution like below, then maximize to get the partial maximum likelihood estimator for β.

$$\prod_{i=1}^{n}[\frac{e^{\boldsymbol{\beta Z_i}}}{\sum_{j\in R(X_i)} e^{\boldsymbol{\beta Z_j}}}]^{\delta_i} \tag{3.8}$$

$R(X_i)$ is the risk set at the failure time of individual i, δ_i is the failure/censoring indicator(1 is death; 0 is censored), and Z_i represents a set of covariate.

Once we have the β, we can obtain the hazard ratio. Log hazard ratio can be expressed like below.

$$log\frac{\lambda_i(t,\mathbf{Z}_i)}{\lambda_0(t)} = \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Z_{3i} + ... + \beta_p Z_{pi} \tag{3.9}$$

As shown above, we can obtain β without any assumption on $\lambda_0(t)$, therefore we can say Cox's proportional hazard's model is semi-parametric. However we do have to keep in mind the model assumes hazard ratios for the covariates remain constant over time. In other words, the hazard functions of different groups would be proportional to each other.

3.1.4 Random Survival Forests

In this section, I explain how random forest can be applied to survival analysis. Random forest is most commonly applied to classification and regression to reduce overfitting hence produces robust outcome to noise and outliers. With random survival forest, we expect this same benefit but in the context of survival analysis. On a high level, random survival forest works like the following [7].

1. We draw bootstrap samples from the training data.

2. For each bootstrap sample, we generate a survival tree with p randomly selected variables. We split each node with the variable that will maximize survival difference between children nodes.

3. We grow each tree until the last node have at lease one unique death.

4. Then we calculate the ensemble cumulative hazard function for each tree, and average those values.

5. Lastly, we use the training data to calculate the prediction error of the ensemble cumulative hazard function.

Mathematically, each tree's cumulative hazard function is as below. $t_{1,h} < t_{2,h} < t_{3,h} ... < t_{N(h),h}$ are distinct event times where individual l is said to have died at time $t_{l,h}$. $d_{l,h}$ and $Y_{l,h}$ are number of deaths and individuals at risk at time $t_{l,h}$ respectively.

$$\hat{H}_h(t) = \sum_{t_{l,h} \leq t} \frac{d_{l,h}}{Y_{l,h}} \tag{3.10}$$

Based on the above equation, w we can calculate the bootstrap ensemble cumulative hazard function like below where B is the number of survival trees.

$$H_e(t|\mathbf{x}_i) = \frac{1}{B} \sum_{b=1}^{B} H_b{}^{(}t|\mathbf{x}_i) \tag{3.11}$$

given $H(t|\mathbf{x}_i) = \hat{H}_h(t)$ if $\mathbf{x}_i \in h$.

3.2 Evaluating Survival Models

The previous section covered different ways of modeling hazard and survival functions. In this section, we explore how to compare and evaluate those different survival models. I will go over the most commonly used Harrell's concordance index, also known as the c-index or c-statistics, as well as Receiver Operating Characteristic (ROC) curve which is useful when comparing survival models on a specific time range. And lastly, I explain Brier score that assesses calibration, which is not assessed with the c-index [13].

3.2.1 Harrell's concordance index

Harrell's concordance index provides insights on the models' discrimination power. In other words, it checks for whether the models generate reliable ranking of survival times based on individual risk scores [11]. The c-index can be calculated like below.

$$\text{C-index} = \frac{\sum_{i,j} 1_{T_j<T_i} \cdot 1_{\eta_j>\eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j<T_i} \cdot \delta_j} \tag{3.12}$$

η_i is the risk score of unit i and if $T_j < T_i$, then $1_{T_j<T_i} = 1$;otherwise 0. Similarly, if $\eta_j < \eta_i$, then $1_{\eta_j>\eta_i} = 1$;otherwise 0. More intuitively, the above formula can be expressed as the following.

$$\text{C- index} = \frac{\text{number of concordant pairs}}{\text{number of concordant pairs} + \text{number of discordant pairs}} \tag{3.13}$$

We call (i, j) is a concordant pair if $\eta_i > \eta_j$ and $T_i < T_j$, and it is a discordant pair if $\eta_i > \eta_j$ and $T_i > T_j$.

The closer the C-index is to 1, the better the model prediction. On the other hand C-index of 0.5 means the prediction is as good as a random prediction.

Although the C-index is intuitive to interpret and calculation is not complicated, there are some drawbacks. First, it tends to be too optimistic when there is more censored data present [10]. Second, it is not helpful in estimating performance if we are interested in event of interest occurring within a specific time range. The first issue can be addressed using an an alternative c-index estimator where we use the c-index for right-censored data based on inverse probability of censoring weights [9]. The second issue can be mitigated by utilizing the receiver operating characteristic curve (ROC curve). With ROC, we can compute how well a model can predict whether subjects will experience the event or not at a given time using sensitivity and specificity. I will cover ROC in more details in the following section.

3.2.2 Time-dependent Area under the ROC

Area under the receiver operating characteristics curve (ROC curve) is a well-known performance estimator for binary classification. Based on the predicted risk score, the ROC curve visualizes the specificity against the sensitivity rate [5]. When applying the ROC curve to survival time in particular, we have to keep in mind the subject's status changes over time. As a consequence, the sensitivity and specificity become dependent on time. Therefore, we calculate the ROC curve for a given time using cumulative cases where subjects experience the event prior or at time $t(t_i \leq t)$ and dynamic controls who are subjects with $t_i > t$ [6]. Then obtain area under curve (AUC) information for each ROC curve, which can be visualized like the following. We have to keep in mind this method is most relevant when we are interested in predicting up to time t instead of a specific point in time.

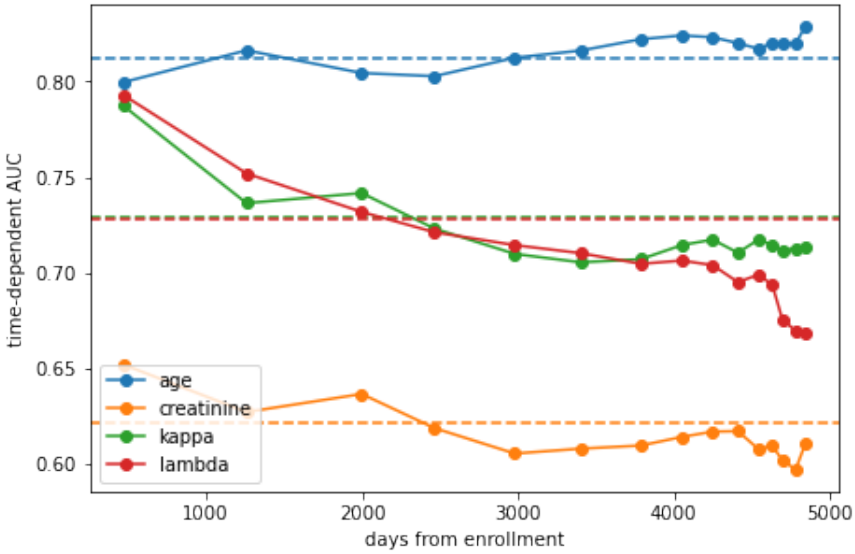

Figure 3.3: Example Plot of Area Under Curve (AUC) for Survival Analysis [9]

3.2.3 Time-dependent Brier Score

Although the concordance index and the ROC curve provide metrics for discrimination, which confirms whether the model's predicted risk scores correctly determine the order of events, they lack in assessing calibration [2]. Fortunately, Brier score works as a metric for calibration as well as discrimination. The Brier score is used to evaluate the accuracy of a predicted survival function at a given time t. It calculates the average squared distances between the observed survival status and the predicted survival probability. It ranges between 0 and 1, with 0 representing the most accurate model. If no right censoring is present in the data, the Brier score can be calculated like below.

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\left(1_{T_i>t} - \hat{S}(t|x_i)\right)^2 \tag{3.14}$$

However, when the dataset has subjects that are right-censored, we must adjust the score by using inverse probability of censoring like below.

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{(0 - \hat{S}(t|\boldsymbol{x_i}))^2 \cdot 1_{T_i\leq t,\delta_i=1}}{\hat{G}(T_i^-)} + \frac{(1 - \hat{S}(t|\boldsymbol{x_i}))^2 \cdot 1_{T_i>t}}{\hat{G}(t)}\right) \tag{3.15}$$

$\hat{G}(t) = P[C > t]$ is the estimator of the conditional survival function of the censoring times calculated using the Kaplan-Meier method, where C is the censoring time. A predictive model will have a Brier score lower than 0.25.

CHAPTER 4

Modeling

In this chapter, I summarize our modeling results. We will look into and compare results of three models: regular cox proportional-hazards model, cox proportional-hazards model with different penalties, and lastly the random survival forest model.

4.1 Regular Cox Proportional-Hazards Model

Covariates	**Feature Importance**	**Coefficient**	$e^{\textbf{Coeffcient}}$
LOS	0.5155	-0.0010	0.9990
ETHCAT	0.5155	0.0292	1.0296
AGE_DON	0.5152	-0.0035	0.9965
EDUCATION	0.5144	-0.0002	0.9998
RT_KI_BIOPSY=Y	0.5115	0.0109	1.0109
GENDER=M	0.5113	0.0674	1.0698
KDPI	0.5112	0.0016	1.0016
HIST_HYPERTENS_DON=Y	0.5108	-0.1281	0.8797
ON_DIALYSIS=Y	0.5094	-0.0006	0.9994
COLD_ISCH_KI	0.5094	-0.0007	0.9993

Table 4.1: Top 10 Covariates based on Feature Importance

Of the 47 variables in the final dataset, I list coefficients of top ten covariates based on their feature importance in table 4.1. To calculate the feature importance, we fit a cox model

to each variable individually and obtain the c-index. Based on the feature importance, we interpret LOS, which represents the length of stay at the hospital after transplant, to have the most predictive power [3].

Furthermore we can use the rankings from above to select which variables to include in our final dataset, but still need to determine the optimal number of covariates. We perform a grid search to select the optimal cut-off. In table 4.2, I list the optimal number of covariates based on their mean test score. Based on the result, we conclude the optimal number of variables to be seventeen.

Number of Covariates	**Mean Test Score**
17	0.521801
15	0.52136
18	0.520827
16	0.52066
13	0.520368

Table 4.2: Mean Test Score based on Number of Covariates

4.2 Penalized Cox Proportional-Hazards Models

The standard Cox Proportional-Hazards model provides great insight into how each covariate affects the hazard function. However when we have to estimate coefficient of many covariates, the standard model may not work since it cannot invert a matrix that becomes non-singular due to correlations among features.

4.2.1 Ridge

Aforementioned mathematical issue can be mitigated by using the ridge penalty, which adds the l_2 term on the coefficients and brings down the coefficients to zero like the equation

below.

$$\arg\max_{\beta} = log\mathrm{PL}(\beta) - \frac{\alpha}{2}\sum_{j=1}^{p}\beta_j^2 \tag{4.1}$$

where $\mathrm{PL}(\beta)$ is the partial likelihood function of the Cox Proportional-Hazards model.

In plot 4.1 below, we can see how the coefficient values decrease as the penalty weight alpha increases. Moreover we see variables such as EXH_VASC_ACCESS=Y, TRTREJ6M_KI=Y, EXH_PERIT_ACCESS=Y, DIAB=998 and DIAB=4 decrease in a steeper fashion than the other coefficients, which shows that these variables are crucial predictors in determining post-transplant kidney failure.

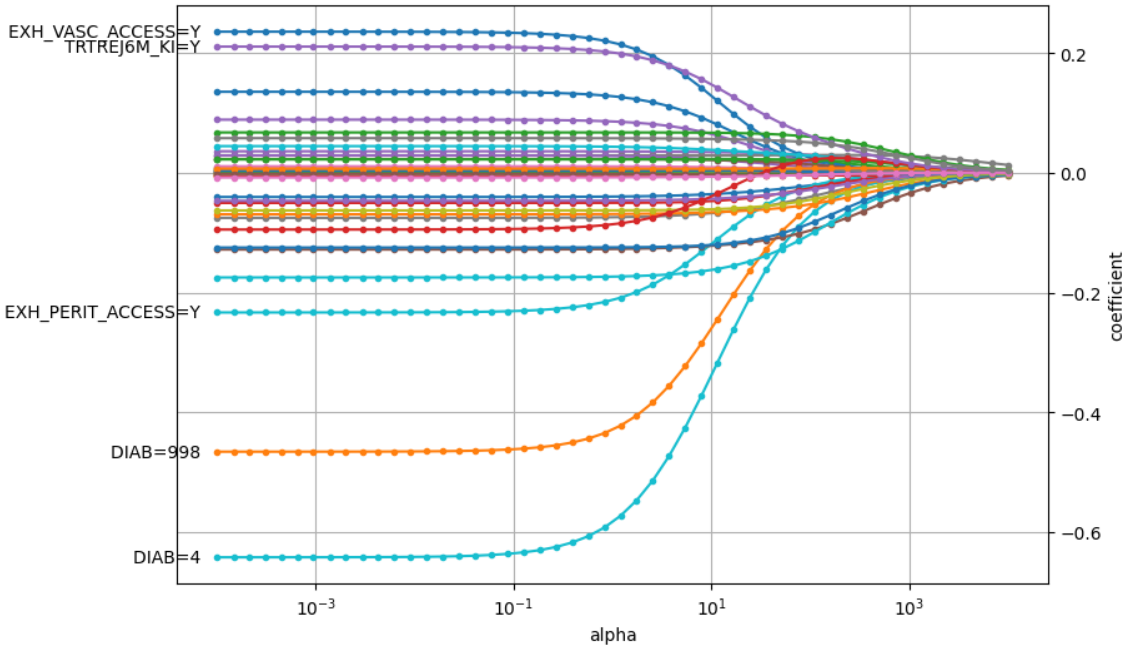

Figure 4.1: Cox Proportional-Hazards model with Ridge penalty

4.2.2 LASSO

Instead of shrinking coefficients to zero like we see with the ridge penalty, Least Absolute Shrinkage and Selection Operator (LASSO) performs a continuous subset selection of vari-

ables, where the selected variables are set to zero and therefore excluded from the model. This allows for reductions in the number of covariates used for prediction. We use the equation below to maximize on the β values.

$$\arg\max_{\beta} = log\text{PL}(\beta) - \alpha \sum_{j=1}^{p} |\beta_j| \tag{4.2}$$

where $\text{PL}(\beta)$ is the partial likelihood function of the Cox Proportional-Hazards model.

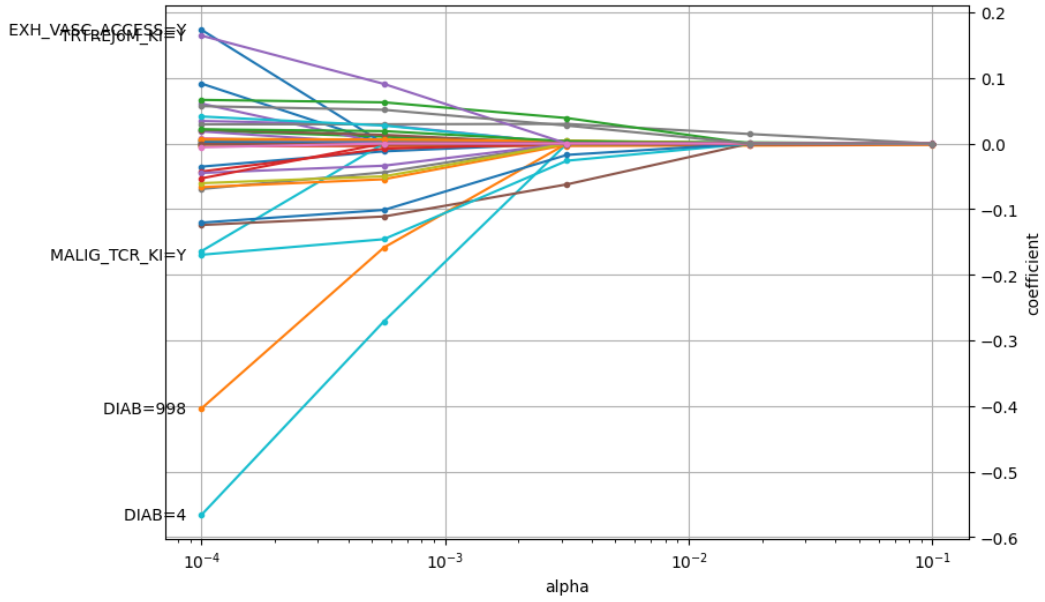

Figure 4.2: Cox Proportional-Hazards model with LASSO penalty

Figure 4.2 shows that the LASSO penalty selects a small subset of covariates when α is bigger. To be more specific, we see only a couple covariates with a non-zero coefficient when α is 0.01. We see similar variables indicated as crucial predictors as we saw in ridge penalty plot. EXH_VASC_ACCESS=Y, TRTREJ6M_KI= Y, , DIAB=998 and DIAB=4 stands out, in addition to MALIG_TCR_KI = Y, which states whether the recipient had previous history of malignancy.

4.2.3 Elastic Net

Elastic Net is a middle ground between ridge and lasso penalty. Generally if we know only a few covariates will be useful for prediction, we prefer LASSO or elastic net over ridge penalty. Between LASSO and elastic net, elastic net is preferred because when there are several strongly correlated features elastic net tend to select all; whereas LASSO chooses one randomly. Elastic Net uses the equation below to maximize on the β values.

$$\arg\max_{\beta} = log\text{PL}(\beta) - \alpha(r\sum_{j=1}^{p}|\beta_j| + \frac{1-r}{2}\sum_{j=1}^{p}\beta_j^2) \tag{4.3}$$

where $\text{PL}(\beta)$ is the partial likelihood function of the Cox Proportional-Hazards model and $r \in [0; 1]$ is the relative weight of the two penalties in the equation above. Usually it is sufficient to give the second penalty a small weight to improve stability of the model. For the elastic net model in this paper, I give second penalty a weight of 0.1.

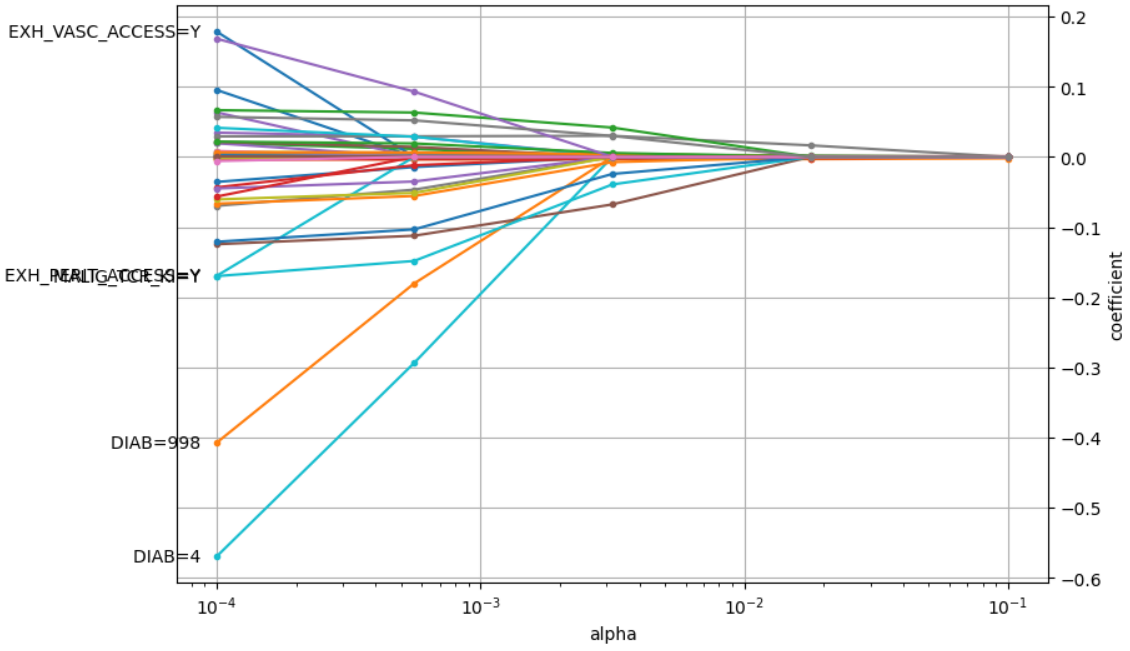

Figure 4.3: Cox Proportional-Hazards model with Elastic Net penalty

As for selecting an appropriate α value, we use GridSearchCV to train our training set

on a range of α values, then apply the compute coefficient values on five different sets of testing data to calculate the mean c-index score. Our result shows α value of 0.0178 results in highest mean c-index score of 0.521845 among the testing data as shown in figure 4.4.

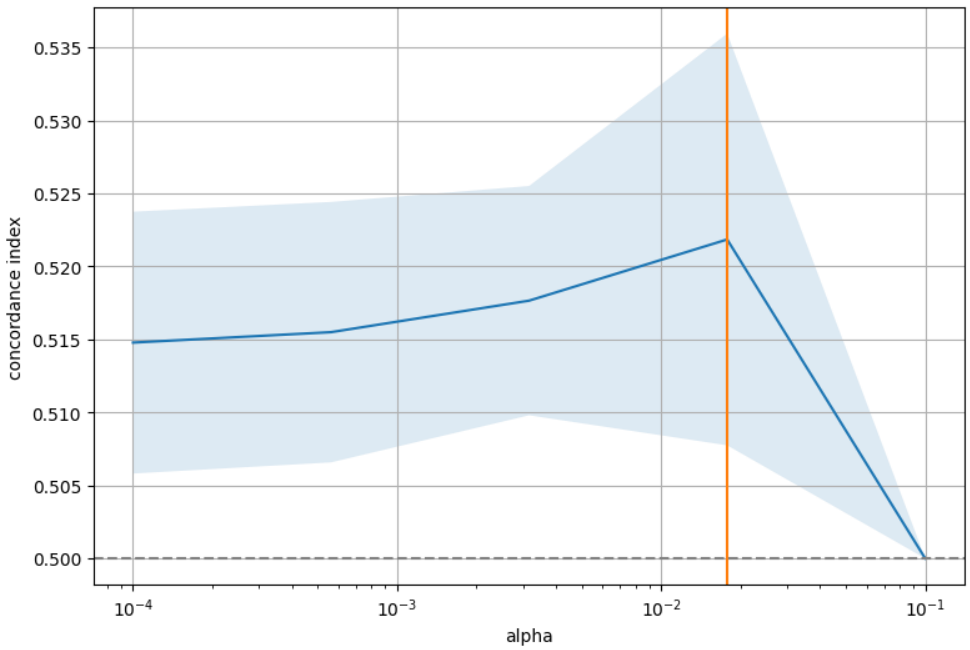

Figure 4.4: Finding α value for Elastic Net Cox Proportional-Hazards

4.3 Random Survival Forest Model

The random survival forest model selects different combination of entries and covariates from the training dataset to train on each instance. On each instance, it splits the training data into different leaves based on the log-rank test, which results in a tree. When we repeat this process numerous times, we end up with multiple trees, which will call a forest. For our training data, I fit a random survival forest with 1,000 trees. After fitting, I use the testing data to c-index score which comes out to 0.5469.

We are curious how much each feature is to the random survival forest model. To estimate this, we use permutation_importance function of scikit-learn package in Python, which

calculates the amount of decrease in log-rank test statistic due to a split in a tree. In table 4.3, I list covariates of top ten mean importance and corresponding standard deviation.

Variables	**Mean of Importance**	σ **of Importance**
LOS	0.017313	0.005098
SERUM_CREAT	0.006445	0.003396
AGE_DON	0.004997	0.00149
ETHCAT	0.004168	0.006387
TOT_SERUM_ALBUM	0.003832	0.001653
END_CPRA	0.003501	0.001523
BMI_TCR	0.002878	0.003264
SHARE_TY=5	0.002269	0.000799
HIST_HYPERTENS_DON=Y	0.00223	0.001842
RT_KI_BIOPSY=Y	0.002021	0.000839

Table 4.3: Mean and Standard Deviation of Importance Test Scores

The result shows that length of stay after transplant at the hospital (LOS) is by far the most important feature. If its relationship to survival time is removed (by random shuffling), the concordance index on the test data drops on average by 0.017313 points.

I would like to stress that features that are deemed to have low importance for a model with low cross-validation score could be more important for a model with high cross-validation score.

CHAPTER 5

Conclusion

5.1 Conclusion

In this paper, we conducted survival analysis with three different models: the standard Cox Proportional-Hazards model, Cox Proportional-Hazards with penalties, and lastly the random survival forest model. Table 5.1 below describes the performance of these models using the concordance index as well as the brier score.

	c-index	**IBS**
RSF	0.527812	0.134962
CPH	0.52254	0.135007
Random	0.5	0.250831
Kaplan-Meier	NaN	0.134953

Table 5.1: Comparison of Model Performance

Random Survival Forest provides the best result in calibration(brier score; IBS) while Kaplan-Meier comes out to be the best for discrimination(c-index). The third line stating "Random" is how a completely random model would perform and listed in the table to provide a benchmark.

5.2 Further Discussion

In further modeling, I plan to improve upon our current models as well as experiment with other models such as Gradient Boosted Models and Survival Support Vector Machine. From our current models, I would like to choose a smaller subset of variables, as well as try out different kinds of transformation, such as log transformation or non-linear terms, and interactions. In the following sections, I describe two new models that I would like to implement on survival analysis of the UNOS Kidney Transplant program.

5.2.1 Gradient Boosted Models

Gradient Boosting is not referring to a particular model, but a framework that optimizes on many loss functions. It combines the predictions of multiple base learners to build a powerful overall model. The base learners are often simple models that might perform a little better than a random model.Theses predictions are put together in an additive manner, where each base model addition provides a boost to the final model.

This method is similar to a Random Survival Forest, since it uses multiple base learners a final prediction, but different in how they are combined. RSF fits each tree independently, then takes the average of all predictions; whereas gradient boosted model combines each prediction sequentially in a greedy stagewise method.

5.2.2 Survival Support Vector Machine

Survival Support Vector Machine's main forte is that it can take in non-linear, or complex relationships between variables utilizing the kernel function. The kernel function implicitly maps the input into high-dimensional space where the survival can be written by a hyper-plane.This makes the Survival SVM vey sophisticated and allows for a wide range of data to utilize the Survival Support Vector Machine. On the other hand the main disadvantage is that predictions cannot be quite related to the survival function and the cumulative hazard

function.

REFERENCES

[1] How to handle missing data. https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4. Accessed: 2023-04-01.

[2] Pysurvival open source package for survival analysis modeling. https://www.pysurvival.io. Accessed: 2023-04-01.

[3] J. Bruin. newtest: command to compute new test @ONLINE, February 2011.

[4] Mitchel Klein David G. Kleinbaum. Survival analysis: A self-learning text, third edition. *Springer New York, NY*, 3(1):1–159, 2011.

[5] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[6] Patrick J Heagerty, Thomas Lumley, and Margaret S Pepe. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344, 2000.

[7] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3), sep 2008.

[8] Tuomas Sandholm John P. Dickerson. Futurematch. *AAA*, 1(1):1–7, 2015.

[9] Sebastian Pölsterl. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020.

[10] Matthias Schmid, Marvin Wright, and Andreas Ziegler. On the use of harrell's c for clinical risk prediction via random survival forests, 2016.

[11] Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117, 2011.

[12] Yanyao Yi, Ting Ye, Menggang Yu, and Jun Shao. Cox regression with survival-time-dependent missing covariate values. *Biometrics*, 76(2):460–471, 2020.

[13] Yan Yuan. Prediction performance of survival models. 2008.