INF 560

Data Informatics Professional Practicum

Liver Transplant Data Analysis

Author 1:
Hyunjun Choi
choi797@usc.edu

Author 2:
Yujia Deng
yujiaden@usc.edu

Instructor:
Dr.Ana Farzindar

May, 2018

USC University of Southern California

Contents

List of Figures

1 Project Information

Project Title: Predict Graft Survival after OLT using Pre-Transplant Features using machine learning and statistical method
Date Started: 1/11/2018
Date Completed: 5/3/2018
Project Sponsor/Champion: Professor Anna Farzindar
Stakeholder: Associate Director Manas Bhatnagar, Data Analytics, Research & Performance, Department of Surgery, University of Southern California

2 Executive Summary

This project aims at the problem of identifying features associated with and predicting graft survival after OLT (orthotopic liver transplantation).

The gathering, study, and analysis of existing data, and presentation to stakeholders lies within the scope of this project. Outside the scope of this project are data collection and data publication.

Project milestones were set and achieved each month. Based on Lean Six Sigma, we built five project phases. The first phase consisted of defining project goals, customer deliverables, and other project details in the form of a project charter and SIPOC diagram, ending in early March 2018. The second phase involved measuring performance through process mapping and statistical analysis, ending in early April. The third phase consisted of analyzing the causes of badly performing or overfitting models and comparing models, ending in mid-April. The fourth phase concerned improving model performance by using other methods and developing other potential solutions for poorly performing models, ending in late April. Last but not least, the fifth phase included controlling project performance by finalizing reports, producing a PowerPoint presentation for stakeholders, and finishing the end product, ending in early May.

The end product is a website containing information about data exploration and prediction, which has been tested multiple times and is only accepted when all functions work properly, including the contents built with Tableau and R. Analysis of the data led us to the conclusion that there are certain variables that are significantly associated with graft survival while some are not, and to the recommendation for patients, doctors, and related experts in the field to use the prediction tool for graft survival after a donor-

recipient pair has been formed.

Tasks completed during the projects include obtaining and studying the dataset and background information, analyzing data consisting of data cleaning, using different approaches for analysis, evaluating the performance of models, and transforming results into easy-to-understand messages in the presentation and the website. Classification models were trained and tested to measure performance while regression and survival analysis models were checked using performance measurements such as R2, AIC, and C-Index. After comparing the information conveyed by the different models, survival analysis was selected for building the prediction tool. With the tool, patients and doctors can benefit from being able to predict graft survival for each new patient at different future timeframes up to about 4,000 days and thus make more informed decisions about the surgery accordingly.

3 Lean Six Sigma Project

3.1 Define Phase

3.1.1 Cost of Poor Quality Statement

According to data from the American Liver Foundation, each year in the U.S., around 6,000 recipients benefit from a liver transplant [1]. The number of liver transplants has been increasing annually; however, in contrast with the rapid increase in the demand for livers, there is a severe shortage of liver donors, and surgical success is not guaranteed. Eight thousand and eighty-two liver transplants were performed in 2017, with 13,869 patients still waiting for a liver as of January 2018. Graft failure accounts for about 20% of deaths within 30 days after surgery from 2010 to 2015 [6].

Therefore, it is extremely important to undergo careful evaluations before deciding on whether to proceed with transplantation to avoid futile outcomes and wasting resources on an operation that has a low chance of success. This project aims to explore the possibility of applying data analysis to help doctors and patients make more informed decisions prior to orthotopic liver transplantation (OLT).

3.1.2 Customer Satisfaction (Voice of the Customer)

A poor-quality end product impacts the customer's usage and therefore, the value of the project. Any website containing data exploration and prediction tools should be aesthetically viewable, conveniently accessible, and easily navigated.

The voice of the customer was tested during the demonstration of the end product and validated when the customer understood how to navigate the website, search for data, and use the prediction tool without issues.

3.1.3 Application of Tools

Python and R were used to build the models. During the process, we learned how to build support vector machine (SVM) and neural network (NN) models for classification in Python, how to use subset selection methods in R, and how to develop a Cox regression model in R. Tableau was used to build the data visualizations, and we have gained knowledge on using the tool. We have also learned how to use Shiny in R to build online applications that can be used for data exploration and prediction. Our learning experience also includes building a website using Google Sites with the ability to embed content from the Tableau and Shiny apps.

3.2 Measure Phase

3.2.1 Process Mapping/Process Visualization

Different levels of process map were developed during the project. A "high-level process" map describes supplier, inputs, process, outputs, and customers, and is abbreviated as SIPOC. A more detailed process map is the "common process map", outlining the main tasks in the project. The "detailed process map" is the most detailed of these maps, including both main processes and their sub-processes. Lastly, a "functional process map" allows users to easily visualize the people responsible for each process and the corresponding function. Please see Appendix A for the process maps.

3.2.2 The Vital Few

This project aims at the problem of identifying features associated with graft survival after OLT. The vital few covariates contributing to patient response

and graft survival were selected using different methods, including boosting the decision tree, subset selection, and purposeful selection based on significance and model performance. More details are outlined in the "Result and System Implementation" section.

3.2.3 Data Collection (Planning and Execution)

Data collection was out of scope for this project; the data used were provided by the Department of Surgery at Keck Medicine of USC. The dataset contained 84,603 observations of 159 variables. The first step in data preprocessing was to remove invalid observations, such as where graft survival time was larger than patient survival time, and irrelevant variables, such as post-transplant features. The process includes checking missing values and determining imputation methods. If a categorical variable had an originally defined 'unknown' category, missing values were assigned to the category; if not, then the proportion of missing values were checked and if larger than 0.05, an unknown category was created for the missing values, if less than or equal to 0.05, missing values were assigned the mode [10]. For continuous variables, predictive mean matching was used in implementing the missing values. Data preprocessing led to a dataset of 42,169 observations of 145 variables useful for regression and survival analysis in addition to graft status (GSTATUS) and graft survival time (GTIME).

For classification, new datasets were subset from the dataset after imputation, based on graft failure time. Then, censored data, including the observations whose graft status was 'alive' but whose graft survival time was less than the time of interest, were removed. The following step was to create a new indicator variable for the time of interest. For example, when creating a dataset to analyze graft survival within one month, observations whose graft status was 'alive' and whose graft survival time was less than 30 days were removed. Then an indicator variable, G30, was created and assigned a value of 1 if the graft survival time was greater than or equal to 30 days and assigned a value of 0 if the graft survival time was less than 30 days. As we are interested in analyzing graft status at one month, two months, three months, six months, nine months, and one year, six subset datasets were created, each with a corresponding new indicator variable.

3.2.4 Measurement System Analysis

Measurement system analysis is performed to analyze the stability, bias, linearity, repeatability, and or reproducibility of a measurement system. In the project, the measurement system measureed the performance of models developed for data analysis; the measurement system analysis focused on measuring the stability of the model's performance. Reference model performance came from current studies, which has a 66.2% accuracy when classifying graft status after the surgery [13].

The measurement system analysis looked at multiple models implemented and recorded the test accuracy for each model, revealing that most of the models achieve reasonable accuracy but were below the reference line.

3.2.5 Tools Application

Flowchart tools were used in the process of creating process maps Google Slide was used to recreate the process maps for them to be presentable. Performance metrics were defined based on the literature and current studies in the field, including AUC-ROC (measures classification accuracy) for classification models, R2 (measures the amount of variation in the dependent variable created by the independent variables), and AIC (measures the amount of information loss) for regression models, and C-Index (analogous to AUC-ROC) for Cox regression models. Knowledge about accessing the performance of machine learning and statistical models was gained throughout the process.

3.3 Analyze Phase

3.3.1 Charts that Help Analyze Data Collected From the Measure Phase

In order to identify the variables related to graft survival after OLT, in addition to focusing on short-term graft survival, it is useful to gain an understanding of the long-term graft survival rate, whose causes were analyzed using a fishbone diagram, as shown in Appendix B.

3.3.2 Additional Charts used by Lean Six Sigma

In addition to the fishbone diagram, a work breakdown structure (WBS) was created to effectively organize the team's work into manageable tasks, as shown in Appendix B.

3.3.3 Y=f(X) Formula

In the process of identifying essential variables associated with graft survival, the outcomes of potential solutions were arranged in a list of variables that were deemed important for graft survival; the inputs to achieve the outcome were all available useful features from the dataset, and the inputs were put into different models to achieve the outcome.

In the classification, the boosted decision tree was used to select top features to be used in the SVM and NN models. The process can be formulated as Selected features=Boosted Decision Tree (Available useful variables).

In the regression, the backward selection was used to select a linear regression model. The process started with GTIME = $B_0 + B_1X_1 + B_2X_2 + \ldots + B_KX_K + \epsilon$ and ended when an optimal model was achieved with the lowest AIC score. The process can be formulated as X, as in selected model = Backward selection (Available useful variables). In Cox regression, variables were selected in a purposeful mechanism. The Cox regression model can be formulated as $\mathrm{h(t\ |X) = h(t\ |X = 0)exp(B_0 + } B_1X_1 + B_2X_2 + \ldots + B_KX_K)$ while the feature selection process can be formulated as X in the final Cox regression model = Purposeful selection (Available useful variables).

3.3.4 Strengths, Weaknesses, Opportunities, and Threats

The strengths of our team include the statistical backgrounds of both authors as we both major in Statistics. We are also adept at creating aesthetic visualizations using various tools and we aim at creating neat, visually enjoyable, and powerful presentations.

As the need for livers is continually increasing, there are numerous opportunities to perform data analysis and use the results to help doctors and patients make better decisions about transplantation surgery, as well as help patients survive longer after the surgery, and ideally to reduce liver-related diseases and therefore reduce the need for a liver transplant.

However, environmental degradation heavily influences people's health. Advances in medicine and technology may help to cure or at least treat the

commonplace diseases and ailments of today, but not all diseases yield to medical intervention. The bad habits people have today also pose threats to their health, making it more difficult to reduce the need for medications or procedures, such as liver transplants.

3.3.5 Root Causes

A fishbone diagram was used to analyze the long-term low graft survival rate. To analyze short-term graft survival, we plotted the graft survival rate at 1 month, 2 months, 3 months, 6 months, 9 months, and 1 year and explored the data by gender, age group, and blood type, in Appendix C. Grafts seem to be able to survive at least 80% of the time for at least 1 year; the trend appears smooth and there are indeed differences between people with different biological characteristics, so the reason for decreasing graft survival might relate to the degradation of the graft's functions as time passes and to biological factors.

3.3.6 Correlation

By examining the scatter plots between graft survival time and each feature, there does not seem to be any obvious correlation between graft survival time and any individual feature. Correlation plots of continuous variables were also created, and no hidden correlation was identified other than repetitive measures such as weight at transplant and weight at registration. Please see Appendix C for the plots.

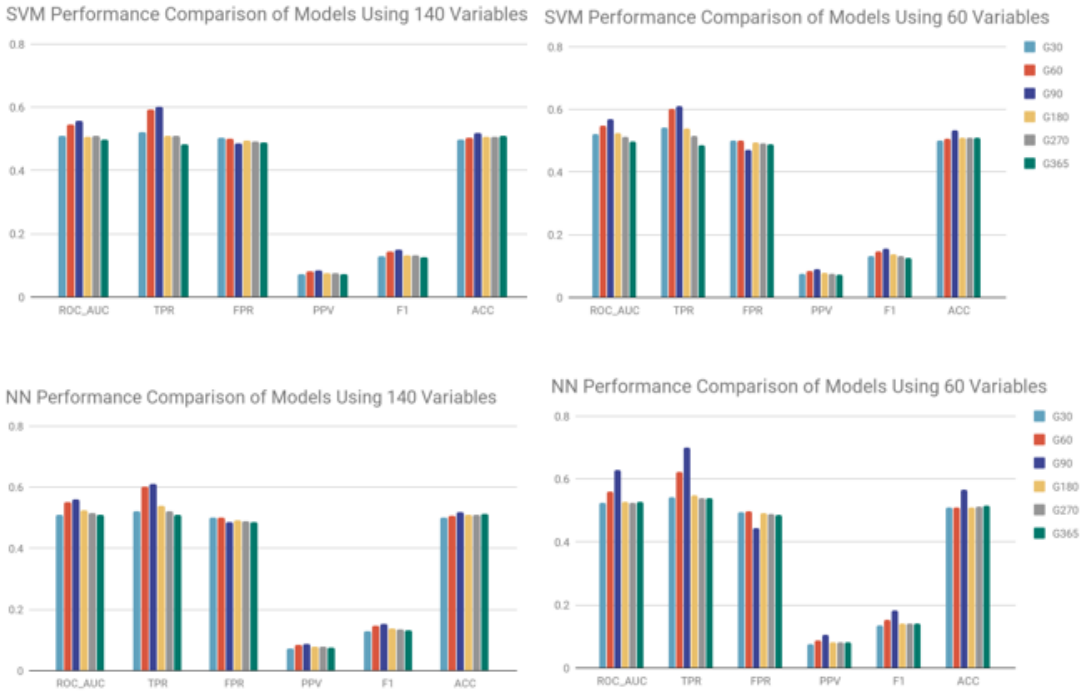
3.3.7 Sources of Variation

Model performance varied as we performed cross-validation or as we tuned models by adding or removing variables in the analysis. Classification of models' performance metrics, such as the area under the receiver operating characteristic curve (AUC-ROC), true positive rate, false-positive rate, positive predictive value, F1-score, R2, AIC, and C-Index all varied as the number of variables included in multivariate models changed.

3.3.8 Potential Solutions

To identify important features related to graft survival, one potential way is to use tree-based methods to select the features. A boosted decision tree

Figure 1: Graphs showing performance metrics of classification models. Variation is visible as the number of variables used changes.



was implemented to select the top features for predicting graft survival in the six different time frames. Among the six selections, many donor-related variables appear multiple times to be on top of the list. Subset selection is another potential solution to identify important variables in a linear regression model attempting to predict graft survival time. Purposeful selection can also be a solution to determine which features relate to graft survival in a Cox regression.

3.4 Improve Phase

3.4.1 Alternative Solutions Considered

For linear regression models, an alternative selection method can be to separate the data by age group so that different analyses can be performed on

the MELD or PELD score.

The selection method in a Cox regression includes fitting univariate models and then using a multivariate model of significant variables in a univariate analysis. To improve upon this, categorical variables with too many levels that might lead to errors in the model should be removed.

Another solution can be in addition to the above steps; multivariate models can be further tuned by adding variables of interest such as blood type, age, gender, and state of residency. Please see Appendix D for the variables selected in classification, linear regression, and Cox regression after performing improvement and comparing the solutions.

3.4.2 Recommended Solution(s)

After comparing the models based on their performance and the amount of information conveyed, the Cox regression model was selected as the recommended solution to build a prediction tool for graft survival. Cox regression can predict the probability of experiencing graft failure at time t+1 assuming the graft has survived until time t. One concern is that there are time-varying variables included in the data set such as the number of the previous transplants and without adjusting the model to account for time-varying effects, the model's accuracy might not be as high. Overall, the model had an accuracy of 0.583.

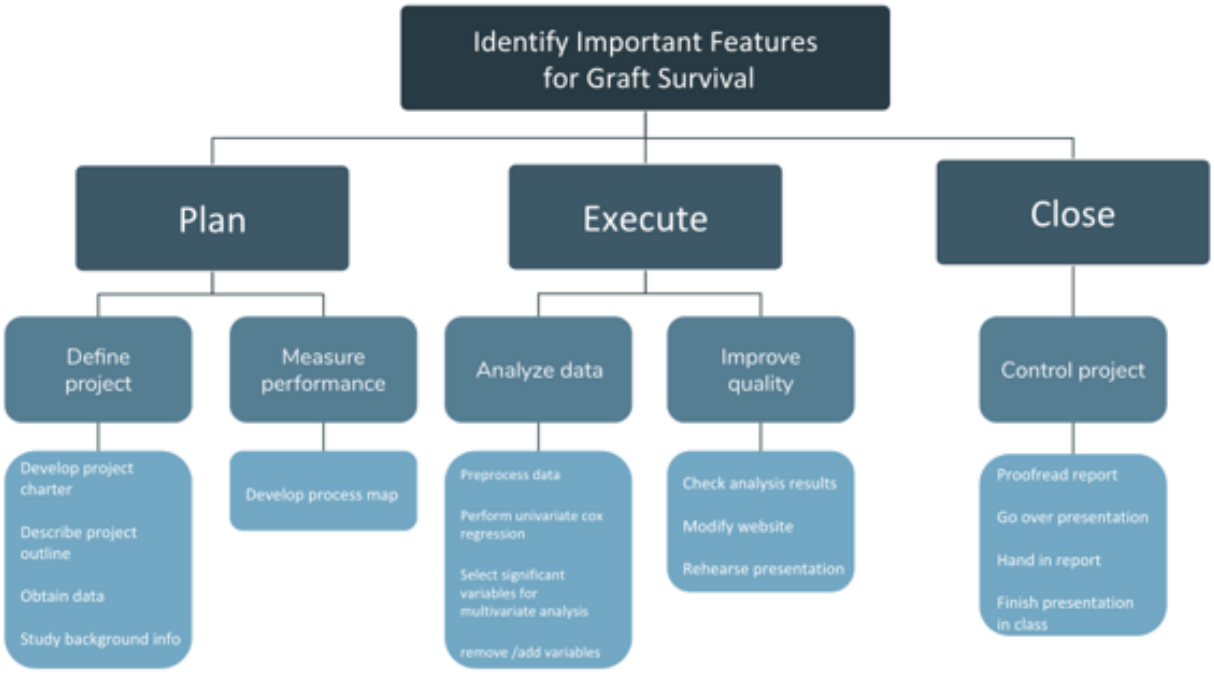
3.4.3 Ways to Pilot the Recommended Solution(s)

The solution was tested by inputting new values for the variables used in the model and checking the prediction. Predicted probability of graft survival should not be higher than 1 or lower than 0. Reasonable values should be close to the survival rate from the data exploration, which shows that the average is about 0.89.

3.4.4 Project Plan

A work breakdown structure (WBS) is explicitly created for the recommended solution.

Figure 2: Work Breakdown Structure for developing prediction tool using cox regression.



3.5 Control Phase

The project was controlled via weekly meetings with the teaching assistant during the latter half of the semester and in regular weekly or bi-weekly team meetings throughout the semester. Day-to-day processes were controlled through live communication between the study authors, updating one another on tasks done or difficulties encountered, and the issues that would be discussed in the next team meeting. A Gantt chart was also used as a control tool to keep up with the schedule, as shown in Appendix B.

One challenge in implementing control is coming up with new ideas and still following the Gantt chart. Therefore, we needed to adjust the Gantt chart to extend the 'Define' and 'Analysis' processes, but we also needed to make sure we will meet the deadline for closing the project.

After closing the project, we will communicate with the customers to

add authorized users as desired. We will also maintain the website and the contents on the site. If the customers have any questions, we will also be available with our email address posted on the website.

4 Result and System Implementation

4.1 Machine Learning Approaches

4.1.1 Neural Network (NN)

Many experts have conducted research on predicting liver survival following transplantation by leveraging machine learning [3, 5, 7, 9, 11, 13, 15]. NN [4], a machine learning algorithm, learns a function by training that function on a dataset. With a set of features and an output, the model can be trained as a classification model. Between input and output, there are hidden layers in the model.

The parameters in NN are the epoch, which is the number of forward and backward passes of all training examples, and the batch size, which is the number of training examples in one forward/backward pass. Different values for the epoch and batch size were checked, and the optimal test performance and computing time were reached at an epoch of 300 and a batch size of 200.

For each of the six subset data sets, a boosted decision tree was implemented to select top features that would be later used in the NN classification model. 20% of data were separated as a test set. Models were trained and validated on the other 80% of the data using 10-fold cross-validation and then tested on the test set. Data were also oversampled to fix the unbalanced class problem, as the proportion of observations with a graft status of 'alive' was much higher than observations with a graft status of 'dead'. Performance metrics were output for comparison. On average, NN was able to reach about 50% of AUC-ROC. More details regarding NN can be found in Appendix E.

4.1.2 Support Vector Machine (SVM)

SVM is a supervised machine learning classifier. Given a set of labeled features and an output, the algorithm models an optimal hyperplane that categorizes new observations [12]. The kernel function was used to study the types of relationships in data. This function needed to be chosen with careful

consideration as the data were not linearly separable [2]. Another parameter is the gamma, which indicates how far the influence of a single training observation can reach [14]. The radial basis function was used as the kernel function. The gamma parameter was set at 10. 20% of data were separated as a test set. Models were trained and validated on the other 80% of the data using 10-fold cross-validation and then tested on the test set. Data were also oversampled to fix the unbalanced class problem, as the proportion of observations with a graft status of 'alive' was much higher than observations with a graft status of 'dead'. Performance metrics were output for comparison. SVM was performed on each of the six subset data sets, and reached AUC-ROC at about 50%. More details regarding NN can be found in Appendix F.

4.1.3 Comparison

Overall, both models achieve an AUC-ROC of about 0.53 on average when predicting graft status over different time frames up to 1 year after the surgery. Compared to SVM, NNs seem to perform better overall in G90. In the G90 prediction case, the AUC of NN was 0.62, while for SVM it was 0.56.

A linear regression was modeled as GTIME $= B_0 + B_1X_1 + B_2X_2 + \ldots + B_KX_K + \epsilon$, k = 145 as there were 145 features available for analysis after data cleaning. Backward selection was used, taking the full model with all features as the start, with one variable removed each time to achieve optimal AIC.

The parameters B_i were estimated using ordinary least squares, which minimized the sum of squares of the differences between the observed GTIME and the predicted values using the linear function. The coefficient estimates of B_i represent the magnitude of the association between the corresponding X and GTIME, when all the other variables were constant.

The baseline model had AIC 570831.1, while the selected model has AIC 569712.5. R^2 was 0.27. New values of the independent variables can be input, and a predicted GTIME will be outputted.

The linear regression model assumed linearity, normality, and homoscedasticity (equal variances). The assumptions were checked using residual plots. The linearity and homoscedasticity assumptions were violated as shown in Appendix G.

Survival analysis was performed using Cox regression, which was modeled

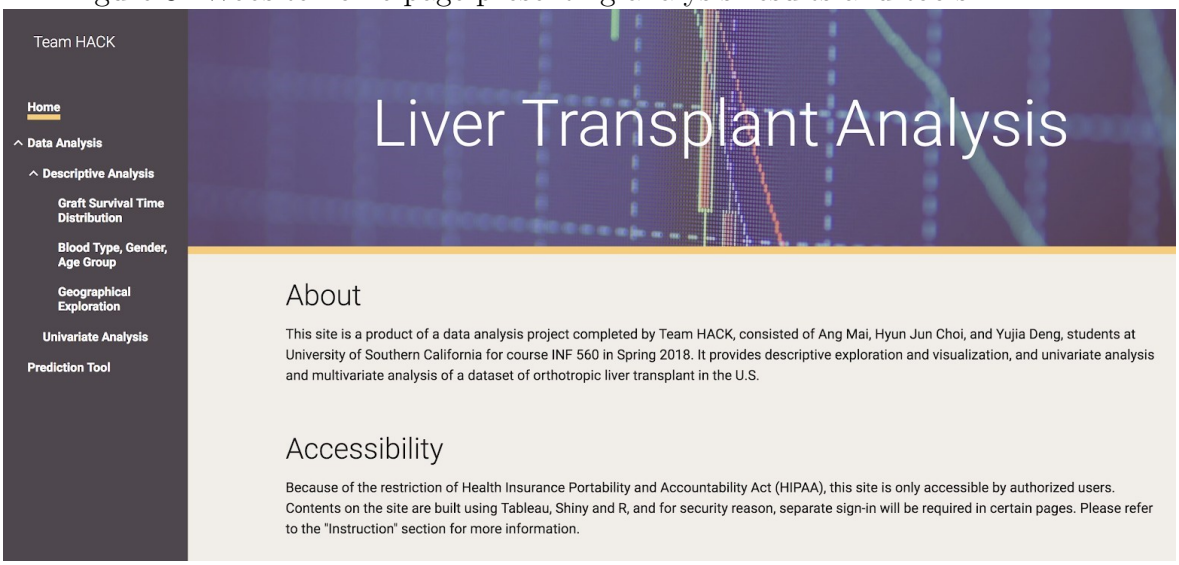
as h(t |X) = h(t |X = 0)exp(B$_0$ + B_1X_1 + B_2X_2 + ...+ B_KX_K). h(t |X) is the hazard function given a set of features X, which can be understood as the probability of an individual experiencing graft failure at time t+1 if the person has survived until time t. In the final model, $k = 17$.

The parameters B_i were estimated using partial maximum likelihood. The exponential of the coefficient estimate represents the effect the corresponding feature has on the ratio of the two hazard functions, when all the other variables were constant. For example, for the continuous variable, the hazard ratio between people whose X_i = x + 1 and people whose X_i = x was exp(B$_i$), meaning that people whose X_i = x + 1 was exp(B_i) times more likely than people whose X_i = x to experience graft failure when all other variables are the same; for categorical variables, the hazard ratio between people of category j and people of baseline model l was exp(B$_i$ - B_l), meaning that people of category j is exp(B$_j$ - B_l) times more likely than people of baseline model l to experience graft failure when all other variables are the same [8]. The model achieved a C-index of 0.583. New values for the independent variables can be input, and the model will output a predicted hazard function. Cox regression assumes proportional hazard and linearity. The assumptions were tested in the cph.zh function in R using the chi-square test. Significant test (p-value¡0.05) indicates non-linearity, and except for one level in the state of residency variable everything else was non-significant, so the model's assumptions can be said to be met, as shown in Appendix H.

4.2 System Implementation

After evaluating the performance, assumptions, and amount of information conveyed by the different models, survival analysis was selected for building the prediction tool as Cox regression seemed to meet the assumptions and was able to predict the hazard function of time, in addition to merely the graft survival time in linear regression or the graft survival status in classification. The selected Cox regression model was incorporated into a Shiny app written in R. Users will be able to input new values for the independent variables in the model in the app. A graph of the hazard function will be plotted with time on the x-axis and survival probability on the y-axis; survival probability at times of interest, including one month, two months, three months, six months, nine months, and one year will be shown, as shown in Appendix H.

Figure 3: Website home page presenting analysis results and tools.



4.3 Prototype and Demo

A prototype of the end product is a website where users can navigate the data analysis tool and prediction tool. Tableau and Shiny in R are used to develop content on the site as well. To ensure the security of information, the website is only accessible by authorized users. To view the contents built with Tableau and Shiny, additional authorizations need to be granted. To fully access the site, three authorizations needed to be granted.

1. Access to the site
2. Access to the Tableau contents in the site
3. Access to certain Shiny contents in the site

Customers can communicate with us to add authorized users. The navigation menu is located on the left. The home page includes information about the website, including instructions on accessing the contents and contact information. Under “Data Analysis,” users can explore the data in the “Descriptive Analysis” section: a distribution of the graft survival time, graft survival statistics by blood type, age group, and gender, and geographical

exploration. To assess the relationship between graft survival time and any other variables, users can click on "Univariate Analysis" to see the scatter plot, the Kaplan-Meier curves, and the survival curve output by the univariate Cox regression model between GTIME and one of the independent features by selecting GTIME as the "Y Feature" and the desired independent feature as the "X Feature". Users can also choose to visualize the correlation between any two variables by selecting the desired "Y Feature" and "X Feature." In the prediction tool, users input values for the variables on the left and on the right; the app will output a survival curve and underneath survival probability at different times of interest.

References

[1] Thomas D Boyer, Arun J Sanyal, Norah A Terrault, and Keith D Lindor. *Zakim and Boyer's hepatology: a textbook of liver disease*. Elsevier Health Sciences, 2016.

[2] André Elisseeff and Jason Weston. A kernel method for multi-labelled classification. In *Advances in neural information processing systems*, pages 681–687, 2002.

[3] Jacob M Feagans, Robert D Gatliff, David Victor, Fredric Regenstein, and Sander S Florman. M1026 predicting survival following liver transplant using artificial neural networks: Optimizing the unos database using machine learning techniques. *Gastroenterology*, 138(5):S–824, 2010.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[5] Wojciech Jarmulski, Alicja Wieczorkowska, Mariusz Trzaska, Michal Ciszek, and Leszek Paczek. Machine learning models for predicting patients survival after liver transplantation. *Computer Science*, 19, 2018.

[6] WR Kim, JM Smith, MA Skeans, DP Schladt, MA Schnitzler, EB Edwards, AM Harper, JL Wainright, JJ Snyder, AK Israni, et al. Optn/srtr 2012 annual data report: liver. *American Journal of Transplantation*, 14(S1):69–96, 2014.

[7] Allison J Kwong and Sumeet K Asrani. Artificial neural networks and liver transplantation: Are we ready for self-driving cars? *Liver Transplantation*, 24(2):161–163, 2018.

[8] WW LaMorte. Cox proportional hazards regression analysis. *Boston: Boston University School of Public Health. Retrieved September*, 27:2018, 2016.

[9] Lawrence Lau, Yamuna Kankanige, Benjamin Rubinstein, Robert Jones, Christopher Christophi, Vijayaragavan Muralidharan, and James Bailey. Machine-learning algorithms predict graft failure after liver transplantation. *Transplantation*, 101(4):e125–e132, 2017.

[10] Katya L Masconi, Tandi E Matsha, Justin B Echouffo-Tcheugui, Rajiv T Erasmus, and Andre P Kengne. Reporting and handling of missing data in predictive research for prevalent undiagnosed type 2 diabetes mellitus: a systematic review. *EPMA Journal*, 6(1):7, 2015.

[11] M Pérez-Ortiz, Manuel Cruz-Ramírez, JC Fernández-Caballero, and César Hervás-Martínez. Hybrid multi-objective machine learning classification in liver transplantation. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 397–408. Springer, 2012.

[12] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[13] Harald Schrem, Anna-Luise Platsakis, Alexander Kaltenborn, Armin Koch, Courtney Metz, Marc Barthold, Christian Krauth, Volker Amelung, Felix Braun, Thomas Becker, et al. Value and limitations of the bar-score for donor allocation in liver transplantation. *Langenbeck's archives of surgery*, 399(8):1011–1019, 2014.

[14] Junping Wang, Quanshi Chen, and Yong Chen. Rbf kernel based support vector machine with universal approximation and its application. In *International symposium on neural networks*, pages 512–517. Springer, 2004.

[15] Kyung Don Yoo, Junhyug Noh, Hajeong Lee, Dong Ki Kim, Chun Soo Lim, Young Hoon Kim, Jung Pyo Lee, Gunhee Kim, and Yon Su Kim.

A machine learning approach using survival statistics to predict graft survival in kidney transplant recipients: a multicenter cohort study. *Scientific reports*, 7(1):8904, 2017.

[16] Jacek M Zurada. *Introduction to artificial neural systems*, volume 8. West publishing company St. Paul, 1992.

A Appendix

A.1 Appendix A

Figure 4: SIPOC.

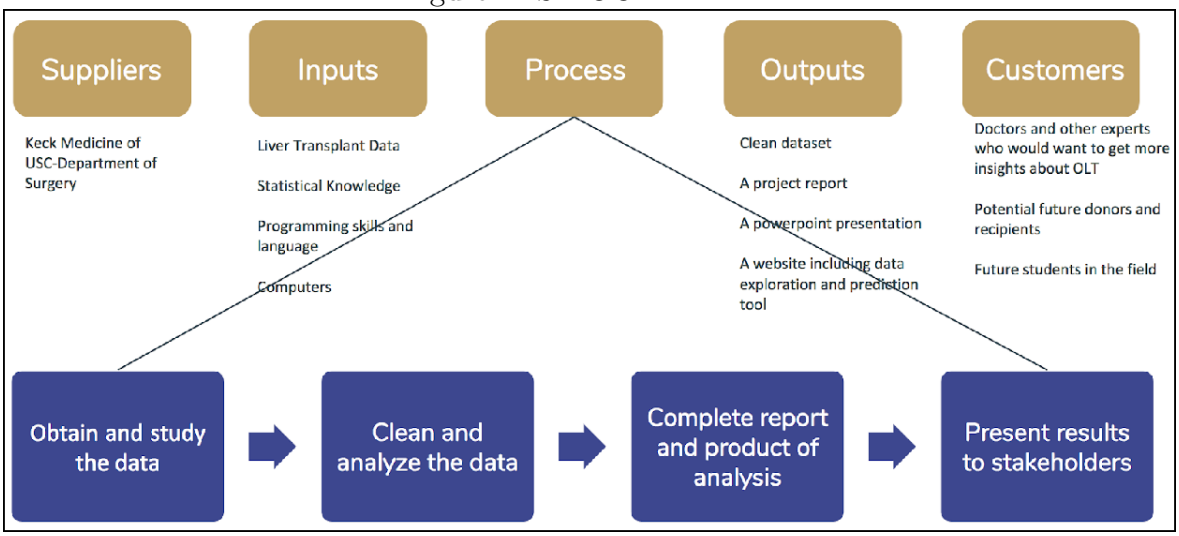


Figure 5: Common Process Map.

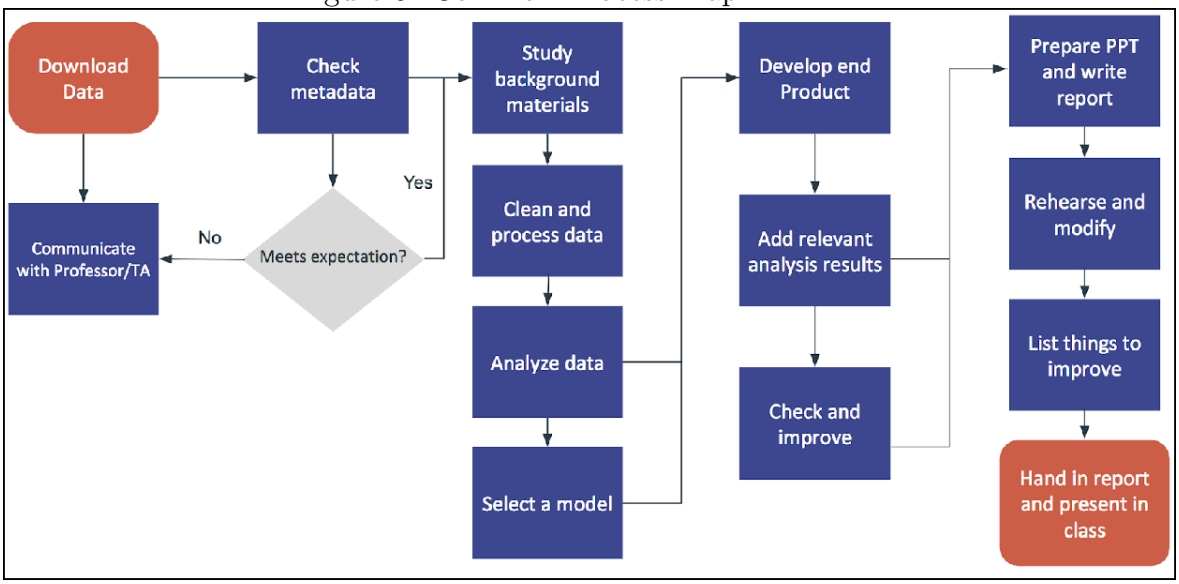


Figure 6: Detailed Process Map.

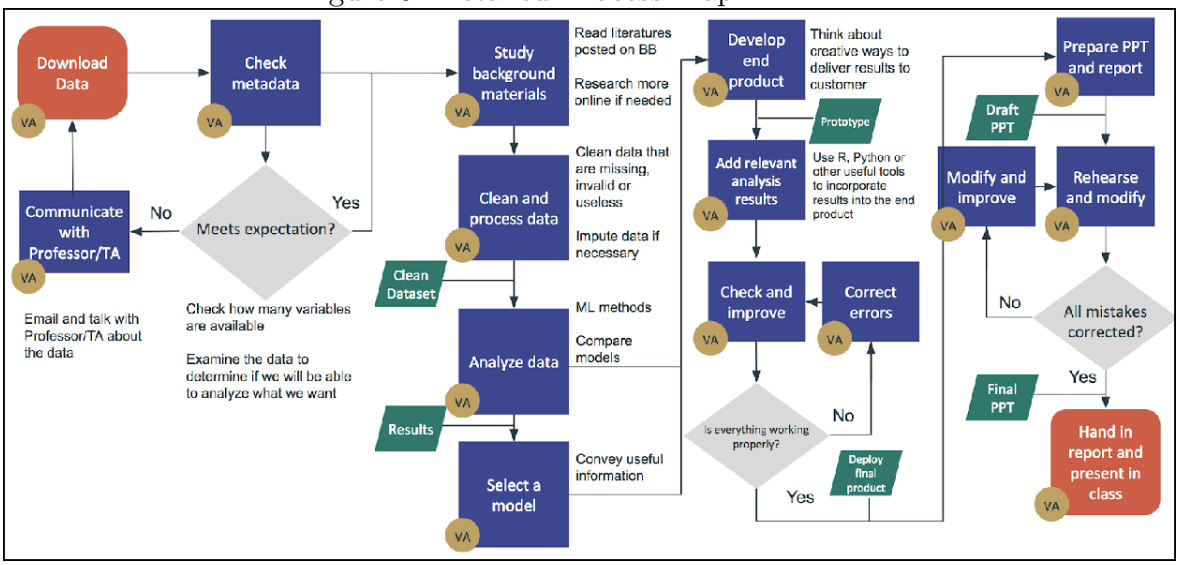
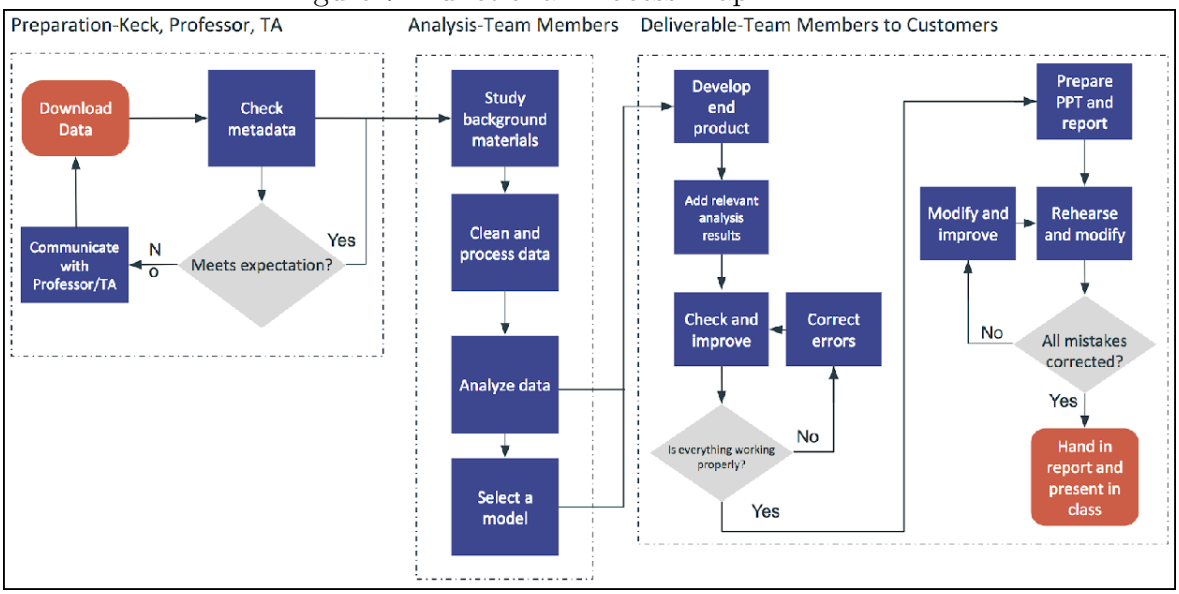


Figure 7: Functional Process Map.



A.2 Appendix B

Figure 8: Fishbone diagram of analysis of long-term graft survival rate.

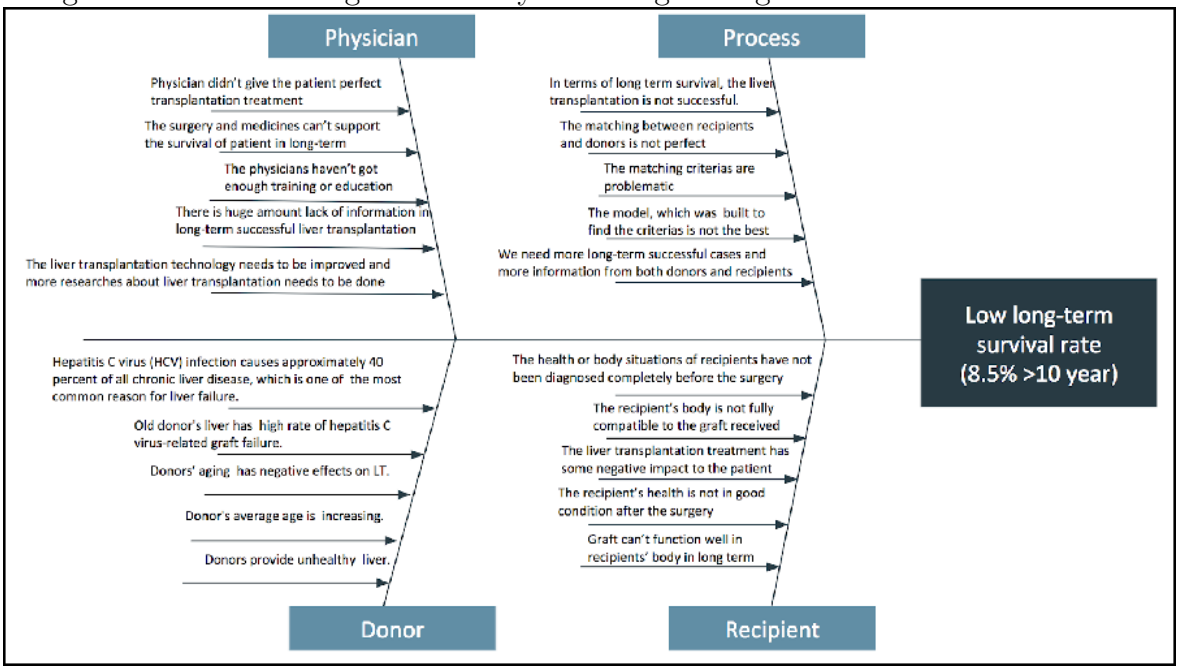


Figure 9: Originally developed WBS for the tasks of the project.

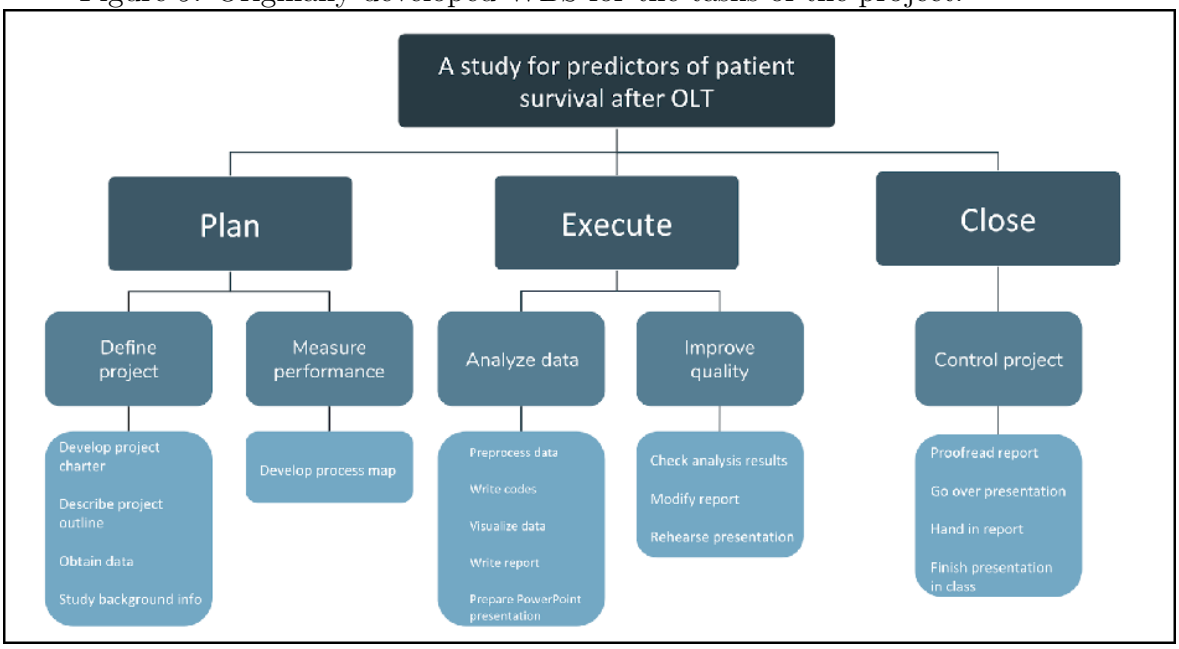
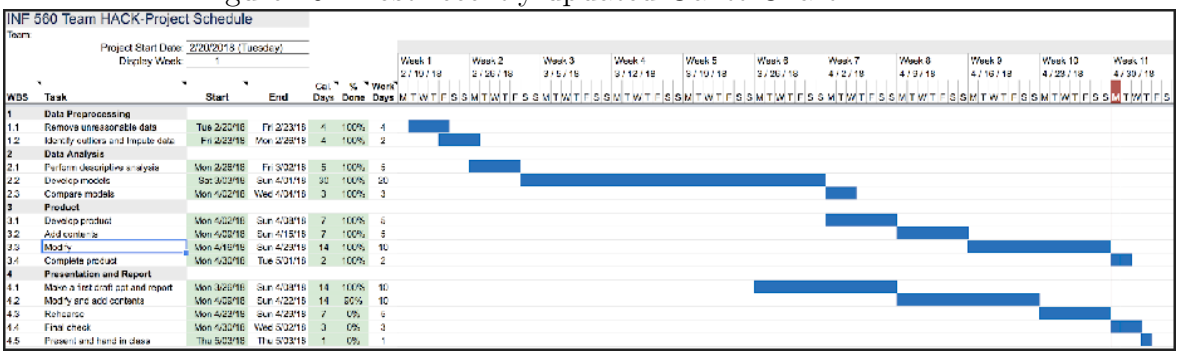


Figure 10: Most recently updated Gantt Chart.



A.3 Appendix C

Figure 11: Correlation matrix among continuous variables.

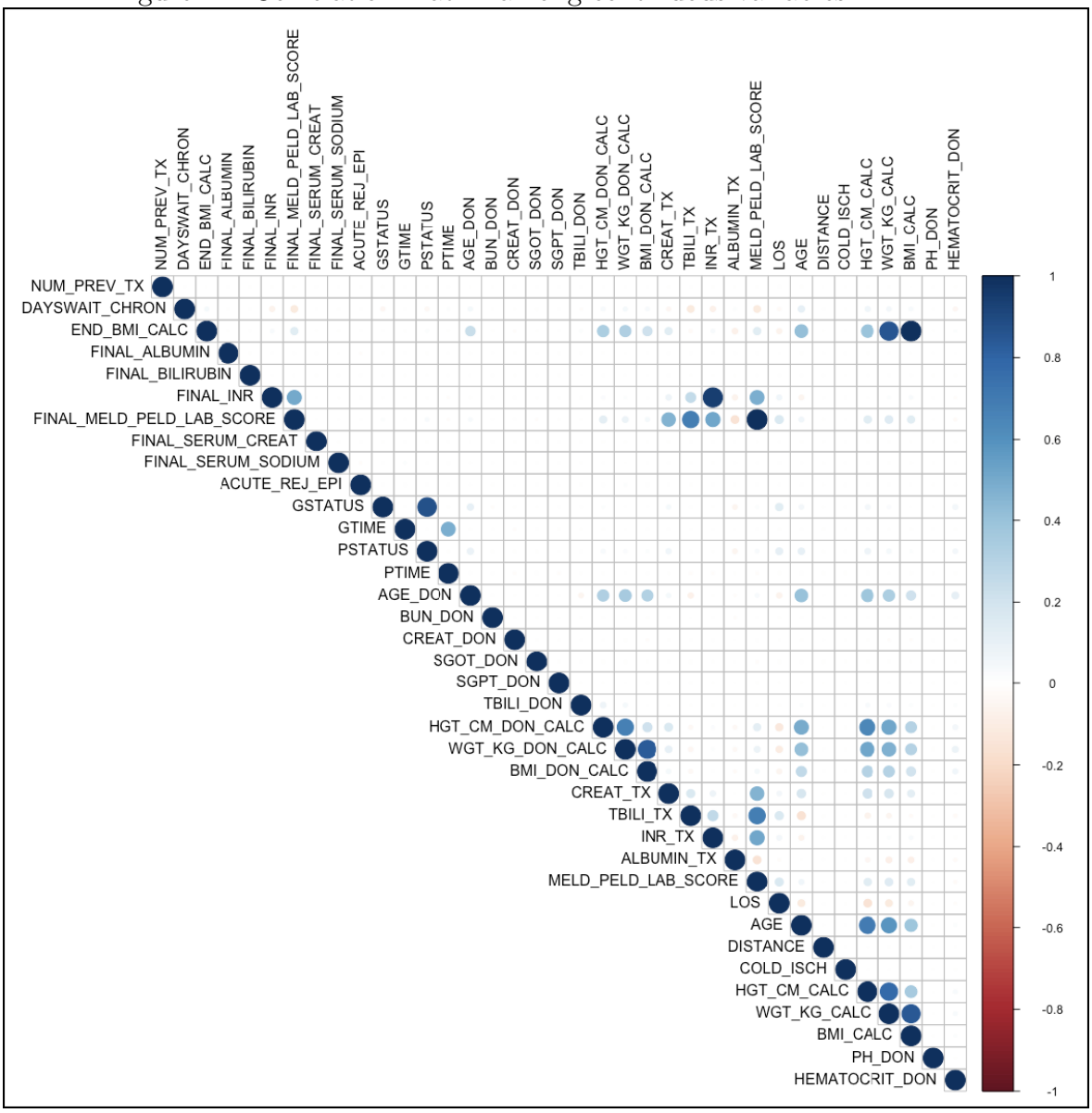


Figure 12: Graft survival rate at different times after surgery.

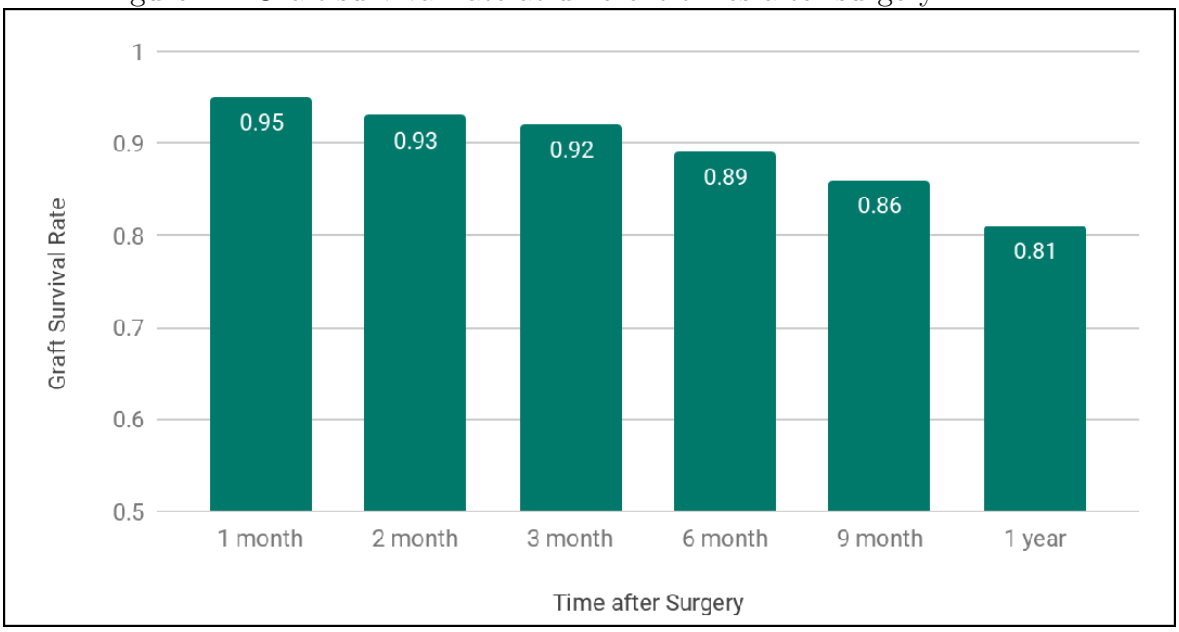
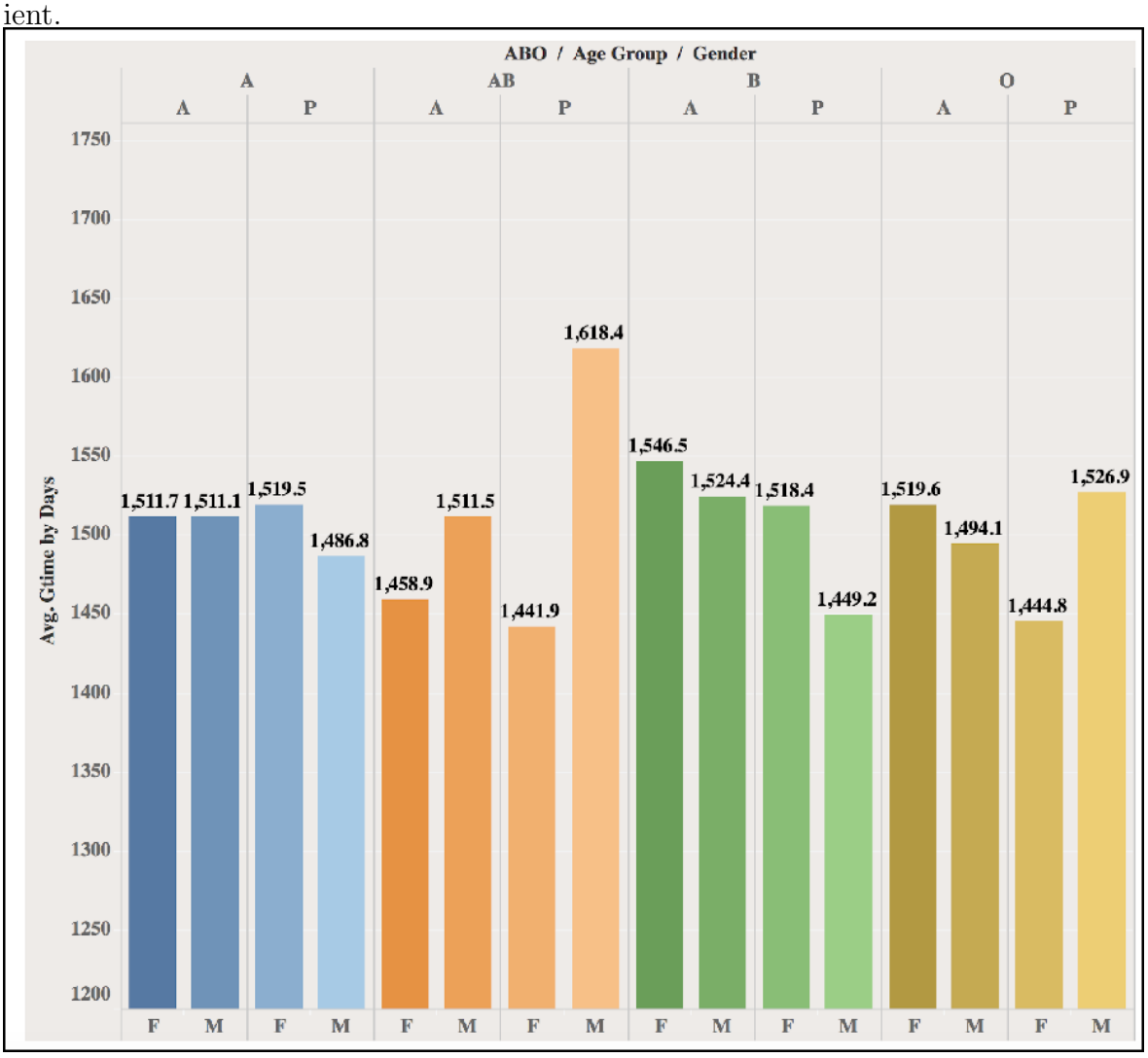
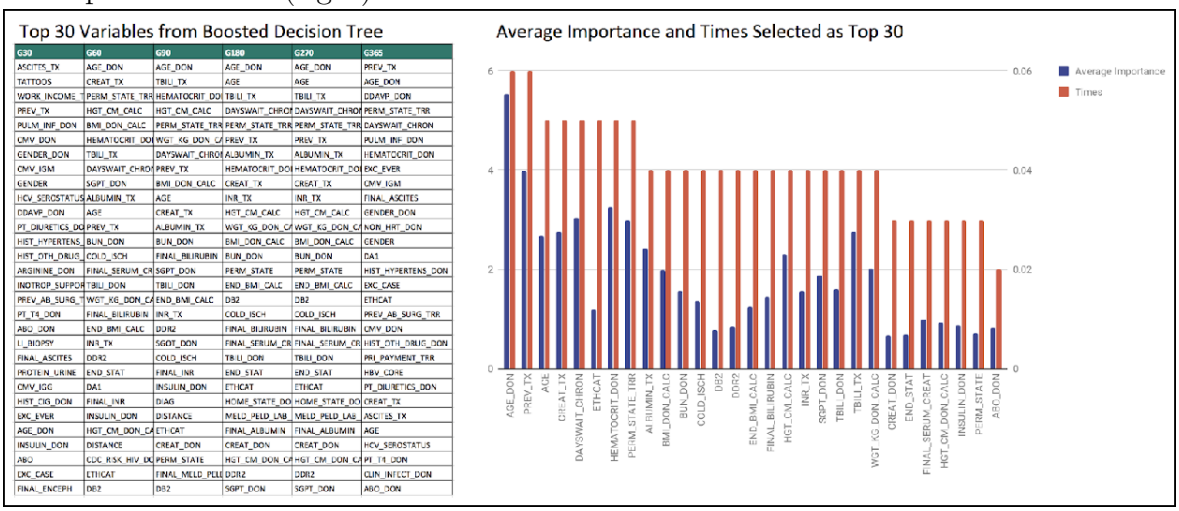


Figure 13: Average graft survival time by ABO/Age group/Gender of recipient.



A.4 Appendix D

Figure 14: Variables selected by the boosted decision tree for classification (left); average importance score and the number of times selected for each of the top 30 variables (right).



Top 30 Variables from Boosted Decision Tree

G30	G60	G90	G180	G270	G365
ASCITES_TX	AGE_DON	AGE_DON	AGE_DON	AGE_DON	PREV_TX
TATTOOS	CREAT_TX	TBIL_TX	AGE	AGE	AGE_DON
WORK_INCOME_T	PERM_STATE_TRR	HEMATOCRIT_DO	TBILI_TX	TBILI_TX	DDAVP_DON
PREV_TX	HGT_CM_CALC	HGT_CM_CALC	DAYSWAIT_CHRO	DAYSWAIT_CHRO	PERM_STATE_TRR
PULM_INF_DON	BMI_DON_CALC	PERM_STATE_TRR	PERM_STATE_TRR	PERM_STATE_TRR	DAYSWAIT_CHRON
CMV_DON	HEMATOCRIT_DO	WGT_KG_DON_CA	PREV_TX	PREV_TX	PULM_INF_DON
GENDER_DON	TBIL_TX	DAYSWAIT_CHRO	ALBUMIN_TX	ALBUMIN_TX	HEMATOCRIT_DON
CMV_IGM	DAYSWAIT_CHRO	PREV_TX	HEMATOCRIT_DO	HEMATOCRIT_DO	EXC_EVER
GENDER	SGPT_DON	BMI_DON_CALC	CREAT_TX	CREAT_TX	CMV_IGM
HCV_SEROSTATUS	ALBUMIN_TX	AGE	INR_TX	INR_TX	FINAL_ASCITES
DDAVP_DON	AGE	CREAT_TX	HGT_CM_CALC	HGT_CM_CALC	GENDER_DON
PT_DIURETICS_DO	PREV_TX	ALBUMIN_TX	WGT_KG_DON_CA	WGT_KG_DON_CA	NON_HRT_DON
HIST_HYPERTENS_	BUN_DON	BUN_DON	BMI_DON_CALC	BMI_DON_CALC	GENDER
HIST_OTH_DRUG_	COLD_ISCH	FINAL_BILIRUBIN	BUN_DON	BUN_DON	DA1
ARGININE_DON	FINAL_SERUM_CR	SGPT_DON	PERM_STATE	PERM_STATE	HIST_HYPERTENS_DON
INOTROP_SUPPOR	TBIL_DON	TBIL_DON	END_BMI_CALC	END_BMI_CALC	EXC_CASE
PREV_AB_SURG_T	WGT_KG_DON_CA	END_BMI_CALC	DB2	DB2	ETHCAT
PT_T4_DON	FINAL_BILIRUBIN	INR_TX	COLD_ISCH	COLD_ISCH	PREV_AB_SURG_TRR
ABO_DON	END_BMI_CALC	DDR2	FINAL_BILIRUBIN	FINAL_BILIRUBIN	CMV_DON
LI_BIOPSY	INR_TX	SGOT_DON	FINAL_SERUM_CR	FINAL_SERUM_CR	HIST_OTH_DRUG_DON
FINAL_ASCITES	DDR2	COLD_ISCH	TBILI_DON	TBILI_DON	PRI_PAYMENT_TRR
PROTEIN_URINE	END_STAT	FINAL_INR	END_STAT	END_STAT	HBV_CORE
CMV_IGG	DA1	INSULIN_DON	ETHCAT	ETHCAT	PT_DIURETICS_DON
HIST_CIG_DON	FINAL_INR	DIAG	HOME_STATE_DO	HOME_STATE_DO	CREAT_TX
EXC_EVER	INSULIN_DON	DISTANCE	MELD_PELD_LAB	MELD_PELD_LAB	ASCITES_TX
AGE_DON	HGT_CM_DON_CA	ETHCAT	FINAL_ALBUMIN	FINAL_ALBUMIN	AGE
INSULIN_DON	DISTANCE	CREAT_DON	CREAT_DON	CREAT_DON	HCV_SEROSTATUS
ABO	CDC_RISK_HIV_DO	PERM_STATE	HGT_CM_DON_CA	HGT_CM_DON_CA	PT_T4_DON
EXC_CASE	ETHCAT	FINAL_MELD_PELD	DDR2	DDR2	CLIN_INFECT_DON
FINAL_ENCEPH	DB2	DB2	SGPT_DON	SGPT_DON	ABO_DON

Figure 15: Variables selected by backward selection for linear regression. The symbol next to the coefficient estimate indicates the p-value of testing whether the estimate is zero. If the p-value is small, the test is significant, and the estimate is said to not equal zero, and therefore there is a significant relationship between the corresponding variable and GTIME (0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1).

Selected Variables in Linear Regression	Estimate
Blood Type	
A	
A1	-4.00
A1B	41.76
A2	359.98
A2B	226.55
AB	-31.09
B	-1.24
O	-12.26
Gender	
Male	1.28
Female	
Education	
1-None	
2-Grade school (0-8)	-0.84
3-High school (9-12)	-0.89
4-Attended college/technical school	-0.01 .
5-Associate/Bachelor degree	-0.93
6-Post-college graduate degree	-0.89
996-Less than 5 years	-0.85
998-Unknown	-0.92
Diabetes	
1-No diabetes	
2-Type 1	0.66 *
3-Type 2	-0.11
4-Type other	-0.29
5-Type unknown	0.74
998-Status unknown	-0.45
Ethnicity	
1-White	
2-Black	-4.07

4-Hispanic	-3.63
5-Asian	-1.57
6-Amer Ind/Alaska Native	6.66
7-Native Hawaiian/other Paci?c Islander	90.34
9-Multiracial	40.42
BMI	-1.40 *
Number of Previous Transplantation	9.62
Days on liver waiting list	-0.0077
Type of Exception Relative to HCC	
HBL	
HCC	282.20
Non-HCC	285.40
Region of Transplantation [11]	
1-Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Eastern Vermont	
2-Delaware, District of Columbia, Maryland, New Jersey, Pennsylvania, West Virginia, Northern Virginia	6.76 24.34
3-Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, Puerto Rico	1.63
4-Oklahoma, Texas	42.24 .
5-Arizona, California, Nevada, New Mexico, Utah	42.45
6-Alaska, Hawaii, Idaho, Montana, Oregon, Washington	50.00 .
7-Illinois, Minnesota, North Dakota, South Dakota, Wisconsin	-2.79
8-Colorado, Iowa, Kansas, Missouri, Nebraska, Wyoming	13.50
9-New York, Western Vermont	17.56
10-Indiana, Michigan, Ohio	23.16
11-Kentucky, North Carolina, South Carolina, Tennessee, Virginia	
State of Residency at Registration	
AK-Alaska	
AL-Alabama	-74.89
AR-Arkansas	-28.05
AS-American Samoa	-149.40
AZ-Arizona	-146.60
CA-California	-112.00
CO-Colorado	-150.90
CT-Connecticut	-68.85
DC-District of Columbia	-33.03
DE-Delaware	85.24
FL-Florida	-97.46
GA-Georgia	-78.92
GU-Guam	-410.00
HI-Hawaii	-88.07
IA-Iowa	-74.26
ID-Idaho	29.85
IL-Illinois	-76.82
IN-Indiana	-162.70
KS-Kansas	-52.96
KY-Kentucky	-102.2
LA-Louisiana	-60.70
MA-Massachusetts	-104.4
MD-Maryland	-84.50
ME-Maine	43.96
MI-Michigan	-129.7
MN-Minnesota	-116.3
MO-Missouri	-88.85
MS-Mississippi	-196.20
MT-Montana	-62.58
NC-North Carolina	-147.50
ND-North Dakota	-229.20
NE-Nebraska	-45.54
NH-New Hampshire	-78.09
NJ-New Jersey	-66.04
NM-New Mexico	-18.41
NV-Nevada	-118.90
NY-New York	-98.82
OH-Ohio	-83.79

OK-Oklahoma	-73.99
OR-Oregon	-109.40
PA-Pennsylvania	-93.32
PR-Puerto Rico	-120.10
RI-Rhode Island	-128.00
SC-South Carolina	-115.10
SD-South Dakota	-42.67
TN-Tennessee	-90.50
TX-Texas	-78.23
UT-Utah	-102.20
VA-Virginia	-120.90
VI-Virgin Islands	-586.80
VT-Vermont	-124.40
WA-Washington	-84.01
WI-Wisconsin	-81.49
WV-West Virginia	-115.50
WY-Wyoming	113.00
ZZ-Unknown	-167.10
Ascites	
1-Absent	
2-Slight	-3.24
3-Moderate	8.57
4-Unknown	15.3
Bilirubin	-0.14
Dialysis prior week	
Yes	-8.49
No	
Encephy	
1-None	
2-1 to 2	15.56 .
3-3 to 4	16.50
4-Unknown	53.58 *
INR	-2.81
Serum Creatinine	3.34
MELD or PELD Score	-0.81 .

Figure 16: Variables used in Cox regression. The symbol next to the coefficient estimate indicates the p-value testing whether the estimate is zero. If the p-value is small, the test is significant, and the estimate is said to not equal zero, and therefore there is significant relationship between the corresponding variable and graft survival (0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1).

Selected Variables in Cox Regression	Exponential of Coe? cient Estimate
Recipient Age	1.00 ***
Recipient Gender M ale Female	 0.094
Recipient Blood Type A A1 A1B A2 A2B AB B O	 0.98 0.56 1.11 -0.000027 1.027 -0.95 . 1.04 *
Recipient BMI	0.93
Recipient ME LD/PE LD Score	1.001
Previous Transplant Yes No	 1.545 ***
State of R esidency Where Transplant will Take Place AK-Alaska	

AL-Alabama	0.60
AR-Arkansas	0.60
AS-American Samoa	0.000055
AZ-Arizona	0.62
CA-California	0.57
CO-Colorado	0.59
CT-Connecticut	0.62
DC-District of Columbia	0.66
DE-Delaware	0.63
FL-Florida	0.67
GA-Georgia	0.59
GU-Guam	0.98
HI-Hawaii	0.44
IA-Iowa	0.61
ID-Idaho	0.51
IL-Illinois	0.69
IN-Indiana	0.82
KS-Kansas	0.67
KY-Kentucky	0.74
LA-Louisiana	0.66
MA-Massachusetts	0.62
MD-Maryland	0.71
ME-Maine	0.85
MI-Michigan	0.75
MN-Minnesota	0.61
MO-Missouri	0.65
MP-Northern Mariana Islands	0.000029
MS-Mississippi	0.56 .
MT-Montana	0.56
NC-North Carolina	0.61
ND-North Dakota	0.46 .
NE-Nebraska	0.74
NH-New Hampshire	0.58 .
NJ-New Jersey	0.64
NM-New Mexico	0.56 .
NV-Nevada	0.81
NY-New York	0.76
OH-Ohio	0.67
OK-Oklahoma	0.70
OR-Oregon	0.57 .
PA-Pennsylvania	0.75
PR-Puerto Rico	0.63
RI-Rhode Island	0.85
SC-South Carolina	0.72
SD-South Dakota	0.67
TN-Tennessee	0.57 .
TX-Texas	0.63
UT-Utah	0.55 .
VA-Virginia	0.62
VI-Virgin Islands	1.03
VT-Vermont	0.61
WA-Washington	0.44 **
WI-Wisconsin	0.65
WV-West Virginia	0.58 .
WY-Wyoming	0.94
ZZ-Unknown	1.00
Days on Waiting List	-0.998 ***
Recipient Ethnicity	
1-White	
2-Black	1.05 .
4-Hispanic	1.09 **
5-Asian	1.08 .
6-Amer Ind/Alaska Native	0.87
7-Native Hawaiian/other Paci?c Islander	0.89
9-Multiracial	1.10

Donor Age	1.009 ***
Donor Gender Male Female	 0.98
Donor Blood Type A A1 A1B A2 A2B AB B O	 1.06 0.56 1.11 0.000027 1.03 0.95 . 1.04 *
Donor BMI	0.78 ***
Donor Type Deceased Alive	 0.91 .
Donor Hematocrit	1.01 ***
Donor Less Than 7 Days Old at Time of Donation Yes No	 7.88 **

A.5 Appendix E

A.5.1 More Details about Neural Network

The target feature is GSTATUS, which is 0 or 1. The input value of the hidden layer S^l can be obtained from the equation:

$$S^l = W^l X^{l-1} + b^l,\ 1 \leq l \leq L.$$

The output of the sigmoid function can be calculated from the equation:

$$X^l = sigmoid\ (S^l) = \frac{1}{1+e^{-S^l}},\ 1 \leq l \leq L.$$

If the loss function is a standard least square error, the error of the output can be calculated as $E(X^L) = \frac{1}{2}(X^L - Y)^2$. A derivative of the error for the output layer is δ^L=$(X^L - Y)$(1-X^L)X^L. For the hidden layer, the derivative of error is

$$\delta_i^{l-1}=x_i^{l-1}(1\text{-}x_i^{l-1})\ \sum_{j=0}^{d^l}\delta_j^l\ W_{i,j},\ 1 \leq l \leq L-1.$$

Then update the weights and bias for the output layer and hidden layer, respectively:

$$W_{i,j}:=W_{i,j}\text{-}\alpha\delta_j^{l-1}x_i^{l-1},\ 1 \leq l \leq L.$$
$$b^l:=b^l\text{-}\eta\delta^l,\ 1 \leq l \leq L.$$

Figure 17: Pseudocode for training the NN

While (the number of iteration <= user specified iteration number)

(1) Compute all $x_j^{(l)}$ in forward direction (Feed Forward Network)

$$x_j^{(l)} = \theta\left(\sum_{i=0}^{d^{(l-1)}} W_{ij}^{(l)} x_i^{(l-1)} + b^{(l)}\right)$$

$$\theta(s) = \frac{1}{1+e^{-s}}$$

(2) Compute all $\delta_i^{(l)}$ in the backward direction (Back Propagation Network)

$$\delta_i^{(l-1)} = x_i^{(l-1)}(1 - x_i^{(l-1)}) \sum_{j=0}^{d^{(l)}} \delta_j^{(l)} W_{ij}^{(l)}$$

(3) Update weight

$$W_{ij}^{(l)} := W_{ij}^{(l)} - \alpha\delta_j^{(l)} x_i^{(l-1)}$$

$$b^{(l)} = b^{(l)} - \eta\delta^{(l)}$$

We used the ADAM optimizer. We trained and tested models with a batch size of 100 and 100 numbers of the epoch. The learning rate of ADAM was 0.01. [16].

To improve our NN model, we considered different NN structures, activation functions, epoch numbers, batch sizes, and optimisers. An epoch number of 300, a batch size of 200, and an ADAM optimizer value of 0.01 gave the best model accuracy.

A.6 Appendix F

A.6.1 More Details about Support Vector Machine

The target feature is graft survival status, labeled with six categories: one month, two months, three months, six months, nine months, and one year. The input features of the model are selected from a boosted decision tree. The output of the SVM model is 0 or 1 in each graft survival label.

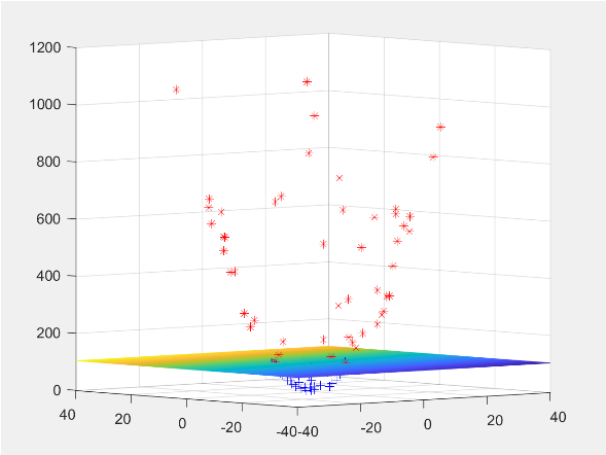
If we use kernel function Ψ(a, b)=(a, b, a^2+b^2), the i th row and j th column of the Q matrix is Y_j Y_j $K(X_i,X_j)$, where X_i means each values of each variable and 20-dimension coordinates of each data point in our model. Y_i represents the label of each graft status, a value of 1 or 0. Because the kernel function is Ψ(a, b)=(a, b, a^2+b^2), $K(X_i,X_j)$ is X_i^T X_j+(X_i^T X_i)(X_j^T X_j).

$$\max_\alpha \sum_{n=1}^{N} \alpha_n - \frac{1}{2}\alpha^T Q\alpha \ (1)$$
$$\text{s.t. for all } n,\ \alpha_n \geq 0, \sum_{n=1}^{N} \alpha_n y_n = 0$$

After performing the steps above, we can get $W = \sum_{n=1}^{N} \alpha_n y_n z_i^T$, where z_i is the nonlinear transformation of X_i. In this case, W is 1 by 3 vector and z is 3 by 1 vector, because if x is (a, b), the nonlinear transformation of x, Ψ(a, b), is (a, b, a^2+b^2), which is in z space. Finally, we can obtain get b from the Equation (2).

$$b^* = (\max_{i:y^i=-1} W^{*T} X^i + \min_{i:y^i=1} W^{*T} X^i) * -1/2 \ (2)$$

Figure 18: Example of SVM in 2-D



The input features of the SVM model can be subsets of the 145 features. For example, if we only use two of the 145 features, each data point and the SVM model can be plotted like in Figure 18.

For the SVM model [12], the input features were selected by a boosted decision tree from the original data set. The kernel is a radial basis function. The gamma value was 10.

A.7 Appendix G

Figure 19: A residual plot of the linear regression model, showing violation of linearity and homoscedasticity. A plot meeting the assumptions should show equal variance above and below 0 across all values on the x-axis.

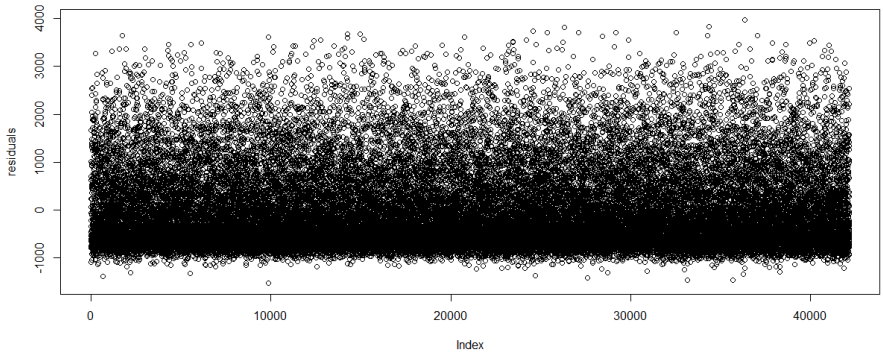
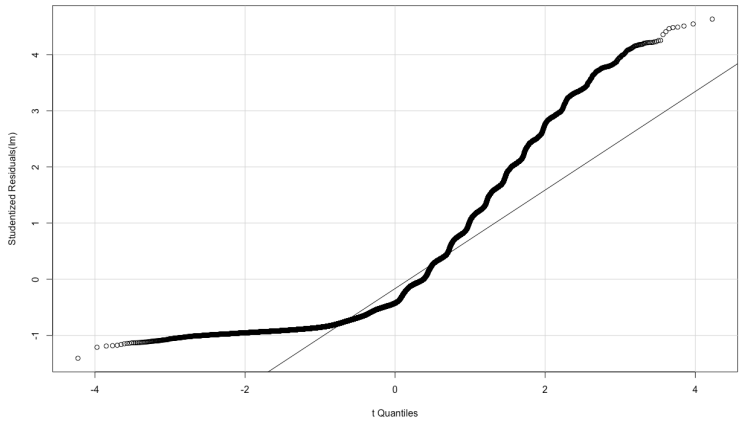


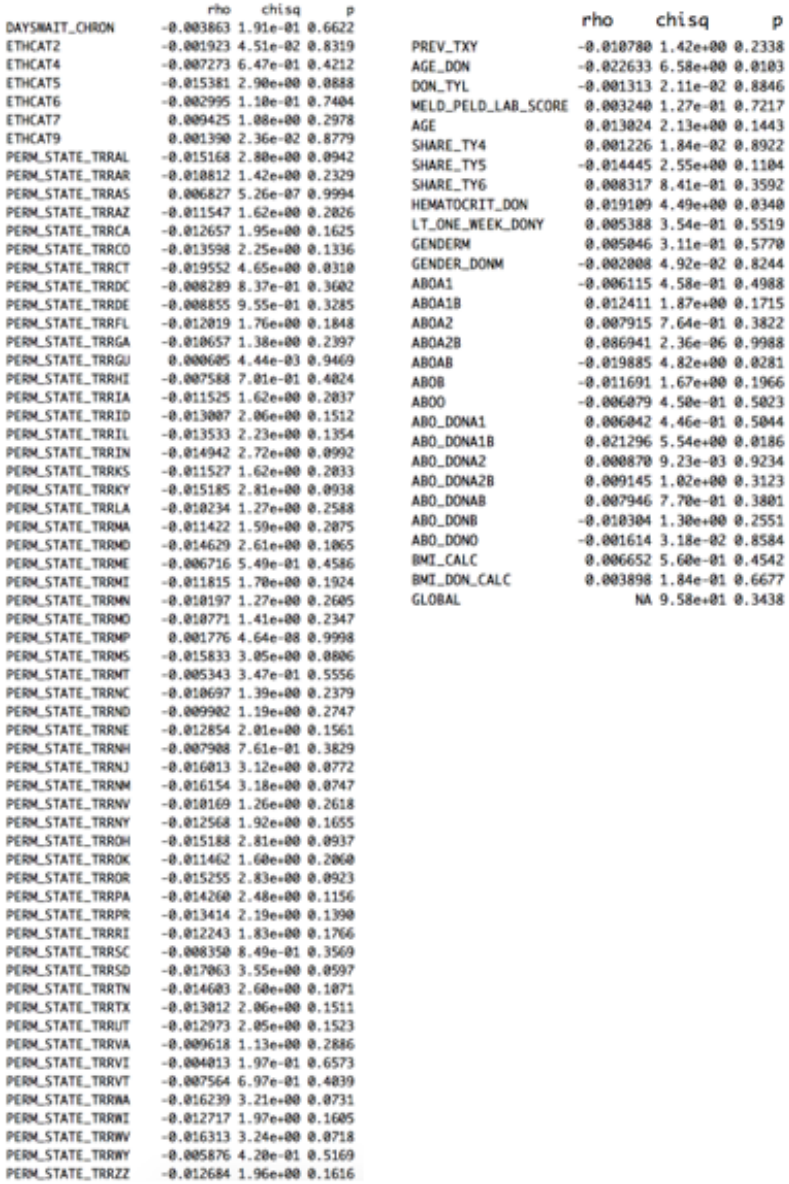
Figure 20: Q-Q plot of linear regression model, showing violation of normality. A plot meeting the assumption should show scatter points closely aligning with the solid line.



A.8 Appendix H

A.8.1 Assessing Assumptions of Cox Regression

Figure 21: Outputs of testing proportional hazard assumption. A p value less than 0.05 indicates violation of the assumption.



	rho	chisq	p
DAYSWAIT_CHRON	-0.003863	1.91e-01	0.6622
ETHCAT2	-0.001923	4.51e-02	0.8319
ETHCAT4	-0.007273	6.47e-01	0.4212
ETHCAT5	-0.015381	2.90e+00	0.0888
ETHCAT6	-0.002995	1.10e-01	0.7404
ETHCAT7	0.009425	1.08e+00	0.2978
ETHCAT9	0.001390	2.36e-02	0.8779
PERM_STATE_TRRAL	-0.015168	2.80e+00	0.0942
PERM_STATE_TRRAR	-0.010812	1.42e+00	0.2329
PERM_STATE_TRRAS	0.000827	5.26e-07	0.9994
PERM_STATE_TRRAZ	-0.011547	1.62e+00	0.2026
PERM_STATE_TRRCA	-0.012657	1.95e+00	0.1625
PERM_STATE_TRRCO	-0.013598	2.25e+00	0.1336
PERM_STATE_TRRCT	-0.019552	4.65e+00	0.0310
PERM_STATE_TRRDC	-0.008289	8.37e-01	0.3602
PERM_STATE_TRRDE	-0.008855	9.55e-01	0.3285
PERM_STATE_TRRFL	-0.012019	1.76e+00	0.1848
PERM_STATE_TRRGA	-0.010657	1.38e+00	0.2397
PERM_STATE_TRRGU	0.000605	4.44e-03	0.9469
PERM_STATE_TRRHI	-0.007588	7.01e-01	0.4024
PERM_STATE_TRRIA	-0.011525	1.62e+00	0.2037
PERM_STATE_TRRID	-0.013087	2.06e+00	0.1512
PERM_STATE_TRRIL	-0.013533	2.23e+00	0.1354
PERM_STATE_TRRIN	-0.014942	2.72e+00	0.0992
PERM_STATE_TRRKS	-0.011527	1.62e+00	0.2033
PERM_STATE_TRRKY	-0.015185	2.81e+00	0.0938
PERM_STATE_TRRLA	-0.010234	1.27e+00	0.2588
PERM_STATE_TRRMA	-0.011422	1.59e+00	0.2075
PERM_STATE_TRRMD	-0.014629	2.61e+00	0.1065
PERM_STATE_TRRME	-0.006716	5.49e-01	0.4586
PERM_STATE_TRRMI	-0.011815	1.70e+00	0.1924
PERM_STATE_TRRMN	-0.010197	1.27e+00	0.2605
PERM_STATE_TRRMO	-0.010771	1.41e+00	0.2347
PERM_STATE_TRRMP	0.001776	4.64e-08	0.9998
PERM_STATE_TRRMS	-0.015833	3.05e+00	0.0806
PERM_STATE_TRRMT	-0.005343	3.47e-01	0.5556
PERM_STATE_TRRNC	-0.010697	1.39e+00	0.2379
PERM_STATE_TRRND	-0.009982	1.19e+00	0.2747
PERM_STATE_TRRNE	-0.012854	2.01e+00	0.1561
PERM_STATE_TRRNH	-0.007988	7.61e-01	0.3829
PERM_STATE_TRRNJ	-0.016013	3.12e+00	0.0772
PERM_STATE_TRRNM	-0.016154	3.18e+00	0.0747
PERM_STATE_TRRNV	-0.010169	1.26e+00	0.2618
PERM_STATE_TRRNY	-0.012568	1.92e+00	0.1655
PERM_STATE_TRROH	-0.015188	2.81e+00	0.0937
PERM_STATE_TRROK	-0.011462	1.60e+00	0.2060
PERM_STATE_TRROR	-0.015255	2.83e+00	0.0923
PERM_STATE_TRRPA	-0.014260	2.48e+00	0.1156
PERM_STATE_TRRPR	-0.013414	2.19e+00	0.1390
PERM_STATE_TRRRI	-0.012243	1.83e+00	0.1766
PERM_STATE_TRRSC	-0.008350	8.49e-01	0.3569
PERM_STATE_TRRSD	-0.017063	3.55e+00	0.0597
PERM_STATE_TRRTN	-0.014603	2.60e+00	0.1071
PERM_STATE_TRRTX	-0.013012	2.06e+00	0.1511
PERM_STATE_TRRUT	-0.012973	2.05e+00	0.1523
PERM_STATE_TRRVA	-0.009618	1.13e+00	0.2886
PERM_STATE_TRRVI	-0.004013	1.97e-01	0.6573
PERM_STATE_TRRVT	-0.007564	6.97e-01	0.4039
PERM_STATE_TRRWA	-0.016239	3.21e+00	0.0731
PERM_STATE_TRRWI	-0.012717	1.97e+00	0.1605
PERM_STATE_TRRWV	-0.016313	3.24e+00	0.0718
PERM_STATE_TRRWY	-0.005876	4.20e-01	0.5169
PERM_STATE_TRRZZ	-0.012684	1.96e+00	0.1616

	rho	chisq	p
PREV_TXY	-0.010780	1.42e+00	0.2338
AGE_DON	-0.022633	6.58e+00	0.0103
DON_TYL	-0.001313	2.11e-02	0.8846
MELD_PELD_LAB_SCORE	0.003240	1.27e-01	0.7217
AGE	0.013024	2.13e+00	0.1443
SHARE_TY4	0.001226	1.84e-02	0.8922
SHARE_TY5	-0.014445	2.55e+00	0.1104
SHARE_TY6	0.008317	8.41e-01	0.3592
HEMATOCRIT_DON	0.019109	4.49e+00	0.0340
LT_ONE_WEEK_DONY	0.005388	3.54e-01	0.5519
GENDERM	0.005046	3.11e-01	0.5770
GENDER_DONM	-0.002008	4.92e-02	0.8244
ABOA1	-0.006115	4.58e-01	0.4988
ABOA1B	0.012411	1.87e+00	0.1715
ABOA2	0.007915	7.64e-01	0.3822
ABOA2B	0.086941	2.36e-06	0.9988
ABOAB	-0.019885	4.82e+00	0.0281
ABOB	-0.011691	1.67e+00	0.1966
ABOO	-0.006079	4.50e-01	0.5023
ABO_DONA1	0.006042	4.46e-01	0.5044
ABO_DONA1B	0.021296	5.54e+00	0.0186
ABO_DONA2	0.000870	9.23e-03	0.9234
ABO_DONA2B	0.009145	1.02e+00	0.3123
ABO_DONAB	0.007946	7.70e-01	0.3801
ABO_DONB	-0.010304	1.30e+00	0.2551
ABO_DONO	-0.001614	3.18e-02	0.8584
BMI_CALC	0.006652	5.60e-01	0.4542
BMI_DON_CALC	0.003898	1.84e-01	0.6677
GLOBAL	NA	9.58e+01	0.3438