A Data Driven Framework to Identify the Critical Variables, Visualize Their Conditional Relations and Predict the Outcomes of U.S. Heart Transplants

by

Ali Dag

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 6, 2016

Keywords: Data Mining, Bayesian Belief Networks, Healthcare Analytics, Medical Decision Making, Transplantation, United Network for Organ Sharing (UNOS)

Approved by

Fadel M. Megahed, Chair, Assistant Professor of Industrial and Systems Engineering
Jorge Valenzuela, Professor of Industrial and Systems Engineering
Richard Sesek, Professor of Industrial and Systems Engineering
Mark Carpenter, Professor of Mathematics and Statistics

Abstract

Predicting the survival of heart transplant patients is an important, yet challenging problem since it plays a crucial role in understanding the matching procedure between a donor and a recipient. Recent studies have shown that data mining models can be used to effectively analyze and extract novel information from large/complex transplantation datasets. The objective of this dissertation is to gain hidden, novel and useful information from these large and complex heart transplant datasets by employing data mining techniques, which helps decision makers to have a better understanding. Specifically, this work: 1) identifies the predictive factors for short-, mid- and long-term survival after the heart transplant, as well as their time-dependent effects on the given follow-up time point. Therefore, it enables us to differentiate the factors whose effect change over time, 2) develops a DSS tool that provides the patient-specific failure risk score based on the values of the relevant preoperative predictors, as well as to investigate the conditional relations among the important predictors of long term survival after heart transplants and 3) is an exploratory study that is still in progress, which evaluates the effect of the newly added variables to the predictability of the survival outcome. Overall, the research goal is to develop mathematical models and tools that present important retrospective findings, which can be the basis for a prospective medical studies.

Acknowledgments

Firstly, I would like to express my sincere thanks to my advisor Dr. Fadel Megahed for providing me continuous support throughout my PhD study. I have learnt how to make a good research that will have a high impact on the society from him. His help, patience and endless support guided me during the research and writing process. Without having him, I could not have completed this process. I would also like to thank to Dr. Richard Sesek., Dr. Carpenter and Dr. Valenzuela for their encouragement and support throughout my PhD.

I also would like to thank my wife Zeynep for her patience and love which has enabled me to complete this challenging process. I would also like to express my thanks to my father (Hasan Dag), mother (Secil Dag), mother-in-law (Nesrin Taspinar) and father-in-law (Tahir Taspinar), my brothers and sisters for their valuable sacrifice and support. Last but not least, I would like to present my appreciation to my primary school teacher (Selma Yilmaz) for her unbelievable effort to support me. Without having had her during those difficult years, I would not have had a PhD in the U.S.

This work was supported in part by Health Resources and Services Administration contract 234-2005-370011C. The content is the responsibility of the authors alone and does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Table of Contents

List of tables

List of Figures

Preface

This dissertation is submitted to the Graduate Faculty of Auburn University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Industrial and Systems Engineering. The completion of this dissertation involves many steps. In the fall semester of 2013, I met with Dr. Megahed to talk about the research concept that I am interested in. The general concept was to apply data mining models to extract useful information from heart transplant datasets. After having consecutive several meetings with him, we came up with a stream of research ideas involving data analytical applications on the concept of survival prediction after heart transplantation.

The first paper in the stream used several data mining models to predict short-, mid- and long-term survival after heart transplantation. He recommended such an idea since it allowed us to differentiate the factors whose importance (on survival) vary over time. After completing the analyses, we have shared our findings with Dr. Serkan Bulur and Dr. Hussam Farhoud, who are cardiologists (MD) and have been doing research in this area. They both provided their expertise in the field which in turn significantly improved the discussion part of the paper. Then, I have presented this study at the *INFORMS 2014 DMA (Data Mining & Analytics) Workshop* in San Francisco, California. Afterwards, we wrote it up as a journal article and submitted it to *Decision Support Systems Journal (DSS).* After completing the revision of this study, we have resubmitted it to the same journal. It is currently under second review. This study is presented in Chapter 2.

After finishing the first paper, I have met with Kazim Topuz, who is currently a PhD candidate at Wichita State University. We have talked about using probabilistic data mining approaches to obtain a patient-specific survival risk score. This idea was well received by Dr. Megahed. He gave a very good suggestion that involved building a graphical user interface (GUI) that can be freely used by the medical experts. Such tool would allow the medical experts to calculate the risk scores (after heart transplant) without having any expertise in the data mining area. The idea was well received by the other team members that also include our medical expert, Dr. Serkan Bulur. Then, I have presented it in *INFORMS 2015 DMA (Data Mining & Analytics) Workshop* in Philadelphia, Pennsylvania. Afterwards, it was accepted for the publication in *Decision Support Systems Journal (DSS).* This paper is presented in Chapter 3.

Finally, in the third paper, the idea was to eluate the effect of the newly added variables to the heart transplant datasets, which Dr. Megahed came up with. Such analysis would allow us to know whether the new variables are contributing in the predictive analytics concept. We have written it up as a journal article and are planning to submit the paper to *Health Care Management Science*.

Completing these three studies enabled me to learn a great deal about not only the field of data mining, but also heart transplantation. I have addressed two major critical research problems that have not been investigated before, which has made my dissertation more valuable.

1.Introduction

1.1 Problem Description and Significance

Heart failure is a very common medical condition in which the heart is weakened and cannot pump enough blood to meet the bodies needs [1]. The reader should note that the term heart failure does not indicate that the heart has stopped working (or is about to stop working). That being said, heart failure is a serious medical condition that affects an estimated 2-3% of the world's adult population, which corresponds to over 26 million people worldwide [2]. In the U.S. there are an estimated 5.8 million people living with heart failure, with an annual estimated incidence rate of over 550,000 [1, 3]. Among these patients, those who have severe end-stage heart failure (meaning all possible treatments except transplant have failed) are selected through a careful process to be placed on the heart transplant waiting list. If a patient is eligible, then she/he is placed on a waiting list for a transplant until a suitable donor heart is found [4]. The current matching process is determined based on a printed out list from the United Network for Organ Sharing (UNOS) computers, which is based on "blood type, body size, UNOS status, and length of time on the waiting list" [6].

The goal of my dissertation is to analyze the large and complex heart transplant datasets by employing data mining techniques. Specifically, my goal is to investigate the critical factors that affect the survival time after heart transplantation, the relations among these critical factors and provide a Decision Support System (DSS) tool that will aid medical decision makers to improve the efficiency of heart transplants. Effective utilization of this critical information can help to improve the prediction of survival, which, in turn, may help to better allocate the donated hearts.

This concern has recently become more important since the demand for heart transplantations has been dramatically increasing for several reasons, such as aging of the population, increasing obesity and diabetes etc. [5] and increasing life expectancy [6]. Currently, in the U.S. there are about 3,000 people waiting for a heart transplant on waiting lists at any one time, while there are only about 2,000 donor hearts available each year [4]. This ever-increasing gap between supply and demand of donated healthy hearts leads to longer waiting times and thereby leaves many to die while on the waiting lists [7]. Although the survival rates increased slightly in the recent years, there are many critical issues that have not yet been investigated, which potentially can be a potential driving force in decision making processes in the heart transplantation domain. Therefore, I address different discussions to analyze the related data and examine the root causes.

1.2 Research Objectives

By employing the proposed analytical methods, the objectives of my research are to:

1) Identify the predictive factors for short-, mid- and long-term survival after the heart transplant, as well as their time-dependent effects on the given follow-up time point. Therefore, we will be able to differentiate the factors whose effect change over time.

2) Develop a DSS tool that provides the patient-specific failure risk score based on the values of the relevant preoperative predictors, as well as to investigate the conditional relations among the important predictors of long term survival after heart transplants

3) Investigate the contribution (to the prediction performance) of recently added variables to the heart transplantation databases.

By developing mathematical models that employ both data mining algorithms and conventional

statistical method, the proposed study will enable researchers to find the hidden patterns embedded in large and complex heart transplantation datasets. The dataset used in this study has been provided by the United Network for Organ Sharing, UNOS, which is a "private, non-profit organization that manages the [United States]'s organ transplant system under contract with the federal government" [8]. This heart transplantation data contains information on all waiting list registrations and heart transplants that have been listed or performed in the U.S. and reported starting from October 1, 1987 until December 31, 2012. The variables in UNOS heart transplantation datasets include clinical and demographic factors related to donors, recipients and the transplant procedure. The variables can be classified into three major groups: a) preoperative factors that include donor/recipient demographics (age, gender, race, etc.), donor/recipient medical history, funding sources, and other factors that are considered prior to an operation; b) intra-operative factors, which describe several medical conditions during the transplant such as: if there is a chronic steroid use at transplant, or if the recipient is on life support at the time of transplant; and c) post-operative factors that include whether the patient died or lived, length of hospitalization after transplant, and information on any other complications. It should be noted that the data has been de-identified by UNOS prior to being received by the research team.

1.3 Dissertation Layout

This dissertation is organized as follows: In Chapter 2, the methodology that was proposed to differentiate the factors whose effect (on survival) vary over time. It should be noted that this paper was submitted to *Decision Support Systems* Journal, and it is currently under second review. Chapter 3 describes the methodology that was employed to calculate patient-specific risk score (after heart transplant). A comprehensive variable selection procedure is adapted in the study. In addition, a decision support systems tool (GUI) is provided for the medical experts to freely use in

their decision making processes. This study was published in *Decision Support Systems Journal (DSS)* in 2016. In Chapter 4, a methodology, by which the effect of newly added variables to the predictability of transplant outcome, has been described. This study will be submitted *to Health Care Management* Science journal. Finally, Chapter is the conclusion part of the dissertation, in which the contributions of the two studies are summarized and some future recommendations are made.

1.4 References

1. *What is Heart Failure ?* 2012 [cited 2014 01/26/2014]; Available from: http://www.nhlbi.nih.gov/health/health-topics/topics/hf/.
2. López-Sendón, J., *The heart failure epidemic.* Medicographia, 2011. **33**(4): p. 363-369.
3. *What Is Heart Transplant?* 2012 01/03/2012 01/31/2014]; Available from: http://www.nhlbi.nih.gov/health/health-topics/topics/ht/.
4. *What to Expect Before a Heart Transplant ?* 2012 [cited 2014 01/26]; Available from: http://www.nhlbi.nih.gov/health/health-topics/topics/ht/before.html.
5. Carmona, M., et al., *Heart failure in the family practice: a study of the prevalence and co-morbidity.* Fam Pract, 2011. **28**(2): p. 128-33.
6. Kong, G.L., et al., *A belief rule-based decision support system for clinical risk assessment of cardiac chest pain.* European Journal of Operational Research, 2012. **219**(3): p. 564-573.
7. Healy, D.G., et al., *Heart transplant candidates: factors influencing waiting list mortality.* Ir Med J, 2005. **98**(10): p. 235-7.
8. *UNOS | About Us*. 2014 2014/6/2]; Available from: http://www.unos.org/contact/index.php.

2 Predicting Heart Transplantation Outcomes through Data Analytics

2.1 Abstract

Predicting the survival of heart transplant patients is an important, yet challenging, problem for researchers and medical practitioners since it plays a crucial role in understanding the matching procedure between a donor and a recipient. Data mining models can be used to effectively analyze and extract novel information from large/complex transplantation datasets. The objective of this study is to predict the 1-, 5-, and 9-year patient's graft survival following a heart transplant surgery via the deployment of analytical models that are based on four powerful classification algorithms (i.e. decision trees, artificial neural networks, support vector machines, and logistic regression). Since the datasets used in this study has a much larger number of survival cases than deaths for 1- and 5-year survival analysis and vice versa for 9-year survival analysis, random under sampling (RUS) and synthetic minority over-sampling (SMOTE) are employed to overcome the data-imbalance problems. The results indicate that the logistic regression when combined with synthetic minority over-sampling achieves the best classification performance for 1-, 5- and 9-year survival prediction, with area-under-the-curve (AUC) values of 0.624, 0.676 and 0. 838, respectively. By applying sensitivity analysis to the data analytical models, the most important predictors and their associated contribution for the 1-, 5-, and 9-year graft survival of heart transplant patients are identified. By doing so, variables, whose importance changes over time, are differentiated. Not only this proposed hybrid approach gives superior results over the literature but also the models

and identification of the variables present important retrospective findings, which can be the basis for a prospective medical study.

2.2 Introduction

Heart failure is a serious medical condition, where the heart is weakened and cannot pump enough blood to meet the body's needs [1]. It affects an estimated 2-3% of the world's adult population, i.e. over 26 million adults worldwide [2]. In the U.S., there are an estimated 5.8 million people living with heart failure, with an annual estimated incidence rate of over 550,000 [1, 3]. The most acute stage of heart failure is called end-stage heart failure since all possible treatments (with the exception of a transplant) have failed [1].

The demand for heart transplantations has been dramatically increasing for several reasons that include: an aging population, increased rates of obesity/diabetes, etc. [4, 5]. Currently, in the U.S. there are about 3,000 people waiting for a heart transplant on waiting lists at any one time, while there are only about 2,000 donor hearts available [6]. This ever-increasing gap between supply and demand of donated healthy hearts leads to longer waiting times and thereby leaves many to die while on the waiting lists [5].

The large data pools recorded during the transplantation procedures include valuable information about donors, recipients and the procedure itself. Effective utilization of this critical information can help to improve the prediction of transplantation outcomes, which, in turn, may help medical decision makers to better allocate donated hearts. Therefore, there is a growing body of literature on predicting organ transplantation outcomes. In our estimation, this literature can be divided into three main streams: A) statistical-based approaches (e.g., [7-12]), B) data mining methods (e.g., [13-17]), and C) studies accounting for the time-dependent effect of covariates on

transplantation outcome (e.g., [14, 18-20]). We present more details on these three streams in the paragraphs below.

Stream A focuses on one of three specific research questions. First, what is the effect of a certain variable on survival? For example, Gupta *et al*. [9] used the *Cox Proportional Hazard Model* (*CPHM*) to examine if an older donor age (>50 years) increases the risk of mortality after a heart transplant. In the second sub-area, the combined effect of several variables on a transplantation outcome is examined [7, 8, 10]. Studies in the third sub-area have mainly focused on identifying risk levels/scores for patients after a transplant (e.g., [10]). In this stream, the examined predictors are primarily based on domain knowledge. This may be somewhat limiting since it assumes that: a) there is a well-defined hypothesis that needs to be investigated (e.g. [7] examined the effect of donor age and organ ischemic time on patient survival post transplantation); and b) a parametric model can be used (i.e., a functional relationship between the predictors and the outcome is assumed).

Stream B attempts to address the aforementioned limitations through the use of data analytic/mining methods, which allows one to extract hidden, novel patterns and non-trivial information from large datasets, without requiring prior knowledge about the data. Based on our review, methods in Stream B attempts to determine: a) what is the best predictive model among a subset of competing models to accurately predict a transplantation outcome at an arbitrary follow-up time (i.e., what is the outcome after *x* years?); and b) what are the significant predictors based on the best model? For example, Oztekin et al. [15] compared the predictive performance of artificial neural networks (ANNs), logistic regression (LR), and decision trees (DT) for determining the survival outcome at 9-years post a thoracic transplant. Their study showed that the ANN and logistic regression models performed similarly, with accuracy rates of 82.4% and

81.9% respectively, and outperformed the decision tree (accuracy rate of 74.9 %). Similar studies are highlighted in Dag et al. [13]. While the results are typically impressive (with the highest reported in [15]), there are several opportunities for improvement:

These method typically focus on the outcome prediction for one time-period only [13, 15-17, 21, 22]. Thus, there is no understanding whether the identified predictors are important in different survival intervals; e.g., is the donor age an important predictor for short-term outcomes only or it also affects long-term survival.

The benchmark approach of [15] used data imputation methods (based on the entire dataset) to handle missing data. While this can improve the model's prediction, its limits the benefits from using data-driven models in health-care. As eloquently stated in [21]: "*Survival analysis using standard statistical tools ... can be considered as population-based models. Predictions are derived on probability or distance from the population estimates. Data mining offers tools for decision-making for an individual patient rather than a population of patients*."

To our knowledge, none of the studies in the literature [13, 14, 21] have considered re-sampling as an approach to handle the class imbalance problem, i.e. the # of survivals and deaths are significantly different, in transplantation datasets. Accordingly, there exists an opportunity to improve the sensitivity of the classifier to the minority class.

Stream C investigates the dynamic effect of significant variables on the transplantation outcome. Most of the work within this stream is based on statistical techniques, with the exception [14] who determined the effect of the predictors on the 1- to 7-year kidney transplantation outcomes by using logistic regression, CPHM, and multiple-output ANNs. It should be noted that the statistical approaches within this stream consider time as a continuous variable instead of

investigating multiple time-points (e.g. [14, 18-20]). Note that there are no methods for examining changes in a predictor's importance over time, for heart transplants, when data mining methods are used.

The main objectives of this study are to develop a data-driven approach to predict the short-term, medium-term, and long-term outcomes for heart transplantation patients, and understand how the importance of a predictor change over time. Based on the literature, we define 1- ([7, 10, 23, 24]), 5- ([20, 22, 23, 25]), and 9-years ([13, 15]) as the time-frame for short-term, medium-term and long-term prediction intervals. To address these objectives, we compare the performance of ANN, DT, SVM (support vector machines), and logistic regression in modeling the relationship between the set of pre-operative predictors and the binary outcome (at each time-point). To overcome the class imbalance problem, two common data balancing techniques for each follow-up time are employed to generate balanced training sets. Note that we compare these two techniques with random sampling (i.e., the classical approach used for balanced datasets) to quantify the improvements gained by applying the two re-sampling techniques. Thus, this paper attempts to: i) address the three opportunities for improvement highlighted in Stream B, and ii) bridge the gap between Streams B & C. The remainder of this paper is organized as follows. In Section 2.3, the overall data analytics methodology used in this study as well as the data source and preparation are described. The results obtained from the actual data analysis are provided and discussed in Section 2.4. In Section 2.5, conclusions are presented and some thoughts on the direction of future research are offered.

2.3 Methodology

In this paper, a hybrid data analytic methodology (as depicted in Figure 2.1) that consists of five sequential phases is proposed. The first phase of data preparation consists of three steps: a) data cleaning, where missing or invalid values for the predictor variables were systematically checked for; b) data inclusion, where three datasets were extracted from the original data to address how the effect of predictor variables change with different survival horizons (1-, 5- and 9-years); and c) application of balancing techniques to prepare the datasets to be processed by the prediction algorithms. Phase 2 represents the process in which data mining prediction algorithms are applied to these balanced datasets, with a 10-fold cross validation procedure. The goal of Phase 2 is to obtain the classification performance of the prediction models and the ranking of the predictor variables based on their importance level., i.e., each model provides its own set of important variables for each of the three time-points. Models with unsatisfactory performance are then discarded in Phase 3. The decision to discard a model is carried out by evaluating the performance of the model to the results from the best deployed model. Models with best performance are then kept to be used in the subsequent phase. In Phase 4, the important-variable sets (obtained from each satisfactory model) are combined with an information fusion method to collect information (ranked variables) from multiple models. This process is repeated for each time point. This enables us to have three different important-variable sets representing the three different time points (1-, 5- and 9-years). In the last phase, the comparison of these three important variable sets allows us to distinguish the variables whose importance change over time (from 1-year to 5- and 9-years). Additional information on each of these phases is provided in the following subsections.

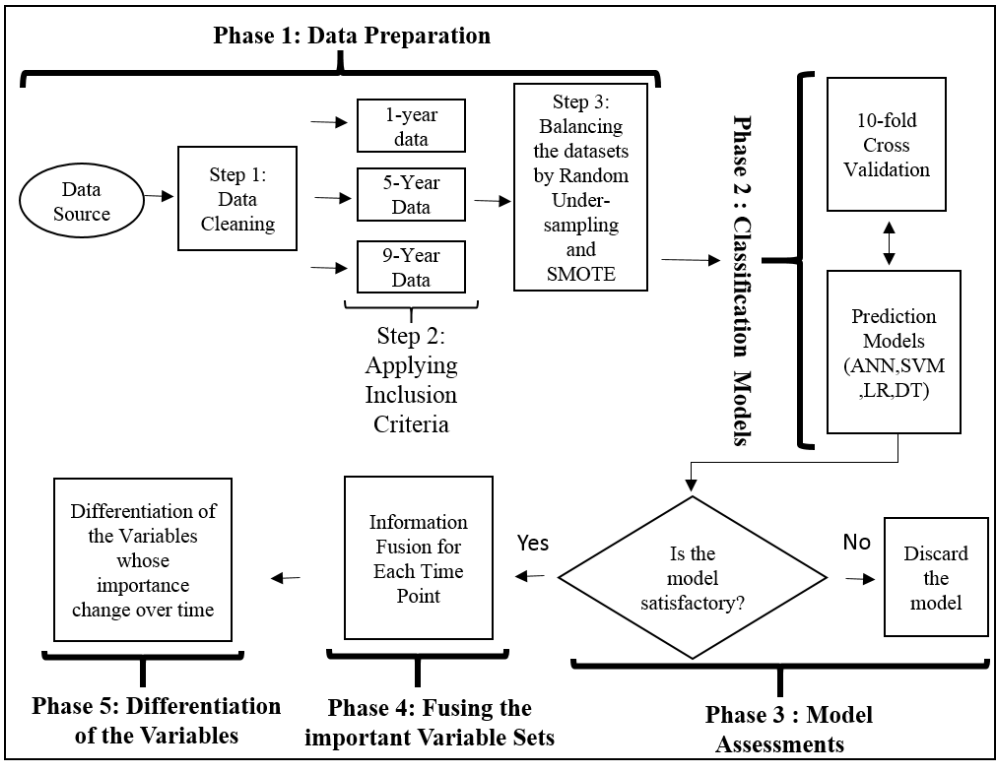

Figure 2.1: An Overview of the Proposed Hybrid Data Analytic Approach

2.3.1 Data Preparation

The dataset used in this study has been provided by the United Network for Organ Sharing, UNOS, which is a "private, non-profit organization that manages the [United States]'s organ transplant system under contract with the federal government" [26]. This heart transplantation data contains information on all waiting list registrations and heart transplants that have been listed or performed in the U.S. and reported starting from October 1, 1987 until December 31, 2012. The variables in UNOS heart transplantation datasets include clinical and demographic factors related to donors, recipients and the transplant procedure. The variables can be classified into three major groups: a) preoperative factors that include donor/recipient demographics (age, gender, race, etc.), donor/recipient medical history, funding sources, and other factors that are considered prior to an operation; b) intra-operative factors, which describe several medical conditions during the

transplant such as: if there is a *chronic steroid use* at transplant, or if the recipient is on life support at the time of transplant; and c) post-operative factors that include whether the patient died or lived, length of hospitalization after transplant, and information on any other complications. It should be noted that the data has been de-identified by UNOS prior to being received by the research team.

As noted earlier, one of the primary objectives of this study is to develop a data-driven model to predict survival through 1, 5 and 9 years. Therefore, four post-operative variables can be used as candidate outcome variables for this study. These variables are *pstatus* (binary variable, denotes whether the person is dead or alive at the last follow-up time), *gstatus* (a similar binary variable, denoting if whether the graft has failed or succeeded at the last follow-up time), *ptime* (referring the time frame from the day of transplant to the recipient's death/last follow-up time, in days) and *gtime* (a similar continuous variable referring to time frame from transplant to graft's failure/last follow-up time). For purposes of this study, the *gstatus* represents the most suitable outcome variable when combined with the *gtime*. It is more informative than the *pstatus* variable since it allows us to distinguish between patients who died solely due to graft compatibility versus those who may have died due to other reasons (e.g. a traffic accident or a medical condition that is not related to the heart transplant). By combining it with gtime, we can determine if the graft failed prior to the 1, 5, and/or 9 year thresholds.

2.3.1.1 Data Cleaning

In this paper, we followed the five-step data cleaning process of [13]. In the first step, we eliminated all the intra- and post-operative factors since our research questions are related to outcome prediction prior to transplant. We cleaned all erroneous and duplicated records using outlier detection algorithms in STATISTICA 11 (www.statsoft.com). In the third and fourth steps, we eliminated variables that will not have any predictive power (e.g., patient ID number), and invariant variables. In the final step, we eliminated missing records from the data since data imputation methods are not suitable for our end-goal as explained in Section 2.2. At the end of this stage, we had 15,580 patient records and 122 variables in the dataset (i.e., same as in [13]).

2.3.1.2 Data Inclusion Criteria

For each of the three discrete time-points, the data is excluded when the *gtime* is less than the number of days required for the time-point analysis and the gstatus indicates that the patient is still alive. The decision (C_i) can be represented mathematically as:

$$C_i = \begin{cases} Yes & \text{if } gtime \leq i \text{ *365 \& } gstatus = 0, \\ \text{No} & \text{otherwise} \end{cases} \tag{2.1}$$

where i =1, 5, or 9 years. Based on Eq. (2.1), the following outcomes can be expected: a) if an observation is excluded at a lower value of i (e.g. i=1), it will be excluded for any higher value of i; b) the total number of records will monotonically decrease as the time-point for analysis increases; and c) the expected gap between survivors (i.e.*gstatus=0*) and graft failures will decrease since transplantation mortality increases over time. After we applied the inclusion criteria to the data, we updated the gstatus to be equal to 0 for those patients whose *gtime* > i *365 since

this means that they expired after our time-horizon. A summary is provided for the distribution of fatalities, survivals, and excluded data over the three time periods in Table 2.1.

Table 2.1: Number of survivals, failures, and excluded observations over the 3 time-points

Time Point	Survivals	Failures	Excluded
1 - Year	12,103	1,617	1,860
5 – Years	5,519	3,081	6,980
9 - Years	1,663	3,837	10,080

2.3.1.3 Sampling Methods

Based on Table 2.1, the number of survivals and failures are significantly different for each time-period. In the data mining literature, there are several sampling methods [27-30] that can be used to improve the prediction outcomes based on imbalanced datasets. The machine learning community has addressed the class imbalance problem in two ways [27]. One way is to assign distinct costs to training examples (e.g., [31]), while more recent approaches attempt to re-sample the original dataset. Re-sampling is typically applied by either over-sampling the minority class and/or under-sampling the majority class [27, 30]. Due to their popularity and excellent results, we have only considered re-sampling approaches in this paper. Note that the literature reports conflicting results on the superiority of over-sampling methods (e.g., [27-32]) vs. under-sampling methods [33]. Therefore, in this paper, we have applied both synthetic minority oversampling technique (SMOTE) and random under-sampling (RUS), which are among the most popular/powerful over-sampling and under-sampling approaches, respectively.

RUS is a systematic process where some of the cases from the majority class are randomly removed from the original training datasets until the remaining number of cases in the two

classification categories becomes approximately equal. On the other hand, SMOTE over-samples the minority class by "*taking each minority class data point and introducing synthetic examples along the line segments joining any or all of the k-minority class nearest neighbors*" [34]. The process repeats until the number of cases from both classes becomes approximately equal. Note that both approaches are implemented in several open-source software (e.g., R Programming Language and Weka). For more details on SMOTE, the reader is referred to [27].

2.3.2 Data Analytics Models

In this study, we apply three popular data analytic models (support vector machines, artificial neural networks, and decision trees) and a conventional statistical method (logistic regression). We selected these four models due to: a) superior performance in several transplantation papers (see e.g., [7, 8, 10, 12, 14, 15, 21, 22, 35]), and b) their superior performance in our preliminary analysis. Since these models are very popular in classification problems, we only provide a brief description for each of them in the subsections below.

2.3.2.1 Support Vector Machines

Support vector machines are a set of supervised learning methods that are used for classification (as in this study, where a known value of the gstatus was used to develop a predictive model for future graft survival or failure), regression and outlier detection [36, 37]. When used for classification purposes, SVMs are extremely powerful since they can be used for linearly and non-linearly separable datasets [38]. For nonlinear cases, the data is typically mapped into a higher-dimensional space so that the new dataset in higher-dimension becomes linearly separable [38]. This mapping procedure can increase the computational complexity. This problem can be handled

efficiently by using one of several Kernel functions (see [38] for more details). In this paper, we use the radial basis kernel function (RBF) since it provided the best results in our preliminary analysis.

2.3.2.2 Artificial Neural Networks

ANNs are widely used in a variety of data mining problems that include classification, pattern recognition, and optimization. An ANN is a computational system that consists of "*a highly interconnected set of processing elements, called neurons, which process information as a response to external stimuli. An artificial neuron is a simplistic representation that emulates the signal integration and threshold firing behavior of biological neurons by means of mathematical equations*" [39]. The information flow between artificial neurons, thereafter referred to as neurons, is determined via connections between peer neurons. The flow of information through each neuron occurs in an input-out manner. In this paper, a multilayer perceptron-based ANN (MLP-ANN) is used as it outperformed other ANN formulations in our preliminary analysis. For more details on the mathematical formulation of the ANN used in this study, the reader is referred to [38, 39].

2.3.2.3 Decision Trees

Decision trees are one of the most understandable and easy to interpret prediction methods. This is one of the main reasons why decision trees are widely used in several data mining and transplantation applications. The procedure starts with splitting the entire dataset into several subsets, which contain more or fewer homogeneous states of dependent variable [40]. At each split in the tree, the impacts of all predictor variables on the dependent variable are evaluated. This procedure takes place successively, until a decision tree is in a stable state. Popular decision tree

algorithms include Quinlan's ID3, C4.5, C5 [41, 42] and C&RT (Classification and Regression Trees) [40]. The C&RT algorithm has been employed in this study due to its favorable performance compared to the other decision tree algorithms obtained in the preliminary analysis.

2.3.2.4 Logistic Regression

Different than the three models explained above, logistic regression is a standard regression technique, used when the dependent variable (i.e. in our case, the *gstatus*) is dichotomous. In such model, the log odds of the outcome are modeled as a linear combination of the predictor variables (i.e., the preoperative variables that are included in the model) [43]. In our study, we applied stepwise logistic regression to select the most relevant predictors.

2.3.3 Sensitivity Analysis of Predictor Variables

After determining the performance of the different predictive models, the relative importance of each of the independent variables is measured using the *sensitivity analysis* (SA). This phase is indispensable to the analyses for several reasons. First, it can suggest the underlying casual factors for any of the prediction models. This is particularly important in understanding and communicating the results of ANNs [44] which are still considered by many to be black box models [for a recent e.g. see 45]. A second major reason for the importance of *sensitivity analysis* is it provides us with a framework to capture the importance of independent variables across different models.

The sensitivity of a specific predictor variable is calculated by taking the proportion of the error of the model that includes this variable to the error of the model when it does not include this specific variable [46]. The importance of a variable is in direct proportion to variance of predictive

error of the classification model in the absence of that specific variable. The same method is followed for all classification models, and is used in ranking the relative importance of the variables of each classification model according to the sensitivity measure defined by Saltelli [47]. Their measure is defined as

$$S_i = \frac{V_i}{V(y)} = \frac{V(E(y \mid x_i))}{V(y)} \tag{2.2}$$

where y is the dichotomous output variable (*gstatus*), and the unconditional output variance is denoted by $V(y)$. The expectation operator is denoted by E, which calls for an integral over all predictor variables except x_i. A further integral operator is implied over x_i by the operator V_i. The importance of a specific variable is then computed as the normalized sensitivity as described by Saltelli *et al.* [48].

2.3.4 Information Fusion

Information fusion (IF) techniques combine information obtained from multiple forecasts (from prediction models) in an attempt to decrease model uncertainty and increase the knowledge gained. The motivation for such techniques also stem from observations in the literature about combining multiple forecasts. Specifically, it has been shown that robustness and accuracy of information can be increased [49], while the uncertainty and bias of individual models can be decreased by combining multiple forecasts [50]. Therefore, there is an increasing deployment of IF techniques in data-mining problems as opposed to the application of a single method [e.g. see 51]. It should be noted, however, that there is no best way of combining predictions in explanatory data analysis situations, as in the case of this heart transplant analysis. This is similar to the application of data

mining techniques, where the best model is problem-specific and often cannot be determined prior to investigation. In such situations, the determination can only be done via trial and error experimentation [52]. In this study, the information fusion model which is presented by Sevim *et al.* [53] is adopted since it allows decision makers to rank the variables in terms of importance order. In other words, it would be more intuitive to explain the findings to the medical practitioners. The mathematical formulation of the IF model is presented in the paragraphs below.

For any of the prediction models (logistic regression, SVM, ANN, and DT), the formulation for the prediction model can be generalized as follows:

$$\hat{y} = g\left(x_1, x_2, ..., x_m\right) \tag{2.3}$$

where g represents a prediction function for the model, y is the dichotomous response variable, x is the value for a specific preoperative variable, and m denotes the number of preoperative variables in the model. Given r predictor models to be combined, the information fusion model can be represented as

$$\hat{y}_{fused} = \psi\left(g_1(\mathbf{x}), g_2(\mathbf{x}), ..., g_r(\mathbf{x})\right) \tag{2.4}$$

In the above equation, $\mathbf{x}$ denotes the vector of predictor variables. In this study, ψ is defined as a linear function, which reduces the IF model to

$$\hat{y} = \sum_{i=1}^{r} \lambda_i g_i(x), \quad \text{where} \quad \sum_{i=1}^{r} \lambda_i = 1 \tag{2.5}$$

The values for the λs are the updated classification AUC measures of a single classification model. That is, the higher the AUC metric of an individual predictor model, the more impact of it on the fused model [53]. The final sensitivity measure of a specific variable (denoted by θ) that is fused by *r* individual models can be obtained by combining Eqs. (2.4) and (2.5). The result is shown in the Eq. below;

$$S_{\theta(\,fused)} = \sum_{i=1}^{r} \lambda i S_{i,\theta} = \lambda_{1\times} S_{1,\theta} + ... + \lambda_{r\times} S_{r,\theta} \tag{2.6}$$

where $S_{i,\theta}$ is the normalized sensitivity measure of the θ^{th} variable in the i^{th} model.

After deriving the importance of each predictor variable dictated from each folds within each classification model, as shown by Eq. (2.6), a union set of important predictors can now be obtained for each method. Finally, by again applying the information fusion method across these four models, the final importance of each variable for each time point is calculated. Therefore, three different sets of important variables and their rankings are constructed for the 1-, 5- and 9-year analyses. This allows us to differentiate the variables whose impact on the outcome changes over time.

2.4 Results and Discussion

Based on the discussion in Section 2.3, four data analytic models (C&RT, ANN, SVM, and Logistic regression) were applied to predict the effect of the preoperative variables on the response/outcome variable *gstatus*. Since the transplantation dataset is imbalanced for all 3 time-periods as illustrated in Table 2.1, two different re-sampling strategies (and a random sampling approach, which we include to show the improvements obtained by re-sampling) on these four

models were used, resulting in 12 different models (for each time frame) that are listed and described in Table 2.2. Note that these results in a total of 36 different models, since there are 3 time-periods where each of the models listed in Table 2.2 was applied. The remainder of this section is organized as follows. In Section 2.4.1, the classification results are presented based on the four performance metrics. In essence, this subsection contains the results from Phases 1-3 of the proposed hybrid method. The results from Phases 4-5 (sensitivity analysis and information fusion phases) are discussed in Section 2.4.2.

Table 2.2: The list of the data analytic models used for each time-period

Model Name	Description
CRT.NO	Classification and regression trees with random sampling
LR.NO	Logistic regression with random sampling
ANN.NO	Artificial neural networks with random sampling
SVM.NO	Support vector machines with random sampling
CRT.S	Classification and regression trees with SMOTE
LR.S	Logistic regression with SMOTE
ANN.S	Artificial neural networks with SMOTE
SVM.S	Support vector machines with SMOTE
CRT.RUS	Classification and regression trees with random under-sampling
LR.RUS	Logistic regression with random under-sampling
ANN.RUS	Artificial neural networks with random under-sampling
SVM.RUS	Support vector machines with random under-sampling

2.4.1 Data Analytical Model Results

Table 2.3 provides the mean and the standard deviation values for the AUC, accuracy, recall, and specificity metrics for all classification models, with the 10-fold cross validation sample. It should be noted that the evaluation criteria (in prediction problems) can be problem-specific (e.g. rare event prediction). In our current study, the AUC criteria is selected as the main evaluation criterion

since there exist imbalances (between survivals and failures) for all of the three time-points. Having said that, recall and specificity values are also considered for a more objective comparison of the models.

As presented in Table 2.3, those models, where re-sampling was not considered (i.e., the NO models), can be considered as poor (with respect to the un-reasonably high difference between recall and specificity) when compared to SMOTE and RUS models even though reasonable AUC values were obtained through no-sampling. For example, the recall for the 1-Yr models and 5-Yr models has not exceeded 13%, and 41%, respectively. This means that the models generated by this data are somewhat naive. As an illustration, consider the 1-YR CRT.NO model. This model essentially forecasts that any patient will survive, as shown in specificity=0.999 and recall=0.024. Note that this trend is reversed for the 9-year data (unsurprisingly, due to Table 2.1), where the NO models do an excellent job in predicting deaths, but a substandard job in predicting survivals. These results justify the need for applying re-sampling techniques for transplantation datasets. Hereafter, we discuss the SMOTE and RUS results.

Second, the AUC metrics for a given model always increase with an increase in the time horizon. The practical implication of this observation is significant since it means that the ability of these classification models in correctly predicting the survival outcome with larger time horizons. This effect has been observed in kidney transplants [14]. One possible explanation is that the time-horizon gets larger, the discrepancies between the operations' start dates decrease.

Table 2.3: Classification results of the eight models for 1-, 5-, and 9-year time-points

Time Frame	Model	AUC	Accuracy	Recall	Specificity
1-Yr Survival	CRT.NO	0.455(0.043)	0.893(0.032)	0.024(0.002)	0.999(0.001)
	CRT.S	0.583(0.021)	0.615(0.027)	0.490(0.056)	0.631(0.033)
	CRT.RUS	0.583(0.015)	0.584(0.029)	0.561(0.039)	0.584(0.030)
	NN.NO	0.595(0.034)	0.865(0.011)	0.118(0.057)	0.965(0.011)
	NN.S	0.569(0.016)	0.691(0.015)	0.412(0.037)	0.728(0.020)
	NN.RUS	0.583(0.021)	0.571(0.033)	0.540(0.053)	0.580(0.044)
	LR.NO	0.630(0.027)	0.881(0.012)	0.128(0.083)	0.988(0.016)
	LR.S	0.624(0.023)	0.626(0.101)	0.543(0.046)	0.636(0.129)
	LR.RUS	0.624(0.031)	0.613(0.191)	0.566(.035)	0.619(0.194)
	SVM.NO	0.614(0.021)	0.847(0.048)	0.119(0.080)	0.988(0.013)
	SVM.S	0.609(0.011)	0.666(0.020)	0.463(0.055)	0.693(0.011)
	SVM.RUS	0.596(0.018)	0.550(0.026)	0.592(0.033)	0.544(0.029)
5-Yr Survival	CRT.NO	0.630(0.019)	0.664(0.020)	0.352(0.024)	0.837(0.036)
	CRT.S	0.641(0.017)	0.609(0.011)	0.479(0.037)	0.682(0.036)
	CRT.RUS	0.634(0.014)	0.609(0.017)	0.600(0.024)	0.615(0.022)
	NN.NO	0.663(0.025)	0.632(0.019)	0.410(0.037)	0.757(0.026)
	NN.S	0.621(0.029)	0.618(0.024)	0.488(0.033)	0.696(0.019)
	NN.RUS	0.628(0.023)	0.594(0.024)	0.581(0.033)	0.600(0.033)
	LR.NO	0.677(0.024)	0.679(0.010)	0.354(0.018)	0.861(0.019)
	LR.S	0.676(0.011)	0.634(0.018)	0.510(0.023)	0.737(0.024)
	LR.RUS	0.671(0.020)	0.632(0.019)	0.616(0.044)	0.641(0.031)
	SVM.NO	0.664(0.013)	0.674(0.010)	0.342(0.016)	0.860(0.015)
	SVM.S	0.673(0.014)	0.654(0.012)	0.541(0.023)	0.702(0.028)
	SVM.RUS	0.669(0.017)	0.616(0.013)	0.640(0.023)	0.602(0.019)
9-Yr Survival	CRT.NO	0.820(0.025)	0.748(0.032)	0.820(0.032)	0.581(0.081)
	CRT.S	0.823(0.014)	0.746(0.017)	0.787(0.021)	0.651(0.043)
	CRT.RUS	0.791(0.019)	0.718(0.020)	0.657(0.030)	0.863(0.021)
	NN.NO	0.825(0.028)	0.747(0.027)	0.837(0.030)	0.538(0.054)
	NN.S	0.814(0.025)	0.742(0.025)	0.808(0.024)	0.583(0.062)
	NN.RUS	0.781(0.017)	0.697(0.014)	0.682(0.018)	0.734(0.021)
	LR.NO	0.840(0.027)	0.748(0.027)	0.820(0.024)	0.593(0.052)
	LR.S	0.838(0.025)	0.754(0.022)	0.777(0.024)	0.712(0.056)
	LR.RUS	0.815(0.012)	0.712(0.014)	0.615(0.019)	0.936(0.027)
	SVM.NO	0.833(0.032)	0.746(0.028)	0.911(0.020)	0.363(0.065)
	SVM.S	0.831(0.027)	0.750(0.022)	0.883(0.021)	0.450(0.070)
	SVM.RUS	0.816(0.011)	0.722(0.013)	0.699(0.019)	0.776(0.021)

A third noteworthy observation is that SMOTE generally outperforms RUS for the AUC metric for all three time-periods (except 4 out of the 24 (S + RUS) models). This result suggests that the use of more computationally intensive sampling methods for heart transplantation datasets may be justified since they result in better performance. The fourth, and potentially the most interesting observation, is that LR provides the best results for the AUC metric for all of the three time periods. As it has been previously discussed, the performances of these analytical models depend on the characteristics of the dataset such as noise, non-linearity, dimension (i.e. feature space and number of samples) etc. One possible conclusion that can be made is that the potential high-level of non-linearity among some features might have been solved via transformation these features in the Logistic Regression.

It is clear that the classification models with SMOTE generally outperform those with RUS (20 out of 24 models) for all three time-periods based on our primary AUC metric. In addition, for those two cases (i.e. NN.RUS for 1- and 5-years), where the RUS models outperform the S models, the AUC difference between the models is 0.014 and 0.007 for 1-year and 5-years, respectively. Having said that, the main goal of this study is not to compare the prediction capabilities of classification models that have been re-sampled by using different sampling algorithms. But it is rather to determine the important variables for survival after different time horizons, which in turn enables us differentiate the ones whose important change over time. Therefore, for the sake of consistency, CRT.S, NN.S, LR.S, and SVM.S classification models are selected to be combined via the IF model explained in Section 2.3.4. This process is repeated for each of the three time-periods. These results are provided and explained in the following subsection.

2.4.2 Information Fusion-based Sensitivity Analysis Results

Using the predictive models as explained earlier, sensitivity analysis (SA) is performed on each model to compute the importance of each variable. A predictive variable is added to the model if its addition results in improving the model's AUC prediction performance. Therefore, for each three time-points, four sets of important variables are identified based on the CRT.S, NN.S, LR.S, and SVM.S models. After obtaining the important variables for each time point, the results are combined via IF analysis (within each time point) as discussed in Section 2.3.4. This procedure provided us with three sets of (combined) variables, each of which represents a different time point. It should be noted that these variable lists are not the important variables of individual predictive models. Rather, they are the important variables obtained by combining the four best prediction models (in terms of AUC measures) for each different follow-up time.

The agreement of the four predictive models (within each time point) is presented in Table 2.4. There are 28, 29 and 26 variables that were found to be important for 1-, 5- and 9-year survival after heart transplants, respectively. Of the 28 variables, 13 were commonly found to be important by all of the four models (i.e. CRT.S, NN.S, LR.S and SVM.S) employed for 1-year survival analysis. These numbers are 14 out of 29 and 7 out of 26 variables for 5- and 9-year survival analysis, respectively. It can be summarized that around 60 % (13+5 = 18 out of 28) of the variables were found to be important for at least 3 of the 4 models employed for 1-year analysis, while around 83 % (14+10=24 out of 29) and 46 % (7 +5= 12 out of 26) of the variables were commonly found to be important by at least 3 of these 4 models, for 5- and 9-year analysis, respectively.

Table 2.4: The agreement of four models on the important variables for each time point

	4-Models	3-Models	2-Models	1-Model	Total
1-Year	13	5	7	3	28
5-Year	14	10	4	1	29
9-Year	7	5	11	3	26

There are forty three (43) variables (in total) that were found to be important for the 1-, 5- and/or 9- year time horizons, as presented in detail in Figure 2.2. Recall that the main advantage of using data mining algorithms is to uncover the hidden, unexpected information in large/complex datasets. Therefore, the findings that are in conflict with the existing literature or have not been yet investigated by the existing studies should be considered for future prospective medical research because of a high potential for new discoveries. The comparison with the existing literature and our findings is limited in that the existing studies have not compared the predictors' level of significance and/or importance throughout different time periods. Rather, they investigated if a particular variable has any significance at a certain time period or not. Having said that, many of the variables selected by the data analytic models can be justified by recent results published in the literature [8, 10-13, 15]. Specifically, it is crucial to emphasize that the medical history of the donor, and recipient play a larger role than the demographic-related information (e.g. age, gender). This is similar to the results obtained by Dag et al. [13] for heart transplantations. For example, *FUNC STAT TRR* (recipient functional status at transplant) has been found to be the most important variable in both studies. One of the main contributions of this paper, however, is identifying how the importance (as measured by the sensitivity measure of Sec. 2.5) of each variable changes over varying time periods of analyses. This has not been examined by any data mining study for transplantation datasets. The changes in the relative importance of a variable can be easily determined from Figure 2.2 by examining the different symbols used for each variable, and their

corresponding value on the y-axis. Note that the absence of a symbol indicates that this specific variable has no importance (i.e. not selected by the IF model) for that specific time-point. The data was also sorted in ascending order with respect to their importance for the 1-, 5-, and 9-year analyses, respectively. This representation allows readers to visualize how the important variables are grouped and uncover interesting patterns (which we detail in the paragraphs below). It should be noted that, hereafter the discussion on these results are mostly about the ones that are found to be important for at least three of the four predictive models (within each time frame).This allows us to better focus on the variables that are identified as "important" by the consensus of models, whose classification analogies are different.

The 43 predictors can be divided into 7 groups (see Figure 2.2). Group 1 contains variables that have some importance for the 1-year prediction, but have no importance for the 5- and 9- year survival. The variables within this group are primarily related to either related with the donor (i.e. *DEATH_MECH_DON*, *GENDER_MATCH*), or the functional status of cardiovascular physiology or the priority/urgency needed for a transplant (*LIFE_SUP_TCR*, *DAYS_STAT1*). Within this group, *GENDER_MATCH* is associated with increased risk perhaps due to effects of both under-sizing (female donor/male recipient) and immune mechanisms (male donor/female recipient) [54]. Singhal *et al.* [55] has investigated the effect of three most common donor mechanism (cause) of death (*DEATH_MECH_DON*) on survival after heart, kidney, liver and lung transplants. Based on the results obtained, it has been found that donors who die from stroke-related reasons have the greatest number of comorbid conditions. Recipient age (*AGE*) have also been studied and found to be associated for 1-month, 1-year and 5-year in different studies [7, 9].

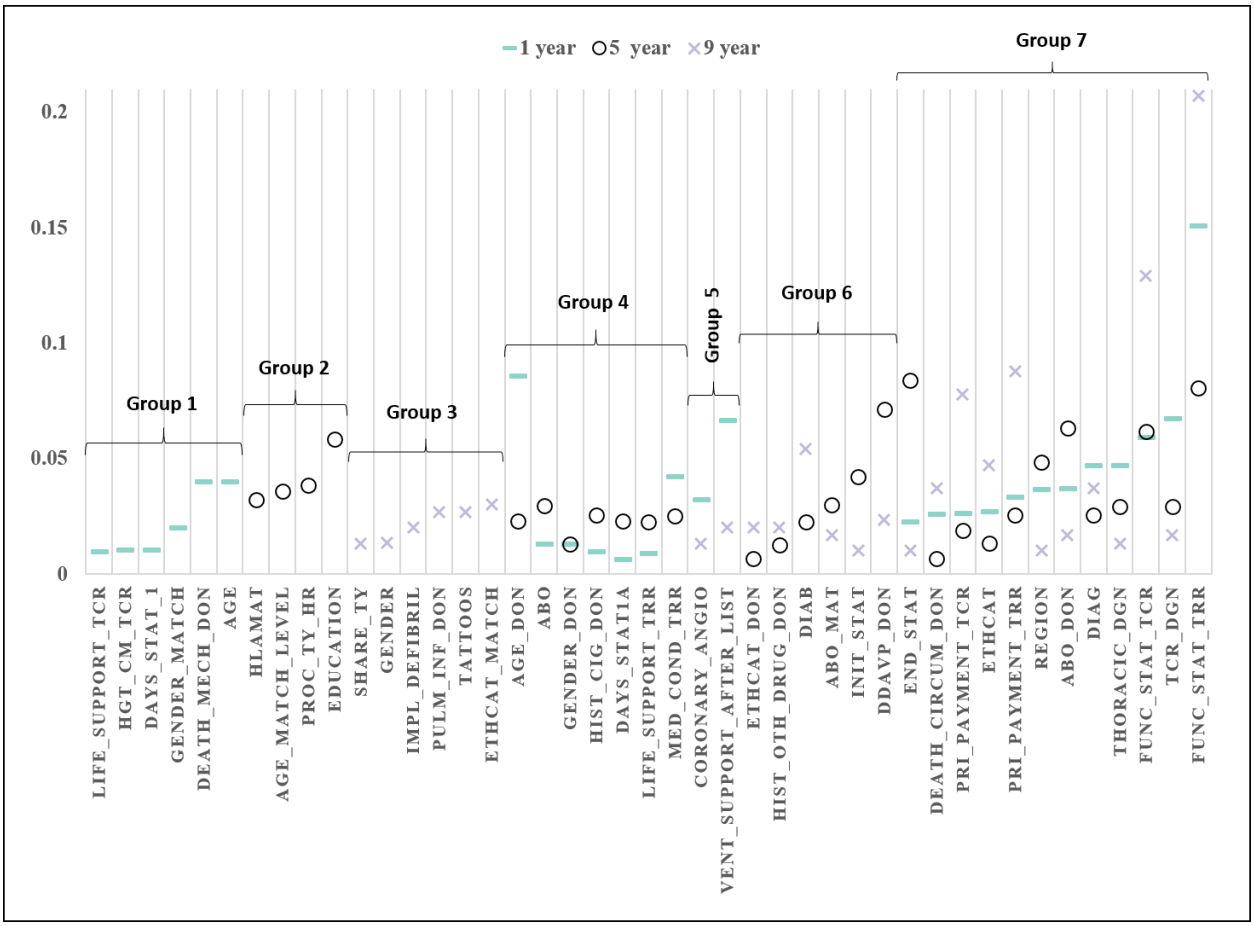

Figure 2.2. The importance of the variables through the three time-periods

Group 2 contains four variables that have some importance for only midterm survival after heart transplant. Regarding with *AGE_MATCH_LEVEL*, It has been observed that a greater negative donor–recipient age difference conferred improved survival [56]. A study, which compared the effect of procedure types of heart transplant, of a meta-analysis on 41 studies that compares bicaval to biatrial anastomoses identified significant benefits for the bicaval technique in terms of early atrial pressure, tricuspid valve regurgitation, return to sinus rhythm, frequency of permanent pacemaker implantation, and perioperative survival [57]. With being slightly different from the existing literature, our study showed that *PROC_TY_HR* (procedure type for heart) has an influence on 5 year survival only but not on short-, and long-term survival after transplant. On

the other hand, in two different independent studies, performed by Shapire et al. [58] and Geller et al. [59], educational level (*EDUCATION*) of the recipients has been found to be associated with the coronary artery disease of the transplanted heart and noncompliance.

Group 3 contains 6 variables, which have importance only for the 9-year prediction, but have no effect on the prediction for the 1- and 5- year time-points. There are 3 out of 6 variables in this group that were identified to be important by at least three out of four prediction models employed in 9-year analysis. These variables have either direct or indirect association with the deceased donor. To exemplify, the association of the existence of infection on deceased donor's body with survival, has been studied by Montoya *et al.* [60]. In their analysis, infectious complications was found to be a major cause of morbidity and mortality for the most common cause of late deaths as well as some of the early deaths. In our analysis, *PULM_INF_DON* (pulmonary related infection on donor's body) was found to be associated with the late death after transplantation. Similarly, the impact of whether the deceased donor had tattoos (*TATTOO*) his/her before death have been discussed by several scientific articles. The consensus of these studies is that tattooed (deceased) donors carry potential risk for skin related (tissue cells) infectious diseases [61-63]. Another important variables in this group; *ETHCAT_MATCH* was created and added to the original UNOS dataset by our research team. It refers to the matching level of recipient and the donor (0 if the donor and recipient belong same race, 1 if not). In this sense, it has direct relation with both recipient- and donor-race.

Regarding the impact of ethnical category of the patients, a recent prospective study by Kilic et al. [23] identified the predictive factors for 10-year survival after orthotropic heart transplant on 22,385 adult patients. The study showed that white race have improved the likelihood

of 10-year survival after OHT. Another study [64] that have investigated the impact of race to post kidney and heart transplant, also found indirect effect of race on survival after transplant.

Group 4 represents the variables that only contribute to short and medium-term predictions of transplantation outcomes. One could see that there is some similarities between these variables and those included in Group 1. Based on our results, LIFE SUP TRR is important in predicting both 1- and 5-year heart transplant survival outcomes. This is consistent with previous studies, which showed that the use of mechanical assist devices do contribute to 1 and 5 year survival prediction [22]. According to [54], increasing donor age (AGE_DON) is associated with progressively worse survival, particularly with donors aged 60 years and older. The effect of donor age is most pronounced in the short-term (1-year post-transplant survival). Medical condition of the recipient before the transplant (MED_COND_TRR) is also one of the variables that were found to be important for survival after thoracic transplants [65]. In our study it has found to be important for both short- and mid-term survival. *DAYS_STAT1*, *DAYS_STAT1A, DAYS_STAT2* are continuous variables that refer to the number of days that the recipient have spent on a certain level of need (urgency) for heart transplant. Therefore, these variables are related with the positive or (usually) negative progress for the urgency of transplant. In this sense, it would be unsurprising to find such variables important, where there are other types of medical condition-related (e.g. LIFE_SUPPORT_TRR, MED_COND_TRR) variables were included in this group.

Group 5 contains 2 variables that are important for 1- and 9-year outcomes, but not for the 5-year outcome. It should be further investigated why predictors may only be important for short- and long-term survival prediction, but not for the time-period in the middle. We believe that the predictors in this group should be investigated further since this can be an indication of: a) over-fitting of our model, and/or b) the existence of multicollinearity in our predictors. Having said that

recipient on ventilator at time of transplantation (VENT_SUPPORT_AFTER_LIST) conferred increased risk for 1, 5, and 15 years (short-, mid- and long-term) based on a related study that was conducted by *Stehlik et al.* [66].

Variables that are only important to medium and long-term outcomes are contained in Group 6. These include; whether the recipient is diabetic or not (DIAB), whether the deceased donor body received synthetic anti-diuretic hormone prior to transplant or not (DDAVP_DON), donor to recipient blood-type match level (ABO_MAT) and the health status of the recipient during the waiting list (INIT_STAT). 3 out of these variables were found to be important for long term survival after heart transplant, based on the analysis conducted by Dag *et al*. It should be noted that their study only identified the variables that have long-term effect post heart transplant. In a similar study that was conducted by Klingenberg *et al. [67]*, DIAB was found to be important for both short- and long-term survival.

Finally, the predictors in Group 7 are important for all of the three follow-up time points. The reader should note that the majority of the variables in Group 11 are highly important (i.e. have the highest fused sensitivity scores) among our predictors. Unsurprisingly, these variables are well studied in the literature and are typically observed in data mining studies involving heart transplantation patients (see e.g. [13, 15]). For example, as it was confirmed by current study as well, recipients who are capable of normal activity by having a high quality functional status (FUNC_STAT_TRR and/or FUNC_STAT_TCR) are more likely to survive longer when compared to others [56]. As can be seen from Figure 2.2, the variables related with the diagnosis for heart transplant at the candidacy stage (i.e. DIAG, TCR_DGN and THORACIC_DGN) play an important role in predicting the survival time after heart transplantations. There are many

reasons for a candidate to be put in the waiting lists such as *refractory heart failure* requiring *continuous inotropic support, cardiogenic shock* requiring mechanical assistance *(e.g., ventilator, intra-aortic balloon pump, and ventricular assist device), Congestive heart failure* with objective evidence of impaired functional capacity, *congenital heart disease* with progressive *ventricular failure and* cardiac tumors confined to the *myocardium.* Such indications have been extensively studied in several published studies [68-77]. It can be summarized that different types of diagnosis (i.e. *refractory heart failure, congestive heart failure, congenital heart disease etc.*) have different association with survival outcomes of heart transplants. The primary payment sources of the patient (PRI_PAYMENT_TCR and PRI_PAYMENT_TRR) was found to be an interestingly prominent factor that has an association with the survival. It should be noted that such variables might have a direct relation with the patient's financial standing, education level, social class etc. In the related literature, there are many researches that have investigated the effect of such factors not only on survivability after heart related problems but also on other medical problems [58, 59, 78-81]. Among these, to specifically exemplify the studies that are related with heart transplantation, Allen et al. [81] had investigated the effect of the payment type and education level of the recipient on long term survival after orthotopic hear transplantation (OHT) in the US. The results obtained showed that both payment (and/or insurance) type and education level of the recipients are associated with the long term survival after OHT. Similarly, Shapiro et al. [58] and, Geller and Connolly [59] conducted studies that investigated the effect of psychosocial factors on survival outcomes after heart transplant. These factors includes variables such as education, social support, living arrangement, etc. It has been found out that psychosocial factors can identify patients with increased risk of postoperative morbidity. In another related study, Gerber et al. [78] investigated the effects of neighborhood income and individual education level on survival after

Myocardial Infarction. Their study showed that both poor neighborhood income and low individual education had association with a worse clinical presentation. In addition, poor neighborhood income was a powerful predictor of mortality even after controlling for a variety of confounding factors. Another predictor that fell into group 7 is REGION. The intuitive rationale behind this could be the difference between the quality and success level of the transplant centers that are located in different regions of the U.S.

A numeric comparison is provided with respect to the agreement of the three IF models on the number of important predictors in Table 2.5. The information is based on data from Figure 2.2. Table 2.5 has only 3 cells since there are only three pairwise comparisons (and therefore, the 5 year IF model versus itself is noted with a not applicable, N/A).

Table 2.5: A numeric comparison of the number of important vs. unimportant variables for each of the three IF models

		5-YEAR		**9-YEAR**	
		Important	**Not Important**	**Important**	**Not Important**
1-YEAR	**Important**	19	9	14	14
	Not Important	10	6	12	4
5-YEAR	**Important**	N/A		18	11
	Not Important			8	7

2.5 Conclusion and Future Recommendations

The main objectives of this paper were to develop an assumption-free data-driven approach to predict the 1- (short-term), 5- (medium-term), and 9-years (long-term) survival for heart transplantation patients, and understand how the effect of important predictor variables change with these three time-points. To achieve these objectives, a five-step data analytic methodology (framework) have been proposed. The proposed approach was used to investigate a large, feature-rich dataset obtained from UNOS, containing all recorded information on heart transplant operations that were performed in the U.S. between October 1, 1987 and December 31, 2012. In this current analysis of the UNOS dataset, the following research questions regarding heart-transplantation have been addressed:

1) Can we develop an individualized patient model, with high AUC performance, for predicting heart transplantation outcomes?
2) What predictive factors contribute to the outcome of a heart transplant for a given follow-up time-point?
3) How does the contribution/importance of each of these factors change over time (based on 1-, 5-, and 9-year time-points for analysis)?
4) Is it possible to group these variables whose effect change over time? If so, how can this be done to potentially provide insights to medical practitioners and/or different stakeholders?

It is important to note that Questions 2-4 have not been addressed previously in the heart transplantation literature. The following medically-relevant results were obtained: A) our proposed approach can predict the 1-, 5-, and 9- year survival outcomes with a mean AUC score of 0.624, 0.676, and 0.838, respectively. Thus, our methodology can potentially assist medical-decision makers in evaluating the suitability of a donor heart for a given patient; and B) we have identified 43 pre-operative variables that contribute to outcome prediction for at least one of the three time-points. We have also grouped these predictors into 7 groups, which reflect how the importance of these predictors change over time.

Based on this work, we believe that are some more general takeaways that can be shared with the business and data analytics communities. First, the importance of \asking the right questions". Most of the previous studies, where data mining methods were applied to transplantation problems, addressed the following question: “What is the best model to predict the outcome at x years post-transplant?" In our opinion, the innovation in these methods is based on using somewhat sophisticated modeling approaches. Second, we believe that any data analytics study/application should address a set of sequential questions. After all, the goal is to uncover hidden patterns and generate insights that are interesting and transforming. As an illustration, in the context of transplantation research, after determining the best model for x years, there are several opportunities for asking some additional questions:

With respect to the best model, are there other models that present somewhat similar results, but can be more advantageous in practice? For example, these models may require less predictors, and/or can be easier to implement. We refer the reader to [13] for an example in the context of heart transplants.

What happens to the model if we change the number of years? This question has two dimensions: a) what happens in terms of the predictive performance? And b) Do the selected /significant predictor variables remain unchanged?

How can we communicate these results (or alternatively, how do uncover patterns from our results)? For example, in this paper, we resorted to a simple graph (i.e., Figure 2.2) to communicate the change in the importance of the variables over time.

If the analysis is implemented in practice, does the model perform as expected. If not, is due to a technology change (e.g., a major breakthrough in transplantation research) or does it reflect some potential issues with the model?

The reader should note that these type of questions are inherent in the data analytic process. For example, the CRISP-DM (CRoss Industry Standard Process for Data Mining) framework emphasizes the sequential nature of data mining applications [82]. However, in our estimation, the literature does not necessarily reflect that.

In summary, this paper demonstrates how data analytic approaches can be used to generate new knowledge (in the context of heart transplantation research). Our approach can be used for applications where the classification problem is temporal (i.e., we want to predict a binary outcome over time). Obvious examples of this include other organ transplantation problems especially 46 since UNOS provides information on other organ transplants. Other examples can include: reliability/maintenance applications (predicting whether a system will fail at different time-points), customer relationship management (i.e. attrition of customers over time), and other health-analytics applications (e.g., detecting cancer in individuals). Finally, it should be noted that the analysis presented in this paper may inform new prospective studies that can test hypotheses based

on the groupings of the variables from the hybrid model.

Acknowledgements

We would like to thank the Samuel Ginn College of Engineering at Auburn University for partially supporting this work. This work was supported in part by Health Resources and Services Administration contract 234-2005-370011C. The content is the responsibility of the authors alone and does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

2.6 References

1. *What is Heart Failure ?* 2012 [cited 2012; Available from: http://www.nhlbi.nih.gov/health/health-topics/topics/hf/.
2. López-Sendón, J., *The heart failure epidemic.* Medicographia, 2011. **33**(4): p. 363-9.
3. *What Is Heart Transplant?* . 2012 [cited 2012; Available from: http://www.nhlbi.nih.gov/health/health-topics/topics/ht/.
4. Carmona, M., et al., *Heart failure in the family practice: a study of the prevalence and co-morbidity.* Family Practice, 2011. **28**(2): p. 128-133.
5. Kong, G., et al., *A belief rule-based decision support system for clinical risk assessment of cardiac chest pain.* European Journal of Operational Research, 2012. **219**(3): p. 564-573.
6. *What to Expect Before a Heart Transplant*. 2012 [cited 2012; Available from: http://www.nhlbi.nih.gov/health/health-topics/topics/ht/before.html.
7. Del Rizzo, D.F., et al., *The role of donor age and ischemic time on survival following orthotopic heart transplantation.* The Journal of Heart and Lung transplantation, 1999. **18**(4): p. 310-319.
8. Drakos, S.G., et al., *Multivariate predictors of heart transplantation outcomes in the era of chronic mechanical circulatory support.* The Annals of Thoracic Surgery, 2007. **83**(1): p. 62-67.
9. Gupta, D., et al., *Effect of older donor age on risk for mortality after heart transplantation.* The Annals of Thoracic Surgery, 2004. **78**(3): p. 890-899.
10. Hong, K.N., et al., *Who is the high-risk recipient? Predicting mortality after heart transplant using pretransplant donor and recipient risk factors.* The Annals of Thoracic Surgery, 2011. **92**(2): p. 520-527.
11. Kilic, A., et al., *What predicts long-term survival after heart transplantation? An analysis of 9,400 ten-year survivors.* The Annals of Thoracic Surgery, 2012. **93**(3): p. 699-704.
12. Kilic, A., et al., *Factors associated with 5-year survival in older heart transplant recipients.* The Journal of thoracic and cardiovascular surgery, 2012. **143**(2): p. 468-474.
13. Dag, A., et al., *A probabilistic data-driven framework for scoring the preoperative recipient-donor heart transplant survival.* Decision Support Systems, 2016. **86**: p. 1-12.

14. Lin, R.S., et al., *Single and multiple time-point prediction models in kidney transplant outcomes.* Journal of Biomedical Informatics, 2008. **41**(6): p. 944-952.
15. Oztekin, A., D. Delen, and Z.J. Kong, *Predicting the graft survival for heart–lung transplantation patients: An integrated data mining methodology.* International Journal of Medical Informatics, 2009. **78**(12): p. e84-e96.
16. Oztekin, A., Z.J. Kong, and D. Delen, *Development of a structural equation modeling-based decision tree methodology for the analysis of lung transplantations.* Decision Support Systems, 2011. **51**(1): p. 155-166.
17. Sheppard, D., et al., *Predicting cytomegalovirus disease after renal transplantation: an artificial neural network approach.* International Journal of Medical Informatics, 1999. **54**(1): p. 55-76.
18. Aydemir, Ü., S. Aydemir, and P. Dirschedl, *Analysis of time-dependent covariates in failure time data.* Statistics in Medicine, 1999. **18**(16): p. 2123-2134.
19. Boschiero, L., et al., *An objective method for detecting time-dependent effects in graft survival.* Transplant International, 2000. **13**(S1): p. S112-S116.
20. Zuckermann, A.O., et al., *Pre-and early postoperative risk factors for death after cardiac transplantation: A single center analysis.* Transplant International, 2000. **13**(1): p. 28-34.
21. Kusiak, A., B. Dixon, and S. Shah, *Predicting survival time for kidney dialysis patients: a data mining approach.* Computers in Biology and Medicine, 2005. **35**(4): p. 311-327.
22. Nakayama, N., et al., *Algorithm to determine the outcome of patients with acute liver failure: a data-mining analysis using decision trees.* Journal of Gastroenterology, 2012. **47**(6): p. 664-677.
23. Brieke, A., et al., *Influence of donor cocaine use on outcome after cardiac transplantation: analysis of the United Network for Organ Sharing Thoracic Registry.* The Journal of Heart and Lung Transplantation, 2008. **27**(12): p. 1350-1352.
24. Stehlik, J., et al., *Interactions among donor characteristics influence post-transplant survival: a multi-institutional analysis.* The Journal of Heart and Lung Transplantation, 2010. **29**(3): p. 291-298.
25. Gasink, L.B., et al., *Hepatitis C virus seropositivity in organ donors and survival in heart transplant recipients.* Jama, 2006. **296**(15): p. 1843-1850.
26. *UNOS | About Us*. 2014 2014/6/2]; Available from: http://www.unos.org/contact/index.php.

27. Chawla, N.V., *Data mining for imbalanced datasets: An overview*, in *Data mining and knowledge discovery handbook*. 2005, Springer. p. 853-867.
28. Guo, X., et al. *On the class imbalance problem*. in *Natural Computation, 2008. ICNC'08. Fourth International Conference on*. 2008. IEEE.
29. He, H. and E.A. Garcia, *Learning from imbalanced data.* Knowledge and Data Engineering, IEEE Transactions on, 2009. **21**(9): p. 1263-1284.
30. Kotsiantis, S., D. Kanellopoulos, and P. Pintelas, *Handling imbalanced datasets: A review.* GESTS International Transactions on Computer Science and Engineering, 2006. **30**(1): p. 25-36.
31. Pazzani, M., et al. *Reducing misclassification costs*. in *Proceedings of the Eleventh International Conference on Machine Learning*. 1994.
32. Batista, G.E., R.C. Prati, and M.C. Monard, *A study of the behavior of several methods for balancing machine learning training data.* ACM Sigkdd Explorations Newsletter, 2004. **6**(1): p. 20-29.
33. Drummond, C. and R.C. Holte. *C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling*. in *Workshop on learning from imbalanced datasets II*. 2003. Citeseer.
34. Das, B., N.C. Krishnan, and D.J. Cook, *RACOG and wRACOG: Two Probabilistic Oversampling Techniques.* Knowledge and Data Engineering, IEEE Transactions on, 2015. **27**(1): p. 222-234.
35. Kaplan, B. and J. Schold, *Transplantation: neural networks for predicting graft survival.* Nature Reviews Nephrology, 2009. **5**(4): p. 190-192.
36. Gunn, S.R., *Support vector machines for classification and regression.* ISIS Technical Report, 1998. **14**.
37. Hodge, V.J. and J. Austin, *A survey of outlier detection methodologies.* Artificial Intelligence Review, 2004. **22**(2): p. 85-126.
38. Han, J., M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2011: Elsevier.
39. Sordo, M., *Introduction to neural networks in healthcare.* Open Clinical Document, 2002.
40. Olshen, L.B.J.H.F.R.A. and C.J. Stone, *Classification and regression trees.* Wadsworth International Group, 1984.
41. Quinlan, J.R., *Induction of decision trees.* Machine Learning, 1986. **1**(1): p. 81-106.

42. Quinlan, J.R., *C4. 5: programs for machine learning*. Vol. 1. 1993: Morgan kaufmann.
43. Hosmer Jr, D.W., S. Lemeshow, and R.X. Sturdivant, *Logistic Regression for Matched Case-Control Studies.* Applied Logistic Regression, Third Edition, 2013: p. 243-268.
44. Davis, G.W., *Sensitivity Analysis in Neural Net Solutions.* Ieee Transactions on Systems Man and Cybernetics, 1989. **19**(5): p. 1078-1082.
45. Molaie, M., et al., *Artificial neural networks: powerful tools for modeling chaotic behavior in the nervous system.* Front Comput Neurosci, 2014. **8**: p. 40.
46. Principe, J.C., N.R. Euliano, and W.C. Lefebvre, *Innovating adaptive and neural systems instruction with interactive electronic books.* Proceedings of the Ieee, 2000. **88**(1): p. 81-95.
47. Saltelli, A., *Making best use of model evaluations to compute sensitivity indices.* Computer Physics Communications, 2002. **145**(2): p. 280-297.
48. Saltelli, A., et al., *Sensitivity analysis in practice: a guide to assessing scientific models*. 2004: John Wiley & Sons.
49. Cang, S. and H.N. Yu, *A combination selection algorithm on forecasting.* European Journal of Operational Research, 2014. **234**(1): p. 127-139.
50. Clemen, R.T., *Combining forecasts: A review and annotated bibliography.* International Journal of Forecasting, 1989. **5**(4): p. 559-583.
51. Graefe, A., et al., *Combining forecasts: An application to elections.* International Journal of Forecasting, 2014. **30**(1): p. 43-54.
52. Ruiz, E. and F.H. Nieto, *A note on linear combination of predictors.* Statistics & Probability Letters, 2000. **47**(4): p. 351-356.
53. Delen, D., R. Sharda, and P. Kumar, *Movie forecast Guru: A Web-based DSS for Hollywood managers.* Decision Support Systems, 2007. **43**(4): p. 1151-1170.
54. Lund, L.H., et al., *The Registry of the International Society for Heart and Lung Transplantation: thirtieth official adult heart transplant report—2013; focus theme: age.* J Heart Lung Transplant, 2013. **32**(10): p. 951-64.
55. Singhal, A., et al. *Impact of donor cause of death on transplant outcomes: UNOS registry analysis*. in *Transplantation proceedings*. 2009. Elsevier.

56. Lund, L.H., et al., *The Registry of the International Society for Heart and Lung Transplantation: thirtieth official adult heart transplant report--2013; focus theme: age.* J Heart Lung Transplant, 2013. **32**(10): p. 951-64.
57. Jacob, S. and F. Sellke, *Is bicaval orthotopic heart transplantation superior to the biatrial technique?* Interactive cardiovascular and thoracic surgery, 2009. **9**(2): p. 333-342.
58. Shapiro, P.A., et al., *Psychosocial evaluation and prediction of compliance problems and morbidity after heart transplantation* Transplantation, 1995. **60**(12): p. 1462&hyhen.
59. Geller, S. and T. Connolly, *The influence of psychosocial factors on heart transplantation decisions and outcomes.* Journal of Transplant Coordination, 1997. **7**(4): p. 173-179.
60. Montoya, J.G., et al., *Infectious complications among 620 consecutive heart transplant patients at Stanford University Medical Center.* Clinical infectious diseases, 2001. **33**(5): p. 629-640.
61. Beele, H., et al., *Physical examination of the potential tissue donor, what does literature tell us?* Cell and tissue banking, 2009. **10**(3): p. 253-257.
62. Scardino, M.K., et al., *The postmortem sociomedical interview: uncertainty in confirming infectious disease risks of young tattooed donors.* Cornea, 2002. **21**(8): p. 798-802.
63. van Wijk, M.J., et al., *Results of the clinical donor case and quality system case workshops of the European Association of Tissue Banks annual meeting 2009.* Cell and tissue banking, 2012. **13**(1): p. 191-202.
64. Hesselink, D.A., et al., *Population pharmacokinetics of cyclosporine in kidney and heart transplant recipients and the influence of ethnicity and genetic polymorphisms in the MDR-1, CYP3A4, and CYP3A5 genes.* Clinical Pharmacology & Therapeutics, 2004. **76**(6): p. 545-556.
65. Delen, D., A. Oztekin, and Z.J. Kong, *A machine learning-based approach to prognostic analysis of thoracic transplantations.* Artificial Intelligence in Medicine, 2010. **49**(1): p. 33-42.
66. Stehlik, J., et al., *The Registry of the International Society for Heart and Lung Transplantation: 29th official adult heart transplant report—2012.* The Journal of Heart and Lung Transplantation, 2012. **31**(10): p. 1052-1064.

67. Klingenberg, R., et al., *Impact of pre-operative diabetes mellitus upon early and late survival after heart transplantation: a possible era effect.* The Journal of heart and lung transplantation, 2005. **24**(9): p. 1239-1246.
68. Blume, E.D., et al., *Outcomes of Children Bridged to Heart Transplantation With Ventricular Assist Devices A Multi-Institutional Study.* Circulation, 2006. **113**(19): p. 2313-2319.
69. Bourge, R.C., et al., *Pretransplantation risk factors for death after heart transplantation: a multiinstitutional study.* Journal of Heart and Lung Transplantation, 1993. **12**(4): p. 549-562.
70. Costard-Jäckle, A. and M.B. Fowler, *Influence of preoperative pulmonary artery pressure on mortality after heart transplantation: testing of potential reversibility of pulmonary hypertension with nitroprusside is useful in defining a high risk group.* Journal of the American College of Cardiology, 1992. **19**(1): p. 48-54.
71. Hsu, D.T., et al., *Heart transplantation in children with congenital heart disease.* Journal of the American College of Cardiology, 1995. **26**(3): p. 743-749.
72. Lamour, J.M., et al., *Outcome after orthotopic cardiac transplantation in adults with congenital heart disease.* Circulation, 1999. **100**(suppl 2): p. II-200-Ii-205.
73. Lamour, J.M., et al., *The effect of age, diagnosis, and previous surgery in children and adults undergoing heart transplantation for congenital heart disease.* Journal of the American College of Cardiology, 2009. **54**(2): p. 160-165.
74. Maron, M.S., et al., *Survival after cardiac transplantation in patients with hypertrophic cardiomyopathy.* Circulation: Heart Failure, 2010. **3**(5): p. 574-579.
75. Murali, S., et al., *Preoperative pulmonary hemodynamics and early mortality after orthotopic cardiac transplantation: the Pittsburgh experience.* American heart journal, 1993. **126**(4): p. 896-904.
76. Reitz, B.A., et al., *Heart-lung transplantation: successful therapy for patients with pulmonary vascular disease.* New England Journal of Medicine, 1982. **306**(10): p. 557-564.
77. Gowdamarajan, A. and R.E. Michler, *Therapy for primary cardiac tumors: is there a role for heart transplantation?* Current opinion in cardiology, 2000. **15**(2): p. 121-125.
78. Gerber, Y., et al. *Neighborhood income and individual education: effect on survival after myocardial infarction*. in *Mayo Clinic Proceedings*. 2008. Elsevier.

79. Hussain, S., et al., *Influence of education level on cancer survival in Sweden.* Annals of oncology, 2008. **19**(1): p. 156-162.
80. Rosso, S., et al., *Social class and cancer survival in Turin, Italy.* Journal of epidemiology and community health, 1997. **51**(1): p. 30-34.
81. Allen, J.G., et al., *Insurance and education predict long-term survival after orthotopic heart transplantation in the United States.* The Journal of Heart and Lung Transplantation, 2012. **31**(1): p. 52-60.
82. Wirth, R. and J. Hipp. *CRISP-DM: Towards a standard process model for data mining*. in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. 2000. Citeseer.

3 A Preoperative Recipient-Donor Heart Transplant Survival Score

3.1 Abstract

Recent research has shown that data mining models can accurately predict the outcome of a heart transplant based on predictors that include patient and donor's health/demographics. These models have not been adopted in practice, however, since they did not: a) consider the interactions between the explanatory variables; b) provide a patient's specific risk of survival (reported results have been primarily deterministic); and c) offer an automated decision tool that can provide some data-driven insights to practitioners. In this study, we attempt to overcome these three limitations through the use of Bayesian Belief Networks (BBN). The proposed BBN framework is comprised of four phases. In the first two phases, the data is preprocessed, and a candidate set of predictors is generated based on employing several variable selection methods. The third phase involves the addition of medically relevant variables to the list. In phase four, the BBN model is applied. The results show that the proposed BBN method provides similar predictive performance to the best approaches in the literature. More importantly, our method provides novel information on the interactions among the predictors and the conditional probability of survival for a given set of relevant donor-recipient characteristics. We offer U.S. practitioners a decision support tool that presents an individualized survival score based on our BBN model (and the UNOS dataset).

3.2 Introduction

Heart failure is a serious medical condition, where a patient's heart is weakened and cannot pump enough blood to meet the body's demands [1]. This condition affects an estimated 2-3 % of the world's adult population [2]. In the U.S., there are over 5.8 million patients living with heart failure, with an estimated annual incidence rate of 550,000 [1, 3]. The majority of these patients can enjoy a full life by managing the condition with medication. However, a certain class of heart failure (end-stage heart failure) cannot be managed with these interventions and can only be overcome by a heart transplant. If a patient is deemed eligible for a transplant, then she/he is placed on a waiting list until a suitable donor heart is found [4]. Currently, in the U.S. there are about 3,000 people on waiting lists for a heart transplant at any one time, while there are only about 2,000 donor hearts available each year [4]. This gap between supply and demand of donated healthy hearts leads to longer waiting times and thus leaves many to die while waiting for a transplant [5].

The current matching process is determined based on a printed out list from the United Network for Organ Sharing (UNOS) computers, which is based on "blood type, body size, UNOS status, and length of time on the waiting list" [6]. There has been a significant amount of research being done to determine the subset of variables that should be included for matching. Much of this work involve data mining techniques since they do not require prior knowledge about the data, nor do they make assumptions about the statistical distribution or properties of the data [7]. In particular, data mining methods have shown great accuracy in determining which subset of variables influence a patient's survival over a pre-specified time period [8-11].

There is extensive research on using data-driven models to predict post transplantation survival time. For any type of transplant, we can classify these models into two streams. The first stream addresses the question of how to accurately predict post transplantation survival for a given time period (i.e. will the patient survive for X amount of years?). In our analysis of the literature, this represents the majority of the work. This question has been addressed for virtually all organ transplants; for example, see the following papers in heart [12-21], kidney [9, 10, 22], and liver [23]. It is important to note that these models are deterministic, i.e., they provide an expected value that is typically a binary survival (after X-years) post-transplantation outcome. On the other hand, the second stream attempts to understand the uncertainty in the prediction as well as identify the conditional dependencies among the predictive factors. Bayesian Belief Networks (BBN) have been used as a framework for reasoning under uncertainty for kidney [24] and liver [25] transplants. Note that none of the papers in the second stream have examined outcomes post a heart transplant. Perhaps more importantly, they have not used the Bayesian framework to provide an individualized survival probability based on the pre-operative variables.

In this paper, we focus on predicting the long-term survival after a heart transplant. The specific research questions that are addressed in this paper are: a) are there interaction effects among the explanatory variables and if so, can they be understood; and b) how can an individualized risk score be developed so that it can reflect the long-term survival probability based on the characteristics of both the recipient and the donor. These questions have not been examined in the heart transplantation literature, yet they provide an important depiction of the mechanics behind heart transplantation's success/failure. We apply BBN to address these two question. The use of BBN requires a discrete dependent variable [26]. We chose to focus on long-term survival since: a) patients would ideally want to live as long as possible; b) the long-term success rate is

close to 50% [3]; and c) the ability to predict long-term survival can provide a platform to explore ethical and socio-economic questions regarding the matching process (an important consideration even though it is not a focus of this paper). We use 9-years as the cut-off for long-term survival so we can compare our results to existing methods in the literature [12, 21] that used the same dataset (but only focused on whether a patient will survive or not).

The remainder of this paper is organized as follows. In Section 3.3, we describe the data analytics methodology, from data gathering to processing and modeling/validation. The results are presented and discussed in Section 3.4. In Section 3.6, we summarize our conclusions and offer some thoughts on future research directions. To facilitate the translation of our model to practice, we provide the description/link for our desktop app that allows practitioners to determine the pre-operative risk score in the Appendix I.

3.3 Proposed Method

In this paper, we propose a hybrid data analytics approach that consists of four main phases, as shown in Figure 3.1. We start with a data preparation phase, where the data is extracted from the UNOS dataset, and pre-processed according to the approach in Dag *et al.* [12]. In Phase 2, we examine several variable selection approaches to generate candidate sets of predictors for the BBN application. Specifically, in Phase 2a, data mining methods (i.e. C&RT and ANNs) are used to identify a set of important variables through three sequential steps. The first step uses 10-fold cross validation to prepare 10 training samples to ensure that the analysis is robust and to minimize the bias that can be associated with data splitting (training vs. testing). In Step 2, the two data mining models on the data samples from Step 1 are used. The third, and final step in Phase 2a, is where information fusion procedure is used to obtain a first set of predictors based on the sets of

important variables obtained from Steps 1-2 via sensitivity analysis (SA), by which the contribution level of each predictor from the C&RT and ANN models is obtained. The second set of predictors is obtained by using Genetic Algorithms (GA) (Phase 2b), while the third set of predictors is obtained by employing Ridge Regression (RR). In phase 3, a fixed set of predictors is obtained by using Cox regression and reviewing the relevant literature. We denote the candidate sets by LR, Cox, DM, GA, and RR based on the outputs from Phases 2a-c and Phase 3, in Figure 3.1. In Phase 4, combinations of these five predictor sets are constructed to present different scenarios that will be examined by the BBN model. Note that we have a total of seven scenarios that represent all combinations of these five sets when LR and Cox are always included. We evaluate the predictive performance of these scenarios by using the BBN framework (i.e. we apply BBN seven times). Using the best scenario and the associated BBN, we can model the interactions between the extracted predictors and provide a patient-specific risk of survival.

3.3.1 Data Acquisition and Preparation

The dataset used in this study has been provided by the United Network for Organ Sharing, UNOS, which is a "private, non-profit organization that manages the [United States]'s organ transplant system under contract with the federal government" [27]. Our heart transplantation data contains information on all waiting list registrations and heart transplants that have been listed or performed in the U.S. and reported starting from October 1, 1987 until December 31, 2012. The variables in UNOS heart transplantation datasets include clinical and demographic factors related to donors, recipients and the transplant procedure. The variables can be classified into three major groups: a) preoperative factors that include donor/recipient demographics (age, gender, race, etc.), donor/recipient medical history, funding sources, and other factors that are considered prior to an operation; b) intra-operative factors, which describe several medical conditions during the

transplant such as: if there is a *chronic steroid use* at transplant, or if the recipient is on life support at the time of transplant; and c) post-operative factors that include whether the patient died or lived, length of hospitalization after transplant, and information on any other complications. It should be noted that the data has been de-identified by UNOS prior to being received by the research team.

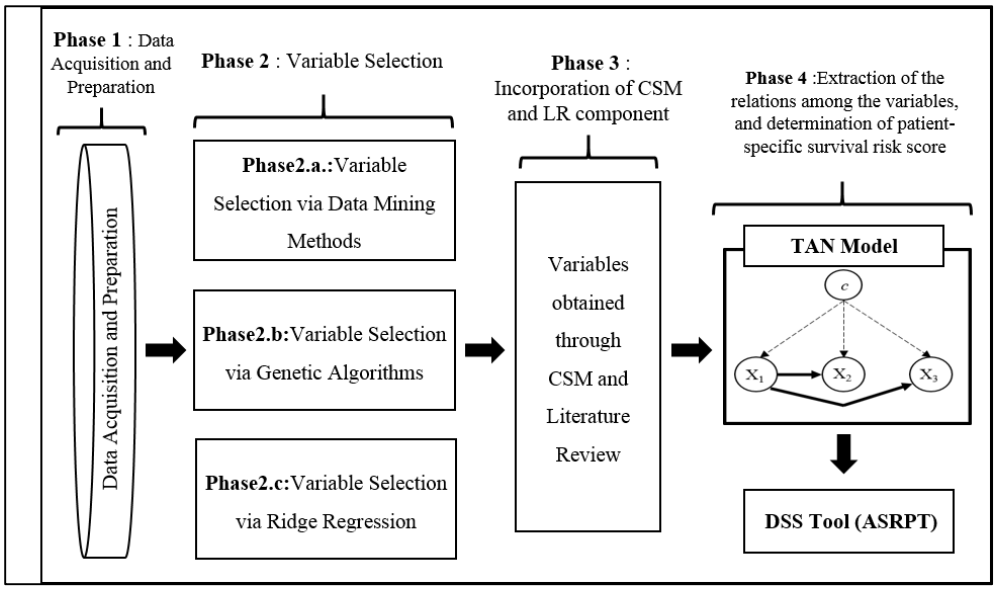

Figure 3.1: An overview of the proposed methodology

Since one of our goals is to predict survival, the dataset contains four candidate dependent variables representing the 9-year survival outcome. These variables are *pstatus* (binary variable, denotes whether the person is dead or alive at the last follow-up time), *gstatus* (a similar binary variable, denoting if the graft has failed or succeeded at the last follow-up time), *ptime* (referring the time frame from the day of transplant to the recipient's death/last follow-up time, in days) and *gtime* (a similar continuous variable referring to time frame from transplant to graft's failure/last follow-up time). The BBN model requires a binary dependent variable, and therefore, for this study, the list of response variables is limited to either *pstatus* or *gstatus*. We use *gstatus* since it allows us to focus on the patients who died solely based on transplant-related reasons versus those who may have died due to other reasons (e.g. traffic accidents or cancer).

The data preparation consists of two steps: data cleaning and data inclusion. For the data cleaning step, we first filtered the data using the variable TX-TYPE (type of transplant) to eliminate the patients who underwent other transplant types. Then we eliminated five sets of variables/records similar to Dag *et al* [12]. The first set of variables included all the intra- and post-operative factors, with the exception of the two outcomes *gstatus* and *gtime*, since they do not contribute to the decision-making process prior to a transplant. The second group included erroneous values and duplicated records, which were detected using outlier detections methods in *STATISTICA* 11 (www.statsoft.com). Variables that do not contribute to the prediction capability of the models such as identification type variables (e.g. patient ID number) and invariant variables are the third and fourth group of excluded variables. The fifth group included records with missing data. Accordingly, the dataset size was reduced to 15, 580 cases and 122 preoperative factors.

For the data inclusion step, we excluded all the records with a *gtime* less than 3,285 (9 years) and a *gstatus* that indicates that the patient is still alive. This inclusion is intuitive since it is impossible to know whether the patient would be alive or dead at our time of analysis of 9 years. Based on this step, the final number of cases included in our analysis has been reduced to 13,720, i.e. 1,860 were excluded. Among the 13,720 observations, the number of survivals and deaths is 12,103 and 1,617, respectively.

3.3.2 Variable Selection Methods

Since our final dataset consisted of 122 preoperative factors, it is necessary to select the subset of these factors that are most relevant. We use five different approaches to determine candidate sets of predictors for our model. Sections 3.3.2.1 to 3.3.2.4 present how these variable selection

approaches were used in this paper. In Section 3.3.2.5, we discuss how combinations of these sets are generated for our analysis.

3.3.2.1 Data Mining –based Variable Selection Models

C&RTs and ANNs have been shown to be suitable for feature selection in the machine learning literature [28-31], with excellent performance in transplantation procedures [8, 10, 32]. We provide a brief description of these two methods and the procedures used to combine the knowledge gained from them in the paragraphs below. For a more detailed introduction to these methods, the reader is referred to [33, 34].

10-fold cross validation:

The *k*-fold cross-validation approach is used to minimize the bias associated with the random sampling of the training and test data samples [35]. The entire dataset is randomly split into *k* mutually exclusive subsets of approximately equal size. The prediction model is tested *k* times by using the test sets. The estimation of the *k*-fold cross validation for the overall performance criteria is calculated as the average of the *k* individual performances as follows [36]:

$$CV = \frac{1}{k}\sum_{i=1}^{k} PM_i \,, \tag{3.1}$$

where *CV* stands for cross validation, *k* is the number of mutually exclusive subsets (folds) used, and *PM* is the performance measure used in the analysis. In our analysis, we use the stratified 10-fold cross validation approach to estimate the performance of the different classification models.

Our choice for *k*=10 is based on literature results [35, 36] that show that 10-folds provide an ideal balance between performance and the time required to run the folds.

Decision trees:

Decision trees are widely used in several data mining and organ transplantation applications since they are easy to interpret algorithms. The modeling procedure starts with splitting the dataset into several subsets each of which consists of more or fewer homogeneous states of the target variable [37]. Then the impacts of each independent variable on the target variable are measured. This procedure takes place successively until the decision tree reaches a stable state. Popular decision tree algorithms include Quinlan's ID3, C4.5, C5 [38, 39] and C&RT (Classification and Regression Trees) [37]. In our data analysis, the C&RT algorithm has been used to select the most important predictor variables due to its superior performance compared to the other decision tree algorithms in the preliminary data analysis.

Artificial neural networks:

ANNs are widely employed in a wide variety of computational data analytics problems that include classification, regression and pattern recognition. In this discussion, we assume that the reader is familiar with ANNs and their construction (otherwise, refer to Sordo [40]). For our analysis, we use the sigmoid function as the activation function for our ANN. We have also used *the Multi-layer Perceptron (MLP) learning model with a back-propagation algorithm* due to its superior performance to the *radial basis function (RBF)* in our preliminary analysis. This result has also been reported by Dag *et al.* [12] in their investigation of short, mid and long-term survival of heart transplant patients based on the same dataset.

Performance Evaluation Metrics:

To evaluate the performance of the data mining procedures, we used six different evaluation metrics: a) *accuracy*, b) *area under the receiver operating characteristic curve* (AUC), c) *F-measure*, d) *G-mean*, e) *recall*, and f) *specificity*. These six are selected since they are suitable for binary classification. A detailed description of these measures can be found in [12, 41]. It should be noted that AUC is our primary criterion in comparing the data mining models used in this study; since our dataset is imbalanced (the number of survivals is much larger than deaths). We use AUC in the information fusion (IF) step to combine the variables selected from both data mining models. In the results section, for the sake of completion, we provide the values for the other metrics when evaluating the performance of the BBN model.

Determining the importance of predictor variables through sensitivity analysis:

After determining the performance of the different predictive models, we measure the relative importance of each of the independent variables using *sensitivity analysis* (SA). This phase is needed for several reasons. First, it can help understand the casual factors for our two data mining models. This is particularly important in understanding and communicating the results of ANNs which are still considered by many to be black box models [e.g. see 42]. A second major reason for the importance of *sensitivity analysis* is it provides us with a framework to capture the importance of independent variables across different models.

The sensitivity of a specific predictor variable is calculated by taking the proportion of the error of the model that includes this variable to the error of the model when it does not include this specific variable [43]. The importance of a variable is in direct proportion to variance of predictive error of the classification model in the absence of that specific variable. The same method is

followed for all classification models, and is used in ranking the relative importance of the variables of each classification model according to the sensitivity measure defined by Saltelli *et al.* [43]. Their measure is defined as:

$$S_i = \frac{V_i}{V(y)} = \frac{V(E(y \mid x_i))}{V(y)}, \tag{3.2}$$

where y is the binary output variable (*gstatus*), and the unconditional output variance is denoted by $V(y)$. The expectation operator is denoted by E, which calls for an integral over all predictor variables except x_i. A further integral operator is implied over x_i by the operator V_i. The importance of a specific variable is then computed as the normalized sensitivity as described by Saltelli et al. [44].

Information Fusion:

Information fusion (IF) techniques combine information obtained from multiple forecasts (from prediction models) in an attempt to decrease model uncertainty and increase the knowledge gained. Studies has shown that robustness and accuracy of information can be increased, while the uncertainty and bias of individual models can be decreased by combining multiple forecasts [45]. Therefore, there is an increasing deployment of IF techniques in data-mining problems as opposed to the application of a single method [12, 46]. In our study, we adopt the information fusion model presented by Delen et al. [47] since it has provided us with good performance in the pilot analysis and it can be easily explained to medical practitioner

3.3.2.2 Genetic Algorithms (GA)

Genetic algorithms are adaptive heuristic search algorithms which are utilized in the situations where exhaustive optimization approaches are not favorable due to high computational time. They initialize by generating a random set of individuals which will represent the domain under consideration. These individuals, limited in set size throughout the optimization process are represented by unique identifiers (genome). The genomes are tested at each iteration based on the specific criterions (fitness functions). The genomes for next iterations are chosen based on their fitness function criteria. Then, new individuals are generated to obtain the result independent of initial solutions formed. The new individuals are generated by two operations: crossover and mutation. Genetic algorithms have also been utilized for attribute selection purposes in medical literature [48, 49]. In our current study, a GA with 150 generations (tournament size of 2 with 2-point crossover and a mutation probability of 0.2) was deployed into the *k-Nearest Neighbor* algorithm (k-NN) (*k=3*) [50]. We have used 150 generations in the GA and k=3 in k-NN models since they provided us with the best AUC and time of execution performance when compared to the other variants that we examined. The reader is referred to [48, 50] for a detailed description of GA and k-NN methods.

3.3.2.3 Ridge Regression

Ridge regression is a linear regularization approach that aims to produce precise estimates of regression coefficients. To do so, a penalty is added on the sum of the squared regression parameters, which reduces the overfitting of data due to the multicollinearity of the covariates [51]. Applying ridge regression (RR) results in small non-zero regression coefficients. The general formula for RR is:

$$RSS_{L2} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\left(\left(\sum_{j=1}^{p}\beta_j^2\right) - C^2\right). \quad (3.3)$$

where y_i is the outcome and, $\hat{y}_i$ is its associated predicted value. The number of predictors and cases are represented by *p* and *n*, respectively. RSS_{L2} corresponds to the second order penalty being used on the parameter estimates and, λ is the penalty parameter that can optimally be determined via trial-and-error based experiments, and C is the upper limit of the sum of squares of the βs. In this paper, we generated a set of candidate variables based on those factors whose coefficient values (βs) are higher than 0.25. Note that we use an arbitrary value of 0.25; however, this was a logical value based on the observed coefficients.

3.3.2.4 Variable Selection through Cox Survival Analysis Regression Model and Literature Review

Survival analysis is a subarea of statistics that focuses on the estimation of the time needed until an event occurs (graft failure in our case). Cox proportional hazard model (CPHM) is widely used in survival analysis to determine the predictors' effects on a dependent variable [52] since it can handle censored data. In such a model, the effect of a unit increase in a predictor on the outcome variable is proportional with the hazard rate [53]. In this sense, the model is similar to *multiple regression analysis*, where the outcome is the hazard function at a given time. In our study, the survival time of each heart-recipient is assumed to have the following exponential hazard function h_i;

$$h_i = h_0.\exp(\mathbf{x_i}.\boldsymbol{\beta}), \quad (3.4)$$

where h_0 represents the baseline hazard function (denoting the probability of failure when all predictors are zero) and $\boldsymbol{x_i}$ denotes the vector of predictor attributes for the i^{th} recipient. The regression coefficients for the predictors are represented by a vector $\boldsymbol{\beta}$ that is assumed to be the same for all recipients. Thus, the effects of covariates on the outcome are determined by measuring the proportional effect in the hazard function.

For a more in-depth coverage of the Cox proportional hazard model, we refer the reader to [2]. In addition to the variables obtained by the Cox model, a list of candidate variables that were found to be important for long-term survival is obtained by reviewing the heart transplantation literature review (LR) [16-18, 54-58].

3.3.2.5 Creating Possible Predictor Sets

Seven possible predictor sets were constructed by using the variable selection methods in Sections 3.3.2.1- 4. In these scenarios, we maintained all variables selected in the LR stage [16-18, 54-58] and those identified by the Cox Model. The rationale behind this decision is: a) the LR variables account for medically relevant information and knowledge gained in previous studies; and b) Cox model is the only approach considered here where survival is analyzed as a continuous variable (thus, it can be the most informative). The seven scenarios were then developed using a simple union operator; LR + Cox Variables with different combinations of the methods explained in Sections 3.3.2.1-4. These seven possible final predictor sets and the methods used in the construction process are shown in Appendix.

3.3.3 Use of Bayesian Belief Networks

A Bayesian Belief Network (BBN) is a directed acyclic graph that represents a probabilistic dependency model. It consists of a set of interconnected variables, where the nodes in the network correspond to the variables (predictors) and the arcs express the conditional dependencies and causal relations between these variables [59]. It is a powerful data mining technique that can be used to help reasoning under uncertainty as well as for modeling complex nonlinear interactions among the attributes [60]. They have been widely utilized in medical domain [61-69] and have become increasingly popular since they are able to handle uncertain knowledge involved in detection of diseases, predicting the outcome of diseases as well as making medical decisions on optimal treatment alternatives in various different areas [70]. Pa_{x_i} is the set of parents for each x_i, the Bayesian Network chain rule can be expressed as [71]:

$$P(\mathrm{x}_1,...,\mathrm{x}_n) = \prod_{i=1}^{n} P(x_i | Pa_{x_i}). \tag{3.5}$$

The Naive Bayes (NB) classification is a simple model that can help to learn the structure. It assumes conditional independence between all predictor variables with the given class/ target. The classification is based on Bayes rule where the probability of class/target value computed for each given attribute variables and then the highest prediction is chosen for the structure [72]. Tree Augmented Naïve Bayes (TAN) method is a relaxation of the NB classifiers where the class variable has no parents, but it is one of the parents of each predictor along with at most one other attribute (Figure 3.2). In such networks, the inferences can be made by looking at arcs and nodes in the presented graph. An arc between two nodes is a sign of dependency between those nodes. As illustrated in Figure 3.2, an arc from the class variable to each predictor means that the class variable is a parent for each predictor and each predictor has at most one more parent along with the class variable:

$$Pa_{x_i} = \left\{ C, x_{\xi(i)} \right\}, \tag{3.6}$$

where $\xi(i)$ is the tree function over $x_1,...,\mathrm{x}_n$, and Pa_{x_i} is the set of parents for each x_i. Class variable (C) has no parents and is defined as:

$$Pa_C = \varnothing. \tag{3.7}$$

The arc between two predictors implies that the contribution of the child node in predicting the outcome (class node C) is dependent on the parent node value. For example, in predicting the class variable C, the contribution of X_3 is dependent on the value of X_1, and the contribution of X_2 is dependent on X_1's value.

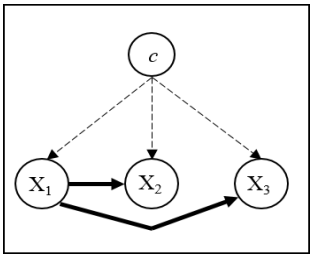

Figure 3.2: Tree Augmented Naïve Bayes Structure

Finding the best tree is an optimization problem where the objectives are to maximize log likelihood of $\xi(i)$, and to construct a maximum likelihood tree to find a maximal weighted spanning tree in a graph [73]. Then, the TAN construction steps can be defined below as shown in [73]:

1. Compute conditional mutual information function for each (i, j) pairs;

$$I_P(x_i : x_j | C) = \sum_{x_i, x_{j,C}} P(x_i, x_j, C) \log \frac{P(x_i, x_j | C)}{P(x_i | C) P(x_j | C)}, i \neq j .$$

   This function tells how much information x_j provides about x_i when the class variable is known.

2. Build a complete undirected graph and use conditional mutual information function to annotate the weight of an edge connecting x_i to x_j.
3. Build a maximum weighted spanning tree.
4. Convert the undirected graph to a directed one by choosing a root variable and setting the direction of all edges to be outward from it.
5. Construct a TAN model by adding a vertex labeled by *C* and adding an arc from *C* to each x_i.

The other critical stage of a BN's construction is to identify the variables that should be incorporated in the model. The selection of the most relevant variables depends on the modeler and the dataset. Constructing a Bayesian network manually may turn out to be quite time consuming in that it requires access to expert knowledge. There are many healthcare related applications in the literature that have been manually constructed by domain experts [74-78]. On the other hand, the network can be learnt

and constructed from the data itself without any explicit knowledge of domain experts as long as the comprehensive dataset is available (as in our case). Therefore, learning from the data has been attracting considerable interest within the research community [79, 80]. Our proposal in the current study is to learn a TAN network, where $x_1, ..., x_n$ can be defined as set of attributes/predictors that represent the union set of predictors which was obtained through merging three different sets of variables (see Section 3.4.2) , and *C* is the target/class variable.

3.4 Results and Discussion

In this section, we only discuss the results from Scenario 1 (LR + Cox + DM) since it provided a competitive AUC performance (≤ 0.01 of the best solution) with ~ half the number of variables. Thus,, we consider this scenario to be the most likely to be adopted by practitioners (see Ockham's razor). For the sake of completion, we provide the comparison among these seven different scenarios in the Appendix.

3.4.1 Variable Selection Results

3.4.1.1 Data Mining-based variable selection results

Table 3.1 provides the *accuracy*, *recall*, *specificity*, *G-mean*, *F-measure* and *AUC* values for each of the 10-fold cross validated samples of the ANN and C&RT models. Based on the AUC values (bolded and underlined), both models have similar performances in predicting the 9-year survival. Additionally, the C&RT model outperforms the ANN model based on the *specificity* and *G-mean* values, while the ANN model is better based on the *recall* and *F*-measures. This means that the C&RT model is more powerful in detecting the patients whose grafts will likely to fail within the 9-year evaluation period, while the ANN is better in predicting those patients who will survive at least 9-years post-transplant. We use the sensitivity analysis (SA) technique of [43, 44] to

understand the relative contribution of the predictor variables selected by C&RT and ANN. The SA method results are shown in Figure 3.3

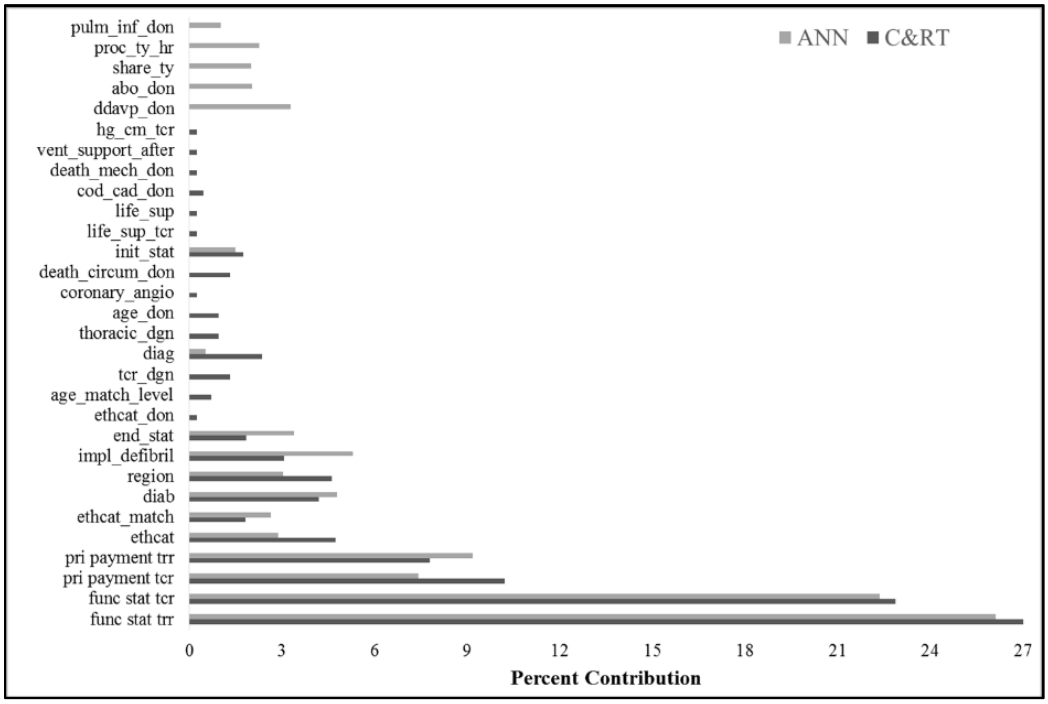

Figure 3.3: Sensitivity Analysis for ML-based Variable Selection Models

Table 3.1: Results of the six evaluation metrics for the C&RT and ANN 10-fold samples

Decision Trees (C&RT)									Artificial Neural Networks (ANN)								
Fold No	TP FP	FN TN	Accuracy	Recall	Specificity	G-mean	F-measure	AUC	Fold No	TP FP	FN TN	Accuracy	Recall	Specificity	G-mean	F-measure	AUC
1	317 67	67 99	0.756	0.826	0.596	0.702	0.826	**0.842**	1	316 82	65 84	0.731	0.829	0.506	0.648	0.811	**0.798**
2	290 52	94 114	0.735	0.755	0.687	0.720	0.799	**0.814**	2	328 81	56 85	0.751	0.854	0.512	0.661	0.827	**0.825**
3	304 65	80 101	0.736	0.792	0.608	0.694	0.807	**0.804**	3	329 83	55 83	0.749	0.857	0.500	0.655	0.827	**0.849**
4	314 82	70 84	0.724	0.818	0.506	0.643	0.805	**0.798**	4	314 82	70 84	0.724	0.818	0.506	0.643	0.805	**0.809**
5	320 50	64 116	0.793	0.833	0.699	0.763	0.849	**0.817**	5	317 57	67 109	0.775	0.826	0.657	0.736	0.836	**0.829**
6	318 80	66 86	0.735	0.828	0.518	0.655	0.813	**0.809**	6	326 89	58 77	0.733	0.849	0.464	0.628	0.816	**0.808**
7	312 93	72 73	0.700	0.813	0.440	0.598	0.791	**0.838**	7	316 78	68 88	0.735	0.823	0.530	0.660	0.812	**0.812**
8	308 81	75 86	0.716	0.804	0.515	0.644	0.798	**0.791**	8	302 83	82 83	0.700	0.786	0.500	0.627	0.785	**0.779**
9	327 57	56 110	0.795	0.854	0.659	0.750	0.853	**0.848**	9	326 65	57 102	0.778	0.851	0.611	0.721	0.842	**0.849**
10	338 69	46 97	0.791	0.880	0.584	0.717	0.855	**0.854**	10	346 71	38 95	0.802	0.901	0.572	0.718	0.864	**0.888**
Mean			0.748	0.820	0.581	0.689	0.820	**0.822**	**Mean**			0.748	0.839	0.536	0.670	0.823	**0.825**
Std Dev			0.034	0.034	0.086	0.052	0.024	**0.022**	**Std Dev**			0.030	0.031	0.059	0.040	0.022	**0.031**

*TP, FP, FN, and TN denote true positives, false positives, false negatives, and true negatives, respectively.

Using the information fusion model of Sec. 2.2.1, we constructed a union set of important predictors and their ranking (see Figure 3.4). Note that the variables are sorted according to their percent contribution, where func_stat_trr (recipient functional status at registration) and func_stat_tcr (recipient functional status at transplant) had the highest contributions of 26.58 % and 22.62 %, respectively. These variables are typically the most significant in predicting a heart transplant's success as in Dag et al. [12]. Based on Figure 3.4, we selected the ordered variables that resulted in a cumulative percent contribution of greater than 90%. These 11 variables, defined in Table 3.2, are our fused set of important predictors from the data mining step. The selection of these variables is also justified by an examination of Figure 3.3. Specifically, these are the only variables whose % contribution was ≥ 1 by both the ANN and C&RT models.

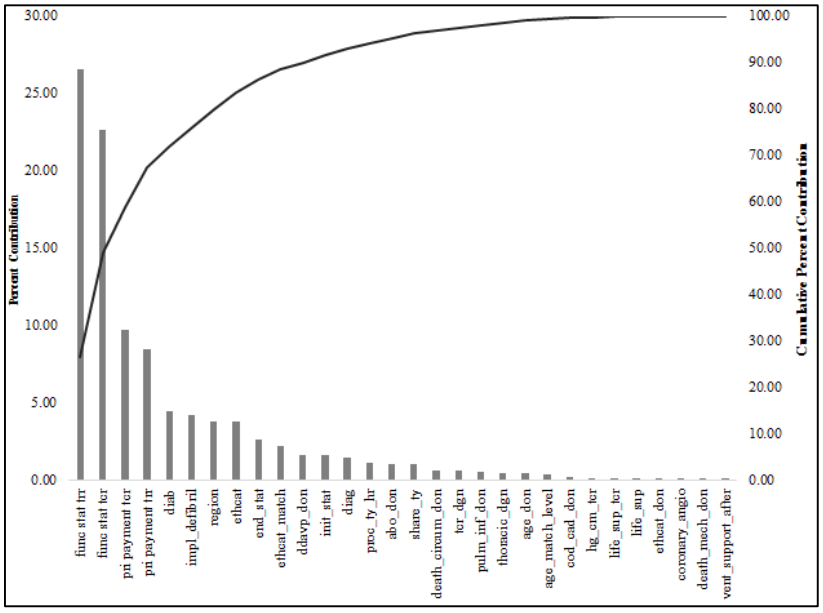

Figure 3.4: The (fused) importance of the union set of predictors based on the IF model

Table 3.2: Data Mining models Variable Set (DMVS)

Variable	Definition
FUNC_STAT_TRR	recipient functional status at registration
FUNC_STAT_TCR	recipient functional status at transplant
PRI_PAYMENT_TCR	recipient primary projected payment type at registration
PRI_PAYMENT_TRR	recipient primary payment source at transplant
DIAB	recipient diabetes at registration
IMPL DEFIBRIL	implantable defibrillator at registration
REGION	UNOS region where transplanted/listed
ETHCAT	recipient ethnicity category
END_STAT	candidate status at transplant offer/removal
ETHCAT_MATCH	donor-recipient ethnicity match level
DDAVP_DON	deceased donor-synthetic anti diuretic hormone

3.4.1.2 Variable Selection Results based on the Cox Model

The second set of predictors was captured based on the Cox survival model. The survivor function was obtained by using a combined stepwise (forward and backward) variable selection in predicting the event (*gstatus*) via the graft time-span that was denoted *gtime*. After running the stepwise variable selection procedure, four variables were found to be significant (*with a p value of* $\leq$0.01), see Table 3.3.

Table 3.3: Cox model variables set

Variable	Sig.	Exp(B)	Definition
AGE_GROUP	0.0000	2.0255	recipient age group
FUNC_STAT_TRR	0.0000	1.0002	recipient functional status at transplant
IMPL_DEFIBRIL	0.0000	0.8127	implantable defibrillator at registration
DDAVP_DON	0.0001	1.1526	deceased donor-synthetic anti diuretic hormone

The effects of individual predictors can be interpreted by using the value of *Exp(*B); it represents the "change in the predicted value" in the hazard function when there is a unit increase in the corresponding predictor variable**.** For categorical predictors, it represents the estimation of the ratio of the "hazard rate in one category" to "the hazard rate in another category". For example, if a

recipient belongs to an older age group, the risk of graft failure increases 2.0255 times (if everything else is held constant). For continuous predictors, it reflects the "change of graft failure risk" for an increase of one unit of that specific predictor.

3.4.1.3 Variable selection based on the literature

The third set of predictors was identified based on the literature (see Section 3.3.2.4). This resulted in 13 additional variables that were not in Tables 3.2 - 3.3. We denote these variables with an asterisk in Table 3.4.

3.4.2 The Union Set of Data Mining, Cox Regression and Domain-Experts Predictors

After obtaining three sets of important/significant variables, we merged them into a final set of variables. Twenty-five unique predictors were selected through using a simple union operator. These variables are defined and presented in Table 3.4. We can broadly categorize these variables into two groups: a) directly biomedically-relevant and; 2) indirectly biomedically-relevant. Since these variables are the inputs to the BBN model, we refer to them as the BBN variables. Only the BBN variables were stored in our database, i.e. the number of predictors was reduced from 122 (initial) to 25 (BBN inputs).

Table 3.4: BBN variables. *Denotes additional variables identified from the literature search

No.	Variable Name	Definition
1	ABO*	recipient blood group at registration
2	ABO_DON*	donor blood type
3	ABO_MAT*	donor-recipient abo match level
4	AGE*	recipient age (years)
5	AGE_DON*	donor age (years)
6	AGE_GROUP	recipient age group
7	DAYS_WAIT_CHRON*	total days on waiting list
8	DDAVP_DON	deceased donor-synthetic anti diuretic hormone
9	DIAB	recipient diabetes at registration
10	DIAG*	recipient primary diagnosis
11	END_STAT	candidate status at transplant offer/removal
12	ETHCAT	recipient ethnicity category
13	ETHCAT_DON*	donor ethnicity category
14	ETHCAT_MATCH	donor-recipient ethnicity match level
15	FUNC_STAT_TCR	recipient functional status at registration
16	FUNC_STAT_TRR	recipient functional status at transplant
17	GENDER*	recipient gender
18	GENDER_DON*	donor gender
19	IMPL_DEFIBRIL	implantable defibrillator at registration
20	MED_COND_TRR*	recipient medical condition pre-transplant at transplant
21	PRI_PAYMENT_TCR	recipient primary projected payment type at registration
22	PRI_PAYMENT_TRR	recipient primary payment source at transplant
23	REGION	UNOS region where transplanted/listed
24	WGT_KG_DON*	calculated donor's weight in kilograms
25	WGT_KG_TCR*	recipient weight (kilograms) at registration

The predictors in the first group include all the LR variables [12-21]. In addition, they include other medically relevant variables; for example, DIAB which represents whether the patient is diabetic or not and has been shown to be medically relevant in non-data mining-based transplantation papers [81, 82]. It should be noted that several of the LR variables were also identified by the data mining and/or the Cox survival model (as discussed earlier in this section). The variables in the second group may not seem to be directly medically relevant; however, they reflect a patient's financial situation, education level and other socio-economic factors that are often predictive of post-surgical outcomes [83-88]. For example, in the context of heart

transplants, Allen et al. [83] had investigated the effect of payment type, education level and the patient's insurance status on long-term survival after an orthotopic hear transplantation (OHT) in the US. Their results showed that payment/insurance type and education level of the recipients are predictors of long-term survival. Moreover, Shapiro et al. [88] and Geller and Connolly [87] investigated the effect of psychosocial factors on survival outcomes after a heart transplant. These factors include variables such as education, social support, living arrangement, etc. Their work revealed that psychosocial factors can identify patients with increased risk of postoperative morbidity. Gerber et al. [85] showed that a poor neighborhood income was a powerful predictor of mortality after a myocardial infraction. Therefore, the inclusion of predictors in the second group provides insights into socio-economic predictors of heart transplantation outcomes. These variables were found to be consistent with the literature that investigated the impact of socio-economic factors on surgical and transplantation outcomes. We did not include them in our LR predictor set, however, since we focused on variables that were agreed upon in all examined papers.

3.4.3 BBN Model Results

By employing the BBN model with 10-fold cross validation, we obtained an average AUC score of 0.840. The detailed results of each of the 10-folds for each of our six metrics are presented in Table 3.5. We should note that the main goal of this study is not to compare the performance of BN model with other models, but rather to uncover the conditional relations between the predictors and understand/calculate the individual survival risks for patients who underwent a heart transplant. With that being said, it is interesting to see that the BBN model outperformed the C&RT and ANN models based on the AUC metric (our main evaluation criterion) as well as the *G-means*. The BBN had a lower *F*-measure, when compared to the two data mining models. We have only

discussed these three metrics here since they are the most suitable for unbalanced datasets [41]. With that being said, the overall results (based on the AUC, *F-measure*, and *G-means*) show that the BBN model is a good predictor for both 9-year survivals and failures of the graft transplant.

As explained in Section 3.3.3, the TAN provides an understanding of the inter-relations between each predictor and the outcome (*gstatus*), and the contribution of inter-variable relations to the probability of each outcome. Since we have employed 10-fold cross validation to develop 10 BBN models, it would not be practical to present and interpret 10 different individual networks in this manuscript. Accordingly, we use the 7^{th} fold to be a representative model since its AUC value, *G-mean*, and *F-measure* are closest to our average values for the 10-folds. The acyclic graph obtained through using seventh fold is provided in Figure 3.5. The numbering of the BBN predictors is based on Table 3. 5. Recall that an arrow from a predictor (parent) to another (child) implies that the impact of the child predictor on the *gstatus* is dependent on the value of the parent. From Figure 3.5, there is only one predictor (*ABO*) who has only one parent, which is the class variable *gstatus*. This means that the contribution of *ABO (recipient blood group at registration*) on predicting the *gstatus* does not change with the change in values of any of the other 24 predictors. This result is interesting since the two other blood type variables (ABO_DON, and ABO_Match) are children of ABO, i.e. their contribution to prediction differs with a change of ABO. However, they do not affect ABO's contribution to the model.

Table 3.5: BBN classification results

Fold No	TP FP	FN TN	Accuracy	Recall	Specificity	*G-means*	*F-measure*	AUC
1	271 25	113 141	0.749	0.706	0.849	0.774	0.797	**0.843**
2	269 30	115 136	0.736	0.701	0.819	0.758	0.788	**0.831**
3	259 10	125 156	0.755	0.674	0.940	0.796	0.793	**0.838**
4	262 26	122 140	0.731	0.682	0.843	0.759	0.780	**0.831**
5	266 23	118 143	0.744	0.693	0.861	0.772	0.790	**0.854**
6	263 29	121 137	0.727	0.685	0.825	0.752	0.778	**0.826**
7	259 19	125 147	0.738	0.674	0.886	0.773	0.782	**0.841**
8	257 33	126 134	0.711	0.671	0.802	0.734	0.764	**0.800**
9	273 22	110 145	0.760	0.713	0.868	0.787	0.805	**0.848**
10	275 13	109 153	0.778	0.716	0.922	0.812	0.818	**0.885**
Mean			**0.743**	**0.692**	**0.862**	**0.772**	**0.790**	**0.840**
Std. Dev.			**0.019**	**0.017**	**0.044**	**0.023**	**0.015**	**0.022**

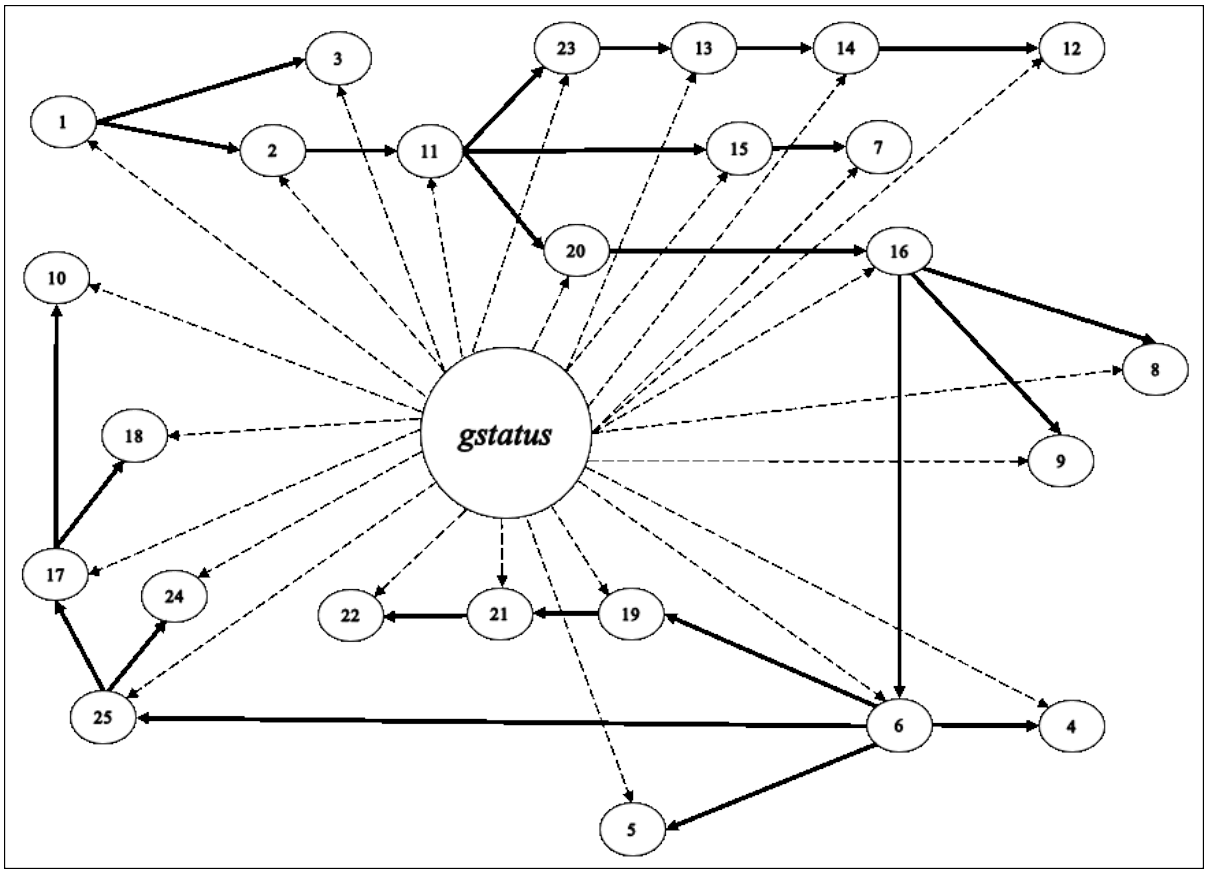

Figure 3.5: TAN structure for the Proposed Method

From Figure 3.5, we can identify 10 nodes that have no influence on the contributions of any other predictors, yet their contribution to predicting the outcome variable is dependent on another predictor. These nodes correspond to: *ABO_MAT*, *AGE*, *AGE_DON*, *DAYS_WAIT_CHRON*, *DDAVP_DON*, *DIAB*, *DIAG*, *ETH_CAT*, *GENDER_DON*, *PRI_PAYMENT_TRR*. By closely examining the parent-child relationships in Figure 3.5, we can identify several relationships that are intuitive. For example, the contribution of predictor 18 (*GENDER_DON*) in predicting the outcome variable depends on predictor 17 (*GENDER*). This means that the interaction between the recipient and donor genders might in some cases have an effect on the survival after the heart transplant. Similarly, the variable 9 (*DIAB*)'s contribution on predicting the *gstatus*, is dependent on predictor 16 (*FUNC_STAT_TRR*). The effect of a

recipient's diabetes on graft survival is dependent on the state of the recipient's functional status (ability to carry out daily activities). Predictor 16 (*FUNC_STAT_TRR*) also influences the contribution of *AGE_GROUP* (Node 6). The information extracted from the BBN can be analyzed in a more detailed manner by domain experts. More specifically, Figure 3.5 (and similar networks) can provide medical practitioners with significant insights on the factors that affect survival and their interactions.

Based on the TAN depicted in Figure 3.5, we can calculate the patient-specific survival risks which we define as how likely (posterior-probability) the transplant will result in a graft failure, within our 9-year horizon, given the values for our 25 predictors. Note that this can be seen as the continuous predictor for the binary outcome predicted variable, *gstatus*. For convenience, we denote the continuous predictor score with *cscore*. When the *cscore* is greater than 0.5, the model predicts that the patient will survive, and when the *cscore* is less than 0.5 the model predicts that the patient will die. Predictions further away from 0.5 (in either direction) are more likely to be realized. We demonstrate this concept in by stratifying and analyzing multiple *cscores* based on the 7^{th} fold of our BBN model. Based on this representation, one could see that the *cscore* improves the prediction of the outcome based on all 6 metrics when the percentage of extreme values in the sample increases (i.e. as one goes down the table).

Table 3.6: Performance of BBN with different cutoffs for the *cscores*, *where* cscore ≠ 0.5 is the model used throughout this paper (∴ first row is identical to 7th fold in Table 3.5)

cscore	# dropped cases	# cases left	# survivals	# failures	TP FP	FN TN	Accuracy	Recall	Specificity	G-mean	F-measure	AUC
≠ 0.5	0	550	166	384	259 19	125 147	0.738	0.674	0.886	0.773	0.782	0.841
∉ [0.4, 0.6]	32	518	156	362	251 13	111 143	0.761	0.693	0.917	0.797	0.802	0.847
∉ [0.3, 0.7]	78	472	136	336	239 4	97 132	0.786	0.711	0.971	0.831	0.826	0.861
∉ [0.2, 0.8]	154	396	97	299	233 3	67 93	0.823	0.777	0.969	0.867	0.869	0.895
∉ [0.1, 0.9]	264	286	40	246	223 1	23 39	0.916	0.907	0.975	0.940	0.949	0.959
∉ [0.05, 0.95]	300	250	22	228	220 1 0	8 21 0	0.964	0.965	0.955	0.960	0.980	0.981

3.5 A Decision Support Tool for Providing Insights to Medical Practitioners

One of the goals of this paper is to provide medical practitioners with insights from data-driven models. These type of insights are expected to play an important role in the future of health-care (see e.g., IBM Watson Health at http://www.ibm.com/smarterplanet/us/en/ibmwatson/health/). The tool is based on the BBN results in Section 3.4. A practitioner can insert the values for the 25 preoperative predictors and obtain a *cscore* for a recipient's survival as shown in Figure 3.6. Note that the user has the option to train the BBN (e.g. by adding records post 2012) or use our training dataset (from 1987-2012).

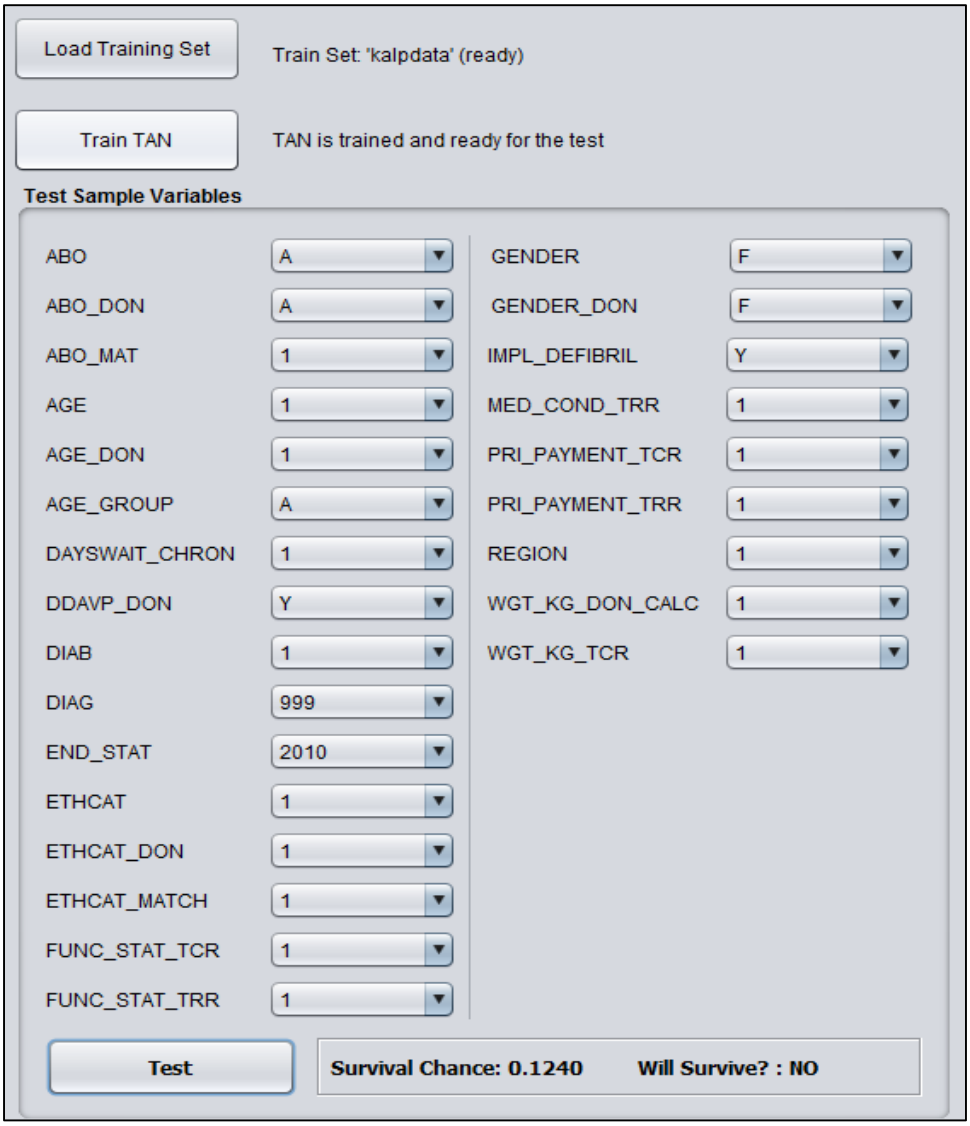

Figure 3.6: The interface of the decision support tool

The tool has been developed in JAVA using Swing components. Since JAVA is a platform-independent programming environment, our tool can be easily installed on any computer having Windows, Linux or MacOS operating systems. On the background, a sophisticated Bayesian Network Classifier JAVA library (see code at: http://jbnc.sourceforge.net/#jBNC_Toolkit), which provides a stable and robust version of the TAN algorithm used in this paper. The link for our tool can be found at this temporary location: https://www.dropbox.com/sh/dv2waunpjw9jpik/AADwRMHTn8c9siCUh7uvZXD6a?dl=0 (a permanent location will be constructed after the review process).

3.6 Conclusions and Future Research Recommendations

The main objectives of this paper were to develop a mathematical model to identify patient-specific survival risk scores and the interactions between the explanatory variables. To achieve these goals, we have proposed a BBN framework that is based on variables selected from data-mining models, statistical models, and the literature. Our approach is used to investigate a large, feature-rich dataset obtained from UNOS, containing all recorded information on heart transplant operations that were performed in the U.S. between October 1, 1987 and December 31, 2012. In our analysis, we have addressed the following questions:

a) What predictive factors contribute to the outcome of a heart transplant for 9-year survival?

b) Can the interactions between these predictors be quantified and visualized?

c) How can data-driven methods be used to construct a probabilistic patient-specific failure risk score based on the values of the relevant preoperative predictors?

The innovation in our framework is based on our ability to address the latter two questions, which were not previously examined within the heart transplantation research community. In addition, we provide a decision support tool that can be used by practitioners to obtain a 9-year survival probability for a patient given attributes extracted from the dead donor.

The results obtained through using the proposed framework indicates that employing a comprehensive variable selection procedure leads to an excellent predictive model as illustrated by our examination of 6 metrics. More importantly, the information obtained using a probabilistic (graphical) approach in the final step has provided the relations among the important predictors as well as the patient-specific failure probabilities. These significant outcomes can be insightful to

medical practitioners and policy-makers since they may provide with the ability to conduct prospective studies.

A number of limitations in our analysis needs to be mentioned. First, there are several variables in the UNOS dataset whose collection has started after October, 1987. For instance, the UNOS team started to collect the information about the variable called "*Alchohol_heavy_don"* on June 30th, 2004. In our data cleaning procedure, most of these variables had to be deleted since these variables have excessive number of missing records regarding to the dates in which they were not considered. A potential remedy was to solve such a problem would be using only the last 10 years of the dataset; however, this would have prevented us from having a large enough dataset to study the factors predicting the long-term survival. A second limitation is that we have only compared two data mining approaches in our variable selection procedure approaches (ANN and DT) for analyses. The results could have been different if other approaches are investigated. Since it is computationally infeasible to apply every possible variable selection model, we chose to use these common approaches. Similarly, the use of a different information fusion method may have resulted in different findings and combinations of important variables. With that being said, we believe that our results are satisfactory when compared to commonly reported values in the data mining literature.

In summary, this paper proposed a novel framework to predict the long-term survival outcomes after heart transplantation. We have shown that extracting the predictive factors from multiple source helps to increase the performance of the prediction model used in this study. Our approach can be easily extended to other organs, especially since UNOS provides information on other organ transplants. Finally, it should be noted that the analysis presented in this paper may inform new prospective studies if the findings are interpreted in detail by medical practitioners.

Acknowledgements

The authors acknowledge the feedback from Dr. Hussam Farhoud, cardiologist at Kansas Medical Center and the Mercy Hospital in Independence KS. We appreciate the help of Mr. Semih Dinc, PhD Student in Computer Science at University of Alabama in Huntsville, for his assistance with the DSS tool. We are also very grateful for the feedback and comments from three anonymous reviewers and the editor that greatly improved our paper. This work was supported in part by Health Resources and Services Administration contract 234-2005-370011C. The content is the responsibility of the authors alone and does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

3.7 References

1. *What is Heart Failure?* 2012 [cited 2014 01/26]; Available from: http://www.nhlbi.nih.gov/health/health-topics/topics/hf/.
2. López-Sendón, J., *The heart failure epidemic.* Medicographia, 2011. **33**(4): p. 363-369.
3. *What Is Heart Transplant?* 2012 01/03/2012 01/31/2014]; Available from: http://www.nhlbi.nih.gov/health/health-topics/topics/ht/.
4. *What to Expect Before a Heart Transplant?* 2012 [cited 2014 01/26]; Available from: http://www.nhlbi.nih.gov/health/health-topics/topics/ht/before.html.
5. Cruz-Ramı´rez, M., C. Hervás-Martı´nez, J.C. Fernández, J. Briceño, and M. de la Mata, *Multi-objective evolutionary algorithm for donor–recipient decision system in liver transplants.* European Journal of Operational Research, 2012. **222**(2): p. 317-327.
6. *Heart Transplant.* 2015 01/26/2015]; Available from: http://my.clevelandclinic.org/services/hic_Surgical_Treatments_for_Heart_Failure/transplant.
7. Levy, A., *A decision-rule for transplanting non-cadaveric organs.* European Journal of Operational Research, 2005. **164**(2): p. 548-554.
8. Greco, R., T. Papalia, D. Lofaro, S. Maestripieri, D. Mancuso, and R. Bonofiglio, *Decisional trees in renal transplant follow-up.* Transplant Proceedings, 2010. **42**(4): p. 1134-6.
9. Kusiak, A., B. Dixon, and S. Shah, *Predicting survival time for kidney dialysis patients: a data mining approach.* Computers in Biology and Medicine, 2005. **35**(4): p. 311-27.
10. Lin, R.S., S.D. Horn, J.F. Hurdle, and A.S. Goldfarb-Rumyantzev, *Single and multiple time-point prediction models in kidney transplant outcomes.* Journal of Biomedical Informatics, 2008. **41**(6): p. 944-52.
11. Taati, B., J. Snoek, D. Aleman, and A. Ghavamzadeh, *Data mining in bone marrow transplant records to identify patients with high odds of survival.* IEEE Journal of Biomedical Health Informatics, 2014. **18**(1): p. 21-7.
12. Dag, A., F.M. Megahed, A. Oztekin, and A. Yucel, *A data analytics approach to predict heart transplant success.* Decision Support Systems (Under Review), 2016.
13. Del Rizzo, D.F., A.H. Menkis, P.W. Pflugfelder, R.J. Novick, F.N. McKenzie, W.D. Boyd, and W.J. Kostuk, *The role of donor age and ischemic time on survival following orthotopic*

heart transplantation. Journal of Heart and Lung Transplantation, 1999. **18**(4): p. 310-9.

14. Drakos, S.G., A.G. Kfoury, E.M. Gilbert, J.W. Long, J.C. Stringham, E.H. Hammond, K.W. Jones, D.A. Bull, M.E. Hagan, J.W. Folsom, B.D. Horne, and D.G. Renlund, *Multivariate predictors of heart transplantation outcomes in the era of chronic mechanical circulatory support.* The Annals of Thoracic Surgery, 2007. **83**(1): p. 62-7.
15. Gupta, D., V. Piacentino, 3rd, M. Macha, A.K. Singhal, J.P. Gaughan, J.B. McClurken, B.I. Goldman, C.A. Fisher, D. Beltramo, J. Monacchio, H.J. Eisen, and S. Furukawa, *Effect of older donor age on risk for mortality after heart transplantation.* The Annals of Thoracic Surgery, 2004. **78**(3): p. 890-9.
16. Al-Khaldi, A., P.E. Oyer, and R.C. Robbins, *Outcome analysis of donor gender in heart transplantation.* Journal of Heart and Lung Transplantation, 2006. **25**(4): p. 461-8.
17. Tjang, Y.S., G.J. van der Heijden, G. Tenderich, D.E. Grobbee, and R. Korfer, *Survival analysis in heart transplantation: results from an analysis of 1290 cases in a single center.* European Journal of Cardiothorac Surgery, 2008. **33**(5): p. 856-61.
18. Kilic, A., E.S. Weiss, T.J. George, G.J. Arnaoutakis, D.D. Yuh, A.S. Shah, and J.V. Conte, *What predicts long-term survival after heart transplantation? An analysis of 9,400 ten-year survivors.* The Annals of Thoracic Surgery, 2012. **93**(3): p. 699-704.
19. Hong, K.N., A. Iribarne, B. Worku, H. Takayama, A.C. Gelijns, Y. Naka, V. Jeevanandam, and M.J. Russo, *Who is the high-risk recipient? Predicting mortality after heart transplant using pretransplant donor and recipient risk factors.* The Annals of Thoracic Surgery, 2011. **92**(2): p. 520-7; discussion 527.
20. Zuckermann, A.O., P. Ofner, C. Holzinger, M. Grimm, R. Mallinger, G. Laufer, and E. Wolner, *Pre- and early postoperative risk factors for death after cardiac transplantation: a single center analysis.* Transplant International, 2000. **13**(1): p. 28-34.
21. Oztekin, A., D. Delen, and Z. Kong, *Predicting the graft survival for heart–lung transplantation patients: An integrated data mining methodology.* International Journal of Medical Informatics, 2009. **78**(12): p. e84-e96.
22. Kaplan, B. and J. Schold, *Transplantation: neural networks for predicting graft survival.* Nature Reviews Nephrology, 2009. **5**(4): p. 190-2.
23. Nakayama, N., M. Oketani, Y. Kawamura, M. Inao, S. Nagoshi, K. Fujiwara, H. Tsubouchi, and S. Mochida, *Algorithm to determine the outcome of patients with acute liver failure: a*

data-mining analysis using decision trees. Journal of Gastroenterology, 2012. **47**(6): p. 664-77.

24. Brown, T.S., E.A. Elster, K. Stevens, J.C. Graybill, S. Gillern, S. Phinney, M.O. Salifu, and R.M. Jindal, *Bayesian modeling of pretransplant variables accurately predicts kidney graft survival.* American Journal of Nephrology, 2012. **36**(6): p. 561-9.
25. Hoot, N. and D. Aronsky. *Using Bayesian networks to predict survival of liver transplant patients*. in *AMIA annual symposium proceedings*. 2005. American Medical Informatics Association.
26. Heckerman, D., *A tutorial on learning with Bayesian networks*, in *Innovations in Bayesian Networks*. 2008, Springer. p. 33-82.
27. *UNOS | About Us*. 2014 2014/6/2]; Available from: http://www.unos.org/contact/index.php.
28. Kohavi, R. and G.H. John, *Wrappers for feature subset selection.* Artificial intelligence, 1997. **97**(1): p. 273-324.
29. Leray, P. and P. Gallinari, *Feature selection with neural networks.* Behaviormetrika, 1999. **26**: p. 145-166.
30. Setiono, R. and H. Liu, *Neural-network feature selector*. Neural Networks, IEEE Transactions on, 1997. **8**(3): p. 654-662.
31. Sugumaran, V., V. Muralidharan, and K.I. Ramachandran, *Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing.* Mechanical Systems and Signal Processing, 2007. **21**(2): p. 930-942.
32. Sheppard, D., D. McPhee, C. Darke, B. Shrethra, R. Moore, A. Jurewitz, and A. Gray, *Predicting cytomegalovirus disease after renal transplantation: an artificial neural network approach.* The International Journal of Medical Informatics, 1999. **54**(1): p. 55-76.
33. Han, J. and M. Kamber, *Data mining: concepts and techniques*. 3rd ed. 2011, Burlington, MA: Elsevier. 703.
34. Hastie, T., R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani, *The elements of statistical learning: Data Mining, Inference, and Prediction*. Vol. 2. 2009, New York, USA: Springer. 745.
35. Kohavi, R. *A study of cross-validation and bootstrap for accuracy estimation and model selection*. in *Ijcai*. 1995.

36. Olson, D.L. and D. Delen, *Advanced data mining techniques*. 2008: Springer Publishing Company, Incorporated.
37. Olshen, L.B.J.H.F.R.A. and C.J. Stone, *Classification and regression trees.* Wadsworth International Group, 1984.
38. Quinlan, J.R., *Induction of decision trees.* Machine learning, 1986. **1**(1): p. 81-106.
39. Quinlan, J.R., *C4. 5: programs for machine learning*. Vol. 1. 1993: Morgan kaufmann.
40. Sordo, M., *Introduction to neural networks in healthcare.* Open Clinical: Knowledge Management for Medical Care, Harvard University [Online], 2002.
41. Mollineda, V.G.J.S.R. and R.A.J. Sotoca, *The class imbalance problem in pattern classification and learning.* 2007.
42. Molaie, M., R. Falahian, S. Gharibzadeh, S. Jafari, and J.C. Sprott, *Artificial neural networks: powerful tools for modeling chaotic behavior in the nervous system.* Frontiers in Computer Neuroscience, 2014. **8**: p. 40.
43. Saltelli, A., *Making best use of model evaluations to compute sensitivity indices.* Computer Physics Communications, 2002. **145**(2): p. 280-297.
44. Saltelli, A., S. Tarantola, F. Campolongo, and M. Ratto, *Sensitivity analysis in practice: a guide to assessing scientific models*. 2004: John Wiley & Sons.
45. Clemen, R.T., *Combining forecasts: A review and annotated bibliography.* International Journal of Forecasting, 1989. **5**(4): p. 559-583.
46. Graefe, A., J.S. Armstrong, R.J. Jones, and A.G. Cuzan, *Combining forecasts: An application to elections.* International Journal of Forecasting, 2014. **30**(1): p. 43-54.
47. Delen, D., R. Sharda, and P. Kumar, *Movie forecast Guru: A Web-based DSS for Hollywood managers.* Decision Support Systems, 2007. **43**(4): p. 1151-1170.
48. Vinterbo, S. and L. Ohno-Machado, *A genetic algorithm to select variables in logistic regression: example in the domain of myocardial infarction.* Proceedings of the AMIA Symposium, 1999: p. 984-988.
49. Sahiner, B., H.P. Chan, D. Wei, N. Petrick, M.A. Helvie, D.D. Adler, and M.M. Goodsitt, *Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue.* Medical Physics, 1996. **23**(10): p. 1671-1684.
50. Lavrač, N., *Selected techniques for data mining in medicine.* Artificial intelligence in medicine, 1999. **16**(1): p. 3-23.

51. Hoerl, A.E. and R.W. Kennard, *Ridge Regression: Biased Estimation for Nonorthogonal Problems.* Technometrics, 2000. **42**(1): p. 80-86.
52. Cox, D.R. and D. Oakes, *Analysis of survival data*. Vol. 21. 1984: CRC Press.
53. Ohno-Machado, L., *Modeling medical prognosis: Survival analysis techniques.* Journal of Biomedical Informatics, 2001. **34**(6): p. 428-439.
54. Gupta, D., V. Piacentino, 3rd, M. Macha, A.K. Singhal, J.P. Gaughan, J.B. McClurken, B.I. Goldman, C.A. Fisher, D. Beltramo, J. Monacchio, H.J. Eisen, and S. Furukawa, *Effect of older donor age on risk for mortality after heart transplantation.* The Annals of Thoracic Surgery, 2004. **78**(3): p. 890-9.
55. Bardage, C. and D.G. Isacson, *Hypertension and health-related quality of life. an epidemiological study in Sweden.* Journal of Clinical Epidemiology, 2001. **54**(2): p. 172-81.
56. Bourge, R.C., D.C. Naftel, M.R. Costanzo-Nordin, J.K. Kirklin, J.B. Young, S.H. Kubo, M.T. Olivari, and E.K. Kasper, *Pretransplantation risk factors for death after heart transplantation: a multiinstitutional study. The Transplant Cardiologists Research Database Group.* Journal of Heart and Lung Transplantation, 1993. **12**(4): p. 549-62.
57. Stehlik, J., L.B. Edwards, A.Y. Kucheryavaya, C. Benden, J.D. Christie, F. Dobbels, R. Kirk, A.O. Rahmel, and M.I. Hertz, *The Registry of the International Society for Heart and Lung Transplantation: Twenty-eighth Adult Heart Transplant Report--2011.* Journal of Heart and Lung Transplantation, 2011. **30**(10): p. 1078-94.
58. Kauffman, H.M., M.A. McBride, and F.L. Delmonico, *First report of the United Network for Organ Sharing Transplant Tumor Registry: donors with a history of cancer.* Transplantation, 2000. **70**(12): p. 1747-51.
59. Pearl, J., Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning. 1985. http://ftp.cs.ucla.edu/tech-report/198_-reports/850017.pdf
60. Anderson, J.R., R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, *Machine learning: An artificial intelligence approach*. Vol. 2. 1986: Morgan Kaufmann.
61. Aronsky, D. and P.J. Haug. *Automatic identification of patients eligible for a pneumonia guideline*. in *Proceedings of the AMIA Symposium*. 2000. American Medical Informatics Association.
62. Bunn, C.C., M. Du, K. Niu, T.R. Johnson, W.S.C. Poston, and J.P. Foreyt. *Predicting the risk*

of obesity using a Bayesian network. in *Proceedings of the AMIA Symposium*. 1999. American Medical Informatics Association.

63. Burnside, E., D. Rubin, and R. Shachter. *A Bayesian network for mammography*. in *Proceedings of the AMIA Symposium*. 2000. American Medical Informatics Association.
64. Burnside, E.S., D.L. Rubin, J.P. Fine, R.D. Shachter, G.A. Sisney, and W.K. Leung, *Bayesian network to predict breast cancer risk of mammographic microcalcifications and reduce number of benign biopsy results: initial experience.* Radiology, 2006. **240**(3): p. 666-73.
65. Hamilton, P.W., R. Montironi, W. Abmayr, M. Bibbo, N. Anderson, D. Thompson, and P.H. Bartels, *Clinical applications of Bayesian belief networks in pathology.* Pathologica, 1995. **87**(3): p. 237-245.
66. Montironi, R., P.H. Bartels, D. Thompson, M. Scarpelli, and P.W. Hamilton, *Prostatic intraepithelial neoplasia (PIN). Performance of Bayesian belief network for diagnosis and grading.* The Journal of pathology, 1995. **177**(2): p. 153-162.
67. Sakellaropoulos, G.C. and G.C. Nikiforidis, *Development of a Bayesian network for the prognosis of head injuries using graphical model selection techniques.* Methods of Information in Medicine, 1999. **38**: p. 37-42.
68. Friedman, N., *Inferring cellular networks using probabilistic graphical models.* Science, 2004. **303**(5659): p. 799-805.
69. Jiang, X., D. Xue, A. Brufsky, S. Khan, and R. Neapolitan, *A new method for predicting patient survivorship using efficient bayesian network learning.* Cancer Informatics, 2014. **13**: p. 47-57.
70. Lee, S.-M. and P.A. Abbott, *Bayesian networks for knowledge discovery in large datasets: basics for nurse researchers.* Journal of Biomedical Informatics, 2003. **36**(4): p. 389-399.
71. Koller, D. and N. Friedman, *Probabilistic graphical models: principles and techniques*. 2009: MIT press.
72. Friedman, N., D. Geiger, and M. Goldszmidt, *Bayesian network classifiers.* Machine learning, 1997. **29**(2-3): p. 131-163.
73. Chow, C.K. and C.N. Liu, *Approximating Discrete Probability Distributions with Dependence Trees.* IEEE Transactions on Information Theory, 1968. **14**(3): p. 462-+.
74. Andreassen, S., C. Riekehr, B. Kristensen, H.C. Schønheyder, and L. Leibovici, *Using probabilistic and decision–theoretic methods in treatment and prognosis modeling.*

Artificial Intelligence in medicine, 1999. **15**(2): p. 121-134.

75. Heckerman, D.E., E.J. Horvitz, and B.N. Nathwani, *Toward Normative Expert Systems .1. The Pathfinder Project.* Methods of Information in Medicine, 1992. **31**(2): p. 90-105.
76. Lucas, P., H. Boot, and B. Taal, *Computer-based decision support in the management of primary gastric non-Hodgkin lymphoma.* change, 1998. **1**: p. 4.
77. Lucas, P.J.F., N.C. de Bruijn, K. Schurink, and A. Hoepelman, *A probabilistic and decision-theoretic approach to the management of infectious disease at the ICU.* Artificial Intelligence in Medicine, 2000. **19**(3): p. 251-279.
78. van der Gaag, L.C., S. Renooij, C.L.M. Witteman, B.M.P. Aleman, and B.G. Taal, *Probabilities for a probabilistic network: a case study in oesophageal cancer.* Artificial Intelligence in Medicine, 2002. **25**(2): p. 123-148.
79. Lucas, P.J., L.C. van der Gaag, and A. Abu-Hanna, *Bayesian networks in biomedicine and health-care.* Artificial Intelligence in Medicine, 2004. **30**(3): p. 201-14.
80. Yet, B., K. Bastani, H. Raharjo, S. Lifvergren, W. Marsh, and B. Bergman, *Decision support system for Warfarin therapy management using Bayesian networks.* Decision Support Systems, 2013. **55**(2): p. 488-498.
81. Mehra, M.R., J. Kobashigawa, R. Starling, S. Russell, P.A. Uber, J. Parameshwar, P. Mohacsi, S. Augustine, K. Aaronson, and M. Barr, *Listing criteria for heart transplantation: International Society for Heart and Lung Transplantation guidelines for the care of cardiac transplant candidates—2006.* The Journal of heart and lung transplantation, 2006. **25**(9): p. 1024-1042.
82. Davidson, J., A. Wilkinson, J. Dantal, F. Dotta, H. Haller, D. Hernandez, B.L. Kasiske, B. Kiberd, A. Krentz, C. Legendre, P. Marchetti, M. Markell, F.J. van der Woude, and D.C. Wheeler, *New-Onset Diabetes After Transplantation: 2003 International Consensus Guidelines.* Transplantation, 2003. **75**(10): p. SS3-SS24.
83. Allen, J.G., E.S. Weiss, G.J. Arnaoutakis, S.D. Russell, W.A. Baumgartner, A.S. Shah, and J.V. Conte, *Insurance and education predict long-term survival after orthotopic heart transplantation in the United States.* The Journal of Heart and Lung Transplantation, 2012. **31**(1): p. 52-60.
84. Hussain, S., P. Lenner, J. Sundquist, and K. Hemminki, *Influence of education level on cancer survival in Sweden.* Annals of Oncology, 2008. **19**(1): p. 156-162.

85. Gerber, Y., S.A. Weston, J.M. Killian, T.M. Therneau, S.J. Jacobsen, and V.L. Roger. *Neighborhood income and individual education: effect on survival after myocardial infarction*. in *Mayo Clinic Proceedings*. 2008. Elsevier.
86. Rosso, S., F. Faggiano, R. Zanetti, and G. Costa, *Social class and cancer survival in Turin, Italy.* Journal of Epidemiology and Community Health, 1997. **51**(1): p. 30-34.
87. Geller, S. and T. Connolly, *The influence of psychosocial factors on heart transplantation decisions and outcomes.* Journal of Transplant Coordination, 1997. **7**(4): p. 173-179.
88. Shapiro, P., D. Williams, A. Foray, I. Gelman, N. Wukich, and R. Sciacca, *Psychosocial evaluation and prediction of compliance problems and morbidity after heart transplantation.* Transplantation, 1995. **60**(12): p. 1462.

4 An Exploratory Study to Evaluate the Effect of Newly Added Variables to Predictability of the Heart Transplant Outcomes

4.1 Abstract

Predicting the survival outcomes of heart transplant patients is a challenging task since there are many factors associated with the outcome. Yet, it is rewarding in that it plays a crucial role in understanding the mechanics of matching procedure. Including additional factors (variables) to the existing datasets might be favorable since it enables to gain more information about the data. On the other hand, it might bring some disadvantages to the predictability of the outcome in that it increases the complexity of the prediction models. The objective of the current study is to investigate the effect of the recently added variables on the predictability of the graft survival after 1-month (acute rejection), 1-year and 5-years. To do so, a powerful probabilistic data mining method (i.e. Tree-augmented Naïve Bayes, TAN) and Logistic Regression have been employed after selecting the features through a comprehensive feature selection procedure, which involves an Information-gain based Fast Feature Selection algorithm, a wrapper method (Random Forests) and review of the existing literature. To overcome the data-imbalance problems, which was caused by the difference between the number of survivals and failures, random under sampling (RUS) is employed for 1-month and 1-year predictions. The results indicate that Logistic Regression models have achieved the best performance for all three different time points, when compared the TAN. When the newly added features are excluded from the analysis, the area-under-the-curve (AUC) values of Logistic regression are 0.68, 0.641 and 0.663 for 1-month, 1-year and 5-year predictions,

respectively. On the other hand, when the newly added features are included in the analysis, the area-under-the-curve (AUC) values of TAN are 0.639, 0.619 and 0.656 for 1-month, 1-year and 5-year predictions, respectively. More importantly, the predictability of the outcome of the 1-month survival has increased with the inclusion of the new variables for all of the three data analytical models employed in the study, with an happened an insignificant increase in the 5-year survival prediction. The findings obtained through data analytical models have been compared with the results obtained through Sensitivity analysis, which is employed in the final phase. The current study presents important retrospective findings, which can be the basis for a prospective medical study.

4.2 Introduction

Heart failure is a very common medical condition in which the heart is weakened and cannot pump enough blood to meet the body's needs [1]. The reader should note that the term heart failure does not indicate that the heart has stopped working (or is about to stop working). With that being said, heart failure is a serious medical condition that affects an estimated 2-3% of the world's adult population, which corresponds to over 26 million people worldwide [2]. In the U.S. the incidence rate is approximately 550,000 [1, 3]. Among these patients, those who have severe end-stage heart failure (meaning all possible treatments except transplant have failed) are selected through a careful process to be placed on the heart transplant waiting list. If a patient is eligible, then she/he is placed on a waiting list for a transplant until a suitable donor heart is found [4]. Reasons such as aging of the population, increasing obesity, diabetes. [5] as well as increasing life expectancy [6] have caused a dramatic increase in the demand for heart transplantation. Currently, there are approximately 3,000 patients waiting for a heart transplant while the annual supply is around only

2,000 [4]. Therefore, there has been a huge gap between supply (donated and healthy grafts) and demand, which brings longer waiting times and thereby leaves many to die on the waiting lists [6].

United Network for Organ Sharing (UNOS) is a "private, non-profit organization that manages the [United States]'s organ transplant system under contract with the federal government" [7]. The system contains datasets regarding heart, lung, liver, kidney, pancreas and intestine organ donations as well as the transplant events occurring in the U.S. since 0ctober 1, 1987. These datasets include clinical and demographical information about both donors and recipients as well as the transplant procedures itself. Therefore, it has become a common source that can potentially serve the interests of researchers and the public. In certain years, UNOS has made significant changes in the number of variables by adding some to its Organ Transplantation Databases, which potentially can bring more information to the existing body of knowledge. For instance, there exist 165, 85 and 105 variables that were added to the thoracic transplant database in 1994, 1998 and 2004, respectively.

Effective utilization of the valuable information contained in UNOS datasets carry a great potential in saving many lives. More specifically, improving the prediction of survival is a critical task, since it is an essential tool in donor/recipient matching procedure. Such improvement could in turn lead to a decrease in the gap between supply and demand. Recent research has shown that the information that exist in these large/complex UNOS datasets can be analyzed effectively in extracting hidden, novel patterns by employing data mining methods [8-13]. Compared to traditional statistical techniques, data mining approaches provide a relatively faster means of finding complex and nonlinear patterns from a large amount of data containing many predictors. In addition, these methods do not require prior knowledge about the data, nor do they make assumptions about the statistical distribution or properties of the data [11].

The existing literature on survival prediction includes a large body of research that uses data-driven approaches in organ transplantations. This work can be divided into three main streams. The first consists of studies that performed conventional statistical methods (e.g. survival analysis and regression models) to predict a patient's survival after a transplant and to identify the most important predictors for survival. This group can also be divided into three subcategories such as; 1) studies examining the effect of a single variable on survival after a transplant [14, 15] , 2) studies investigating the effect of more than one variable by employing conventional statistical approaches [16-19] and, finally 3) studies that aims at creating risk scores/levels by using the effect of multiple variables on the outcome after transplants .

Studies in the second stream utilize data mining (analytics) methods to predict survival. The main difference between these two streams can be seen as a philosophical difference between both streams [20]. Statistical methods focus on the efficient collection of data for the purpose of answering specific questions. On the other hand, the data analytic approaches aim at finding unsuspected relationships and patterns from collected data. To exemplify, Kusiak *et al. [9]* compared the results of two rule-based data mining techniques, decision trees and rough sets, for predicting the survival time of kidney dialysis patients. Decision trees (DT) were also used by Nakayama *et al.* [21] to predict the prognosis of acute liver failure in an effort to improve the indication criteria for liver transplants. Kaplan and Schold [22] compared the power of an artificial neural network (ANN) and a statistically derived nomogram in predicting the 5-year graft survival after kidney transplants based on demographic, clinical and pharmaceutical data. Oztekin *et al.* [11] compared the predictive performance of ANN, logistic regression, support vector machines (SVM) and decision tree for long-term survival after a thoracic transplantation. Recently, Dag et *al.* [8] has employed a probabilistic data mining approach (i.e. Tree –augmented Naïve Bayes

(TAN) algorithm) to; 1) obtain patient-specific survival risk score for the patients who underwent heart transplant, and 2) to investigate the conditional interdependencies between the predictors. In general, it can be concluded that all the aforementioned data-driven studies have uncovered useful hidden patterns and information embedded in the large transplantation datasets.

The third, and the final, research stream investigated the dynamic effect of significant variables on the survival outcome. Most of the work within this stream is based on statistical techniques [23] [24] [25], with the exception of the study performed by Lin *et al.* [10] who determined the effect of the predictors on the 1- to 7-year survival by using logistic regression, CPHM, and multiple-output ANNS.

As can be inferred from the brief discussion on the literature, existing studies have employed data analytical approaches (both machine learning and conventional statistical methods) to predict the survival outcome, with including all of the variables (that exist in the dataset) in their analysis. A potential improvement area would be to investigate the effect of the newly added variables within the predictive analytics frame. Therefore, the main goal of this study is to evaluate the contribution of these newly added variables (that are added to the system in 2004) in predicting the survival outcome after heart transplants. Particularly, a data analytical methodology is employed that aims at investigating; 1) which of these new variables are selected by powerful variable selection algorithms, 2) whether there is any significant and consistent improvement in the predictability of the outcomes, when these variables are included in the prediction analysis, and 3) the contribution of these variables in predicting the outcome by applying sensitivity analysis. The remainder of this paper is organized as follows. In Section 4.3, the overall data analytics methodology used in this study as well as the data source and preparation are described. The results obtained from the actual data analysis are provided and discussed in Section 4.4. In

Section 4.5, conclusions are summarized and some thoughts on the direction of future research are offered.

4.3 Methodology

In this study, a comprehensive data analytical methodology that consists of four sequential phases is offered. The first phase includes 3 steps, where in the first step, the transplants that happened before 2004 are discarded from the dataset. The rationale behind that is; the overall goal of the proposed study is to measure the effect of the variables that were added to the UNOS dataset in 2004. This leads us to perform two separate analysis; 1) with including, and 2) excluding the variables that were added in 2004. (Throughout the paper, we call these group of variables "combined" variables and "old" variables, respectively, for simplicity). After obtaining the dataset, the dataset was cleaned in a systematic way in step 1. In step 2, the entire set of variables are differentiated as; 1) old and 2) combined variables. In step 3, three different datasets were extracted from each dataset to evaluate the effect of the new variables in different time horizons (1- month, 1- and 5-years) via inclusion criteria. In Phase 2, well-known and suitable feature selection algorithms were applied to select the most important variables to both datasets. Phase 3 represents the process in which different data mining classification algorithms were applied by deploying the most important features that were selected in the previous phase. In addition, the performance of these models were assessed and the unsatisfactory ones were discarded in the same phase. In the final phase (Phase 4), the contribution of the new variables were measured via sensitivity analysis. This process is repeated for both 1-month, 1- and 5-year analysis. More detailed explanation about the procedure is provided in the following related sections.

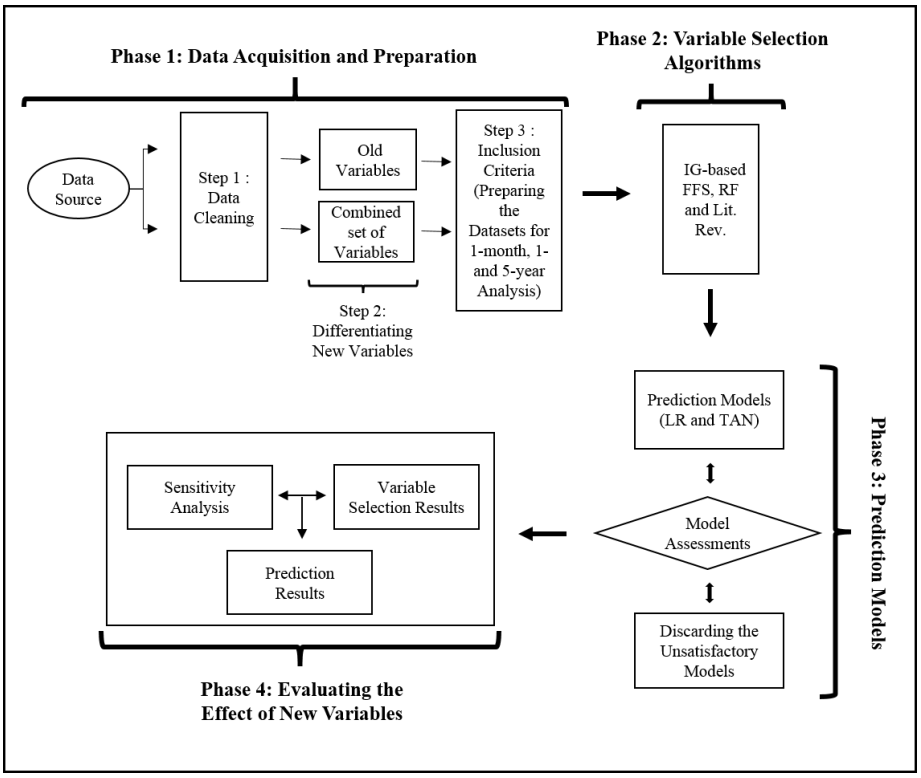

Figure 4.1: An Overview of the Proposed Methodology

4.3.1 Data Acquisition and Preparation

The UNOS dataset contains information about the heart transplantation cases that have been performed between 0ctober 1, 1987 and December 31, 2012. It includes clinical and demographical variables about both donors and recipients as well as the transplant procedure itself. These variables can be grouped into three main categories such as; 1) pre-operative variables which includes *age, gender, blood type* etc., 2) the surgical procedural variables which includes if there is a *chronic steroid use* at transplant, or if the recipient is on any type of life support at the time of transplant and, 3) post-operative variables such as if the patient has *acute rejection* or *length of hospital stay* after the operation etc. The dataset was de-identified by UNOS prior to being received by our research team.

4.3.1.1 Data Cleaning and differentiating the newly added variables

UNOS has made vital changes in the number of variables in its Organ Transplantation Databases since October 1, 1987, in order to store more information that might in turn can be used by medical research community. For instance, there are 165, 85 and 105 variables, which have been added to the Thoracic Transplant database in 1994, 1998 and 2004, respectively. As discussed earlier, our end goal in this proposed study is to evaluate the effect of these changes (in the number of variables) in predicting the survival outcome after heart transplants. In order to meet such research goal, 2004 has been selected by the research team in order to focus one of these three different time points (i.e. 1994, 1998 and 2004). The rationale behind that is that the surgical technology that are being used for heart transplant procedures, and thus the post-transplant survival time have changed in the recent years [4]. Under such circumstances, it would be naïve to compare the prediction performances of the classification models on the survival outcomes of the transplant cases that were performed in 1994 with the ones performed in 2012. That would be more objective to compare them on the cases that were performed in a narrower time frame. Therefore, the transplants that were performed before June 30, 2004 have been eliminated from the dataset.

In this study, these datasets were cleaned in a systematic way in which the invariant variables, outliers and the variables that have no effect on the survival outcome such as patient ID, transplant ID, transplant date have been checked for and eliminated. It should be noted that the UNOS dataset has excessive number of missing records. In addition, the dataset includes excessive number of cells that are filled with "*U*" meaning "*Undefined*", which also can be considered as a missing record from the data analytic point of view. Therefore, these (*U*) values have been converted to missing values. One should be very careful in dealing with the situation where there are a lot of missing records, which in turn might lead the classification models to suffer. In addition,

the two common data cleaning approach namely *Case-wise deletion* and *Column-wise deletion* have their own disadvantages for such situations. The former one would cause a significant decrease in the number of samples, which would in turn cause the data mining models to not to be trained well enough. While the latter one would leave us with few variables, which would cause a significant information loss. Therefore, we have built a data cleaning algorithm that can be adapted for the current situation. More specifically, the algorithm checks the *columns* first (for the number of missing values for each variable in the dataset) and deletes the variable with the maximum number of missing records. Similarly, in the second step, the *rows* are checked to find and delete the *case* that has the maximum number of missing records. This procedure is repeated until there is no missing record left in both datasets.

As noted earlier, the primary objective of this study is to evaluate/measure the effect of the newly added variables in predicting the survival outcome after the heart transplants. Therefore, four post-operative variables can be employed as candidate survival outcomes of the proposed study. These variables are *gstatus* (a binary variable denoting if the graft has survived or failed at the last follow-up time), *pstatus* (a binary variable denoting if the patient has survived or failed at the last follow-up time), *gtime* (a continuous variable denoting the time range between the transplant date and graft's last follow-up/failure time, in days) and *ptime* (a continuous variable denoting the time range between the transplant date and person's last follow-up/death time, in days). Since our purpose to focus on the patients who solely died due to heart-transplantation related reasons, the *gstatus* variable along with *gtime* are the suitable outcomes for our study. The *gstatus* is selected as our main target, however *gtime* will also be incorporated throughout the study in that it enables is to apply the data inclusion criteria, which is described in the subsequent section in detail. Therefore, the *pstatus* and *ptime* variables are also eliminated from the datasets. Having

applied such cleaning process have left us with **15,263** cases and **155** variables of which **12** are new and **143** are old variables.

Table 4.1: 12 variables added to UNOS heart transplant databases after 2004 (after cleaning)

Variable Name	Description
ARGININE_DON	Deceased donor-was donor given arginine vasopressin
BMI_TCR	Bmi at listing
CDC_RISK_HIV_DON	Does the deceased donor meet cdc guidelines for high risk for an organ donor
HEMATOCRIT_DON	Hematocrit:
INHALED_NO_TCR	Tcr patient on life support://inhaled no
INSULIN_DON	Deceased donor-was donor given insulin?
PCO2_DON	Pco2
PH_DON	Blood ph
PROSTACYCLIN_TRR	Candidate prostacyclin inhalation
TRANSFUS_TERM_DON	Number of transfusions during this (terminal) hospitalization:
VAD_DEVICE_TY_TCR	Candidate type of vad device at listing
VAD_DEVICE_TY_TRR	Vad_device type

4.3.1.2 Data Inclusion criteria

In this study, three different time points are determined (i.e. 1-month, 1- and 5-years), in order to model the classification algorithms. There are several reasons for us to choose these three specific time points. One of the primary reasons is that these time points are commonly studied in the existing heart transplant literature [18, 23, 26-29]. In addition, the reason that we have not included larger time horizons such as 6-, 7-, 8-, 9- or 10-years analysis is due to the limitation on the sample size, since we only include the transplants performed after 2004. Another important reason to study especially 1- and 5-year time points is that they are considered as the reference time points in providing the common survival rate statistics after the cardiac transplant [4].

For each time point that is used in this study (1-month, 1- and 5-year), some of the cases in the dataset are excluded with the following procedure. For a case to be excluded, the patient

should be alive on the last follow-up date and the last follow-up date should be less than the number of days required (1* 365 = 365 days for 1 year analysis) for the analysis. The reason for those patients to be eliminated from the analysis is that it is not known whether they will still be alive or not, after the required day. In addition, the *gtatus* values of the patients who have died after the number of days required should be changed from "*1*" to "*0*". The reason behind that is, the patient have died sometime after the threshold value. However, from a classification point of view, she/he has succeeded by expiring the predetermined time point. This inclusion procedure can be represented by the following pseudo code;

If $(gstatus = 0 \ and \ gtime < \#\ of\ \ days\ required)$ **then** Exclude the case
Elseif $(gstatus = 1 \ and \ gtime > \#\ of\ \ days\ required)$**then** gstatus = 0

After applying the inclusion procedure, a summary of distribution of survivals, failures and excluded cases are represented in Table 4.2. It should be noted that the sum of survivals, failures and *excluded observations* for each row is **15,263**, as it was discussed in Section 4.3.1.1.

Table 4.2: Number of survivals, failures, and excluded observations over the time-points

Time Point	Included		Excluded
	Survivals	Failures	
1-month	14,160	683	420
1-year	11,054	1,601	2,608
5-year	3,148	2,761	9,354

4.3.2 Variable Selection

As discussed in Section 4.3.1.1, after the cleaning procedure, the dataset includes 155 in total (i.e. 143 existing and 12 new variables). Deploying hundreds of variables (as in this case) into any prediction model would both computationally be expensive and carry a high risk of overfitting. In such situations, feature selection methods can be employed in selecting the suitable subset of the existing variables. Existing feature selection methods can broadly be categorized into two groups i.e. filters and wrappers [30, 31]. Wrapper methods are specifically dedicated to a specific type of prediction method. They rely on the performance of one type of classifier to evaluate the quality of a set of features. On the other hand, the filter methods can be considered as agnostic models in that they rely on general characteristics of the training data to select some features without involving any learning algorithm. The filter methods rank features according to their individual predictive power, which can be estimated by various means such as Fisher score, *Kolmogorov-Smirnov test*, *Pearson correlation* or *mutual information* [32]. Both approaches have their own advantages, which depend on the number of features, samples and the existence of complex nonlinear relations among the variables. In our analysis, we applied a hybrid variable selection approach which consists of both a wrapper and filter-based feature selection method to extract the most important features. In addition to the selected variables (by feature selection algorithms), the variables, which have been found important in the existing literature, have also been added to the final set of predictors. It should be noted that the selected variables will be employed in the classification algorithms in Phase 3. What follows is a brief description of the variable selection algorithms.

4.3.2.1 Fast Feature Selection (FFS) via Information Gain Analysis

In prediction models, a feature can generally be considered as "good" if it is relevant to the target but is not redundant to any of the other predictors in the dataset. If the correlation between two variables is adopted as a "goodness" measure, it can be concluded that a feature can be selected if it has high enough correlation with the target (which makes it "predictive of" the class) and a low enough correlation with any other variables (which is not above a certain threshold level). In this sense, a suitable measure of correlation between features needs to be employed. The existing correlation measures can be mainly classified into two groups as; 1) based on classical linear correlation and, 2) based on information theory [33]. However, in the situations where there are hundreds of variables (as in our case), it would be very naïve to assume linear correlation between features. Therefore in this study, a correlation measure that is based on the information-theoretical concept of entropy is adopted. The entropy of a variable *x* is defined as

$$H(x) = -\sum_i P(x_i).\log_2(P(x_i)) \tag{4.1}$$

and the entropy of a variable *(x)* can be defined as

$$H(x) = -\sum_i P(x_i).\log_2(P(x_i)) \tag{4.2}$$

Where the prior probabilities of variables are defined as P(x), and posterior probabilities are defined as P*(xi |yi*) for given values of *y* . The additional information gained through *y* is reflected by the amount of decrease in the entropy value of x , which in turn is called; information gain (Quinlan, 1993), given by

$$IG(X \mid Y) = H(X) - H(X \mid Y) \tag{4.3}$$

It can be concluded from such measure that, a feature *y* is regarded more correlated to feature *x* than to feature *z*, if $IG(x \mid y) > IG(z \mid y)$. In other words, the features are selected based on the increase in the amount of information (gained) by knowing the value of the attribute. The reader is referred to [33] for a more detailed explanation of the fast feature selection algorithm that is employed in this study.

4.3.2.2 Random Forests

Random Forests method is a data mining algorithm that have been commonly used for both classification and regression problems [34] in many different fields as well as for variable selection purposes in the recent literature.

The prediction decision for the random forests is made based on either the *mode* (for classification) or the *mean* (regression) of the aggregated votes that are obtained by a collection of decision trees. Each tree in the forest is trained via a bootstrap sample of individuals. The construction procedure of each tree can be summarized through the following steps [34];

Suppose there are *I* individuals and *M* explanatory attributes in the dataset,

1.) A random sample whose size is equal to the training set is selected from the entire data.
2.) At each node in the tree, *m* random attributes out of *M (*entire set of attributes in the data*)* is selected, where *M/m is* relatively much greater than *1*.
3.) The best split is then chosen from the subset of *m* attributes selected in the previous step.
4.) Second and third steps are iterated until there is fully grown tree is built (no pruning).

4.3.3 Prediction Models

In this study, two popular machine learning-based classification algorithms (i.e. SVM and TAN) along with LR have been employed to predict the binary outcomes. These classification algorithms have been selected due to their superior performance in our preliminary analysis as well as their superior performance in the recently published transplantation literature.

4.3.3.1 Tree Augmented Naïve (TAN) Bayesian Belief Network

A Bayesian Belief Network is a directed acyclic graph by which a stochastic dependency model is represented. In such graph, the nodes represent the predictors while the conditional dependencies among the predictors are represented by the arcs [35]. It is commonly used for the situations where there might be complex nonlinear interactions among the predictors [36]. Tree Augmented Naïve Bayes (TAN) is a branch of Bayesian family where the target does not have any parents, while it is one of the parents of each predictor along with another variable. More specifically, a predictor variable might have at most two parents of which one of them is the target/outcome [37]. In TAN model, an arc from X_1 to X2 means that, the contribution of X2 in predicting the target outcome is dependent on the value that X1 gets. The classification decision is made via Bayes rule where, for each predictor, the outcome probabilities are calculated. In the final phase, among these probabilities, the highest one is chosen for the structure [38]. For a more detailed description on TAN, we refer to reader to Dag et *al*. [8].

4.3.3.2 Logistic Regression

Logistic regression is a standard regression method that are employed for either binary or multinomial classification problems. In such setting, the log odds of the dependent variable are modeled by linear combination of the independent predictors [39]. Logistics regression has been

very commonly used for classification problems in the relevant medical literature [10, 40, 41] due to its "easy to interpret" structure.

4.4 Results and Discussion

4.4.1 Variable Selection Results

Based on the discussion in Section 4.3.2, two feature selection models (i.e. IG-FFS and RF) were applied to select a suitable subset of both existing and newly added preoperative variables, separately. Therefore, the first and second set of predictors were selected via an Information gain-based fast feature selection algorithm and random forest algorithm, respectively. Also, variables that are selected through literature review analysis have been added to this subset. In essence, this subsection contains the results from Phase 2 of the proposed hybrid method. It should be noted that three different time points (i.e. 1-month, 1-year and 5-years) have been investigated to see whether the newly added variables have any significant contribution in predicting the *gstatus* in different time horizons. The number of the features that were found to be "good" via the feature selection algorithms as well as the variables that were found to be (potentially) important in the existing literature are presented in Table 4.3.

Table 4.3: The number of the features selected through variable selection methods and literature review

	IG-FFS		**RF**		**Lit. Rev**		**Total**	
	Old	**New**	**Old**	**New**	**Old**	**New**	**Old**	**New**
1- month	4	2	60	6	10	6	65	11
1-year	3	2	58	6	38	7	70	12
5-year	9	2	38	5	30	8	55	9

Based on Table 4.3, it can be seen that there are 2,6 and 6 variables (14 in total) selected through IG-FFS, RF and literature review analysis, respectively. Because of the commonality among these variable set there are only 11 variables (in total) to be deployed to prediction algorithms. Similar pattern can be recognized for 1- and 5-year analysis as well. It should be noted that, there are 76 (65 + 11), 82 ((70+12) and 64 (55+9) variables to be deployed in the final 1-month, 1-year and 5-year prediction models, respectively.

Table 4.4: Variables that are selected through different time points

	1- month				1- year				5- year			
	IG-FFS	RF	Lit. Rev	Inclusion	IG-FFS	RF	Lit. Rev	Inclusion	IG-FFS	RF	Lit. Rev	Inclusion
ARGININE_DON	×	×	√	√	×	×	√	√	×	×	√	√
BMI_TCR	×	×	×	×	×	×	√	√	×	√	√	√
CDC_RISK_HIV_DON	×	×	√	√	×	×	√	√	×	×	√	√
HEMATOCRIT_DON	×	√	×	√	×	√	×	√	×	√	×	√
INHALED_NO_TCR	×	×	√	√	×	×	√	√	×	×	√	√
INSULIN_DON	√	×	×	√	√	×	×	√	×	×	×	×
PCO2_DON	×	√	×	√	×	√	×	√	×	×	×	×
PH_DON	×	√	×	√	×	√	×	√	×	×	×	×
PROSTACYCLIN_TRR	×	×	√	√	×	×	√	√	×	×	√	√
TRANSFUS_TERM_DON	×	√	×	√	×	√	×	√	√	√	√	√
VAD_DEVICE_TY_TCR	×	√	√	√	×	√	√	√	×	√	√	√
VAD_DEVICE_TY_TRR	√	√	√	√	√	√	√	√	√	√	√	√

In Table 4.4, the new variables, which were selected through the comprehensive selection procedure, are presented for each time point in detail. The rationale behind presenting only the new variables (but not the old ones) in Table 4.4 is that the main goal of this study is to specifically focus on the newly added variables. In this table × and √ denote whether the variable is selected through the associated variable selection model or not (with × representing "No" and √ representing "Yes"). It should be noted that as long as a feature is selected through any of the

selection methods, it should be included in the prediction models (in the next phase). Therefore, column "Inclusion" denotes whether the variables will be included in the prediction models or not. As it also can be confirmed by Table 4.3, there are 11, 12, and 9 variables that will be included (selected through at least one of the feature selection methods) in the 1-month, 1-year and 5-year time predictions, respectively. In the next subsection, the prediction (classification) results obtained through employing the selected (old + new) variables for three different time points.

4.4.2 Prediction Results

Based on the discussion in Section 4.3.3, TAN and Logistic R. were employed to predict the *gstatus* (outcome) for three different time points (i.e. 1-month, 1-year and 5-years) by both excluding and including the newly added variables in the prediction models by also employing 5-fold cross validation concept. The rationale behind employing cross validation concept is to decrease the bias in the outcomes as well as to increase the robustness of the classification models [45]. Therefore, 2 (classification models) X 2 (for old and combined set of variables) = 4 different models that will be employed for three different time points. The description of each models are provided by Table 4.5.

Table 4.5: The list of the prediction models used for each time-period

Model Name	Description
Old.LR	Logistic regression by including only the old variables
Combined.LR	Logistic regression by including both new and old variables
Old.TAN	TAN by including only the old variables
Combined. TAN	TAN by including both new and old variables

.

The prediction results that were obtained through using these models for each time period are presented in Table 4.6. *AUC (area under the, accuracy*, *recall*, and *specificity* metrics were presented for all models as the evaluation criteria. For a more detailed description of these criteria, the reader is referred to Genc et *al. [37]*. *AUC* will be used as the main evaluation criterion among these metrics since it is considered to be an objective criterion specifically in evaluating the classification performances for imbalanced datasets [46, 47] (as in our case).

The highest mean for a given metric is underlined in Table 4.6..One of the several interesting observations that can be made is that Logistic Regression models have outperformed the TAN models for both inclusive (combined) and exclusive (old) models, in almost all of the three different time points (e.g. Old_LR > Old_TAN, and Combined_LR > Combined_TAN). Another interesting pattern is that the AUC metrics for both 1-month and 5-years have increased with the inclusion of the newly added variables (e.g. Combined_TAN > Old_TAN and Combined_LR > Old_LR). Same pattern does not exist in the 1-year survival prediction.

Table 4.6: Prediction results obtained through including and excluding the newly added variables

		# of variables	AUC	Accuracy	Recall	Specificity
1_Month	**Old_LR**	65	0.662	0.643	0.590	0.645
	Combined_LR	76(65+11)	**0.680**	**0.650**	**0.619**	**0.651**
	Old_TAN	65	0.621	0.632	0.529	0.637
	Combined_TAN	76(65+11)	0.639	0.629	0.551	0.633
1_Year	**Old_LR**	70	**0.645**	**0.646**	0.544	**0.661**
	Combined_LR	82(70+12)	0.641	0.644	0.542	0.659
	Old_TAN	70	0.622	0.618	**0.550**	0.628
	Combined_TAN	82(70+12)	0.619	0.616	0.536	0.627
5_years	**Old_LR**	55	0.658	0.612	**0.595**	0.627
	Combined_LR	64 (55+9)	**0.663**	**0.623**	0.520	**0.713**
	Old_TAN	55	0.647	0.606	0.594	0.616
	Combined_TAN	64 (55+9)	0.656	0.618	0.540	0.686

As discussed in the previous section, one of the main goals of this study investigate whether the newly added variables increase the prediction power of the algorithms or not. Therefore, the prediction powers of the models (in terms of *AUC*) can be compared in order to come up with a reasonable answer for such question. As can be seen from Table 4.6, both 1-month and 5-year survival prediction capability (AUC) of the both models have increased. In 1-month survival prediction, the amount of increase in the AUC values for Logistic Regression and TAN models are both 0.018., while the difference between the prediction capabilities are less in the 5-year survival prediction. On the other hand, a slight decrease exists in the performances of both LR (0.645 to 0.641) and TAN models (0.622 to 0.619) in 1-year survival prediction.

Hereafter, Logistic Regression will be used as a reference model for further analysis in the next subsection (sensitivity analysis) since it has outperformed not only the TAN model (in our analysis) but also other classification algorithms employed in the preliminary analysis of tour study.

4.4.3 Sensitivity Analysis Results

In this subsection, sensitivity analysis is performed on the Logistic regression model (only) since it has been selected as the reference model due to its superior performance over TAN, in Section 4.4.2. Sensitivity analysis is an alternative common measure (in data mining community) [48] that enables to discover the relative contribution of each variable towards predicting the outcome variable (*gstatus*). In such setting, a predictive variable is found to be important if its addition results in improving the model's prediction performance, as measured by the *AUC* metric. According to the sensitivity measure defined by Saltelli [48], the sensitivity measure is defined as

$$S_i = \frac{V_i}{V_y} = \frac{V(E(y/x_i))}{V(y)} \tag{4}$$

where *y* is the dichotomous output variable (*gstatus*), and the unconditional output variance is denoted by *V(y)*. The expectation operator is denoted by *E*, which calls for an integral over all predictor variables except x_i. A further integral operator is implied over x_i by the operator V_i. The importance of a specific variable is then computed as the normalized sensitivity.

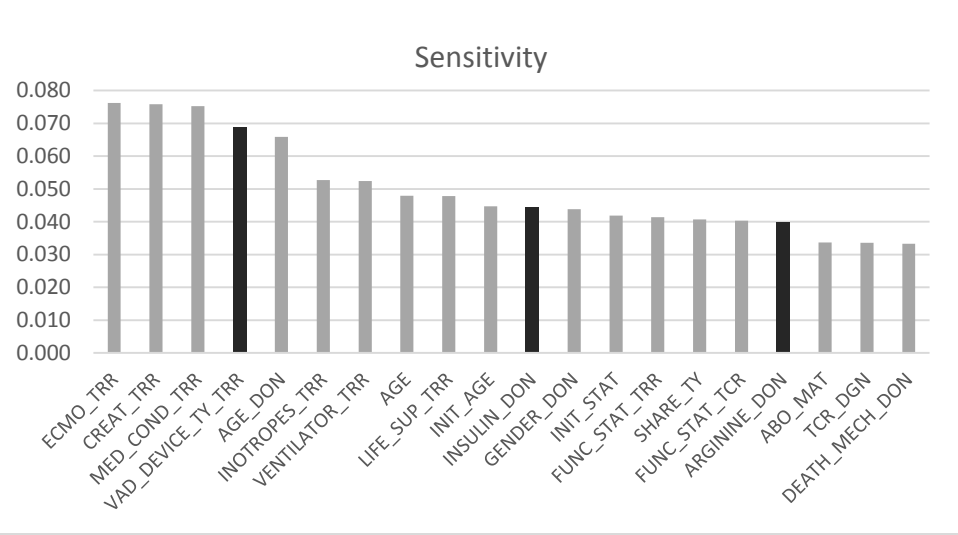

Figure 4.2. Top 20 most contributory predictors for 1-month survival prediction

Figure 4.2 represents the 20 variables that are found to be the most contributory ones in predicting the 1- month survival after cardiac transplant. The variables that are highlighted with black (in the chart below) refers to the variables that have been added to the UNOS dataset after 2004(new variables). As can be seen from Figure 4.2, 3 out of 20 and 1 out of 5 variables are the newly added ones. In addition *VAD_DEVICE_TY_TRR*, which is among the most important 5 variables were selected by the agreement of all of the three variables selection methods employed

(Section 4.4.1.) in our study (i.e. *IG-based FFS, RF* and *Literature Review*), while *INSULIN_DON* and *ARGININE_DON* were selected by only one variable selection method.

Figure 4.3 represents the 20 variables that are found to be the most contributory ones in predicting the 1-year survival. As can be inferred from Figure 4.3, *VAD_DEVICE_TY_TRR and INSULIN_DON* were selected identified to be among top 10 and top 20 variables, respectively. It should also be noted that, *VAD_DEVICE_TY_TRR* were selected by the consensus of all of the three variables selection mechanisms while *INSULIN_DON* was selected by only IG-FFS.

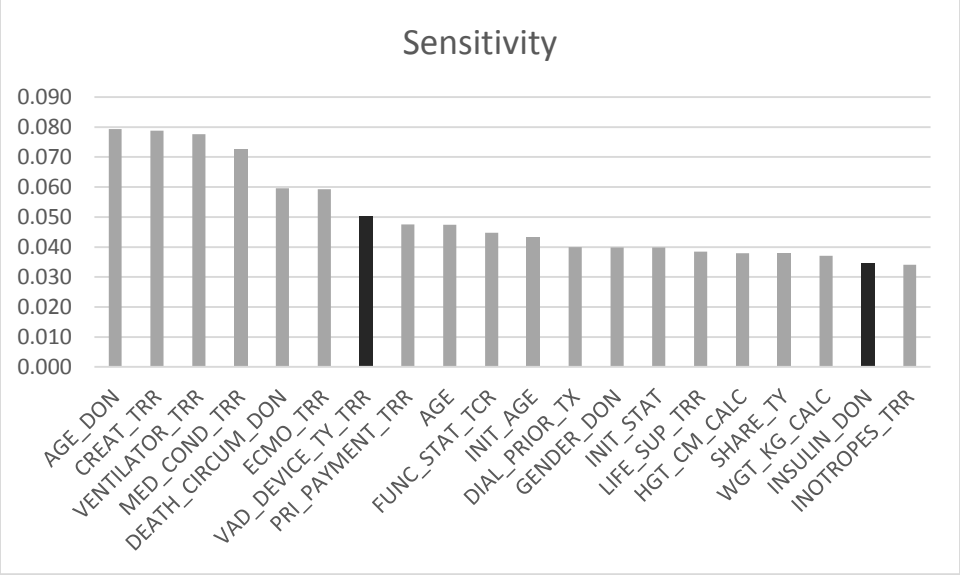

Figure 4.3. Top 20 most contributory predictors for 1-year survival prediction

Similarly, Figure 4.4 represents the 20 variables that are found to be the most contributory ones in predicting the 5-year survival. As can be inferred, *VAD_DEVICE_TY_TRR and INHALED_NO_TCR* were selected to be among top 10 and top 20 variables, respectively. It should also be noted that, *VAD_DEVICE_TY_TRR* were selected commonly by all of the three variables selection methods while *INHALED_NO_TCR* was selected through reviewing the literature.

Based on the sensitivity analysis employed for 1- month, 1- and 5-year analysis, it can be concluded that, the newly added variables have more contribution in predicting the 1-month than the 5-year survival. The rationale behind that may be *VAD_DEVICE_TY_TRR was* found to be the fourth most important variable in predicting the outcome. Therefore, including such important variable in the prediction procedure might have boosted the AUC rate for 1-month prediction, while same importance level have not been witnessed in 1- and 5-year survival prediction. This finding somewhat matches with the results presented in Table 4.6, in Section 4.4.2. Recall that, 1-month time frame is the only one among three different time points, where the AUC values of Logistic regression of all of the four evaluation metrics increase when the newly added variables are included in the prediction algorithm.

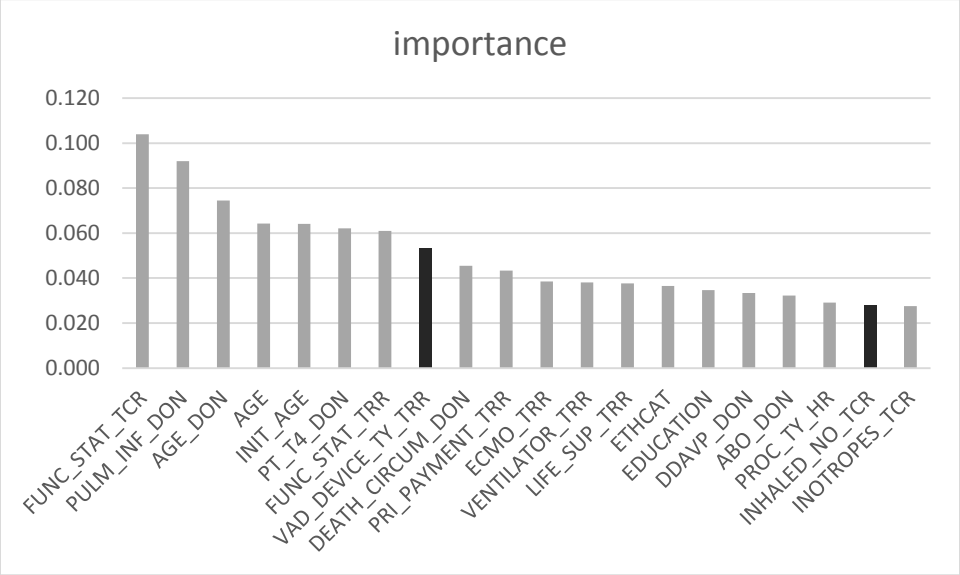

Figure 4.4. Top 20 most contributory predictors for 5-year survival prediction

4.5 Conclusion and Future Recommendations

The main objectives of this paper was to evaluate the effect of the variables to the UNOS heart transplant databases after June 30, 2004. Our methodology has 5 phases, which includes data preparation, feature selection procedure, prediction models and assessment, and evaluation phase. In essence, the ultimate goal of the current study can be reached via answering the following critical questions, which provide us with different perspectives to evaluate the effect of these new variables;

a) Which of the newly added variables were found to be important by the feature selection methods?

b) Do the power of prediction models (prediction performance) improve by including the new variables that were selected through the variable selection mechanism?

c) What is the contribution of each of these (new) variables in predicting the *gstatus* for different time horizons?

Therefore, three main perspectives have been constructed to answer the above research questions as 1) Feature selection perspective, which employs both Information gain theory and data mining based wrapper method, 2) Prediction improvement perspective, which includes two data analytical approaches and 3) Predictor contribution perspective, which employs a sensitivity analysis approach.

Based on the proposed framework, three sets of newly added variables (for three different time points) are selected through feature selection algorithms to be deployed into data analytical models. Then, a popular probabilistic data mining-based approach and a logistic regression have been employed to see whether there is any significant improvement in the prediction results. In the

final component, the selected features are also investigated through sensitivity analysis to measure their individual contribution to the prediction model.

In summary, a novel data analytical framework is proposed in this study to evaluate the effect of the variables that have been added to the UNOS heart transplant datasets after June 30, 2004. In this framework, three different perspectives have been built; 1) Feature selection, 2) Prediction perspective and 3) Sensitivity analysis perspective. An important outcome of this comprehensive data analytical approach can be extended to other organ transplant dataset, which are provided by UNOS.

Acknowledgements

This work was supported in part by *Health Resources and Services Administration* contract 234-2005-370011C. The content is the responsibility of the authors alone and does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government

4.6 References

1. *What is Heart Failure ?* 2012 [cited 2012; Available from: http://www.nhlbi.nih.gov/health/health-topics/topics/hf/.
2. López-Sendón, J., *The heart failure epidemic.* Medicographia, 2011. **33**(4): p. 363-9.
3. *What Is Heart Transplant?* . 2012 [cited 2012; Available from: http://www.nhlbi.nih.gov/health/health-topics/topics/ht/.
4. *What to Expect Before a Heart Transplant*. 2012 [cited 2012; Available from: http://www.nhlbi.nih.gov/health/health-topics/topics/ht/before.html.
5. Carmona, M., et al., *Heart failure in the family practice: a study of the prevalence and co-morbidity.* Family Practice, 2011. **28**(2): p. 128-133.
6. Healy, D., et al., *Heart transplant candidates: factors influencing waiting list mortality.* Irish Medical Journal, 2004. **98**(10): p. 235-237.
7. *UNOS | About Us*. 2014 [cited 2014.
8. Dag, A., et al., *A probabilistic data-driven framework for scoring the preoperative recipient-donor heart transplant survival.* Decision Support Systems, 2016. **86**: p. 1-12.
9. Kusiak, A., B. Dixon, and S. Shah, *Predicting survival time for kidney dialysis patients: a data mining approach.* Computers in Biology and Medicine, 2005. **35**(4): p. 311-327.
10. Lin, R.S., et al., *Single and multiple time-point prediction models in kidney transplant outcomes.* Journal of Biomedical Informatics, 2008. **41**(6): p. 944-952.
11. Oztekin, A., D. Delen, and Z.J. Kong, *Predicting the graft survival for heart–lung transplantation patients: An integrated data mining methodology.* International Journal of Medical Informatics, 2009. **78**(12): p. e84-e96.
12. Oztekin, A., Z.J. Kong, and D. Delen, *Development of a structural equation modeling-based decision tree methodology for the analysis of lung transplantations.* Decision Support Systems, 2011. **51**(1): p. 155-166.
13. Sousa, F., et al. *Application of the intelligent techniques in transplantation databases: a review of articles published in 2009 and 2010*. in *Transplantation Proceedings*. 2011. Elsevier.
14. Al-Khaldi, A., P.E. Oyer, and R.C. Robbins, *Outcome analysis of donor gender in heart transplantation.* The Journal of Heart and Lung Transplantation, 2006. **25**(4): p. 461-468.

15. Gupta, D., et al., *Effect of older donor age on risk for mortality after heart transplantation.* The Annals of Thoracic Surgery, 2004. **78**(3): p. 890-899.
16. Del Rizzo, D.F., et al., *The role of donor age and ischemic time on survival following orthotopic heart transplantation.* The Journal of Heart and Lung transplantation, 1999. **18**(4): p. 310-319.
17. Drakos, S.G., et al., *Multivariate predictors of heart transplantation outcomes in the era of chronic mechanical circulatory support.* The Annals of Thoracic Surgery, 2007. **83**(1): p. 62-67.
18. Kilic, A., et al., *Factors associated with 5-year survival in older heart transplant recipients.* The Journal of thoracic and cardiovascular surgery, 2012. **143**(2): p. 468-474.
19. Tjang, Y.S., et al., *Survival analysis in heart transplantation: results from an analysis of 1290 cases in a single center.* European Journal of Cardio-Thoracic Surgery, 2008. **33**(5): p. 856-861.
20. Hand, D.J., *Data mining: Statistics and more?* The American Statistician, 1998. **52**(2): p. 112-118.
21. Nakayama, N., et al., *Algorithm to determine the outcome of patients with acute liver failure: a data-mining analysis using decision trees.* Journal of Gastroenterology, 2012. **47**(6): p. 664-677.
22. Kaplan, B. and J. Schold, *Transplantation: neural networks for predicting graft survival.* Nature Reviews Nephrology, 2009. **5**(4): p. 190-192.
23. Zuckermann, A.O., et al., *Pre-and early postoperative risk factors for death after cardiac transplantation: A single center analysis.* Transplant International, 2000. **13**(1): p. 28-34.
24. Boschiero, L., et al., *An objective method for detecting time-dependent effects in graft survival.* Transplant International, 2000. **13**(S1): p. S112-S116.
25. Aydemir, Ü., S. Aydemir, and P. Dirschedl, *Analysis of time-dependent covariates in failure time data.* Statistics in Medicine, 1999. **18**(16): p. 2123-2134.
26. Brieke, A., et al., *Influence of donor cocaine use on outcome after cardiac transplantation: analysis of the United Network for Organ Sharing Thoracic Registry.* The Journal of Heart and Lung Transplantation, 2008. **27**(12): p. 1350-1352.
27. Gasink, L.B., et al., *Hepatitis C virus seropositivity in organ donors and survival in heart transplant recipients.* Jama, 2006. **296**(15): p. 1843-1850.

28. Sarris, G.E., et al., *Cardiac transplantation: the Stanford experience in the cyclosporine era.* The Journal of thoracic and cardiovascular surgery, 1994. **108**(2): p. 240-51; discussion 251-2.
29. Stehlik, J., et al., *Interactions among donor characteristics influence post-transplant survival: a multi-institutional analysis.* The Journal of Heart and Lung Transplantation, 2010. **29**(3): p. 291-298.
30. Das, S. *Filters, wrappers and a boosting-based hybrid for feature selection*. in *ICML*. 2001. Citeseer.
31. Kohavi, R. and G.H. John, *Wrappers for feature subset selection.* Artificial intelligence, 1997. **97**(1): p. 273-324.
32. Fleuret, F., *Fast binary feature selection with conditional mutual information.* Journal of Machine Learning Research, 2004. **5**(Nov): p. 1531-1555.
33. Yu, L. and H. Liu. *Feature selection for high-dimensional data: A fast correlation-based filter solution*. in *ICML*. 2003.
34. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
35. Pearl, J., *Bayesian networks: A model of self-activated memory for evidential reasoning*. 1985: University of California (Los Angeles). Computer Science Department.
36. Michalski, R.S., *Understanding the nature of learning: Issues and research directions.* Machine learning: An artificial intelligence approach, 1986. **2**: p. 3-25.
37. Genc, O. and A. Dag, *A Bayesian network-based data analytical approach to predict velocity distribution in small streams.* Journal of Hydroinformatics, 2015: p. jh2015110.
38. Friedman, N., D. Geiger, and M. Goldszmidt, *Bayesian network classifiers.* Machine learning, 1997. **29**(2-3): p. 131-163.
39. Hosmer Jr, D.W., S. Lemeshow, and R.X. Sturdivant, *Logistic Regression for Matched Case-Control Studies.* Applied Logistic Regression, Third Edition, 2013: p. 243-268.
40. Khedmat, H., et al. *A logistic regression model for predicting health-related quality of life in kidney transplant recipients*. in *Transplantation proceedings*. 2007. Elsevier.
41. Tollemar, J., et al., *Variables predicting deep fungal infections in bone marrow transplant recipients.* Bone marrow transplantation, 1989. **4**(6): p. 635-641.
42. Gunn, S.R., *Support vector machines for classification and regression.* ISIS Technical Report, 1998. **14**.

43. Hodge, V.J. and J. Austin, *A survey of outlier detection methodologies.* Artificial Intelligence Review, 2004. **22**(2): p. 85-126.
44. Han, J., M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2011: Elsevier.
45. Refaeilzadeh, P., L. Tang, and H. Liu, *Cross-validation*, in *Encyclopedia of database systems*. 2009, Springer. p. 532-538.
46. Chawla, N.V., *Data mining for imbalanced datasets: An overview*, in *Data mining and knowledge discovery handbook*. 2005, Springer. p. 853-867.
47. Kotsiantis, S., D. Kanellopoulos, and P. Pintelas, *Handling imbalanced datasets: A review.* GESTS International Transactions on Computer Science and Engineering, 2006. **30**(1): p. 25-36.
48. Saltelli, A., et al., *Sensitivity analysis in practice: a guide to assessing scientific models*. 2004: John Wiley & Sons.

5 Conclusions and Summary of Dissertation Contributions

The data analytic studies as well as the decision support tool that are provided in this dissertation fill critical gaps in the heart transplantation field. It makes several contributions to the transplantation literature which include: 1) Identifying the predictive factors for short-, mid- and long-term survival after the heart transplant, as well as their time-dependent effects on the given follow-up time point, which in turn allows us differentiate the factors whose effect change over time , 2) developing a decision support tool that provides the patient-specific failure risk score based on the values of the relevant preoperative predictors, as well as to investigate the conditional relations among the important predictors of long term survival after heart transplants and 3) investigating the effect of recently added variables (to the UNOS dataset) in predicting the survival outcome after heart transplant.

Therefore, the overall goal of this dissertation is to fill a critical knowledge gap in heart transplantation field. It consists of two components. In the first component, which is introduced in Chapter 2, a hybrid data analytical framework has been developed to investigate the factors whose importance change over the time. In the second component (introduced in Chapter 3), a decision support tool is introduced. Such tool can be used by medical practitioners to calculate the patient specific survival risk score of transplant patient, without having any knowledge about the data mining field. In Chapter 4, the effect of the new variables in the predictability of the survival outcome after heart transplantations have been investigated. The procedure that are employed in all of the three components can also be applied for other type of organ transplants.