DATA MINING-BASED SURVIVAL ANALYSIS AND SIMULATION MODELING FOR LUNG TRANSPLANT

By

ASIL OZTEKIN

Bachelor of Science in Industrial Engineering
Yildiz Technical University
Istanbul, Turkey
2004

Master of Science in Industrial Engineering
Fatih University
Istanbul, Turkey
2006

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
December, 2010

DATA MINING-BASED SURVIVAL ANALYSIS AND

SIMULATION MODELING FOR LUNG TRANSPLANT

Dissertation Approved:

Dr. Zhenyu James Kong

Dissertation Adviser

Dr. Dursun Delen

Co-Adviser

Dr. Camille DeYong

Dr. William J. Kolarik

Dr. Leva K. Swim

Dr. Mark E. Payton

Dean of the Graduate College

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my academic advisor, Dr. Zhenyu (James) Kong, for his enthusiasm in academic guidance for me, inspirations, continuous motivation, and also financial support he provided me with. He has been and will continue to be a great role model for me in terms of balancing research and teaching. I also would like to extend my gratitude to my co-advisor, Dr. Dursun Delen, for his creative ideas to support my research, invaluable guidance, assistance, and friendliness. This dissertation would not be complete without their aspirations. I have always felt lucky to have worked with them. I present my since thanks to my dissertation committee member, Dr. William J. Kolarik, for his inspiring questions and suggestions throughout my research. He, as the department head, trusted in me from the day he sent the admission letter for my doctoral study in our department and supported me at all stages of my stay at OSU. I am also thankful to Dr. Camille DeYong for her patience, friendliness, and kindness, which made me feel comfortable at OSU. The last but not the least, I want to emphasize the contribution that Dr. Leva K. Swim provided for this research. She helped me get in touch with medical experts whose advice was addressed at various stages of this research. My heartfelt thanks are also to the Director of Integris Nazih Zuhdi Transplant Institute, Dr. Nicolas Jabbour, and the Chief of Lung Transplant and Pulmonary Hypertension Division, Dr. Remzi Bag. My extreme appreciation goes to Dr.

Mark Gavin and Dr. Ricki Ingalls for teaching me the methods used in this research and for their time to discuss the details of some unclear points.

I also want to express my gratitude to all of my friends in the department and in Stillwater. Without their friendship and encouragement at the time of difficulties, this study would not be completed. I will always remember this friendship and support.

Finally, special thanks go to my sister, Secil, and my parents, Ulku and Ali, for their strong encouragement, love and understanding, patience and sacrifice. Despite the distance, I always felt them overcoming the difficulties together with me. Therefore, I dedicate this dissertation to my father and mother.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

1.1. Research Motivation

As the economies in developed countries are shifting from a manufacturing base toward a service orientation, the role of the service industry has gained greater importance [1]. The healthcare sector is one of the most critical sectors in the service industry since it is life-crucial and any mistakes can cause inevitable and incurable results [2]. Improper resource allocation has been one of the perennial problems in the healthcare service industry [3]. Particularly, the allocation of “scarce” organs for organ transplantation has been one of the most critical problems faced in the healthcare service. Although organ allocation is the sole viable therapy for various end-stage diseases, often times the number of donor organs unfortunately does not meet the need [4]. Therefore, the organ-waiting patients are lined up in waiting lists whereas some of the donor organs are wasted due to suboptimal match between the donor and the recipient.

Long organ waiting lists can mainly be attributed to the following two reasons. (1) Since the success rate in the organ transplantation has increased due to the advancements in the medical field, today there are more patients asking for a transplant. On the other hand, the accessibility of the patients to transplant centers is easier than ever due to the drastic increase of these centers throughout the US. While there were only four transplant centers thirty years ago, as of 2008 there are 249 centers in US [5]. (2) Although there has been some increase in the number of donated organs, it has never reached the level of the increase in demand, which results in a shortage of donor organs. This increasing gap between the organ waiting patients and donor supplies has caused increased waiting times, which in turn led to the death of patients while waiting on the list (c.f. Figures 1.1-1.2) [6].

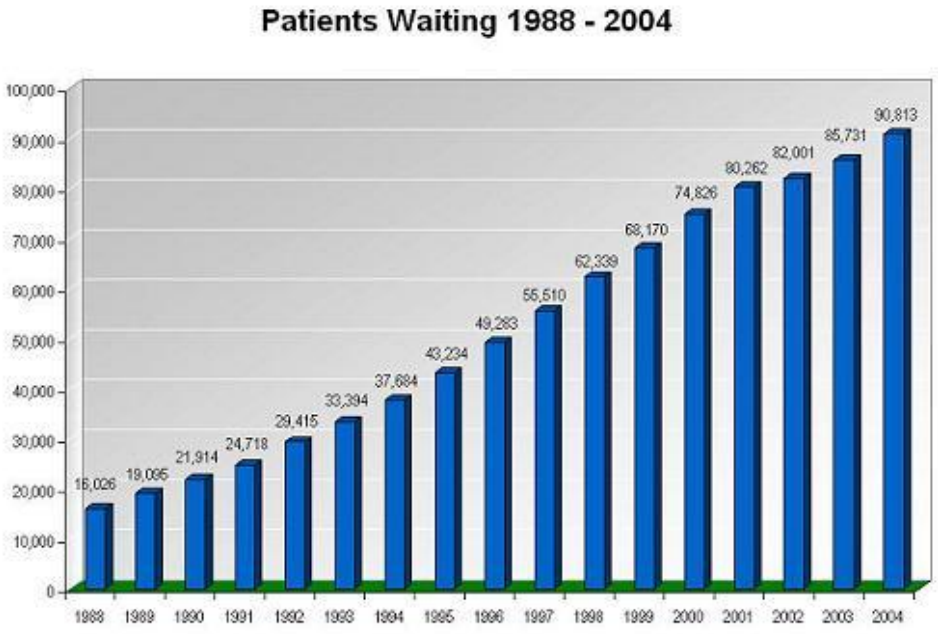


Figure 1.1 Number of patients on the waiting list per year in US [6]

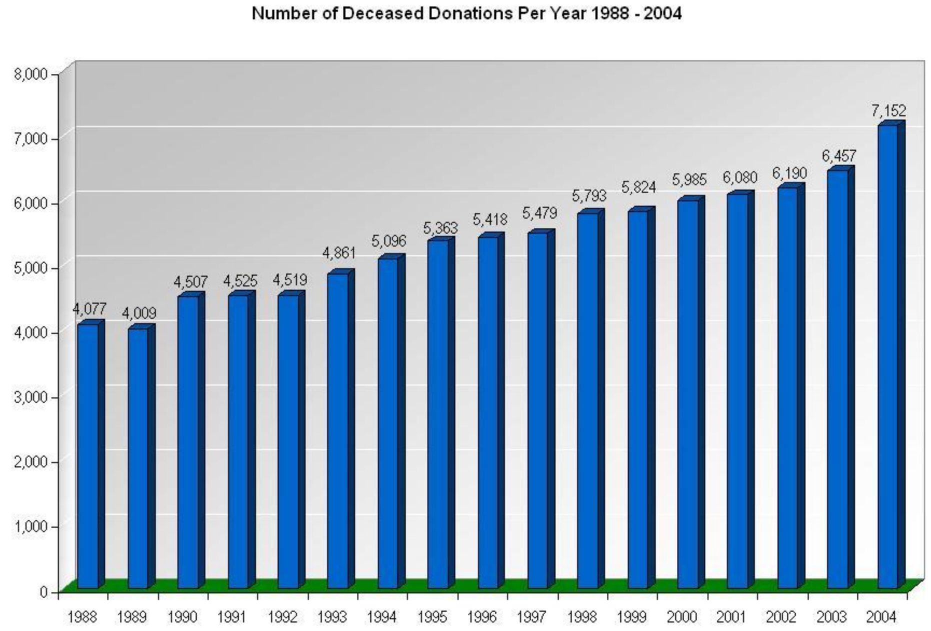


Figure 1.2 Number of donor organs per year in US [6]

1.2. Problem Statement, Research Goal, and Research Objectives

Organ transplantation is a vital treatment for the chronic failure of major organs. Survival analysis, which is defined as the surviving time after a patient receives transplantation surgery, has been the primary evaluation method for the effectiveness of such an operation. The primary objective of this research is to develop an integrated data mining methodology to accurately predict the survivability and to analyze the prognostic factors for different risk groups of transplant patients in order to discover novel patterns to augment clinical and biological studies. By incorporating the findings of these data mining-based survival and prognostic analyses, a simulation model will be developed to search for more efficient and effective scenarios of matching and allocation of organs. In doing so, we propose to use very large data sets with hundreds of determinative variables regarding the donors, the potential recipients, and transplantation procedures. While the main research goal can be summarized as “to improve the effectiveness and efficiency of the organ transplantation procedures”, the specific objectives in this research can be listed as follows:

(1) Develop an integrated data mining methodology to build accurate predictive models for survivability, and use these models to investigate the fundamental relationship between predictor variables and survivability in order to identify the factors that have the most significant impact on survivability;

(2) Create a comprehensive prognostic index related to lung organ transplantation, and determine risk groups of patients based on their survivability quantified using the developed prognostic index, and identify the optimal setting so as to achieve better survivability;

(3) Develop a composite scoring approach-based matching index in which the survival-critic variables are hierarchically integrated in order to rank the potential candidate organ recipients and match them with the organ donors so that the survivability and quality of life (QoL) regarding the organ transplant procedures can be simultaneously predicted; and

(4) Develop discrete event simulation models to validate (and to conduct sensitivity analysis on) the patterns identified by the abovementioned data mining methodologies. Various simulation models will be developed and executed to better analyze the validity and significance of the composite scoring scheme in order to improve

organ allocation policies in terms of various performance measures such as average waiting time on the list.

REFERENCES

[1] L.S. Lee, K.D. Fiedler, J.S. Smith, Radio frequency identification (RFID) implementation in the service sector: A customer-facing diffusion model, *International Journal of Production Economics* 112 (2008) 587–600.

[2] B. Kaplan, The medical computing ‘lag’: Perceptions of barriers to the application of computers to medicine, *International Journal of Technology Assessment in Health Care* 3 (1987) 123–126.

[3] S.H. Lee, A.W. Ng, K. Zhang, The quest to improve Chinese healthcare: some fundamental issues, *International Journal of Health Care Quality Assurance* 20 (2007) 416-428.

[4] S.M. Shechter, C.L. Bryce, O. Alagoz, J.E. Kreke, J.E. Stahl, A.J. Schaefer, D.C. Angus, M.S. Roberts, A Clinically Based Discrete-Event simulation of end-stage Liver Disease and the Organ Allocation Process, *Medical Decision Making* 25 (2005) 199-209.

[5] http://www.unos.org

[6] http://www.gsds.org/

CHAPTER II

LITERATURE REVIEW

In this chapter, four main research streams in organ transplantation area are summarized. In Section 2.1, survival analysis in organ transplantation is presented. Prognostic index devising is followed in Section 2.2. State-of-the art research regarding composite scoring approaches to develop an index to measure the quality of life is provided in Section 2.3. Finally, in Section 2.4 simulation modeling for organ transplantation procedures is introduced.

2.1 Related Research in Survival Analysis for Organ Transplantation

A large body of research exists for data-driven analytics in various organ transplantation cases. Kusiak et al. [1] conducted a study which compared two rule-based data mining techniques, i.e., decision trees and rough sets, for predicting survival time of kidney dialysis patients. Their study presented not very high but considerable prediction accuracy rates. The main limitation of the study was the utilization of a small dataset with 188 patients in total and many patient-related parameters were ignored. Hong et al. [2]

presented a survival analysis of liver transplant patients in Canada by considering only some of the determinative factors such as age, blood type, donor type (cadaveric or alive), race and gender of recipient and donors. Having limited the variables with this scope, in their study they also admit that the clinical information used in the study lacks many details. Specifically focused on thoracic transplantation, Jenkins *et al.* [3] and Fernandez-Yanez *et al.* [4] had a rich pool of dependent variables for survivability prediction. They employed the Kaplan-Meier method of survival analysis with the Mantel-Haenszel log-rank test which are fundamental statistical survival analysis techniques. These studies have two major limitations: First, they lack an enhanced data-mining perspective which would utilize machine learning and artificial intelligence tools (which are independent of the nonlinearity and multicollinearity assumptions of traditional linear modeling techniques) to reveal the previously-unknown potentially-useful patterns. Secondly, the variables were selected based on the experiences and intuitions of the analysts who conducted the study. A more recently held study has the same drawbacks, which was proposed by Tjang *et al.* [5]. Based on their experience, they adopted some newer explanatory variables such as body mass index, waiting time on the list, and previous cardiac surgery to determine the survivability in heart transplantation. However, similar to the aforementioned studies they also utilized only statistical techniques such as the Chi-Square test, the Fisher’s test, non-parametric Kruskal-Wallis rank test, and the Kaplan-Meier survivorship function. Similar limitations exist also in some other studies related to lung transplantation [6]-[8] which renders them disqualified to be considered a detailed data mining study.

2.2 Related Research in Devising a Prognostic Index

Prognostic index (PI) provides compact prognosis information regarding a specific patient based on the results of a Cox proportional hazards model [9]. The Cox proportional hazards model helps identify variables of prognostic importance and hence the prognostic index can be used to define groups of individuals at different risk categories. Even though the prognostic index is a convenient tool to measure how well the patients are doing after the transplantation, its use in the organ transplantation area has been limited mostly due to the lack of follow-up data. Some existing studies related to devising a PI in transplant area are summarized as follows.

In the study conducted by Christensen *et al.* [10], it is mentioned that primary biliary cirrhosis requires a liver transplantation operation at the end stage. However, a very critical issue is the timing for transplantation: neither too early nor too late. Based on the prognosis analysis with and without transplantation, it will be easier to decide whether or not the transplantation is required, and if so, when. To achieve this goal, corresponding PI's and thence probabilities of surviving are computed for transplantation and non-transplantation cases. Using these, a Cox regression model was created for 6-month survival which also confirms some variables used in the literature previously and their model brings new significant variables. As a result, the gain from transplantation starts to become positive around 8 months prior to death (this is when PI=2.5). The gain of transplantation is defined to be the difference between survival probability with transplantation and without transplantation. If it gives a negative value out, transplantation should not be performed and vice versa. The predicted gain from transplantation starts to become clinically important when PI reaches about 2.5, corresponding to a predicted 6-month survival of about 0.85. The consequence of this is the following: If PI>=2.5, transplantation should be done within the following 6 months. Yoo *et al.* [11] developed a similar index and sought to answer whether or not socioeconomic status affects the survivability in liver transplantation. They handled the survivability in both cases for patients and grafts. The study revealed that socioeconomic status does not influence patient or graft survival that undergoes liver transplantation at their institute. Deng *et al.* [12] conducted a study with a national dataset in Germany, which discovers the effect of receiving a heart transplant for the patients in a waiting list. The results indicate that cardiac transplant is associated with a survival benefit only for patients with a predicted high risk of dying on the waiting list. Ghobrial *et al.* [13] performed a study to determine prognostic factors for overall survival in 107 adult patients with post-transplantation lymphoproliferative disorders (PTLDs). It is validated that in discriminating the low and high scored patients the proposed prognostic scoring significantly performs better than the International Prognostic Index for the subset of the patients (56 out of 107) with lactate dehydrogenase.

The common limitation in all of these studies is similar to the limitations of the studies summarized in Section 2.2. Namely, they directly devise a prognostic index without determining if the variables used in prognostic index devising phase are necessary and sufficient. This motivates a machine learning-based initial step of variable selection procedure. Because, if the critical predictive factors are not captured effectively due to the intuition- and experience-based selection, the resulting prognostic indices developed based on the selected variables would be inaccurate and, in turn, related risk groups of patients would be deviated from the real classes. This may cause mistakes for decision makers in making organ transplantation policies.

2.3 Related Research in Composite Scoring to Measure Quality of Life

Voluminous data has been collected from transplant procedures and analyzed to evaluate the organ allocation process [14]. Attempts to analyze organ transplants with this huge amount of data have focused on identifying the characteristics of thoracic transplant recipients and their associated post-transplant outcomes [15]. Molhazn *et al.* [16] examined the perceived quality of life (QOL) of patients with end stage renal disease by incorporating patients' medical characteristics, their health status, functional status, ability to work, and ability to perform activities. Significant direct effects of these variables on QOL were determined. Smith *et al.* [17] conducted a survey to reveal whether or not quality of life and health status are distinct constructs. Using three functioning domains (mental, physical, and social) they found out that these two are different measures and hence, should be analyzed through separate questionnaires. Devins *et al*. [18] devised a novel scale named the Illness Intrusiveness Ratings Scale (IIRS) by pooling responses from separate studies concerning quality of life in renal, heart, liver, and lung transplants among many others. The study was aimed at investigating the factor structure underlying the IIRS. By using exploratory and confirmatory factor analyses in a step-by-step fashion, they first identified the factor structure and then confirmed it against various patient groups (i.e. renal, heart, lung transplants and etc.). Two more recent studies [19]-[20] have analyzed the quality of life after liver transplant as well as chronic heart failure, respectively. Castaldo *et al*. [19] examined the effect of preoperative and postoperative factors on the model for end-stage liver disease (MELD score), which is mainly used for organ allocation decisions by predicting short-term mortality of patients. This research has revealed that increasing MELD score can be attributed to improved physical health-related QOL (HRQOL) whereas it does not have an association with mental HRQOL. On the other hand, Faller *et al*. [20] focused on the chronic heart failure patients with the question whether depression affects only the psychological domain of patients' HRQOL or it is broader and may affect the physical domain of HRQOL. The analysis results suggest that depression has an independent impact on both physical and psychological domains of HRQOL in patients with chronic heart failure while the heart failure severity affects only physical HRQOL.

Although the abovementioned studies reveal very useful initial knowledge for the organ transplantation field based on the classical statistical assumptions adopted, they still have some limitations as follows: (1) They implicitly ignore the fact that the predictive variables may not necessarily be independent of each other. On the contrary, they often do affect each other. These predictor variables can be categorized into higher level classes as a group which they refer to. (2) Following the first explanation, grouped and aggregated variables may have a nonlinear relationship and/or additive interaction effects with the outcome measures of the transplant. However, such features cannot be revealed through the existing methods. (3) In the state-of-the-art, the transplant performance measure is evaluated based on a single metric. The transplant success may not be a single metric to be predicted (e.g. only survival time) to satisfy various benefits of the organ allocation. Instead, this study assumes that it should be the combination of various metrics (e.g. survival time, quality of life, and etc.).

2.4 Related Research regarding Simulation of Organ Transplant Procedures

The vast majority of analytics-driven organ transplantation research involves simulation studies and has been studied since mid 80's specifically in a simulation modeling standpoint pioneered by Ruth *et al.* [21]. It was further developed by Pritsker and his students [22]-[23]. Their study provided a useful simulation tool which utilized UNOS liver allocation data hence named as ULAM (UNOS Liver Allocation Model). UNOS stands for United Network for Organ Sharing which is a tax-exempt, medical, scientific, and educational organization that operates the national Organ Procurement and Transplantation Network (OPTN) under contract to the Division of Organ Transplantation (DOT) of the Department of Health and Human Services (DHHS). ULAM is a simulation proposed to compare various liver allocation policies. It uses either historical or simulated data for patient listings and donor arrivals. Patients are modeled in a dynamic fashion, namely they can change medical urgency status or be removed from the list due to death. Once they are transplanted, patients might die, relist, or survive. It adopted a policy that patients will be ranked based on the waiting time and blood type compatibility with the donor, using four ranks: 1 showing the most urgent. The main components of ULAM are listed as follows: initial waiting list, recipient stream, patient medical urgency status change process, donor stream, allocation policy, liver offer/acceptance process, post-transplant relisting/mortality, outputs.

Implementation of ULAM revealed the fact that the drawbacks in the previously adopted policy (sickest patient first allocation policy) could be overcome with a new policy. The new policy suggested distributing livers to patients in local, regional and then, national areas. In each of these areas, the sickest status group patients were prioritized first. The comparison study showed that 1500 more transplants would be achieved if the new policy was adopted. Additionally, 1626 fewer post-transplant deaths would occur.

Based on the successful findings of ULAM, UNOS requested Pritsker to create another simulation tool for kidney transplant procedures, which gave birth to the UNOS Kidney Allocation Model (UKAM) [24]. Inclusion of 256 nation-wide transplant centers in UKAM enhanced its reliability in estimates at the national level. For the ULAM outputs, some key measures are determined by the transplant community and are assumed to be the most valuable in evaluating policy changes [25].

Simulation studies have existed and been mostly helpful to adopt an organ transplantation policy. Some other studies can be listed as follows: (1) McEwan *et al.* [26] focused on evaluating the cost-effectiveness of sirolimus compared with cyclosporine in UK for post surgical management of renal transplant recipients. It is based on an evaluation of both cost-effectiveness and cost utility by using a discrete event stochastic simulation. (2) Thompson *et al.* [27] proposed a more sophisticated simulation tool which can handle various organ transplantation scenarios, namely heart-lung, liver and kidney. (3) Roberts *et al.* [28] developed a simulation model to compare and contrast various organ allocation policies for liver transplants.

Table 2.1 Summary of the literature review

Study held by:	Survivability analysis	Devising a prognostic index	Modeling Quality of Life	Simulation modeling
Trigt *et al.*, 1996	x			
Tringali *et al.*, 1996	x			
Knoll *et al.*, 1997	x			
Schnitzler *et al.*, 1997	x			
Cope *et al.*, 2001	x			
Mehra *et al.*, 2004	x			
Kusiak *et al.*, 2005	x			
Fu *et al.*, 2006	x			
Boin *et al.*, 2007	x			
Aguero *et al.*, 2007	x			
Bleyer *et al.*, 1996	x	x		
Christensen *et al.*, 1999	x	x		
Deng *et al.*, 2000	x	x		
Esparrach *et al.*, 2001	x	x		
Yoo *et al.*, 2002	x	x		
Ghobrial *et al.*, 2005	x	x		
Johnson *et al.*, 2008	x	x		
Molhazn *et al.*, 1996	x		x	
Smith *et al.*, 1999	x		x	
Devins *et al.*, 2001	x		x	
Castaldo *et al.*, 2009	x		x	
Faller *et al.*, 2009	x		x	
Ruth *et al.*, 1985	x			x
Pritsker *et al.*, 1995	x			x
Pritsker *et al.*, 1996	x			x
Baldwin *et al.*, 2000	x			x
Harper *et al.*, 2000	x			x
Taranto *et al.*, 2000	x			x
Ratcliffe *et al.*, 2001	x			x
Roberts *et al.*, 2004	x			x
Thompson *et al.*, 2004	x			x
McEwan *et al.*, 2005	x			x
Shechter *et al.*, 2005	x			x

In addition to using a UNOS liver transplant-related dataset, they incorporated a large transplant center's (i.e. University of Pittsburgh Medical Center) data to model the disease progression while waiting on the list. They mainly assessed the effect of using a single national waiting list as opposed to the current allocation strategy with the combination of regional waiting lists. The simulation model accurately captured the pattern in waiting time and survival rate after transplant. However, the model results were far different than the UNOS in predicting the number of deaths on the waiting list. This discrepancy was explained by the fact that the disease progression on the waiting list was determined using a local transplant center's data instead of the national one due to the lack of the latter's. The study concluded that the switch to a national waiting list for liver transplant would decrease the number of deaths on the waiting list and increase the overall survival rate, but it would also increase the graft failures and increase the median waiting time. To conclude, the state-of-the-art about organ transplantation efforts can be summarized as in Table 2.1.

2.5 Research Gap and Challenges

The main drawback of the aforementioned studies is that they do not give satisfactory results at the local or regional levels whereas they validate well against the national level since the datasets are mostly retrieved from national sources. This refers to a major gap in the modeling of transplantation procedures. Besides, the outcome measure that drove many of the original allocation debates, waiting time, was found to be a poor measure of differences in access to transplantation and not a good indicator of medical urgency or priority. This refers to another major gap supposed to focus on other measures of equity and justice such as pre-transplant mortality [25]. The former issue is, in fact, partially a consequence of the latter. Therefore, if a well-established decision support system that would determine good indicators of medical urgency/priority through data mining-based survival and prognostic analyses can be developed; there will be a linkage to better simulation scenarios at all potential levels of organ transplantation. Also, such a comprehensive methodology could help incorporating more outcome measures (in addition to waiting time on the list). This research study is intended to cover these gaps and overcome its challenges as explained in Chapter III.

REFERENCES

[1] A. Kusiak, B. Dixon, S. Shah, Predicting Survival time for Kidney Dialysis Patients: A Data Mining Approach, *Computers in Biology and Medicine* 35 (2005) 311-327.

[2] Z. Hong, J. Wu, G. Smart, K. Kaita, S.W. Wen, S. Paton, M. Dawood, Survival Analysis of Liver Transplant Patients in Canada, in *Transplantation Proceedings* (2006) 2951-2956.

[3] P.C. Jenkins, M.F. Flanagan, K.J. Jenkins, J.D. Sargent, C.E. Canter, R.E. Chinnock, R.N. Vincent, A.N.A. Tosteson, G.T. O'Connor, Survival Analysis and Risk Factors for Mortality in Transplantation and Staged Surgery for Hypoplastic Left Heart Syndrome, *Journal of the American College of Cardiology* 36 (2000) 1178-1185.

[4] J. Fernandez-Yanez, J. Palomo, E.G. Torrecilla, D. Pascual, G. Garrido, J.J.G. de Diego, M. Dominguez, J. Almendral, Prognosis of Heart Transplant Candidates Stabilized on Medical Therapy, *Rev. Esp. Cardiol.* 58 (2005) 1162-1170.

[5] Y.S. Tjang, G.J.M.G. Heijdan, G. Tenderich, D. Grobbee, R. Korfer, Survival Analysis in Heart Transplantation: Results from an Analysis of 1290 Cases In

A Single Center, *European Journal of Cardio-Thoracic Surgery* 33 (2008) 856-861.

[6] H.M. Lin, H.M. Kaufmann, M.A. McBride, D.B. Davies, J.D. Rosendale, C.M. Smith, E.B. Edwards, O.P. Daily, J. Kirklin, C.F. Shield, L.G. Hunsicker,Center- Specific Graf and Patient Survival Rates: 1997 UNOS Report, *JAMA* 280 (1998) 1153-1160.

[7] J.T. Cope, A.K. Kaza, C.C. Reade, K.S. Shockey, J.A. Kern, C.G. Tribble, I.L. Kron, A Cost Comparison of Heart Transplantation Versus Alternative Operations for Cardiomyopathy, *Annual Thoracic Surgery* 72 (2001) 1298-1305.

[8] J. Aguero, L. Almenar, L.Martinez-Dolz, J. Moro, M.T. Izquierdo, O. Cano, A. Salvador, Differences in Clinical Profile and Survival After Heart Transplantation According to Prior Heart Disease, *Transplantation Proceedings* 39 (2007) 2350-2352.

[9] M.K.B. Parmar, D. Machin, *Survival Analysis: A Practical Approach* (John Wiley & Sons, Cambridge, UK, 1996).

[10] E. Christensen, B. Gunson, J. Neuberger, Optimal timing of Liver Transplantation for Patients with Primary Biliary Cirrhosis: Use of Prognostic Modeling, *Journal of Hepatology* 30 (1999) 285-292.

[11] H.Y. Yoo, V. Galabova, D. Edwin, P.J. Thuluvath, Socioeconomic Status does not affect the Outcome of Liver Transplantation, *Liver Transplantation* 8 (2002) 1133-1137.

[12] M.C. Deng, M.J. DeMeester, J.M.A. Smiths, J. Heinecke, H.H. Scheld, Effect of Receiving a Heart Transplant: Analysis of a National Cohort entered on to waiting List, Stratified by Heart Failure Severity, *British Medical Journal* 321 (2000) 540-545.

[13] I.M. Ghobrial, T.M. Habermann, M.J. Maurer, S.M. Geyer, K.M. Ristow, T.S. Larson, R.C. Walker, S.M. Ansell, W.R. Macon, G.G. Gores, M.D. Stegall, C.G. McGregor, Prognostic Analysis for Survival in Adult solid Organ Transplant Recipients with Posy-Transplantation Lymphoproliferative Disorders, *Journal of Clinical Oncology* 23 (2005) 7574-7582.

[14] R.N. Pierson, M.L. Barr, K.P. McCulluough, T. Egan, E. Garrity, M. Jessup, S. Murray, Thoracic Organ Transplantation, *American Journal of Transplantation* 4 (2004) 93-105.

[15] F.L. Grover, M.L. Barr, L.B. Edwards, F.J. Martinez, R.N. Pierson, B.R. Rosengard, S. Murray, Thoracic Transplantation, *American Journal of Transplantation* 3 (2003) 91-102.

[16] A.E. Molhazn, H.C. Northcott, L. Hayduk, Quality of Life of Patients with End Stage Renal Disease: A Structural Equation Model, *Quality of Life Research* 5 (4) (1996).

[17] K.W. Smith, N.E. Avis, S.F. Assman, Distinguishing between Quality of Life and Health Status in Quality of Life Research: A meta-analysis, *Quality of Life Research* 8 (1999).

[18] G.M. Devins *et al.*, Structure of Lifestyle Disruptions in Chronic Disease: A Confirmatory Factor Analysis of the Analysis of the Illness Intrusiveness Ratings Scale, *Medical Care* 39 (10) (2001).

[19] E.T. Castaldo *et al.*, Correlation of Health-Related Quality of Life After Liver Transplant with the Model for End-Stage Liver Disease Score, *Arch. Surg.* 144 (2) (2009).

[20] H. Faller *et al.,* Depression and Disease Severity as Predictors of Health-Related Quality of Life in Patients with Chronic Heart Failure-A Structural Eqaution Modeling Approach, *Journal of Cardiac Failure* 15 (4) (2009).

[21] R.J. Ruth, L. Wysezewianski, G. Herline, Kidney Transplantation: A Simulation Model for Examining Demand and Supply, *Management Science* 31 (1985) 515-526.

[22] A.A.B. Pritsker, O.P. Daily, J.R. Wilson, J.P. Roberts, M.D. Allen, J.F. Burdick, Organ Transplantation Policy Evaluation, in: Proceedings of the Winter Simulation Conference (1995) 1314-1323.

[23] A.A.B. Pritsker, O.P. Daily, K.D. Pritsker, Using simulation to Craft a National Organ Transplantation Policy, in: Proceedings of the Winter Simulation Conference (1996) 1163-1169.

[24] S. E. Taranto, A.M. Harper, E.B. Edwards, J.D. Rosendale, M.A. Bridge, O.P. Daily, D. Murphy, B. Poos, J. Reust, B. Schmeiser, Developing a National Allocation Model for Cadaveric Kidneys, in: Proceedings of the Winter Simulation Conference (2000) 1971-1977.

[25] A.M. Harper, S.E. Taranto, E.B. Edwards, O.P. Daily, An Update on A Successful Simulation Project: The Unos Liver Allocation Model, in: Proceedings of the Winter Simulation Conference (2000) 1955-1962.

[26] P. McEwan, K. Baboolal, P. Conway, C.J. Currie, Evaluation of the Cost-Effectiveness of Sirolimus Cyclosporin for Immunosuppression after Renal Transplantation in the United Kingdom, *Clinixal Therapeutics* 27 (2005) 1834-1846.

[27] D. Thompson, L. Waisanen, R. Wolfe, R.M. Merion, K. McCullough, A. Rodgers, Simulating the Allocation of Organs for Transplantation, *Health Care Management Science* 7 (2004) 331-338.

[28] M.S. Roberts, D.C. Angus, C.L. Bryce, J. Stahl, S. Stehl, O. Alagoz, L. Weissfeld, A. Schaefer, Modeling the Consequences of Alternative Organ Allocation Policies for Liver Transplantation, Harvard Medical School, Department of Industrial Engineering University of Pittsburgh, January 8, 2004. Available from: http://www.ie.pitt.edu/~schaefer/Papers/evaluatingorganallocationpolicies.pdf

[29] P.V. Trigt, D. Davis, G.S. Shaeffer, J.W. Gaynor, K.P. Landolfo, M.B. Higginbotham, R.M. Ungerleider, Survival Benefits of Heart and Lung Transplantation, *Annals of Surgery* (1996) 576-584.

[30] I.F.S.F. Boin, L.V. Almeido, E.Y. Udo, R.S.B. Stucchi, A.R. Cardoso, C.A. Caruy, M.I. Leonardi, L.S. Leonardi, Survival Analysis of Obese Patients Undergoing Liver Transplantation, *Transplantation Proceedings* (2007) 3225-3227.

[31] M.A. Schnitzler, R.S. Woodward, D.C. Brennan, E.L. Spitznagel, W.C. Dunagan, T.C. Bailey, The Effects of Cytomegalovirus Serology on Graft and Recipient Survival in Cadaveric Renal Transplantation: Implications for Organ Allocation, *American Journal of Kidney Diseases* 29 (1997) 428-434.

[32] G.A. Knoll, M.R. Tankersley, J.Y. Lee, B.A. Julian, J.J. Curtis, The Impact of Renal Transplantation on Survival in Hepatisis C-Positive End-Stage Renal Disease Patients, *American Journal of Kidney Diseases* 29 (1997) 608-614.

[33] P. Fu, M.J. Laughlin, H. Zhang, Comparison of survival Times in a Transplant Study of Hematologic Disorders, *Contemporary Clinical Trials* (2006) 174-182.

[34] J.T. Cope, A.K. Kaza, C.C. Reade, K.S. Shockey, J.A. Kern, C.G. Tribble, I.L. Kron, A Cost Comparison of Heart Transplantation versus Alternative Operations for Cardiomyopathy, *The Society of Thoracic Surgeons* (2001) 1298-1305.

[35] R.A. Tringali, P.T. Trzepacz, A. DiMartini, M.A. Dew, Assessment and Follow-up of Alcohol-Dependent Liver Transplantation Patients, *General Hospital Psychiatry* (1996) 70-77.

[36] J. Aguero, L. Martinez-Dolz, M.T. Izquierdo, O. Cano, A. Salvador, Differences in Clinical Profile and Survival after Heart Transplantation according to Prior Heart Disease, *Transplantation Proceedings* (2007) 2350-2352.

[37] M.R. Mehra, P.A. Uber, S. Potluri, H.O. Ventura, R.L. Scott, M.H. Park, Usefullness of an Elevated B-Type Natriuretic Peptide to Predict Allograft

Failure, Cardiac Allograft Vasculopathy, and Survival after Heart Transplantation, *American Journal of Cardiology* (2004) 454-458.

[38] Bleyer AJ, Tell GS, Evans GW, Ettinger WH, Burkart JM. Survival of Patients Undergoing Renal Replacement therapy in One Center with Special Emphasis on Racial Differences. American Journal of Kidney Diseases, 1996; 28 (1): 72-81.

[39] G.F. Esparrach, A.S. Fueyo, P. Gines, J. Uriz, L. Quinto, P.J. Ventura, A. Cardenas, M. Guevara, P. Sort, W. Jimenez, R. Bataller, V. Arroyo, J. Rodes, A Prognostic Model for Predicting Survival in Cirrhois with Ascites, *Journal of Hepatology* (2001) 46-52.

[40] E.S. Johnson, M.L. Thorp, R.W. Platt, D.H. Smith, Predicting the Risk of Dialysis and Transplant among Patients with CKD: A Retrospective Cohort Study, *American Journal of Kidney Diseases* (2008) Article in Press.

[41] L.P. Baldwin, T. Eldabi, R.J. Paul, A.K. Burroughs, Using Simulation for the Economic Evaluation of Liver Transplantation, in: Proceedings of the Winter Simulation Conference (2000) 1963-1970.

[42] J. Ratcliffe, T. Young, M. Buxton, T. Eldabi, R. Paul, A. Burroughs, G. Papatheodoridis, K. Rolles, A Simulation Modeling Approach to Evaluating Alternative Policies for the Management of the Waiting List for Liver Transplantation, *Healthcare Management Science* 4 (2001) 117-124.

CHAPTER III

RESEARCH METHODOLOGY

To address the aforementioned issues in the state-of-the-art as summarized in Chapter II, in this research we propose to apply an integrated data mining method to model the complex relationship between predictor variables and survivability at the first step (Task 1). Then a prognostic index will be developed in Task 2 and used to group the differing risk groups of organ recipients. In accordance with the outputs of Task 2, a new matching index (composite score) and a scheme which would be composed with the consideration of various criteria of organ transplant would be created in Task 3. This index would be used in the following simulation study (Task 4). The simulation will conduct what-if analyses to validate and fine-tune the weights of the new matching index via response surface methodology. A pictorial representation of the overall methodology is illustrated in Figure 3.1. These four tasks are further explained in the following sub-sections.

3.1 Task 1: Data Mining and Model Integration

In this task, by assigning the output as the survival time of the patients after transplant takes place and input as feature-rich dataset (patient-, donor-, and transplant-

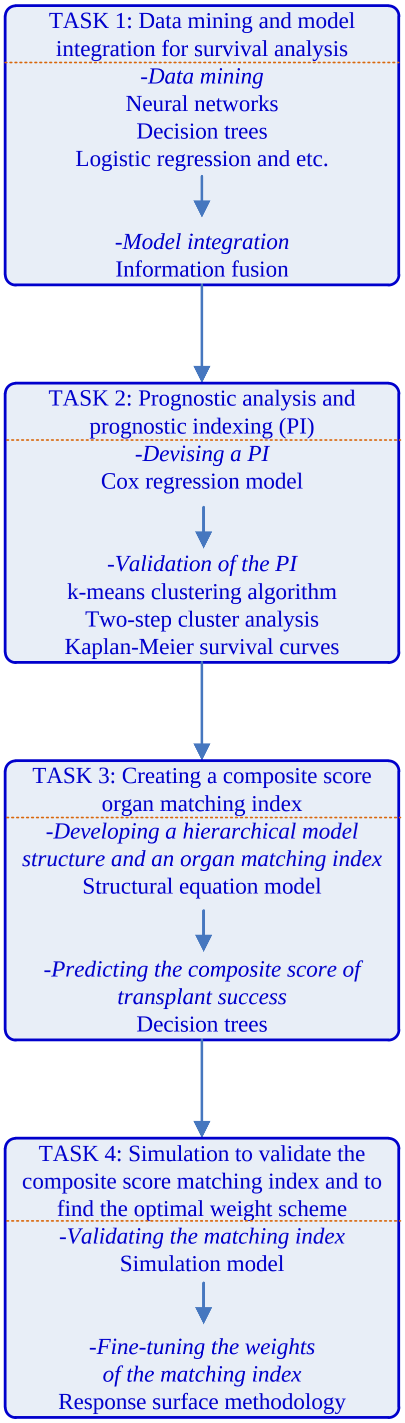

TASK 1: Data mining and model integration for survival analysis
-Data mining
Neural networks
Decision trees
Logistic regression and etc.
-Model integration
Information fusion
TASK 2: Prognostic analysis and prognostic indexing (PI)
-Devising a PI
Cox regression model
-Validation of the PI
k-means clustering algorithm
Two-step cluster analysis
Kaplan-Meier survival curves
TASK 3: Creating a composite score organ matching index
-Developing a hierarchical model structure and an organ matching index
Structural equation model
-Predicting the composite score of transplant success
Decision trees
TASK 4: Simulation to validate the composite score matching index and to find the optimal weight scheme
-Validating the matching index
Simulation model
-Fine-tuning the weights of the matching index
Response surface methodology

Figure 3.1 Framework for the research methodology

related) we will deploy an integrated data mining method to reveal the underlying relation between the output and input variables as well as the relationships among input variables themselves. Based on the integration of the results in terms of accuracy, it is possible to rank the predictor variables, considering their importance contributing to the graft status prediction. The data mining models used in this research are introduced in the following section and the overall process of developing prediction models is depicted in Figure 3.2.

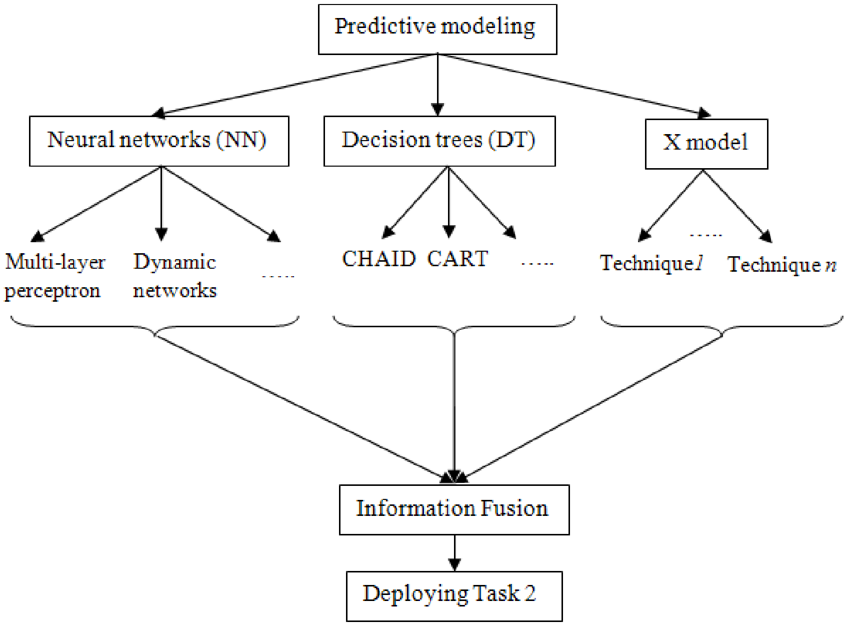


Figure 3.2 Illustration of the integrated data mining and model integration

3.1.1 Task 1.1: Predictive Modeling

Since the dependent variable here was a binary variable (graft status: with 0 representing survived and 1 representing not-survived), the problem refers to a *classification type prediction* problem. This task is to apply various predictive models to predict the graft status. For the modeling purposes, one prediction model from the statistical field (logistic regression) and two models from the machine learning field (neural networks and decision trees) will be used. These models are selected to be included in the study due to their popularity in the literature. Neural networks (NN) have been the most popular artificial intelligence-based data modeling algorithm used in clinical medicine due to their good predictive performance [1]. Multi-layer perceptron (MLP) has been the most commonly used and well-studied NN architecture in almost all fields.

On the other hand, compared with other machine learning methods (e.g. NNs), decision trees have the advantage in that they are not black box models and hence can easily be explained as rules. This advantage has made them widely usable in medicine [2].

3.1.2 Task 1.2: Hierarchal Model Integration Method

Much research has focused on developing procedures to select a single “best” model. These procedures often neglect the uncertainty inherent in the model selection process. Choosing only one model for prediction comes with inherent risk. When multiple possible models fit the observed data similarly well, it is risky to make inferences and predictions based on only a single model. In this case, predictive performance suffers, because standard statistical inference typically ignores model uncertainty. Information fusion is an approach to combine the prediction information received from various data mining models. As illustrated in Fig. 3.2 an “information fusion” technique will be used to combine the models together to further improve the accuracy of them and rank the importance of the critical factors accordingly.

3.2 Task 2: Devising and Validating a Prognostic Index

Having constrained hundreds of predictor variables to a manageable extent by means of Task 1, Task 2 will devise a prognostic index that categorizes the organ recipients by the Cox regression model.

3.2.1 Task 2.1: Determining the Candidate Sets of Predictor Variables

This subtask is to determine which predictor variables to be used in devising a prognostic index in subtask 2.2. Task 2.1 will eliminate the insignificant variables and minimize the crowded set of predictor variables. It consists of three candidate predictor variables sets. The first set is composed of predictive model-selected variables. The predictive models adopted in Task 1 can rank the predictor variables based on their importance level in predicting the graft survival. In this way, a union set of predictive variables would be constructed which is named as the first set of predictive variables. The second set of predictive variables is obtained by considering the common-sense domain knowledge. This set includes variables which are logically related to lung transplantation such as donor’s history of cigarette usage. The third set of predictive variables is compiled from the literature research conducted. This set essentially consists of the variables which have been commonly repeated in previous studies in the organ transplantation area. The second and third sets of predictive variables can be referred as the expert input to the variable determination stage of the proposed methodology. These sets (Set 2 and Set 3) provide one more chance to the next step in Task 2.2 -Cox model- to evaluate the variables that might have importance in the survival analysis although they may be determined as insignificant by the predictive models in Task 1.

3.2.2 Task 2.2: Prognostic Index Devising

This subtask takes all the three sets of predictive variables in Task 2.1 and applies the Cox regression to model the graft survivability and filter out the candidate predictive variables which do not have a survival effect. Hence, in Task 2.2, the final critical predictive variables can be determined by the Cox regression. The Cox regression model also enables us to devise a prognostic index to categorize the patients into differing risk groups such as low, medium, and high.

The Cox regression model is a semi-parametric model which is extensively used in survival analysis [3]-[4]. One important application of Cox regression model is to help identify variables which may be of prognostic importance [5]. Once identified, knowledge from these variables may be combined and used to define a prognostic index, which in turn defines groups of organ recipients at differing risk. To use the prognostic index, key patient characteristics are recorded and a score is derived from these. This score gives an indication of whether for example; the particular patient has a good, intermediate or bad prognosis for the disease [5].

3.2.3 Task 2.3: Determining Risk Groups of Lung Recipients

A major issue arises in Task 2.2: how many risk groups to classify the patients into? In this task, k-means algorithm [6] and two-step cluster analysis [7] are applied to reveal the answer to this question. The findings of these two algorithms would be compared and contrasted against the widely-used medical expert-based heuristics. As a statistically and pictorial verification of the number of groups determined by these algorithms, Kaplan-Meier survival analysis [4] would be adopted and hence corresponding survival curves would be generated.

3.3 Task 3: Creating a Composite Score for Modeling Transplant Success

Task 3 develops a hierarchical structure to model the transplant success in a state-by-step fashion. By means of adopting the structural equation modeling technique [8], it first determines the measurement models and, in turn, determines the composite scores for the latent variables which are attributed to the prediction of *transplant success*. Then these composite scores are used for matching the donor organ and the recipient. After the matching process, decision trees are employed to predict the overall transplant success, which would also be a combination of two performance measures (i.e. graft survival time and a kind of quality of life metric). The integration of structural equation modeling and decision trees would hypothetically provide more transparency (interpretability) to the medical experts and more prediction accuracy.

3.4 Task 4: Simulation to Validate the Matching Index and Optimize its Weights

The main objective of simulation modeling is to gain invaluable insight into the dynamics of complex systems, which is the focus in this research. Simulation models of complex systems consist of numerous input variables, linked together by logical relationships. The process of determining the set of input variables that produce the optimal output has often posed the greatest challenge during simulation studies. In recent years, the ability to integrate optimization technology into simulation models has significantly improved this process. To effectively utilize optimization technology, however, modelers must define optimization variables [9]. Task 2 and Task 3 outputs would determine these optimization variables in our research. That is, the prognostic index and the *weights* of the composite score matching index of the donor and the recipient would be the target to fine-tune and optimize. However, since the organ-recipient match is too complex to optimize we propose to implement a simulation study for a satisfying solution through the usage of response surface methodology. By considering the utility function along with the efficiency of the process through the simulation, a more sensitive matching index could hopefully be derived.

REFERENCES

[1] R. Bellazzi, B. Zupan, Predictive Data Mining in Clinical Medicine: Current Issues and Guidelines, *International Journal of Medical Informatics* 77 (2008) 81-97.

[2] S. Dreiseitl, L. Ohno-Machado, Logistic Regression and Artificial Neural Network Classification models: A Methodology Review, *Journal of Biomedical Informatics* 35 (2002) 352-359.

[3] D.R. Cox, Analysis of Survival Data (Chapman&Hall, London, 1984).

[4] E. Kaplan, P. Meier, Nonparametric Estimation from Incomplete Observations, *Journal of the American Statistical Association* 53 (1958) 187-220.

[5] M.K.B. Parmar, D. Machin, Survival Analysis: A Practical Approach (John Wiley & Sons, Cambridge, UK, 1996).

[6] J. B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Symposium on Math, Statistics, and Probability, University of California Press, Berkeley, CA, USA (1967), pp. 281-297.

[7] T. Chiu, D. Fang, J. Chen, Y. Wang, C. Jeris, A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment, in: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining (2001), 263.

[8] K.G. Jöreskog, A General Method for Analysis of Covariance Structure, Biometrika 57 (1970), 239-51.

[9] T.F. Brady, E. Yellig, Simulation Data Mining: A New Form of Computer Simulation Output, Proceedings of the Winter Simulation Conference (2005) 285-289.

CHAPTER IV

SURVIVAL ANALYSIS OF LUNG ORGAN TRANSPLANTS

Predicting the survival of lung transplant patients has the potential to play a critical role in understanding and improving the matching procedure between the recipient and graft. Although voluminous data related to the transplantation procedures is being collected and stored, only a small subset of the predictive factors has been used in modeling lung transplantation outcomes. The main objective of this study is to improve the prediction of outcomes following the lung transplantation by proposing an integrated data-mining methodology. A large and feature-rich dataset (16,604 cases with 283 variables) is used to (1) develop machine learning based predictive models; and (2) extract the most important predictive factors. Then, using three different variable selection methods, namely, i) machine learning methods driven variables—using decision trees, neural networks, logistic regression, ii) literature review-based expert-defined variables, and iii) common sense-based interaction variables, a consolidated set of factors is generated and used to develop Cox regression models for lung graft survival.

4.1 Motivation and Background

In many circumstances, organ transplantation is the preferred treatment, sometimes the only permanent treatment, for the chronic failure of the major organs. For example, dialysis can be an option for survival (for months or even years) for a kidney patient, whereas for a lung-awaiting patient, there is no option other than transplantation. There has been considerable success in the field of organ transplantation, and further improvements in the outcome of transplantation procedures are in prospect [1]. The main challenge in organ transplantation is the shortage of donated organs. Additionally, a significant number of organs are being rejected due to a suboptimal match between the graft and the patient. The demand for organ transplantation is increasing while the number of donors remains the same, resulting in longer lists of patients waiting for transplantation [2]. In such a setting, outcome prediction is becoming increasingly important in medicine. But when a resource is scarce the need for accurate prediction becomes acute [3]. Especially prediction of survival is a clinically important but challenging problem [4]. Therefore, optimization of the system necessitates sophisticated procedures for the selection of optimal organ recipients since currently it is impossible to satisfy all organ demands. On the other hand, there are competing principles in hand to satisfy such as utility, justice, and equity principles. Namely, the likelihood of satisfactory outcomes must be jointly optimized with the urgency of need. To be able to achieve this level of sophistication, the first step is to reveal the underlying knowledge in the large amount of data that is recorded in organ transplantation procedures. The main idea would be to maximize the survival rate for transplantation in the light of hundreds of determinative variables captured and stored in databases. These databases include variables regarding the donor/graft and the potential recipient. The proposed method would simultaneously optimize the utility, justice and equity principles as well. Until now, the main focus has been only on some specific factors although there might be many more to be taken into account. The findings in our study will provide a new insight into the aforementioned three principles. For example, Kirklin *et al.* [5] defines utility as "an allocation policy that maximizes patient and graft survival". Here come two questions in mind: "1) Based on what determinative variables can the patient and graft survival be maximized?" and "2) How can these critical determinative variables be objectively specified and combined in a methodological manner?" It would be naïve to assume that a decision maker can take all of the independent factors into account to optimize the solution, due to the bounded rationality of human beings, and attempting to do that would be extremely time-consuming, resulting in some trivial information being inferred and acted upon in the process. Therefore, the abovementioned two questions are essentially addressed in our study and a data-driven variable selection methodology is provided for an effective solution for these two main questions.

Organ transplantation consists of kidney/pancreas, liver, and thoracic transplantation. Thoracic transplantation refers to heart, lung, and simultaneous heart-lung organ transplantation procedures. It has become an established form of therapy for patients with end-stage heart and lung disease since its first clinical introduction in the 1960s [6]. The number of heart transplant operations performed annually in the United States has grown from 2,108 in 1990 to 2,192 in 2006 (a marginal increase) while the number of lung transplants has grown from 18 in 1987 to 1,400 in 2006 (a dramatic increase) [7]-[9]. Thoracic transplantation is significantly different from other organ transplantation procedures in that it requires transplantation faster and is more vital to patient survival. For example, a kidney transplantation awaiting patient might survive for extended periods of time by using a dialysis machine, while for a patient awaiting thoracic transplantation does not have this choice - at least not at the same comfort and cost level. A huge amount of data is complied for thoracic organ transplant patients and is analyzed to assess the importance of patient demographics, risk factors, and mortality [10]. These analyses have focused on identifying the characteristics of thoracic transplant recipients and their associated post-transplant outcomes, namely survival [11].

These previous studies have mainly focused on applying statistical techniques to a small set of factors selected by the domain-experts in order to reveal the simple linear relationships between the factors and survival. The collection of methods known as ‘data mining’ offers significant advantages over conventional statistical techniques in dealing with the latter’s limitations such as normality assumption of observations, independence of observations from each other, and linearity of the relationship between the observations and the output measure(s). There are statistical methods that overcome these limitations. Yet, they are computationally more expensive and do not provide fast and flexible solutions as do data mining techniques in large datasets.

4.2 An Integrated Data Mining-based Methodology

Organ transplantation procedures involve a large number of variables that may have a significant impact on the survival of the graft and/or the patient. However, as explained in Section 4.1, existing studies on lung transplantation procedures rely heavily on some specific variables derived from expert knowledge and experience rather than data-driven analytical methodologies. The omission of the vast majority of the variables may hinder the discovery of underlying relationships between survival and the related factors. In such approaches the complete information underlying the transplantation datasets cannot be revealed effectively. This may cause non-optimal policy adoptions. The further steps (e.g., donor-recipient match) would also be ineffective since they build on the very first step, namely, determination of significant variables, which would indicate to which patient an organ should be allocated based on what criteria.

In this study, we adopt an integrated data mining methodology to overcome the aforementioned limitations and more effectively reveal the underlying relations between survival and predictive factors. We chose the dependent variable as graft survival (which is a binary variable with 0 representing survived and 1 not-survive). Thus, the problem refers to a classification problem. However, the relationship between the dependent variable and the independent/predictor variables are not known in advance. Therefore, as a first step of the methodology, various data mining techniques (specifically binary classifiers here) which can conduct classification are implemented to predict the graft survival. The classification models explain the relationship between the dependent variable and independent variables, some explicitly like decision trees and some as a black box like the neural networks. They also rank the predictor variables based on their importance level in predicting the survival. This step would help determine the common variables in all classification models, which will be kept as the first set of critical predictive variables. The second set of predictive variables is obtained by considering the common-sense domain knowledge. This set consists of variables which are logically related to heart and lung transplantation and also some interaction terms which are transformed from the variables provided by UNOS data files. For example, a variable might be created to answer the question of “Is it important if the recipient and the donor are from the same ethnicity?” The third set of predictive variables is determined from the published studies and is referred as the expert input to the variable determination phase of our methodology. This set consists of the variables which have been commonly repeated in previous studies in the published literature. The last step would take these three sets of variables and deploy Cox regression modeling to predict the survival time by determining the significant covariate. Cox regression model is the main survival prediction technique used in this study.

The first set of predictive variables would enable the analysis to model all existing determinative factors as a whole in aspect of the modeling. Hence, the interference of possibly biased human thoughts is eliminated, which would be later incorporated in the analysis through the second and third sets of predictive variables. Well-established expert opinions should not be ignored either. Therefore, these perspectives are integrated in different stages in a way that one’s effect does not overshadow the other.

4.2.1 Data Source and Data Preparation

The proposed methodology could be applied for any type of organ transplantation procedure. In this study, the data source that was used to validate the methodology was thoracic organ transplant data provided by UNOS, which is a tax-exempt, medical, scientific, and educational organization that operates the national Organ Procurement and

Transplantation Network under the contract to the Division of Organ Transplantation of the Department of Health and Human Services [12]. The data files were obtained from UNOS using a formal data requisition procedure (which includes submission of specific data needs, purpose of the study, and a data use agreement). These data files are named as UNOS Standard Transplant Analysis and Research (STAR) files for thoracic transplants (heart, lung, and simultaneous heart-lung transplants. Each transplant STAR file consists of information on all thoracic transplants that had been performed in the US and reported to UNOS since October 1, 1987. It includes both deceased- and living-donor transplants. None of the files include any specific patient or transplant hospital identifiers due to the privacy and security issues. However, there is a patient identification number, unique to each patient, which allows tracking of the patient. Considering these features, UNOS is perceived to be the most comprehensive data available in any single field of medicine and for organ transplantation in US [13].

The complete dataset consists of 443 variables and 61,391 records. These variables include the socio-demographic and health-related factors of both the donor and the recipients. There are also procedure-related factors among the dataset. To assign as an output (dependent variable), there are four possible variables which are called pstatus, ptime, gstatus, and gtime. These variables have the following meanings: whether or not the patient died after transplantation occurred (referring to pstatus, with dead=1 and alive=0). A very similar variable was gstatus, referring to whether or not graft has failed (1 denoting failed and 0 denoting succeeded). The variable ptime denoted patient follow-up time (in days) from transplant to death/last follow up time. Similarly, gtime is explained as graft lifespan from transplant to death/last follow up time. For most of the records gtime and ptime had the same value and so gstatus and pstatus. Since the goal of this study is to develop models to predict the survival of lung recipients solely based on lung-related causes of death, the variables TX_TYPE (type of transplant), COD (recipient primary cause of death) and COD_OSTXT (recipient contributory cause of death) were used to filter out the lung recipients and discriminate the patients who died solely due to the lung graft incompatibility from the ones who died from any other reason. In UNOS thoracic files the dependent variable was assigned as gstatus with 9-year survival after transplantation and used that way in this study. Therefore, the rest of the potential dependent variables (pstatus and ptime) were eliminated from the dataset.

Considering the gstatus as the categorical dependent variable, the records for the patients who were not entered the corresponding value for gstatus were removed from the dataset. Data set also included some identification variables (e.g., Donor ID) which would track the recipient patient anonymously, track the transplant procedure, or link records from multiple data files to each other. Since these identification type variables do not have any information content to enhance the prediction capability of the models, they were also excluded from the analysis dataset. Moreover, the name of transplantation type was recorded in the dataset as a variable named Dataset which had one value (TH referring to thoracic) and the date of data processing is recorded as a variable named Date of Run which are useful for data integration purposes but have no bearing on the prediction of survival are also excluded from the analysis dataset. Similarly, other variables having only one possible value for all records in the dataset are also eliminated from the predictive modeling. UNOS STAR files also include some post transplant variables (such as length of stay and ischemic time) which have a substantial effect on survival. However, these variables would not be available before transplantation takes place. Therefore, they were excluded from the candidate sets of variables as well.

This dataset had an excessive number of missing values which render most of the records and variables seemingly insignificant. However, in data mining studies one should be very reluctant to remove the candidate predictor variables while trying to avoid artificial data imputation procedures. There is an obvious tradeoff here. As a rule of thumb, for column (variable) deletion, we were cautious to remove any variable from the analysis and assumed that if a variable has more than 95 % missing values, only then it should be regarded as not having significant information content and hence be deleted. Next step was to handle the missing values where we followed the general convention: for the categorical variables we filled the missing values with some heuristic values such as E (referring to empty) or NR (referring to not reported), and for the continuous variables we imputed the missing valued with the average of the existing records. After adopting these data preparation strategies, the final dataset was reduced to 283 cleansed independent variables and one dependent variable (gstatus) with the total record count of 16,604.

4.2.2 Data Mining Prediction Models for Survival Analysis

In this study, two popular classification models from the machine learning field were adopted, namely neural networks and decision trees. The preliminary studies were conducted to determine which models perform better in terms of classification accuracy and these two model types appeared to be the best. In a recently published survivability study, these model types were found to be among the top survivability predictors [14]. In building the prediction models, we used SPSS Clementine® [15] and SAS Enterprise Miner® [16], two if the most popular data mining toolkits. The next sub-sections provide brief descriptions of the classification models used in this study.

1. Neural Networks

Neural networks (NNs) have been utilized to model complex relationships among the predictor variables and the dependent variable such as nonlinear functions and multicollinearity [17]. Formally defined, NNs are highly sophisticated analytic techniques capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process of so-called "learning" from existing data [18]. NNs were up until the most popular artificial intelligence-based data modeling algorithm used in clinical medicine due to their good predictive performance [19]. We used a popular NN architecture called multi-layer perceptron (MLP) with back-propagation (a supervised learning algorithm). The MLP is known to be a powerful and robust function approximator for prediction and classification problems. It is arguably the most commonly used and well-studied NN architecture. Our experimental runs also proved the notion that for this type of classification problems MLP performs better than other NN architectures such as radial basis function (RBF), recurrent neural network (RNN), and self-organizing map (SOM). In fact, Hornik *et al.* [20] empirically showed that given the right size and the structure, MLP is capable of learning arbitrarily complex nonlinear functions to arbitrary accuracy levels. The MLP is essentially the collection of nonlinear neurons (a.k.a. perceptrons) organized and connected to each other in a feed-forward multi-layer structure.

2. Decision Trees

Decision trees recursively split the data in branches according to a preset criterion (e.g. information gain) to maximize the prediction accuracy resulting in a tree-like structure [21]. To achieve this, they use mathematical algorithms (such as information gain, Gini index, and Chi-squared test) to identify a pair of variables and its threshold that splits the input observation into two or more subgroups. This step is repeated at each leaf node until the complete tree is constructed. The objective of the splitting algorithm is to find a variable-threshold pair that maximizes the homogeneity (order) of the resulting two or more subgroups of samples. Popular decision tree algorithms include Quinlan's ID3, C4.5, C5 [21]-[22], and Breiman *et al.*'s CART [23]. Compared with other machine learning methods, decision trees have the advantage that they are not black box models and hence can easily be explained as rules. This advantage makes them widely usable in medicine [24]. Based on the favorable prediction results we have obtained from the preliminary runs, in this study we chose to use C5 algorithm as our decision tree method, which is an improved version of C4.5 and ID3 algorithms.

3. Logistic Regression

Logistic regression is a generalization of linear regression [24]. It is used primarily for predicting binary or multi-class dependent variables. Because the response variable is discrete, it cannot be modeled directly by linear regression. Therefore, rather than predicting point estimate of the event itself, it builds the model to predict the odds of its occurrence. In a two-class problem, odds greater than 50% would mean that the case is assigned to the class designated as “1” and “0” otherwise. While logistic regression is a very powerful modeling tool, the modeler, based on his or her experience with the data and data analysis, must choose the right inputs and specify their functional relationship to the response variable.

4.2.3 Cox Regression Modeling

The Cox regression model is a semi-parametric model which is extensively used in survival analysis [25]. It assumes a parametric form of the impacts of the predictor variables but such an assumption is not required for the survival function. Another major assumption for Cox model is that the hazards for the different groups are proportional [26]. The hazard function of each patient is assumed to follow the hazard function ($h_i(t)$) given by Eq. 4.1 as follows:

$$h_i(t) = h_0(t)\exp(x_i.\beta(t)) \tag{4.1}$$

where $h_0(t)$ is the baseline hazard function, x_i is the vector of predictor variables for the ith patient, and $\beta(t)$ is the vector of regression coefficients for the predictor variables. $\beta(t)$ is a function of time and is assumed to be same for all patients. By eliminating the time effect on it, namely assuming it to be constant over time, the effects of the predictor variables would be the same for long-term and short-term survival rates. This is known as proportional hazard rate which is an assumption that should be supported by goodness-of-fit statistics [27].

4.3 Performance Criteria for Model Evaluation

4.3.1 k-fold Cross-Validation

In order to minimize the bias associated with the random sampling of the training and holdout data samples in comparing the predictive accuracy of two or more methods, researchers tend to use k-fold cross-validation [28]. In k-fold cross-validation, also called rotation estimation, the complete dataset (D) is randomly split into k mutually exclusive subsets (the folds: $D_1, D_2, \ldots, D_k$) of approximately equal size. The classification model is trained and tested k times. Each time ($t \in \{1, 2, \ldots, k\}$), it is trained on all but one fold (D_t) and tested on the remaining single fold (D_t). The cross-validation estimate of the overall accuracy is calculated as simply the average of the k individual accuracy measures as follows,

$$CV = \frac{1}{k}\sum_{i=1}^{k} A_i \tag{4.2}$$

where *CV* stands for cross-validation accuracy, *k* is the number of folds used, and *A* is the accuracy measure of each fold.

Since the cross-validation accuracy would depend on the random assignment of the individual cases into *k* distinct folds, a common practice is to stratify the folds themselves. In stratified *k*-fold cross-validation, the folds are created in a way that they contain approximately the same proportion of predictor labels as the original dataset.

Empirical studies showed that stratified cross-validation tend to generate comparison results with lower bias and lower variance when compared to regular *k*-fold cross-validation [28].

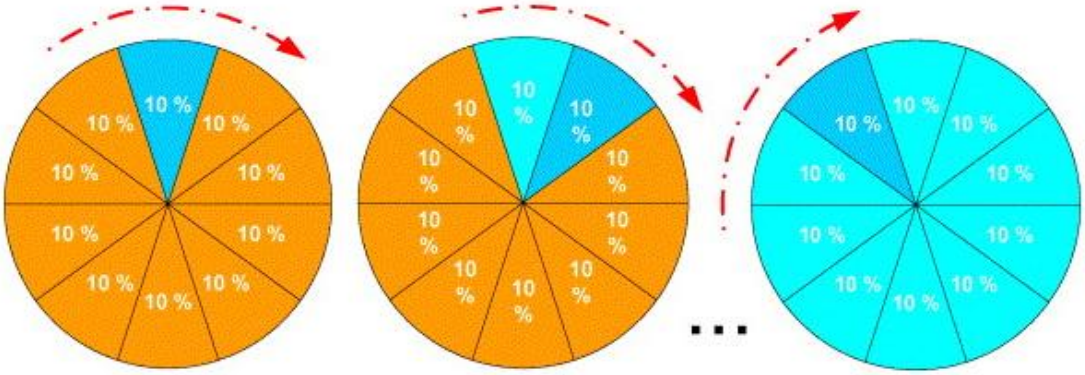
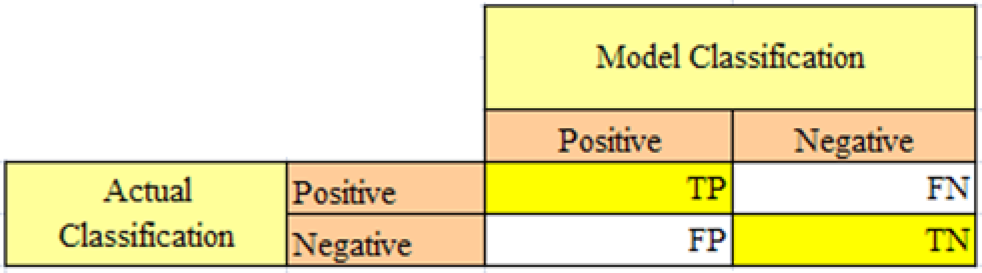


Figure 4.1 Graphical representation of 10-fold cross-validation [29]

In this study, to estimate the performance of classifiers a stratified 10-fold cross-validation approach is used. Empirical studies showed that 10 seem to be an optimal number of folds (that optimizes the time it takes to complete the test while minimizing the bias and variance associated with the validation process) [28]. In 10-fold cross-validation the entire dataset is divided into 10 mutually exclusive subsets (or folds) with approximately the same class distribution as the original dataset (stratified). Each fold is used once to test the performance of the classifier that is generated from the combined data of the remaining nine folds, leading to 10 independent performance estimates (Figure 4.1).

4.3.2 Accuracy, Sensitivity, and Specificity by Confusion Matrix

A confusion matrix (as shown in Figure 4.2) is a matrix representation of the classification results. In a two-class classification problem (as in our case), the upper left cell denotes the number of samples classified as true while they were true in the actual classification (also called true positives), and lower right cell denotes the number of samples classified as false while they were actually false (a.k.a. true negatives). The upper right cell represents the number of samples classified as false while they were actually true (a.k.a. false negatives) and the lower left cell represents the number of samples classified as true while they were actually false (a.k.a. false positives).



		Model Classification	
		Positive	Negative
Actual Classification	Positive	TP	FN
	Negative	FP	TN

Figure 4.2 A confusion matrix representation for two-class classification problem

To compare the classification models, three performance criteria are adopted as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.3}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4.4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4.5}$$

where *TP, TN, FP, FN* denote true positive, true negative, false positive, and false negative, respectively. Accuracy, shown by Eq. (4.3), measures the proportion of correctly classified test examples, therefore predicting the overall probability of the correct classification. Sensitivity and specificity, shown by Eqs. (4.4) and (4.5) respectively, measure the model's ability to *recognize* the patients of a certain group. For example, if the grafts are in case, sensitivity is a probability that a graft which has failed in reality is also classified as failed and specificity is a probability that a succeeding graft is classified as succeeding [29]-[30].

4.3.3 Information Fusion

There is no apply-to-all generic model which would give the best prediction results in predictive modeling. Based on the case study and the data set to be used on hand; the best model can only be determined via several trial-and-error steps [31]. Therefore, rather than relying on the results received from one of the prediction models developed it is suggested combining information received from various models to further improve the prediction accuracy [32]. Such a sophisticated forecast combination would hypothetically render the information more accurate and unbiased. A sample information fusion algorithm was developed by Delen *et al.* [33] which can be summarized as follows:

A prediction model *(f)* can be formulated as in Eq. (4.6) with an output (dependent) variable, *y*, and the input (independent) variables (x_1, x_2, …, x_n)

$$\hat{y} \ = f(x_1, x_2, ..., x_n) \tag{4.6}$$

To exemplify the prediction model, *f*, take into account a single-neuron artificial neural network model which would be written as in Eq. (4.7)

$$f(x_1, x_2, ..., x_n) = \phi(w_0 + \sum_{j=1}^{n} w_j x_j) \tag{4.7}$$

where ϕ is the transfer function and w_i's are the weights for x_i's. With *m* number of prediction models, the information fusion model can be written as in Eq. (4.8)

$$\hat{y}_{fused} = \psi(\hat{y}_{individual,i}) = \psi(f_1(x), f_2(x), ..., f_m(x)) \tag{4.8}$$

If the multi-model fusion algorithm, ψ, is a linear function, then Eq. (4.9) can be rewritten as

$$\hat{y}_{fused} = \sum_{i=1}^{m} \omega_i f_i(x) = \omega_1 f_1(x) + \omega_2 f_2(x) + ... + \omega_m f_m(x) \tag{4.9}$$

where $\sum_{i=1}^{n} \omega_i \ = 1 \cdot$

The values for ω's refer to the weighing coefficient for each prediction model and are the normalized prediction accuracy measure of the individual prediction model (e.g. *accuracy* as calculated in Eq. (4.3)). In other words, a higher weight is assigned to the information provided by a prediction model, which achieves a higher accuracy on the testing (hold-out) dataset [33].

4.3.4 Gains Chart

To measure the performance of Cox regression models, the gains chart analysis is widely used in comparing alternative techniques [34]. It is an application of the Lorenz curve of incremental expenditure to the database marketing setting [35]. Cumulative gains charts always start at 0% and end at 100% while going from left to right. For a good model, the gains chart will rise steeply toward 100% and then level off. A model that provides no information will follow the diagonal from lower left to upper right. The y-axis shows what percentage of cases/observations are captured correctly by the model, given the corresponding percentage of cases/observations handled, indicated on the x-axis. For example, the point (30%, 55%) on a gains chart would indicate that 55% of total cases can be expected to be captured by the target selection model, when 30% of the cases are randomly selected.

4.4 Results and Discussion

4.4.1 Classification Results

Following the methodology proposed in Section 4.2, preliminary analysis showed that neural networks, decision tree, and logistic regression models gave satisfactory high prediction accuracy results in terms of performance measures. Hence, these three models were employed for classification on the dependent variable gstatus. Tables 4.1 shows the confusion matrices for all three models. Based on the confusion matrix, accuracy, sensitivity, and specificity of each fold were calculated using the method presented in Section 4.3. Table 4.1 reveals that neural networks and logistic regression showed similar levels of accuracy while outperforming decision tree model. The results are noteworthy that a statistical technique (i.e. logistic regression) could predict the graft survival as well as a machine learning technique (i.e. neural network). Note that the accuracy level for all three models (are better than any other study reported in the existing literature. Moreover, none of the reported studies used the voluminous lung transplant procedure dataset, and none applied data-mining methodology. These three machine learning models were kept as a modeling technique and some other statistical binary classifier models such as discriminant analysis were eliminated since their accuracy rates were not observed to be satisfactory in our preliminary trials. The cutoff value for success was to adopt a general rule of thumb [36] which claimed that the model should be able to predict the classes 25% better than random chance.

For our case which has 47% and 53% of each class of dependent variables, a “good enough” model should exceed the random chance of 59 % and 66 %, respectively. Hence, neural networks, decision trees, and logistic regression were kept to be utilized to sort out the first set of candidate predictor factors as further explained in Section 4.4.2.

4.4.2 Determination of Candidate Covariates for Cox Regression Modeling

Since the results in Section 4.4.1 were received by 10-fold cross-validation, they are reliable and independent of the random assignment of the testing and training datasets. In the conventional approach, independent variables are identified as “significant” by invoking a variable selection procedure. Subsequent prediction uses the single best model which outperforms the others. Apparently, any such procedure ignores the importance of model uncertainty. They underestimate the uncertainty about the parameters and overestimate the confidence in relying on a specific model to be correct and hence lead to poor predictive ability [37]. Therefore, in our approach, the first set of predictive variables which were commonly utilized in all three classification models (i.e. MLP, C4.5 and logistic regression) were determined through the accuracy metric-based information fusion which was explained in Section 4.3.3 and listed as in Table 4.2. As a rule of thumb we adopted the following assumption: The variable was decided to be important and deserved to be in the first set of potential predictors as presented in Table 4.2. if it was utilized in all three models (MLP, C4.5 and logistic regression) for more than one fold out of all 10 folds of prediction.

Table 4.1 10-fold cross validation results for prediction models

	Neural Networks (MLP)					**Decision Tree (C4.5)**					**Logistic Regression**				
Fold No	Confusion Matrix		Accuracy	Sensitivity	Specificity	Confusion Matrix		Accuracy	Sensitivity	Specificity	Confusion Matrix		Accuracy	Sensitivity	Specificity
1	710	120	0.856	0.858	0.855	612	159	0.794	0.770	0.816	671	115	0.855	0.842	0.866
	118	710				183	704				126	746			
2	670	117	0.853	0.842	0.864	612	159	0.794	0.770	0.816	664	103	0.858	0.833	0.880
	126	745				183	704				133	758			
3	657	111	0.859	0.843	0.874	621	165	0.795	0.781	0.809	666	107	0.857	0.837	0.876
	122	767				174	697				130	754			
4	680	110	0.864	0.855	0.872	619	171	0.791	0.779	0.802	669	121	0.850	0.840	0.859
	115	752				176	691				127	740			
5	668	114	0.854	0.839	0.868	606	158	0.791	0.762	0.817	672	110	0.858	0.843	0.872
	128	748				189	705				125	751			
6	669	106	0.860	0.841	0.877	600	159	0.786	0.755	0.816	674	112	0.859	0.847	0.870
	126	756				195	703				122	749			
7	674	121	0.853	0.847	0.860	609	154	0.795	0.766	0.822	665	108	0.855	0.834	0.875
	122	741				186	709				132	753			
8	672	104	0.863	0.845	0.879	636	188	0.791	0.800	0.782	668	109	0.856	0.838	0.873
	123	759				159	675				129	752			
9	675	108	0.863	0.849	0.875	636	188	0.790	0.799	0.782	667	113	0.853	0.837	0.869
	120	755				160	674				130	748			
10	676	115	0.859	0.850	0.867	630	181	0.791	0.792	0.790	663	107	0.855	0.832	0.876
	119	748				165	682				134	754			
Mean			**0.859**	0.847	0.869			**0.792**	0.777	0.805			**0.856**	0.838	0.872
Std. Dev.			0.004	0.006	0.008			0.003	0.016	0.015			0.003	0.005	0.006

Table 4.2 Variables determined as significant by the classification models

Variables	**Explanation**
Sternotomy_Tcr	Events occurring prior to listing: Sternotomy
Angina_Cad	Recipient angina/cad at registration
Pulm_Inf_Don	Deceased donor-infection pulmonary source
Func_Stat_Tcr	Recipient functional status at registration
Death_Circum_Don	Deceased donor-circumstance of death
Age	Recipient age (yrs)
Cig_Use	History of cigarette use of the recipient

The second set of predictive variables consisted of the ones which are not in the literature but are thought to have importance in lung transplantation. This set also includes the interaction terms which were not in the dataset but were created by us. These binary variable terms are as follows: GINT, the interaction term between gender of recipient and gender of donor; and EINT, the interaction term between the ethnicity (race) of the donor and recipient to see if being in the same gender/race has an influence on survival. The second set of candidate variables are listed in Table 4.3.

The third set of predictive variables was complied by considering the existing studies as mentioned in Chapter II. This set can be referred as the expert component input of our methodology. This set consists of the variables which have been commonly used in literature. The third set of variables is summarized in Table 4.4. The variable names and explanations listed here are provided by UNOS.

Table 4.3 Variables determined due to common-sense domain knowledge

Variables	**Explanation**
*Cancer_Free_Int_Don	Deceased donor-cancer free interval (years)
Cig_Use	History of cigarette use of the recipient
Contin_Cig_Don	Deceased donor-history of cigarettes in past and > 20 pack yrs
*Contin_Cocaine_Don	Deceased donor-history of cocaine use + recent 6 mo. Use
Contin_IV_Drug_Old_Don	Deceased donor-history of iv drug use + recent 6 mo. Use
Contin_Oth_Drug_Don	Deceased donor-history of other drugs in past + recent 6 mo. Use
$^\#$EINT	Ethnicity interaction between donor and recipient
$^\#$GINT	Gender interaction between donor and recipient
Hist_Alcohol_Old_Don	Deceased donor-history of alcohol dependency
Hist_Cancer_Don	Deceased donor-history of cancer
Hist_Cig_Don	Deceased donor-history of cigarettes in past and > 20 pack yrs
Hist_Cocaine_Don	Deceased donor-history of cocaine use in past
Hist_Diabetes_Don	Deceased donor-history of diabetes, incl. Duration of disease
Hist_Hypertens_Don	Deceased donor-history of hypertension
*Hist_Insulin_Dep_Don	Deceased donor-insulin dependent diabetes
Hist_IV_Drug_Old_Don	Deceased donor-history of iv drug use in past
Oth_Tobacco	Other tobacco use
Pack_Yrs	If history of cigarette use, number of pack years

$^\#$*not existing in UNOS dataset but created in this study*
**could not be included in the analysis due to excessive missing values*

The variables listed in Table 4.3 and marked with an asterisk (*) could not be included in any of the analyses (neither in classification nor in Cox modeling) since they had excessive missing values. These variables had less than 5% valid records in the original (not imputed) dataset. Apart from those, the union set of Tables 4.2, 4.3, and 4.4 were used in Cox model as candidate variables to predict the status of the graft.

Table 4.4 Variables determined based on literature research

Variables	**Explanation**
ABO	Recipient blood group at registration
ABO_Don	Donor blood type
ABO_Mat	Donor-recipient ABO match level
Age	Recipient age (yrs)
Age_Don	Donor age (yrs)
Dayswait_Chron	Active days on waiting list
Ethcat	Recipient ethnicity category
Ethcat_Don	Donor ethnicity category
Gender	Recipient gender
Gender_Don	Donor gender
Hbsab_Don	Deceased donor Hbsab test result
Med_Cond_Tcr	Recipient medical condition at registration
Wgt_kg_Don	Donor weight (kg)
Wgt_kg_Tcr	Recipient weight (kg) at registration

4.4.3 Deployment of Cox Regression Modeling

A combined stepwise selection (forward and backward) in Cox regression model was utilized to obtain the survivor function by predicting the gstatus through the time-related variable gtime (graft lifespan). Note that the variable gtime was eliminated in classification models in order to not overshadow the other variables' effect, but here it was needed for Cox modeling. The union set of candidate predictor variables were assigned and stepwise variable selection procedure was run. The variables found significant are listed along with their corresponding statistics in Table 4.5. The rest were eliminated due to their insignificance in the prediction.

Table 4.5 Variables kept in the equation (only the last step, Step 12, is shown)

Variables	DF	Standard Error	Chi-Square Value	Sig.	Exp(β)
Wgt_kg_Tcr	1	2.801E-01	96.906	<.0001	1.997
Func_Stat_Tcr	1	1.492E-02	477.610	<.0001	1.000
Eint	1	2.323E-01	32.549	<.0001	1.142
Sternotomy_Tcr	1	9.576E-03	24074.169	<.0001	0.999
Wgt_kg_Don	1	3.926E-01	56.521	<.0001	1.003
Dayswait_Chron	1	1.284E-02	3060.772	<.0001	1.001
Age_Don	1	3.537E-01	523.719	<.0001	1.008
ABO_Mat	1	1.088E-01	4.664	0.031	0.977
Gint	1	9.247E-03	19.341	<.0001	1.047
Death_Circum_Don	1	9.265E-03	427.637	<.0001	1.000
Med_Cond_Tcr	1	8.051E-06	1026.859	<.0001	1.294
Hist_Diabetes_Don	1	5.049E-01	10.238	0.001	0.998

The effects of individual predictors are represented by the parameter estimates, Exp (β), and can be interpreted as follows: for a categorical variable, say Gint, the value of Exp (β) (1.047) implies that if the donor and the recipient are not of the same gender the risk of graft failing is 1.047 times the failure risk if they are of the same gender. For a continuous variable, say Age_Don (the age of the donor), the risk of graft failing is increased by 1.008 for each increase in one unit change of the donor's age. Note that these selected variables by Cox model include the potential predictor variable elements from all sets we defined earlier. Chi-Square statistics was performed to determine whether or not a specific variable would be kept in the Cox model. The results summarized in Table 4.5 also represent that the significance level (Sig.) of the corresponding variables. The standard errors associated with the corresponding parameter estimates and the degrees of freedom for each test (DF) are included in Table 4.5 for each variable as well.

In order to make a comparison between the proposed methodology and the existing literature, gains charts were utilized in terms of performance measure evaluation. The variables in Table 4.4 were assigned in Cox regression model as a benchmark representative of the state-of-the-art. This Cox model was named Cox-LR (meaning Cox modeling by the variables only from the literature review). Our proposed methodology with aggregation of all variable sets in Tables 4.2, 4.3, and 4.4 was called Cox-PM (meaning Cox modeling by the proposed methodology in this study). Figures 4.3 and 4.4 illustrate the gains charts for Cox-LR and Cox-PM, respectively. Note that the gains chart for Cox-PM has superiority over Cox-LR. The term *gstatus* represents how well the Cox model has done where the best possible prediction would be as *best-gstatus.* The closer the *gstatus* line to *best-gstatus* line, the better the Cox model has performed. Hence, these charts illustrate that the proposed methodology has brought more information to predict the dependent variable *gstatus* and can be proposed as a validation of this study.

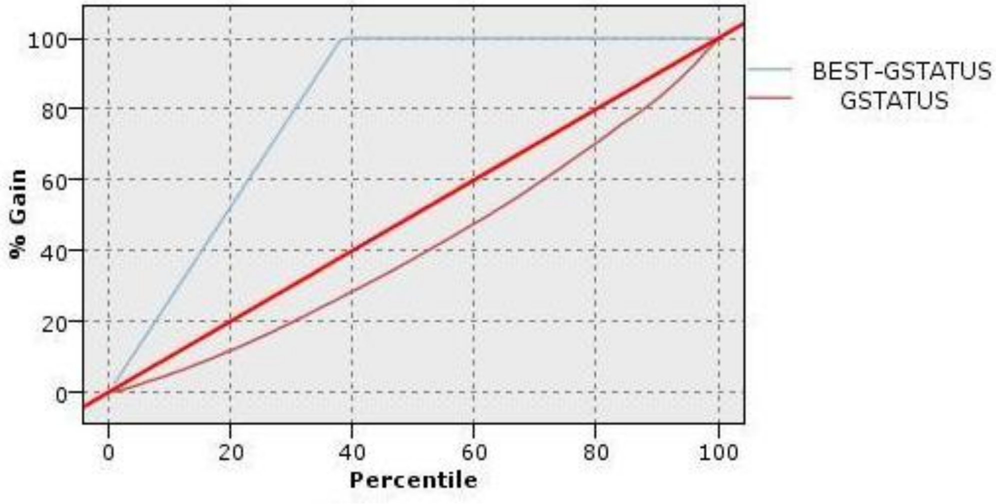


Figure 4.3 Gains chart for Cox-LR

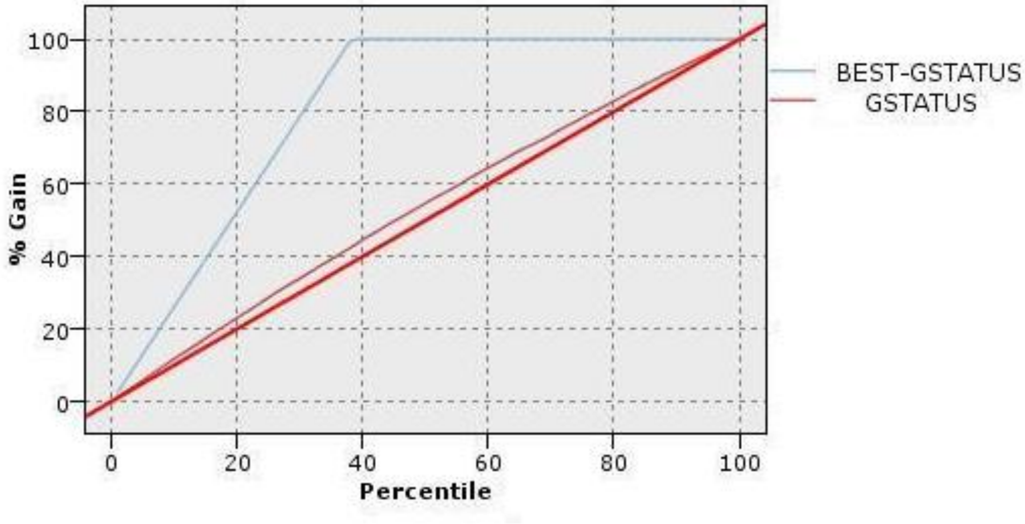


Figure 4.4 Gains chart for Cox-PM

On the other hand, Akaike information criteria (AIC) is a measure for goodness-of-fit of an estimated model and a tool for model selection among competing models

[38]. The smaller the AIC, the better the model has performed. The AIC value for Cox-PM has been received as 1374012.3 and 1465300.2 for Cox-LR. This is another numeric validation of our proposed methodology.

4.5 Conclusions

This study suggests that when modeling lung transplantation procedures a data-mining-driven methodology should be used to augment the variable selection process rather than focusing on mere expert-selected predictor variables. The human expert's input cannot be ignored in modeling lung transplantation (nor can be in any area of medicine) but should be (and as shown in this study, could be) strengthened with the knowledge that can be discovered from data. In order to make use of voluminous datasets, it may be useful to apply the data mining models to extract previously unknown patterns and relationships among the predictor variables. Thus, a small set of effective variables (predictors) could be identified for analysis instead of the original large number of variables, which enables more effective and efficient analyses. This study proposes that the data mining models select the significant variables as the first step. Thereafter, potential variable sets from domain experts will be integrated in the process. In the subsequent analysis, the medical experts should especially be referred to interpret the results that this methodology reveals in lung transplantation. The medical experts are to evaluate the patterns and the newly-introduced predictor variables as to their significance and if they bring new actionable and logical directions in transplant area. An example is the GINT variable which was created in this study, which is shown to be important in predicting graft survival. The medical professionals who have years of experience in the lung transplant area will be expected to decide if it is medically important to assign an organ to a recipient who is the same gender as the donor.

Because of its ability to model highly-complex data-rich phenomenon, predictive data mining is destined to become an essential instrument for researchers in medical informatics. Due to the increasingly more effective and efficient data collection and storage mechanisms in a variety of medical fields coupled with the enormity of ever more complex problems, data mining applications will continue to gain popularity. Future research efforts will involve extension of the data mining analysis for UNOS thoracic dataset along with the follow-up datasets. This perspective will hopefully open a new window to observe patients' medical condition after the lung transplant has been performed. A critical prognostic index can be devised, which categorizes the transplant patients in terms of various risk groups, namely low, medium, and high.

REFERENCES

[1] S. Daar, D.R. Salomon, R.M. Ferguson, J.H. Helderman, P. Macchiarini, New directions for organ transplantation Nature 392 (1998) 11-12.

[2] R.J. Ruth, L. Wysezewianski, G. Herline, Kidney Transplantation: A Simulation Model for Examining Demand and Supply. Management Science 31 (1985) 515-526.

[3] D. Sheppard, D. McPhee, C. Darke, B. Shretha, R. Moore, A. Jurewitz, A. Gray, Predicting Cytomegalovirus Disease After Renal Transplantation: An Artificial Neural Network Approach, International Journal of Medical Informatics 54 (1999) 55-76.

[4] R.S. Lin, S.D. Horn, J.F. Hurdle, S. Goldfarb-Rumyantzev, Single and Multiple Time-Point Prediction Models in Kidney Transplant Outcomes, Journal of Biomedical Informatics (2008) In Press.

[5] J.K. Kirklin, S.V. Pambukian, D.C. McGiffin,, R.L. Benza, Current Outcomes Following Heart Transplantation, Thoracic and Cardiovascular Surgery (2004) 395-403.

[6] P.V. Trigt, D. Davis, G.S. Shaeffer, J.W. Gaynor, K.P. Landolfo, M.B. Higginbotham, V. Tapson, R.M. Ungerleider, Survival Benefits of Heart and Lung Transplantation, Annals of Surgery (1996) 576-584.

[7] L.E. Bennett, B.M. Keck, O.P. Daily, R.J. Novick, J.D. Hosenpud, Worldwide thoracic organ transplantation: a report from the UNOS/ISHLT International Registry for Thoracic Organ Transplantation, Clin. Transpl. (1995) 35-48.

[8] Heart Transplant Statistics – [cited 2008 October 28th]. Available from: http://www.americanheart.org/presenter.jhtml?identifier=4588

[9] Respiratory Disorders: Lung Transplantation – [cited 2008 October 28th]. Available from: http://www.healthsystem.virginia.edu/UVAHealth/peds_respire/lungstran.cfm

[10] R.N. Pierson, M.L. Barr, K.P. McCulluough, T. Egan, E. Garrity, M. Jessup, S. Murray, Thoracic Organ Transplantation, American Journal of Transplantation 4 (2004) 93-105.

[11] F.L. Grover, M.L. Barr, L.B. Edwards, F.J. Martinez, R.N. Pierson, B.R. Rosengard, S. Murray, Thoracic Transplantation, American Journal of Transplantation 3 (2003) 91-102.

[12] A.M. Harper, S.E. Taranto, E.B. Edwards, O.P. Daily, An Update on A Successful Simulation Project: The Unos Liver Allocation Model, in: Proceedings of the Winter Simulation Conference (2000) 1955-1962.

[13] S.A. Cupples, L. Ohler, Transplantation Nursing Secrets, Hanley & Belfus Publication, 2002.

[14] D. Delen, G. Walker, A. Kadam, Predicting Breast Cancer Survival: A Comparison of Three Data Mining Methods, Artificial Intelligence in Medicine 34 (2006) 113-127.

[15] SPSS Inc. (2007) Clementine Data Mining Toolkit, Version 12.0, http://www.spss.com/clementine/.

[16] SAS Institute Inc. (2006), Enterprise Miner Data Mining Toolkit, Version 5.3, http://www.sas.com/technologies/analytics/datamining/miner/.

[17] T. Mitchell, Machine Learning, McGraw-Hill, (1997).

[18] S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice Hall, New Jersey, (1998).

[19] R. Bellazzi, B. Zupan, Predictive Data Mining in Clinical Medicine: Current Issues and Guidelines, International Journal of Medical Informatics 77 (2008) 81-97.

[20] K. Hornik, M. Stinchcombe and H. White, Universal Approximation of an Unknown Mapping and its Derivatives Using Multilayer Feed-forward Network, Neural Networks 3 (1990) 359–366.

[21] J. Quinlan, Induction of decision trees, Machine Learning 1 (1986) 81–106.

[22] J. Quinlan, C4.5: programs for machine learning, Morgan Kaufmann, San Mateo, CA, (1993).

[23] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, Classification and regression trees, Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA, (1984).

[24] S. Dreiseitl, L. Ohno-Machado, Logistic Regression and Artificial Neural Network Classification models: A Methodology Review, Journal of Biomedical Informatics 35 (2002) 352-359.

[25] D.R. Cox, Analysis of Survival Data, Chapman&Hall, London, (1984).

[26] L. Ohno-Machado, Modeling Medical Prognosis: Survival Analysis Techniques, Journal of Biomedical Informatics 34 (2001) 428-439.

[27] P. Grambsch, T. Therneau, Proportional Hazards Rates and Diagnostics Based on Weighted Residuals, Biometrika 81 (1994) 515-526.

[28] R. Kohavi, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: S. Wermter, E. Riloff and G. Scheler, Editors, in : The Fourteenth International Joint Conference on Artificial Intelligence (IJCAI) Montreal, Quebec, Canada, Morgan Kaufman, San Francisco, CA, (1995) 1137–1145.

[29] D.L. Olson, D. Delen, *Advanced Data Mining Techniques*, Springer-Verlag, Berlin-Heidelberg, 2008.

[29] J. Demsar, B. Zupan, N. Aoki, M.J. Wall, T.H. Granchi, J.R. Beck, feature Mining and Predictive Model construction from Severe Trauma Patient’s Data, International Journal of Medical Informatics 63 (2001) 41-50.

[30] D.D. Lewis, Evaluating and Optimizing Autonomous Text Classification Systems, in: Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval, Seattle, WA, (1995) 246–254.

[31] E. Ruiz and F.H. Mieto, A note on linear combination of predictors, Statistics & Probability Letters 47 (2000), pp. 351–356.

[32] R. Batchelor and P. Dua, Forecaster diversity and the benefits of combining forecasts, *Management Science* **41** (1995), pp. 68–75.

[33] D. Delen, R. Sharda, P. Kumar, Movie forecast Guru: A web-based DSS for Hollywood managers, Decision Support Systems 43 (2007) 1151-1170.

[34] P. Furness, New pattern analysis methods for database marketing, Journal of Database Marketing 1 (3) (1994) 220–232.

[35] J. Thompson, Targeting for response value and profit, Journal of Targeting, Measurement and Analysis for Marketing 3 (1994) 133–146.

[36] J.F. Hair, R.E. Anderson, R.L. Tatham, W. Black, Multivariate Data Analysis, Prentice Hall, (1998).

[37] C. Chatfield, Model Uncertainty, Data Mining, and Statistical Inference, Jour. R. Stat. Soc. 158 (1995) 419-466.

[38] K.P. Burnham, D.R. Anderson, Model Selection and Multimodel Inference: A Practical-Theoretic Approach, 2nd ed. Springer-Verlag, (2002).

CHAPTER V

PROGNOSTIC ANALYSIS OF LUNG ORGAN TRANSPLANTS

The prediction of survival time after organ transplantations and prognosis analysis of different risk groups of transplant patients are not only clinically important but also technically challenging. The current studies, which are mostly linear modeling-based statistical analyses, have focused on small sets of disparate predictive factors where many potentially-important variables are neglected in their analyses. Data mining methods, such as machine learning-based approaches, are capable of providing an effective way of overcoming these limitations by utilizing sufficiently large data sets with many predictive factors to identify not only linear associations but also highly complex, non-linear relationships. Therefore, this study is aimed at exploring risk groups of lung recipients through machine learning-based methods.

5.1 Motivation and Background

Thoracic (heart and lung) transplantation has been accepted as a viable treatment for end-stage cardiac and pulmonary failure. The increased experience in cardiac and pulmonary transplantation, improvements in patient selection, organ preservation, and preoperative support have significantly reduced the early threats to patient survival [1]. Over the past decade, the thoracic transplant waiting time for a listed patient has markedly increased, but the number of transplants performed has declined. In addition, the research also found that there is a perceived inequity in access to organs. The organ allocation system need to be improved since it may become a major factor negatively influencing the survivability of thoracic transplant [2].

The survivability prediction is becoming increasingly more important in medicine. When a resource is scarce the need for accurate prediction becomes acute [3]. Especially prediction of survival time and prognosis prediction of medical treatments are clinically important and challenging problems [4]. Scarceness of organs necessitates the development of effective and efficient procedures to select the most optimal organ receiver since demand for organs of all patients might not be satisfied. To achieve this, one critical step is to reveal the knowledge underlying huge amount of data collected and stored from organ transplantation procedures performed in the past. The objectives are (1) to maximize the patients' survival time after the organ transplantation surgery, and (2) to optimize the prognosis for the organ recipients. These can be potentially achieved by discovering the knowledge that may be contained in large dataset consisting of more than hundreds of determinative variables regarding the donors, the potential recipients, and transplantation procedures. Therefore, in this study a data mining method is proposed to process large amount of transplantation data obtained from UNOS to identify the important factors as well as their relationships to the survival of the graft and the patient.

Thereafter, a prognostic index [5]-[6] is developed to classify the patients into different risk groups for better understanding of the transplantation phenomenon. In short, this study will address the following questions: (1) what are the most important variables to be included in an effective prognostic index related to lung organ transplantations? (2) what are the most coherent risk groups that can be formed based on the prognostic index? Predicting the lung survivability and classifying the patients (potential lung organ receivers) into different classes of risks would help decision makers in determining patients' priority for transplantation source assignment.

5.2 Proposed Method for Prognostic Analysis and Risk Group Determination

Section 2.2 shows that the most of the existing studies for organ transplantation procedures utilize conventional statistical approaches such as Kaplan-Meier function and log-rank test along with expert-selected variables to predict the survivability. However, organ transplantation procedures consist of a large number of variables (several hundred) that may have nontrivial impact on modeling the prognosis of the grafts/patients. Using a somewhat comprehensive variable list may help discriminate patients from each other by placing them into proper risk groups. Unintentional omission of the important variables may lead to inaccurate classification of patient risk groups, which may, in turn, lead to less than optimal organ allocation policies and ineffective treatments.

This study is aimed at overcoming the abovementioned shortcomings by employing both machine learning techniques as well as statistical methods to identify the most critical factors affecting the survivability of lung transplant patients.

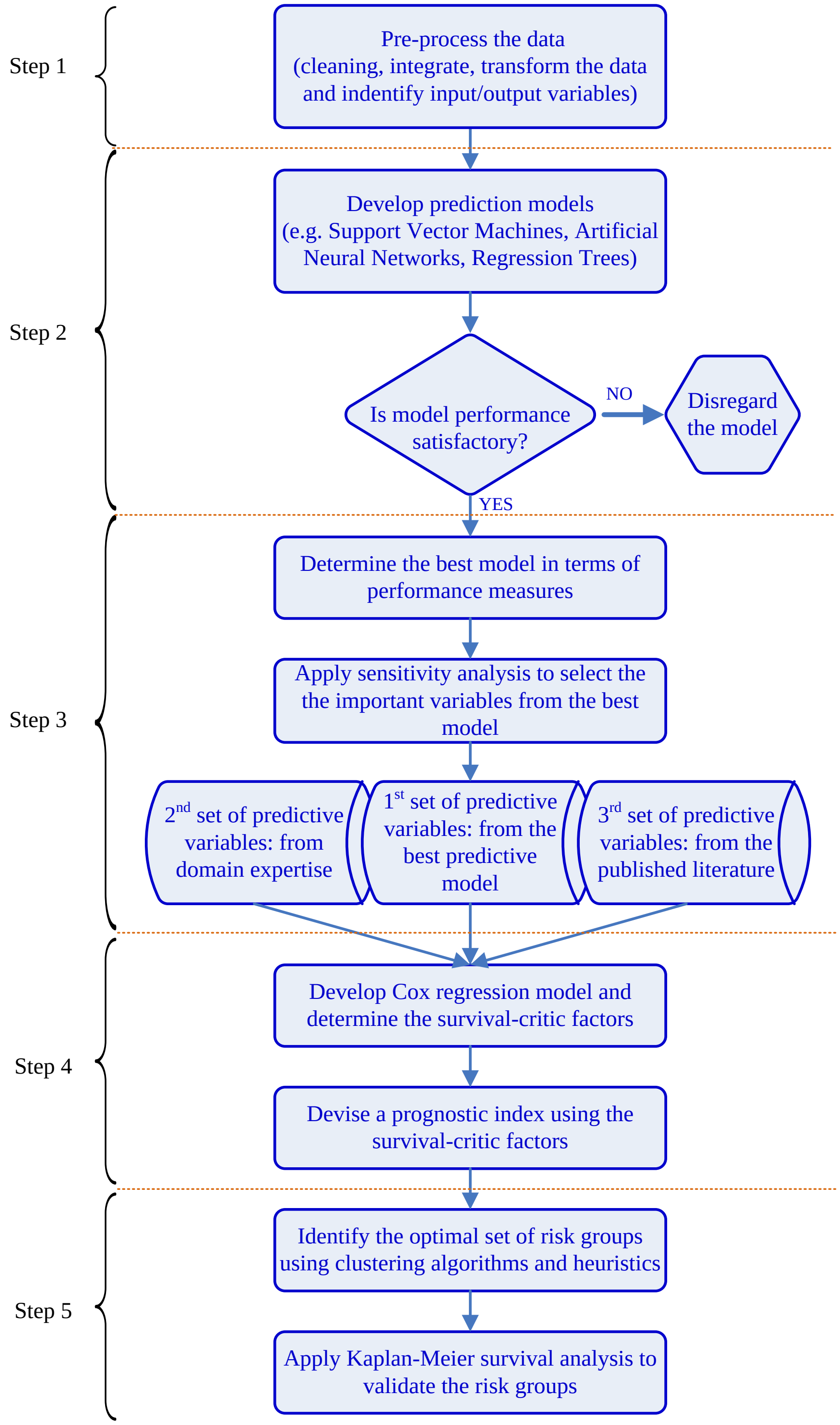

Step 1
Pre-process the data
(cleaning, integrate, transform the data
and indentify input/output variables)
Step 2
Develop prediction models
(e.g. Support Vector Machines, Artificial
Neural Networks, Regression Trees)
Is model performance
satisfactory?
NO
Disregard
the model
YES
Step 3
Determine the best model in terms of
performance measures
Apply sensitivity analysis to select the
the important variables from the best
model
2nd set of predictive
variables: from
domain expertise
1st set of predictive
variables: from the
best predictive
model
3rd set of predictive
variables: from the
published literature
Step 4
Develop Cox regression model and
determine the survival-critic factors
Devise a prognostic index using the
survival-critic factors
Step 5
Identify the optimal set of risk groups
using clustering algorithms and heuristics
Apply Kaplan-Meier survival analysis to
validate the risk groups

Figure 5.1 A flowchart representation of the proposed method

To achieve this goal, this study proposes adopting a 5-step approach illustrated in Figure 5.1. Step 1 involves data understanding and preparation, which is arguably the most time demanding step in the process. Step 2 employs various predictive modeling techniques such as support vector machines, artificial neural networks, and regression trees to develop survival time prediction models and to extract the most important variables by means of sensitivity analysis through the best performing model. Step 3 determines the consolidated candidate set of critical predictor variables. Step 4 develops a Cox regression model using the consolidated set of predictor variables and also devises a prognostic index. The last step, Step 5, classifies the patients into various risk categories by comparing and contrasting the clustering performance of algorithm-based and manually calculated groups. Then the resulting risk categories are validated by using the Kaplan-Meier survival curves.

5.2.1 Data Source and Data Preparation

There are two datasets involved in our study, which are regular dataset and follow-up dataset. The regular dataset contains all information of donors and recipients before transplantation occurred, and the follow-up dataset provides all information of donors and recipients after the transplantation. The TRR_ID variable (transplant identifier) is the common variable between these two datasets and the one which is proposed by UNOS to merge and integrate these two datasets. Therefore, these two datasets were combined in a relational database environment using the link (a.k.a. primary key) of TRR_ID. Overall, the complete dataset consists of 310,773 records and

565 variables. Since the goal of this study is to develop models to predict the survivability solely based on lung transplant, the dependent variable was assigned as *gtime*. This assignment was done to discriminate the patients who died solely due to the lung graft incompatibility from the ones who died from any other reasons. Therefore, the rest of the potential dependent variables (pstatus and ptime) were eliminated from the dataset. Besides, *gstatus* was kept inactive up to the stage where Cox regression model was implemented (Step 4 in Figure 5.1). After adopting various data preparation strategies, the final dataset was reduced to 372 cleansed independent variables and one dependent variable (*gtime*) with the total record count of 106,398.

5.2.2 Predictive Modeling for Prognostic Analysis

Since the dependent variable herein was a continuous variable (graft survival time, which is the number of days from transplant to death or last follow-up), the problem refers to a prediction (or regression) problem (as opposed to a classification problem). Since the relationships between the dependent variable and the independent variables were not known in advance, this step was to develop various predictive models for graft survival time using all of the available independent variables. It is also required to check whether the models have passed the pre-specified threshold values of performance measures, specifically the R^2 and mean square error (MSE), to determine the best model that explains these unknown relationships between dependent and independent variables by ranking them according to these measures. The model which is deemed to be the most successful one would be kept for further modeling steps to determine the importance of the independent variables. Since neural networks and decision trees were already described in Chapter IV, here we briefly define support vector machines only.

Support vector machines (SVMs) are supervised learning methods that generate input–output mapping functions from a set of training data. They belong to a family of generalized linear models which achieve a classification or regression decision based on the value of the linear combination of features. They are also said to belong to the kernel methods [7]. The mapping function in SVMs can be either a classification function (used to categorize the data) or a regression function (used to estimate the numerical value of the desired output, as is the case in this study). Nonlinear kernel functions are often used to transform the input data (inherently representing highly complex nonlinear relationships) to a high dimensional feature space in which the input data become more separable (i.e. linearly separable) compared to the original input space. Then, maximum-margin hyperplanes are constructed to optimally separate the classes in the training data. Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data by maximizing the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes, the better the generalization error of the prediction would be.

5.3 Performance Measures of Model Evaluation

To compare the abovementioned prediction models, two performance criteria are considered: mean squared error (MSE) of the model on testing dataset and R-square value between the actual observation for the target variable (Y_t) and the predicted value by the model (F_t). MSE which is given by the Eq. (5.1) does not have a rule-of-thumb threshold cut-off value for acceptable models. It is a relative criterion to select the best model, namely the smaller the value the better the model has performed [8].

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(Y_t - F_t)^2 \qquad (5.1)$$

On the other hand, R-square ($R^2{}_{F_t,Y_t}$ or shortly R^2) which is given by Eq. (5.2) can be considered as both an absolute measure and a relative measure to determine and rank the satisfactory models [9]. Unlike the MSE, the higher the R^2, the better the performance for the compared models.

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(F_t - Y_t)^2}{\sum_{t=1}^{n}(Y_t - \overline{Y}_t)^2} \qquad (5.2)$$

After selecting the best prediction model based on the performance criteria (i.e. MSE and R^2), it is required to determine the importance of the independent variables. In machine learning algorithms, sensitivity analysis is a method for extracting the cause and effect relationship between the inputs and outputs of a trained model [10]. In the process of performing sensitivity analysis, after the model is trained the learning is disabled so that the network weights are not affected. The fundamental idea is that the sensitivity analysis measures the predictor variables based on the change in modeling performance that occurs if a predictor variable is not included in the model. Hence, the measure of sensitivity of a specific predictor variable is the ratio of the error of the trained model without the predictor variable to the error of the model that includes this predictor variable [10]. The more sensitive the network is to a particular variable, the greater the performance decrease would be in the absence of that variable, and therefore the greater the ratio of importance. This method is followed in support vector machines and artificial neural networks to rank the variables in terms of their importance according to the sensitivity measure defined in Eq. (5.4) [12].

$$S_i = \frac{V_i}{V(F_t)} = \frac{V(E(F_t|X_i))}{V(F_t)} \qquad (5.4)$$

where $V(F_t)$ is the unconditional output variance. In the numerator, the expectation operator E calls for an integral over X_{-i}; that is, over all input variables but X_i, then the variance operator V implies a further integral over X_i. Variable importance is then computed as the normalized sensitivity. Saltelli *et al.* [13] show that Eq. (5.4) is the proper measure of sensitivity to rank the predictors in order of importance for any combination of interaction and non-orthogonality among predictors. As for the decision trees, variable importance measures were used to judge the relative importance of each predictor variable. Variable importance ranking uses surrogate splitting to produce a scale which is a relative importance measure for each predictor variable included in the analysis. Further details on this procedure can be seen in Breiman *et al.* [14].

5.4 Determining Candidate Sets of Predictor Variables

Step 3 is to determine which predictor variables to be used in devising a prognostic index in Step 4. This step helps eliminate the insignificant variables and improves the accuracy of the model by optimizing the predictor variables list. The potential input variables to this step consist of three candidate sets of predictor variables. The first set is composed of variables selected by the predictive models. The predictive models rank the predictor variables based on their importance level in predicting the graft survival time. The predictive variables selected by the sensitivity analysis of the best-performing model (ranked in terms of R^2 and MSE) are chosen as the first set of predictive variables. The second set of predictive variables is obtained by considering the expert domain knowledge. This set includes variables which are logically related to heart and lung transplantation such as donor's history of cigarette usage. The third set of predictive variables is selected from the related literature. This set consists of the variables which have been commonly and repeatedly used in previous studies in the organ transplantation area. The second and third sets of predictive variables provide more comprehensive information for the next step, the Cox regression model, by including the variables that might have importance in the survival analysis but were determined to be insignificant by the predictive models in step 2.

5.5 Prognostic Index Devising for Lung Transplants

Step 4 takes all the three sets of predictive variables identified in Step 3, and then applies Cox regression to model the graft survivability and filter out the candidate predictive variables which do not have significant survival effect. Hence, in Step 4 the final critical predictive variables are determined by the Cox regression model. Cox regression model also enables devising a prognostic index to categorize the patients into various groups with different levels of risks. One important application of Cox regression model is to identify variables which may be of prognostic importance [5]. Once identified, knowledge from these variables will be combined and used to define a prognostic index, which in turn defines groups of organ recipients with different levels of risk. To use the prognostic index, key patient characteristics are recorded, from which a score is derived. This score gives an indication of whether a particular patient has high, intermediate or low levels of prognosis for the disease [5]-[15]. Recalling Eq. (4.1), the prognostic index (PI) for each patient can be calculated by Eq. (5.5):

$$PI = x_1.\beta_1 + x_2.\beta_2 + ... + x_n.\beta_n \tag{5.5}$$

Note that PI in Eq. (5.5) represents the exponent portion in Eq. (4.1). Therefore, the smaller the PI, the smaller the hazard function value, and hence the smaller the risk associated with a particular recipient.

An important question following Step 4 is “How many risk groups should the patients be classified into?” In Step 5, k-means clustering algorithm, two-step cluster analysis, and conventional heuristics-based approaches are used to answer to this question. As a statistical and/or pictorial verification mechanism for the number of groups determined by the best performing abovementioned clustering approaches, finally the Kaplan-Meier survival analysis [16] is adopted and corresponding survival curves are generated.

k-means method is an extensively used, arguably the most popular clustering algorithm that searches for a nearly optimal partition with fixed number of clusters represented by the parameter k [17]. It proceeds by assigning k initial centroids to the multi dimensional datasets. Each record in the dataset is allocated to the centroid which is nearest and hence forming a cluster. Each cluster centroid is then updated to be the center of its members, followed by a new assignment of records to the nearest centroids to re-construct the clusters. The algorithm converges when there is no further change in allocation of members to clusters or some predefined time-based stopping criteria is satisfied [18].

Another popular clustering algorithm is two-step cluster analysis (TSCA) [19]-[20]. It has two steps: (1) to pre-cluster the cases (or records) into many small sub-clusters, and (2) to cluster the sub-clusters resulting from pre-cluster step into the desired number of clusters. The pre-cluster step uses a sequential clustering approach. It scans the data records one by one and decides if the current record should be merged with the previously formed clusters or starts a new cluster based on the distance criterion. Then the cluster step takes sub-clusters resulting from the pre-cluster step as input, and groups them into the desired number of clusters. Since the number of sub-clusters is much less than the number of original records, the traditional clustering methods can be used effectively. This step uses the agglomerative hierarchical clustering method [19]-[20]. Although there are several other clustering algorithms (e.g. Kohonen networks) they do not allow the modeler to specify a desired number of clusters at the beginning of the clustering algorithm. k-means and TSCA algorithms overcome this issue. The modeler can predefine a specific number of clusters to group the variables and compare them according to their clustering performances. Since this is the main focus of our study, we utilized k-means and TSCA algorithms for clustering the PIs and thus identify the risk groups of lung patients.

The Kaplan–Meier analysis is a non-parametric technique used to test the statistical significance of differences between the survival curves associated with two different circumstances [16]. The analysis expresses the distribution of patient survival times in terms of the proportion of patients still alive up to a given time. On the other hand, the Kaplan-Meier survival curves plot the proportion of patients surviving against time which has a characteristic decline. In biostatistics, a typical application of Kaplan-Meier survival curves involves grouping patients into risk groups such as low, medium, and high risks.

5.6 Results and Discussion

5.6.1 Prediction Model Results

To reveal the initially unknown relationship between the lung input/independent variables and the continuous output/dependent variable (gtime), due to the high computational time required for 10-fold cross validation of each model we only used two most popular models from each family of machine learning techniques. Radial basis function (RBF) and polynomial functions as Kernel methods in support vector machine were deployed. We used multi layer perceptron (MLP) and RBF type of network structures for ANNs. The most recent algorithms C&RT and M5 were utilized for prediction with the decision trees. The 10-fold averaged prediction results in terms of MSE and R^2 for each model are tabulated in Table 5.1.

Table 5.1 Comparison of machine learning prediction model results

Performance Measures **Prediction Models**	**MSE**	**R^2**
Support Vector Machine		
RBF	0.023	0.879
Polynomial	0.793	0.643
Artificial Neural Network		
MLP	0.031	0.847
RBF	0.146	0.835
Decision Tree		
M5	0.324	0.785
C&RT	0.578	0.766

The acceptance of predictive models is first evaluated based on their coefficient of determination (R-square) values. It is widely accepted that if R-square is higher than 0.6, the predictive model has performed fairly well [21]-[22]. Therefore, we set this as a threshold value for the model sufficiency. Since all the models have passed this threshold, we kept the one with the highest R^2 and the smallest MSE for further analyses, which came out to be the support vector machine model with radial basis Kernel function in this case study.

5.6.2 Determination of Candidate Covariates for Prognostic Analysis

Step 3 in the proposed method provides three different sets of candidate covariates to be used in the Cox model.

Table 5.2 The 1st set of candidate covariates generated from RBF-SVM

Variables	**Explanation**
Citizenship	Recipient citizenship @ registration
Contin_alcohol_old_don	Deceased donor-history of alcohol dependency+ recent 6mo use
Contin_iv_drug_old_don	Deceased donor-history of iv drug use+recent 6mo use
Creat2_old	Most recent creatinine > 2.0 mg/dl y/n
Da2	Donor a2 antigen
Dantiarr_old	Deceased donor given antiarrythmics 24 hours prior to cross clamp
Dayswait_chron	Active days on waiting list
Dobut_don_old	Deceased donor-dobutamine w/in 24 hrs pre-cross clamp
Education	Recipient highest educational level @ registration
Ethcat_don	Donor ethnicity category
Func_stat_tcr	Recipient functional status @ registration
Func_stat_trr	Recipient functional status @transplant
Gender	Recipient gender
Hbsab_don	Deceased donor hbsab test result
Hemo_pa_dia_tcr	Most recent hemodynamics pa (dia) mm/hg @ registration
Hemo_pa_mn_tcr	Most recent hemodynamics pa (mean) mm/hg @ registration
Heparin_don	Deceased donor management - heparin
Hgt_cm_tcr	Recipient height @ registration
Hist_alcohol_old_don	Deceased donor-history of alcohol dependency
Htlv2_old_don	Deceased donor-antibody to htlv ii result
Impl_defibril_after_list	Implantable defibrillator inserted between listing and transplant
Inotrop_agents	Deceased donor- three or more inotropic agents at time of incision
Inotrop_support_don	Deceased donor inotropic medication at procurement (y/n)
Med_cond_tcr	Recipient medical condition @ registration
Med_cond_trr	Recipient medical condition pre-transplant @ transplant
Physical_capacity_tcr	Physical capacity at listing
Pretreat_med_don_old	Deceased donor medication(s) from brain death to 24 hrs prior to procurement
Prior_lung_surg_tcr	Recipient prior lung surgery (non-transplant) at listing
Pt_t4_don	Deceased donor-thyroxine-t4 b/n brain death w/in 24 hrs of procurement
Sternotomy_tcr	Events occurring prior to listing: sternotomy
Sternotomy_trr	Events occurring between listing and transplant: sternotomy
Steroid	Chronic steroid use y/n/u @ transplant
Trtrej1y	Treated for rejection within 1 year
Trt_pulm_sepsis	IV treated pulmonary sepsis y/n/u @ registration
Vad_tah_tcr	Recipient on life support - ventilator @ registration (1=yes, 0=no)

Since the best performing model to explain the relationships of independent and dependent variables was found to be RBF-SVM, the sensitivity analysis as explained in

Section 5.3 by Eq. (5.4) was conducted on the predictor variables to rank them in terms of their importance in predicting the gtime (c.f. Table 5.2). The second set of predictor variables were selected by the authors through brainstorming sessions with medical professionals as summarized in Table 4.3. The third set of candidate covariates was determined through the recent literature [23]. This set includes the variables commonly used in the previously published studies related to organ transplantation. The third set of candidate covariates are shown in Table 4.4. The second and third set of candidate covariates can be perceived as the expert component of the method.

If the predictive models in Step 3 do not reveal some very critical predictor variables (such as the age of the recipient in our case study), the method proposes to *force* the Cox model once more to review the significance of this kind of predictor variables.

5.6.3 Devising the Lung Prognostic Index

All the candidate covariates as determined in Section 5.6.2 were assigned to Cox regression model at this step. The stepwise variable selection procedure was applied with 0.05 for entry and 0.1 for removal as significance threshold criteria. The predictor variables determined to be significant by Cox regression model are listed along with their corresponding statistics in Table 5.3.

The rest of the variables (which were in Tables 5.2, 4.3, or 4.4 but not in Table 5.3) were eliminated since they were found to be insignificant by Cox regression model. As listed in Table 5.3, 9 of the variables had prognostic value which are determined by the Cox model as significant and kept in the Cox equation. Therefore, they were used to calculate the PIs by means of Eq. (5.5). The PI values received here were ranging between 0 and 3.

Table 5.3 The variables kept in the Cox regression model

Variable	SE	Chi_square Test	DF	Significance	exp(β)	95% CI for exp(β)	
						Lower	Upper
Eint	0.0178	56.9447	1	<.0001	0.844	0.844	0.905
Gint	0.0183	11.8644	1	0.0006	0.906	0.906	0.973
Age_Don	0.0006	247.3162	1	<.0001	1.009	1.009	1.011
Wgt_kg_Tcr	0.0004	5.5091	1	0.0189	0.998	0.998	1.000
Wgt_kg_Don	0.0005	21.3483	1	<.0001	0.997	0.997	0.999
Citizenship	0.0554	5.5538	1	0.0184	0.787	0.787	0.978
Dayswait_Chron	0.0002	7.5318	1	0.0061	1.000	1.000	1.000
Med_Cond_Tcr	0.0109	75.6231	1	<.0001	1.076	1.076	1.123
Vad_Tah_Tcr	0.0077	48.9955	1	<.0001	1.040	1.040	1.072

5.6.4 Clustering the Prognostic Indices and Creating the Risk Groups

Once the prognostic indices (PIs) for each recipient calculated, the next step was to cluster the recipients through these PIs. However, the problem of defining these clusters and deciding which value to cut off and categorize the recipients should be solved first. Two commonly used clustering algorithms as described in Section 5.5, namely k-means and TSCA were used to determine these clusters. We also compared these algorithm-based clusters to conventional PI devising methods in medicine. Two potential ways to do the clustering are constructing equal-width PIs and equal-percentile PIs in this research domain. In the former one, the PIs are separated in groups so that the increments of PI in each group are equal whereas the latter method focuses on allocating the patients equally to each group. The algorithms k-means and TSCA were run by changing the value for k (number of clusters to be formed).

Table 5.4 The comparison of results for clustering algorithms and heuristics

		By Clustering Algorithms					
		k-means algorithm			**two-step cluster analysis**		
Number of Clusters	**Risk Group**	**Prognostic Index**	**Number of patients**	**Intraclass inertia**	**Prognostic Index**	**Number of patients**	**Intraclass inertia**
Cluster 1	Low	0-0.69	21163 (58%)	12.4*10^-8	0-1.09	34199 (94%)	866.30*10^-8
Cluster 2	High	0.70-3	15262 (41%)		1.1-3	2226 (6%)	
Cluster 1	Low	0-0.56	13766 (38%)	1.68*10^-8	0-1.04	33529 (92%)	2.20*10^-8
Cluster 2	Medium	0.57-0.91	5834 (16%)		1.05-1.83	2807 (7.7%)	
Cluster 3	High	0.92-3	16825 (46%)		1.84-3	89 (0.3%)	
Cluster 1	Low	0-0.49	15227 (42%)	11.2*10^-8	0-0.41	6410 (17%)	445.39*10^-8
Cluster 2	Low-Medium	0.50-0.77	1764 (5%)		0.42-0.70	15163 (42%)	
Cluster 3	High-Medium	0.78-1.12	9542 (26%)		0.71-1.04	11892 (33%)	
Cluster 4	High	1.13-3	9892 (27%)		1.05-3	2960 (8%)	
Cluster 1	Very Low	0-0.44	13266 (36%)	3.02*10^-8	0-0.36	2960 (8%)	720.71*10^-8
Cluster 2	Low	0.45-0.69	451 (1%)		0.37-0.53	10475 (29%)	
Cluster 3	Medium	0.70-0.95	4449 (12%)		0.54-0.73	10815 (29%)	
Cluster 4	High	0.96-1.39	7814 (22%)		0.74-1.04	4674 (13%)	
Cluster 5	Very High	1.40-3	10445 (29%)		1.05-3	7501 (21%)	

		By Heuristics-based Calculation					
		with equal PI widths			**with equal percentiles**		
Number of Clusters	**Risk Group**	**Prognostic Index**	**Number of patients**	**Intraclass inertia**	**Prognostic Index**	**Number of patients**	**Intraclass inertia**
Cluster 1	Low	0-1.5	36154 (99%)	713.68*10^-8	0-0.64	18212 (50%)	7.02*10^-6
Cluster 2	High	1.6-3	271 (1%)		0.65-3	18213 (50%)	
Cluster 1	Low	0-0.9	32571 (89%)	1678.65*10^-8	0-0.53	12142 (33.5%)	2.01*10^-6
Cluster 2	Medium	1-1.9	3794 (10%)		0.54-0.76	12141 (33%)	
Cluster 3	High	2-3.0	60 (1%)		0.77-3	12142 (33.5%)	
Cluster 1	Low	0-0.7	26087 (72%)	12961.43*10^-8	0-0.47	9106 (25%)	2755.48*10^-6
Cluster 2	Low-Medium	0.8-1.5	10153 (28%)		0.48-0.64	9106(25%)	
Cluster 3	High-Medium	1.6-2.3	162 (0.4%)		0.65-0.82	9106 (25%)	
Cluster 4	High	2.4-3	23 (0.06%)		0.83-3	9107 (25%)	
Cluster 1	Very Low	0-0.5	15605 (43%)	457.67*10^-8	0-0.43	7285 (20%)	3.16*10^-6
Cluster 2	Low	0.6-1.1	19608 (54%)		0.44-0.58	7285 (20%)	
Cluster 3	Medium	1.2-1.7	1109 (3%)		0.59-0.71	7285 (20%)	
Cluster 4	High	1.8-2.3	80 (0.2%)		0.72-0.87	7285 (20%)	
Cluster 5	Very High	2.4-3	23 (0.06%)		0.88-3	7285 (20%)	

The value of k with 2, 3, 4, and 5 were tried because it was considered that having clusters more than 5 would not provide logical risk groups to categorize and would probably not be easy to name and interpret medically afterwards. The results for each run are represented in Table 5.4. The performance of these entire four approaches with different number of clusters (k=2, 3, 4, 5) was compared using intraclass inertia as the performance measure to decide which one to adopt. It is a measure which shows how compact each cluster is. Intraclass inertia is the average of the distances between the means and the observations in each cluster. Eq. (5.6) indicates this value for given k number of clusters [24].

$$F(k) = \frac{1}{n}\sum_{k}\sum_{i \in C_k}\sum_{P=1}^{m}(X_{iP} - \mu_{kP})^2 \tag{5.6}$$

where n is the number of total observations, C_K is the set of k^{th} cluster, X_{iP} is the value of the attribute P for observation i and μ_{kP} is the mean of the attribute P in the k^{th} cluster. Note that in our case there is only one attribute which is PI, and hence m=1. The intraclass inertia values for each possible cluster are also summarized in Table 5.4. Prognostic indices were clustered best with *k=3* with k-means clustering algorithm in our case as seen in Table 5.4 considering its low intraclass inertia value. As seen in Table 5.4, this classification not only gives the lowest intraclass inertia value but also provides a even distribution of the patients for our nation-wide dataset (38%, 16%, and 46% for low, medium, and high risk groups of patients respectively). Although 5 clusters with k-means algorithm and 3 clusters in two-step cluster analysis perform very close to k-means algorithm with 3 clusters, neither of them provides such an even distribution of patients.

Note that in addition to considerably higher inertia scores, heuristic calculation with equal-width PIs distribute the nation-wide patients highly skewed to lower tails of risk groups for all five potential cluster formations. Therefore, we conclude that the k-means algorithm based clustering performs better than the other potential groupings in terms of both objective and subjective aspects.

5.6.5 Validation of Risk Groups by Kaplan-Meier Survival Analysis

To validate the established prognostic indices with 3 clusters in Section 5.6.4 and hence the various risk groups, Kaplan-Meier survival analysis was conducted. The corresponding PI clusters were matched with the patients and their predictor variables from Table 5.3. In Kaplan-Meier survival analysis the predictor variables were used as explanatory variables and the PI-based clusters were used as the strata variable to label the patients with different risks. The main objective here was to compare survivor functions for different risk groups of lung recipients. If the survivor function for one risk group is always higher than the survivor function for another risk group, than the first group clearly lives longer than the second one. The less the survivor functions cross, the better the discrimination of the patients would be. Figure 5.2 shows this clear distinction for k-means algorithm-based PIs.

In order to show that there is a statistically significant difference among these three risk groups, the test of equality over strata was also conducted. Test of equality over strata contains rank and likelihood-based statistics for testing homogeneity of survivor functions across strata. The rank tests with the log-rank test and Wilcoxon test indicate a significant difference between the risk groups. These results are also supported by likelihood-based statistics. These statistical test results are summarized in Table 5.5.

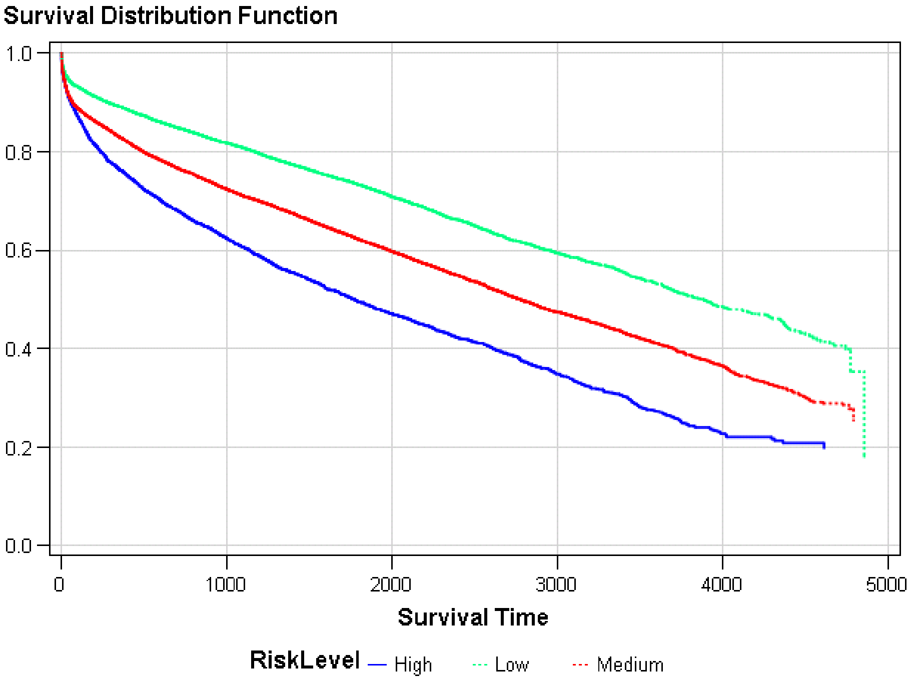


Figure 5.2 Kaplan-Meier survival curves for three PIs

Table 5.5 Tests of equality over risk groups for k-means based three PI cluster

Test	Chi-Square	DF	Pr>Chi-Square
Log-Rank	1002.6135	2	<.0001
Wilcoxon	939.7492	2	<.0001
-2Log(LR)	1013.3153	2	<.0001

5.7 Conclusions

This study demonstrates that machine learning-based methodology for selecting predictor variables in survivability and prognostic modeling of lung organ transplantation is superior to the approaches adopting only expert-selected variables. The study showed that of the comprehensive list of predictors, some have been included in the previous studies (such as gender and age of the recipient, his/her medical condition at registration) while some others (which are found to be critical) have been absent from the related literature. These variables (e.g. such as recipient length of stay post transplant and the interaction of gender and ethnicity between the recipient and the donor) should be combined with the factors identified in previous studies to better understand and improve the organ transplantation process.

The study revealed that based on k-means clustering algorithm the lung organ recipients should be allocated into an optimal number of "three" risk groups, namely low, medium, and high. This finding confirms the conventional medical discrimination commonly used in this field of study. However, it also proves that this grouping should be better performed through a data mining perspective rather than a heuristics-based approach because the latter one gives more skewed distribution of patients for our US nation-wide dataset. This is the point where the medical professionals should be advised to handle the problem in the future.

Some of the research extensions to the study reported in this manuscript includes analysis of other organ types as well as the analysis of multi organ scenarios where the correlations among the organs coming from the same donor are also included in the formulation of the problem. Another potential further research direction of this study is to validate the patterns obtained from the data mining models with a comprehensive simulation model of the organ transplantation process. Using actual cases, a comprehensive discrete-event simulation model can be developed and be used as a testbed where the potential benefits and limitations of these novel patterns are tested and validated lengthy period of time in the computer simulation environment.

REFERENCES

[1] P.V. Trigt, D. Davis, G.S. Shaeffer, J.W. Gaynor, K.P. Landolfo, M.B. Higginbotham, V. Tapson, R.M. Ungerleider, Survival Benefits of Heart and Lung Transplantation, Annals of Surgery 223 (1996) 576-584.

[2] R.N. Pierson, M.L. Barr, K.P. McCullough, T. Egan, E. Garrity, M. Jessup, S. Murray, Thoracic organ transplantation, American Journal of Transplantation 4 (2004) 93-105.

[3] D. Sheppard, D. McPhee, C. Darke, B. Shretha, R. Moore, A. Jurewitz, A. Gray, Predicting Cytomegalovirus Disease After Renal Transplantation: An Artificial Neural Network Approach, International Journal of Medical Informatics 54 (1999) 55-76.

[4] R.S. Lin, S.D. Horn, J.F. Hurdle, S. Goldfarb-Rumyantzev, Single and Multiple Time-Point Prediction Models in Kidney Transplant Outcomes, Journal of Biomedical Informatics (2008) In Press.

[5] M.K.B. Parmar, D. Machin, Survival Analysis: A Practical Approach (John Wiley & Sons, Cambridge, UK, 1996).

[6] D.R. Cox, Analysis of Survival Data (Chapman&Hall, London, 1984).

[7] N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, (Cambridge University Press, 2000).

[8] S. Makridakis, S.C. Wheelwright, R.J. Hyndman, *Forecasting: Methods and Applications* (John Wiley and Sons, 1998).

[9] B.S. Everitt, *Cambridge Dictionary of Statistics*, (2002).

[10] G. Davis, Sensitivity Analysis in Neural Net Solutions, IEEE Transactions on Systems, Man, and Cybernetics 19 (1989) 1078-1082.

[11] J.C. Principe, N.R. Euliano, W.C. Lefebvre, Neural and Adaptive Systems, (John Wiley and Sons, 2001).

[12] A. Saltelli, Making best use of model evaluations to compute sensitivity indices, Computer Physics Communications 145 (2002) 280–297.

[13] A. Saltelli, S. Tarantola, F. Campolongo, M. Ratto, Sensitivity Analysis in Practice – A Guide to Assessing Scientific Models (John Wiley and Sons, 2004).

[14] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, *Classification and regression trees* (Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA, 1984).

[15] E. Christensen, Multivariate Survival Analysis Using Cox's Regression Model, *Hepatology 7* (1987) 1346-1358.

[16] E. Kaplan, P. Meier, Nonparametric Estimation from Incomplete Observations, *Journal of the American Statistical Association 53* (1958) 187-220.

[17] J. B. MacQueen, Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Symposium on Math, Statistics, and Probability,* University of California Press, Berkeley, CA, USA (1967), pp. 281-297.

[18] K. Krishna, M.N Murty, Genetic k-Means Algorithm, *IEEE Transactions on Systems*, Man, and Cybernetics-Part B: Cybernetics 29 (1999) 433-439.

[19] T. Chiu, D. Fang, J. Chen, Y. Wang, C. Jeris, A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. In: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining (2001), 263.

[20] Z.H. Li, P. Luo, Statistical Analysis Lectures of SPSS for Windows (Beijing: Publishing House of Electronics Industry, 2004).

[21] J.F. Hair, R.E. Anderson, R.L. Tatham, W. Black, *Multivariate Data Analysis* (Prentice Hall, 1998).

[22] D.E. Johnson, *Applied Multivariate Methods for Data Analysts* (Duxbury Press, Pacific Grove, CA, 1998).

[23] A. Oztekin, D. Delen, Z.J. Kong, Predicting the graft survival for heart–lung transplantation patients: An integrated data mining methodology, *International Journal of Medical Informatics* 78 (2009), e84-e96.

[24] P. Michaud, Clustering Techniques, *Future Generation Computer Systems 13* (1997) 135-147.

CHAPTER VI

DEVELOPING A COMPOSITE SCORE OF MATCHING INDEX FOR LUNG TRANSPLANTS

Thoracic (heart-lung) transplantation has a vital role among all organ transplant procedures since it is the only accepted optimal treatment for the end-stage cardiac and pulmonary failure. There have been several research attempts to model the performance of lung transplants. Yet, they either lack model predictive capability by relying on strong statistical assumptions or provide adequate predictive capability but suffer from less interpretability to the medical professionals.

The proposed method in this chapter is focused on overcoming the abovementioned limitations by providing a structural equation modeling-based decision tree for lung transplant performance evaluation. Specifically, partial least squares-based path modeling is used for the structural equation modeling part. The proposed method is validated through a US nation-wide dataset obtained from United Network for Organ Sharing (UNOS). The results are promising in terms of both prediction and interpretation capabilities and are superior to the existing techniques. Hence, a proposed method-based decision support system can bridge the gap between the large amount of available data and in-depth understanding of the lung transplant procedure.

6.1 Motivation and Background

Organ transplantation is regarded as a viable treatment for the chronic failure of major organs and is an inevitable option for the end-stage cardiac and pulmonary failure, namely thoracic (heart/lung) patients [1]. Although lung transplantation is the accepted optimal treatment for eligible patients, the shortage of organs seriously limits this option. Additionally, a significant number of organs are rejected due to a suboptimal match between the donor and the recipient. Benefit-driven organ allocation schemes, where post-transplant outcome is taken into account as a performance criterion, are very attractive approaches because they are targeted at ensuring that organs are not wasted on patients who would not benefit from them [2]. Recently, the demand for organ transplantation has drastically increased whereas the number of donors has remained almost the same, which, in turn, caused longer lists of patients waiting for transplantation [3]. Therefore, outcome prediction (i.e. transplant success) has emerged as a critical issue in organ transplantation. Moreover, when a resource (the donor organ in this case) is scarce, the need for an accurate outcome prediction becomes acute [4]. Especially *prediction of survival* and the *quality of life* are clinically important but challenging problems [5]. However, the design of such schemes is very complex, even more difficult to validate and to control the outcome of the transplantation [2]. Therefore, modeling such a system necessitates effective procedures for the selection of optimal organ recipients since currently it is not possible to satisfy all organ demands.

Voluminous data has been collected from lung transplant procedures and analyzed to evaluate the organ allocation process [6]. Attempts to analyze lung transplants with this huge amount of data have focused on identifying the characteristics of lung transplant recipients and their associated post-transplant outcomes [7]. However, they have not analyzed the allocation procedure in a cause-and-effect relationship perspective. To fill this gap, our study handles benefit-driven organ allocation schemes in terms of "causality" perspective because such a methodology would give clearer *interpretability* as well as a better *prediction accuracy* of the transplant success. While the former is extremely important to the medical professionals, the latter is critical to establish a satisficing optimal allocation scheme [8].

6.2 Proposed Method for Deriving a Composite Score of Organ Matching Index

The related research work summarized in Chapter II studied the organ transplant success in a cause-and-effect relationship and analyzed the predictor factors as both independent and dependent variables. However, most of the data used in the literature is obtained by conducting surveys on the patients. Such an approach broadens the scope of voluminous datasets to a small set of predictors by bringing the previous data collection efforts to naught. Also, the relationships among the aggregated constructs are limited to linearity which may not hold true in reality. Although structural equation modeling presents a clear depiction of causality, it lacks prediction accuracy since it is a model-testing approach rather than a prediction method. There is a trade-off between model prediction accuracy and interpretability. In order to better handle this trade-off, we propose a structural equation modeling-based decision tree construction. Such a method can identify the causality with high prediction accuracy. Thus, it would satisfy the medical professionals by its clear interpretability and predictability for a benefit-driven allocation scheme considering the expected transplant success.

In this study, considering the recent literature [9]-[10] we chose 27 variables from the UNOS database to predict the transplant success. Among these 27 variables, GTIME (graft lifespan from transplant to death/last follow-up) and FUNC_STAT_TRF (functional status at last follow-up) reflect the success rate of the organ transplantation. Hence, these two variables will be combined to create the performance measure, namely *transplant success* for organ transplantation in this study. This relationship is shown by Model 1 in Figure 6.1. The rest 25 variables are considered as the *causal indicators*, which are associated with the 3 main decision variables used by the medical professionals to model the organ transplantation. These 25 variables are listed in Table 6.1 along with their brief explanations. Three decision variables include (1) *recipient's profile*, (2) *donor's profile*, and (4) *match level*. Although these 3 decision variables cannot be found from the UNOS database directly, they are related to the 25 variables chosen from the database. The mapping between these 3 decision variables and the 25 variables from the UNOS database is constructed based on the medical knowledge in organ transplantation. Their quantitative relationship can be obtained using formative modeling which would be explained in Section 6.3.1. Thus, we also call these 3 decision variables as composite (or latent) variables. We can consider that these 3 decision variables are in fact the latent/composite variables hidden behind the 25 causal indicators. These latent/composite variables and their underlying causal indicators are pictorially summarized as Model 2 in Figure 6.1. Note that in this study *latent/composite variables* are always written in lower case to discriminate them from their corresponding causal indicators (all of which are written in upper case). To model the underlying causal relationship between these 3 decision variables and their corresponding 25 causal indicators and between transplant success and its items (GTIME and FUNC_STAT_TRF), we use partial least squares (PLS) path modeling technique because it allows to construct the formative models (as well as reflective relations), both of which are required in this study. In formative modeling, the causal indicators affect on their corresponding composite variable as shown in Model 2 of Figure 6.1. In other words, in formative modeling the composite variable would be determined by its causal indicators. In contrast, in reflective modeling the latent variable drives its indicators. To exemplify, referring to Model 1 in Figure 6.1 if the transplant has been conducted successfully (referring to *transplant success*), the patient would live for a long time (referring to GTIME) with a high quality of life (referring to FUNC_STAT_TRF). On the other hand, referring to Model 2 of Figure 6.1, for example *recipient's profile* can be determined by considering his/her age, weight, medical condition before the transplant and etc. Finally, the model that discloses the relationship between the 3 decision variables with the organ transplantation performance variable (i.e. transplant success) is developed

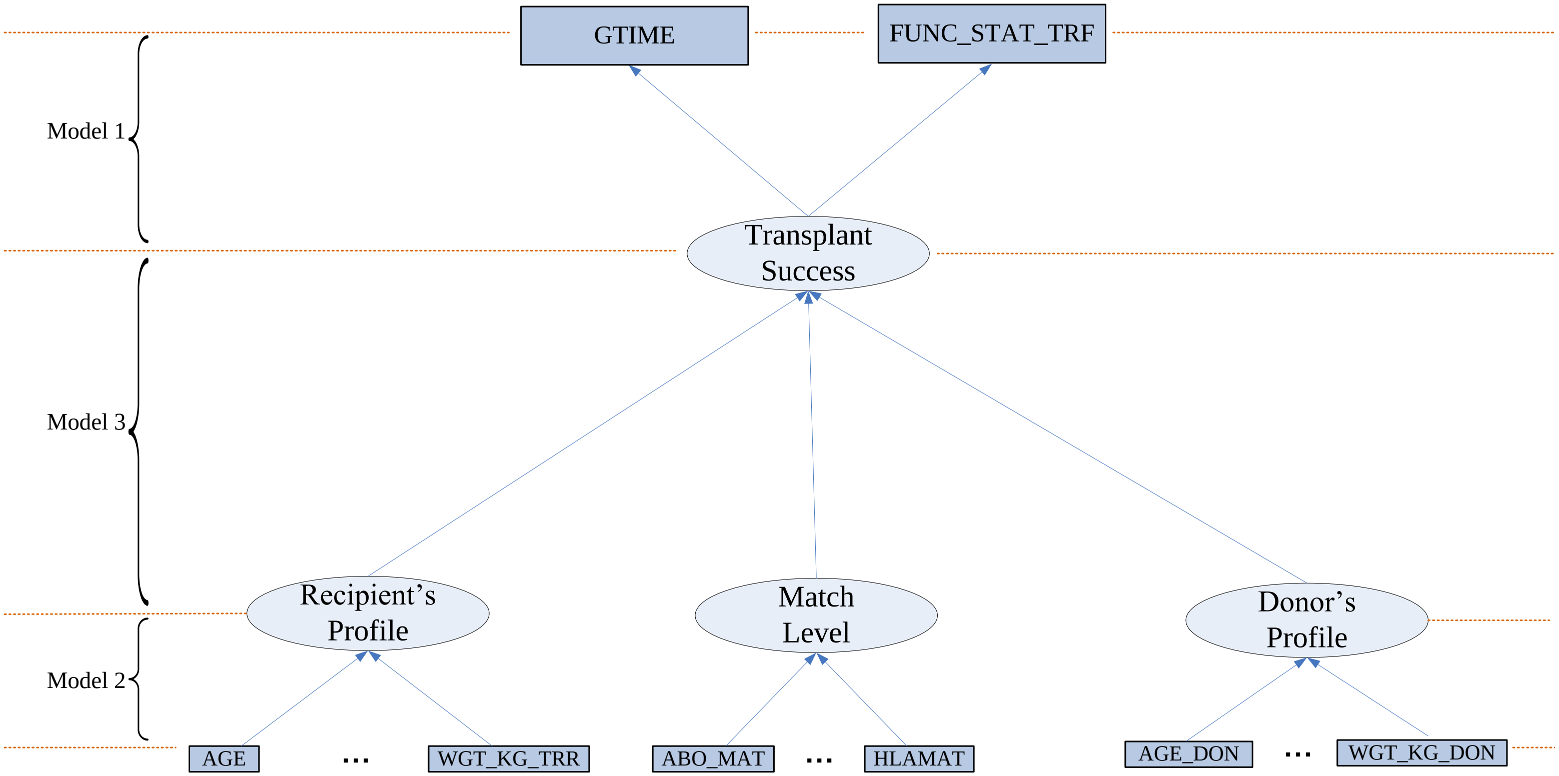


Figure 6.1 Causal relationships diagram for the proposed modeling

using the decision tree predictive approach, which is shown as Model 3 in Figure 6.1. The details regarding these relations through Model 1 and Model 2 are presented in Section 6.3.1. Using Model 2 results as inputs and Model 1 result as output, Model 3 would construct a decision tree prediction model, which is explained in Section 6.3.2. Based on the three models shown in Figure 6.1, we propose a 5-step approach which is depicted in Figure 6.2 to achieve interpretability and predictability simultaneously.

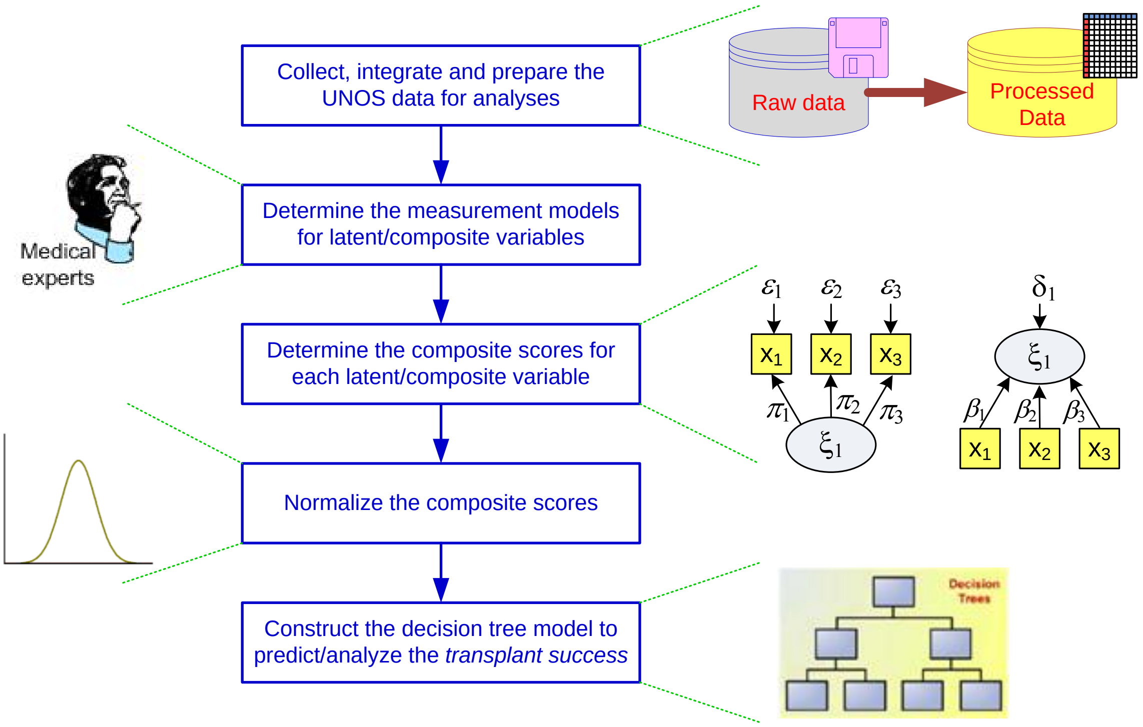

Collect, integrate and prepare the UNOS data for analyses
Raw data
Processed Data
Medical experts
Determine the measurement models for latent/composite variables
Determine the composite scores for each latent/composite variable
ε_1 ε_2 ε_3
x_1 x_2 x_3
π_1 π_2 π_3
ξ_1
δ_1
ξ_1
β_1 β_2 β_3
x_1 x_2 x_3
Normalize the composite scores
Construct the decision tree model to predict/analyze the *transplant success*
Decision Trees

Figure 6.2 Flowchart of the proposed method

The first step in the methodology is to prepare the dataset to be used in further modeling. The second step is to create the measurement model explaining the cause-and-effect relation between the latent/composite variables and their corresponding indicators as shown in Figure 6.1 (Model 1 and Model 2). In the second step, medical experts' opinion is also consulted. Then the composite scores for each latent/composite variable can be calculated through the measurement models as a third step. These scores are then normalized to an interval of [0-1] as the fourth step. The fifth step is to implement decision tree construction by using the composite scores of the latent/composite variables as the predictors/inputs and the performance variable, namely transplant success, as output. These steps are presented in Section 6.3 in detail.

Regarding the dataset, we used the same dataset in Chapter V, namely the UNOS thoracic regular and follow-up datasets merged into one file. To be able to claim that a transplant has been conducted successfully, namely a satificing match has been performed, not only the length of survival after transplant but also how well the recipient feels after the transplant should be considered. This is referred to as functional status (i.e. ability to work and ability to perform activities of daily living) or as quality of life [11]. Hence, in addition to GTIME we incorporated FUNC_STAT_TRF variable (functional status at last follow-up, which is an ordinal variable) as a causal indicator of the transplant success. The causal indicators of the other latent/composite variables and their definitions in UNOS dataset are tabulated in Table 6.1. This dataset had excessive number of missing values which render some of the records and variables seemingly unusable. *Case-wise deletion method* excludes all records (cases) that have missing data in at least one of the selected variables [12].

Table 6.1 Explanation of indicators and their corresponding composite variables

Indicators	Explanation	Variable Type	Composite Variable
AGE	Recipient's age (years)	Continuous	Recipient's Profile
DAYSWAITCHORN	Active days on waiting list	Continuous	
FUNC_STAT_TRR	Recipient functional status @ transplant	Ordinal	
HGT_CM_TRR	Recipient height @ transplant	Continuous	
MED_COND_TRR	Recipient medical condition pre-transplant @ transplant	Ordinal	
STERNOTOMY_TRR	Events occurring between listing and transplant: sternotomy	Ordinal	
WGT_KG_TRR	Recipient weight (kg) @ transplant	Continuous	
ABO_MAT	Donor-recipient ABO match level	Ordinal	Match Level
AMAT	A locus match level	Ordinal	
BMAT	B locus match level	Ordinal	
DRMAT	DR locus match level	Ordinal	
EINT	Ethnicity interaction between donor and recipient	Binary	
GINT	Gender interaction between donor and recipient	Binary	
HLAMAT	HLA match level	Ordinal	
AGE_DON	Donor age (years)	Continuous	Donor's Profile
HGT_CM_DON	Donor height (cm)	Continuous	
HIST_ALCOHOL_OLD_DON	Deceased donor-history of alcohol dependency	Binary	
HIST_CANCER_DON	Deceased donor-history of cancer	Binary	
HIST_CIG_DON	Deceased donor-history of cigarettes in past	Binary	
HIST_COCAINE_DON	Deceased donor-history of cocaine use in past	Binary	
HIST_DIABETES	Deceased donor-history of diabetes	Binary	
HIST_HYPERTENS_DON	Deceased donor-history of hypertension	Binary	
HIST_IV_DRUG_DON	Deceased donor-history of IV drug use in past	Binary	
HIST_MI	Deceased donor-history of previous Myocardial Infarction	Binary	
WGT_KG_DON	Donor weight (kg)	Continuous	

We applied this method considering the 27 indicators as our reference in hand, which ended up with 6512 records. This technique was implemented here mainly because a sample size of 6512 is satisfactory considering the fact that the PLS-based path analysis can be conducted with relatively small sample sizes [13]. As a general rule of thumb, Chin and Newsted [14] suggested using a minimum sample size of ten times the maximum number of paths aiming at any latent/composite variable in the PLS path model, which renders 6512 records far beyond this heuristic threshold value for our model.

6.3 Structural Equation Modeling-based Decision Tree Construction

The structural equation model can be described by two models: (1) a measurement (a.k.a outer) model explaining the relationship between the observed variables (already existing variables in the UNOS database for our case) and their corresponding composite/latent variables. (2) a structural (a.k.a inner) model explaining the relationship between some (or all) of the composite/latent variables (i.e. the decision variables to predict transplant performance such as recipient's profile) with other composite/latent variables (i.e. the transplant performance variable, transplant success). What follows next in Section 6.3.1 is a short description of these models with a partial least squares path modeling algorithm summarized from Tenenhaus *et al.* [15].

6.3.1 The Measurement (Outer) Model

A latent variable ξ is an unobservable variable (a.k.a construct, component or composite variable) which is indirectly described by a set of observable variables (x_h) (a.k.a. indicators). There are two ways of explaining the relationship between the latent/composite and observable variables: reflective and formative models.

1. Reflective modeling of transplant success (Model 1 in Figure 6.1)

In reflective model, the latent variable is assumed to underlie or cause its related causal indicators (observable variables). Each is attributed to its latent variable by a simple linear regression as in Eq. (6.1).

$$x_h = \pi_{h0} + \pi_h \xi + \varepsilon_h \tag{6.1}$$

where x_h is the causal indicator, π_{h0} is the constant intercept, π_h is the item loading, ξ is the latent variable, and ε_h is the measurement error/residual. The index h refers to the h^{th} causal indicator which would be related to its latent variable. Here ξ has a mean of m and a standard deviation of one. It is interpreted as each causal indicator x_h reflects its latent variable ξ. Eq. (6.1) is solved based on the main assumption that the residual ε_h has a zero mean and is uncorrelated with the latent variable ξ. This is called the predictor specification condition and shown by Eq. (6.2).

$$E(x_h|\xi) = \pi_{h0} + \pi_h \xi \tag{6.2}$$

In this study, the latent variable *transplant success* affects its causal indicators, namely GTIME and FUNC_STAT_TRF. Considering Figure 6.1, these reflective models can be formed and predicted as shown by Eqs. (6.3) and (6.4).

$$GTIME = \pi_{10} + \pi_1 * TransplantSuccess + \varepsilon_1 \tag{6.3}$$

$$FUNC\_STAT\_TRF = \pi_{20} + \pi_2 * TransplantSuccess + \varepsilon_2 \tag{6.4}$$

2. Formative modeling for the predictors of the transplant success (Model 2 in Figure 6.1)

In this model, it is assumed that the composite variable is formed or caused by its causal indicators. The composite variable is a linear function of its causal indicators plus a residual term as shown in Eq. (6.5).

$$\mathrm{\xi_k} = \sum_h \mathrm{\beta_{kh} x_{kh}} + \delta_k \tag{6.5}$$

where β_{kh} is the regression weight and δ_k is the residual error. The subscripts *kh* refer to the k^{th} composite variable with its h^{th} causal indicator in sequence. Note that for formative models 'composite variable' is the preferred generic term instead of 'latent variable'.

Eq. (6.5) is solved under the assumption that the residual vector δ_k has a zero mean and is uncorrelated with the indicators x_h. This assumption is called the predictor specification condition and hypothesized by Eq. (6.6).

$$E\left(\xi \middle| x_1, \dots, x_{p_j}\right) = \sum_h \beta_h x_h \tag{6.6}$$

Referring to Figure 6.1 and Table 6.1, the five formative models of this study corresponding to the composite variables *recipient's profile*, *match level,* and *donor's profile* can be constructed as in Eqs.(6.7)-(6.9), respectively.

$$Recipient's\ Profile = \beta_{11} * AGE + \cdots + \beta_{17} * ETHCAT + \delta_1 \tag{6.7}$$

$$Match\ Level = \beta_{21} * AMAT + \cdots + \beta_{27} * BMAT + \delta_2 \tag{6.8}$$

$$Donor's\ Profile = \beta_{31} * AGE\_DON + \cdots + \beta_{3(11)} * HIST\_MI + \delta_3 \tag{6.9}$$

Since the indicators were on different measurement scales, we implemented a scale transformation as in Eq. (6.10) so as to have an interpretable reference scale to compare the individual scores to each other.

$$x_{h(new)} = \frac{x_{h(old)} - x_{\min(old)}}{x_{\max(old)} - x_{\min(old)}} \left(x_{\max(new)} - x_{\min(new)}\right) + x_{\min(new)} \tag{6.10}$$

The next step is to estimate the *standardized* composite variables y_k (given by $y_k = \xi_k - m_k$). The unstandardized composite variable ξ_k and its mean m_k are estimated by Eqs. (6.11) and (6.12), respectively.

$$\widehat{\xi}_\mathrm{k} = \sum \tilde{\beta}_{kh} x_{kh} \tag{6.11}$$

$$\widehat{m}_k = \sum \tilde{\beta}_{kh} \bar{x}_{kh} \tag{6.12}$$

where $\tilde{\beta}_{kh}$ refers to the estimated regression weight between the k^{th} composite variable with its h^{th} causal indicator and $\bar{x}_{kh}$ is the mean of the h^{th} causal indicator that loads onto the k^{th} composite variable. By using Eqs. (6.11) and (6.12), the *standardized* composite variables y_k can then be estimated as combinations of their causal indicators as shown in Eq. (6.13).

$$y_k = \sum \tilde{\beta}_{kh}(x_{kh} - \bar{x}_{kh}) \tag{6.13}$$

6.3.2 The Structural (Inner) Model (Model 3 in Figure 6.1)

In conventional structural equation modeling, the structural model is composed of linear equations relating the latent variables with other latent variables. This is formulated as in Eq. (6.14).

$$\xi_j = \beta_{j0} + \sum_i \beta_{ji}\xi_i + v_j \tag{6.14}$$

where ξ_j is the latent/composite variable that has a path from another latent/composite variable, i.e. ξ_i. β_{j0} is the constant intercept, β_{ji} is the path coefficient from ξ_i to ξ_j, and v_j is the residual error. Although this modeling approach is very powerful in terms of causality explanation, it relies on a strong assumption that the relationships among the latent/composite variables are linear. Therefore, for the structural model part we propose to employ decision trees which are effective nonlinear data mining techniques. The composite scores of the latent/composite variables can be calculated by the reflective and formative modeling as explained in Section 6.3.1. These normalized composite scores are then used to construct the decision tree to predict the transplant success. This proposed structural model with decision tree-based construction should hypothetically be more effective than a linear regression-based structural model since it is capable of revealing the nonlinear relationships.

Decision trees recursively split the data in branches according to a preset criterion (e.g. information gain) to maximize the prediction accuracy resulting in a tree-like structure [16]. To achieve this, they use mathematical algorithms (such as information gain, Gini index, and Chi-squared test) to identify a pair of variable and its threshold that splits the input observation into two or more subgroups. This step is repeated at each leaf node until the complete tree is constructed. The objective of the splitting algorithm is to find a variable-threshold pair that maximizes the homogeneity (order) of the resulting two or more subgroups of samples. Popular decision tree algorithms include Quinlan's ID3, C4.5, C5, M5 [16]-[18], Breiman *et al.*'s CART [19], and CHAID introduced by Kass [20]. Compared with other machine learning methods, decision trees have the advantage that they are explicit models (as opposed to black box models) and hence can easily be interpreted and summarized as rules. This advantage makes decision trees widely used in medicine [21]. If the dependent (output) variable is categorical or ordinal, the decision tree is specifically called classification tree; if the dependent variable is continuous (as in our case) the resulting decision tree is called regression tree. Regression trees are one of a group of relatively flexible and computer-intensive statistical techniques [22]. These methods use repeated re-sampling of the data to develop empirical sampling distributions of the relevant statistics in place of the more restrictive distributional assumptions in classical statistical methods. Popular regression trees are CART, CHAID, and M5 all of which can be used as classification and regression trees. Based on the favorable prediction results we have obtained from preliminary runs in our case study, we chose to use CART algorithm as the regression tree method.

6.4 Universal Structure Modeling: Bayesian neural networks-based PLS path modeling

As a benchmark to our methodology, we compare and contrast our case study results with the universal structure modeling (USM) which was developed by Buckler and Hennig-Thurau [23]. The reason that USM was chosen in this study as a benchmark is that it does capture the nonlinearity perfectly and hence achieves high prediction accuracy, yet it lacks interpretability since it uses a black-box model, namely neural networks. Additionally, it requires high computational time to reveal potential nonlinear and latent variable interaction effects on each other through the bootstrapping method. What follows next is a short description of USM. Similar to our approach, the USM also limits the nonlinear relations *only* to the structural model and it assumes that the measurement model part is linear. In other words, measurement model portion of USM is the same as described in Section 6.3.1. As for the structural model, USM substitutes the linear least squares regression with Bayesian neural networks. This enables the model to discover unproposed structural paths, nonlinearity, and interaction effects. The estimator $\hat{\xi}$ of the latent variable ξ is defined as the output of multilayer perceptron (MLP) architecture and shown as in Eq. (6.15).

$$\widehat{\xi^j} = f_{Act2}\left(\sum_{h=1}^{H} w_h . f_{Act1}\left(\sum_{i=1}^{I} w_{ih} . S_i^j . \xi^i + b_{1h}\right) + b_2\right) \tag{6.15}$$

where f_{Act1} is the activation function of the hidden neural units and f_{Act2} is the output neural unit. H is the number of hidden neural units, I is the number of latent input variables ξ, *w*'s are the weights and *b*'s are the biases for the neural network. S_i^j is the apriori likelihood that a variable *i* influences another variable *j*. To prevent the overfitting in the neural network model, USM minimizes the error function *E* for each latent variable *i* of the structural model. *E* refers to the overall error of the respective variable's neural network and shown as in Eq. (6.16).

$$E_i = \beta . \sum_{n=1}^{N} (\widehat{\xi_{t-1,n}^{i}} - \widehat{\xi_{t,n}^{i}})^2 + \sum_{h=1}^{H} \alpha_{t,h} . \sum_{p=1}^{P} w_{ph}^2 \qquad (6.16)$$

where *n* refers to the individual cases, *N* is the total number of cases, and *p* is the index for the weights, *w*. On the other hand, ξ_t^i is the conditional estimate of the latent variable *i* in the current estimation step, *t*, calculated from the structural model by the Bayesian neural network, and $\xi_{t-1,n}^i$ is the estimate of the previous iteration for the same latent variable. If the case is the first step for this estimation, $\xi_{t-1,n}^i$ would then refer to the initial composite score received from the measurement model. The hyperparameters α and β prevent overfitting of the neural network model. They are updated in every iteration of the learning process and are given by Eqs. (6.17) and (6.18).

$$\alpha_h = \frac{\gamma}{2\sum_{n=1}^{N} w_{nh}^2} \qquad (6.17)$$

$$\beta = \frac{N - \gamma}{2\sum_{n=1}^{N}(\xi_n^{i} - \widehat{\xi_n^{i}})^2} \tag{6.18}$$

where N is the total number of records and $\gamma = \sum_{p=1}^{P} \frac{\lambda_p}{\lambda_p + \alpha_{LI-1}}$. λ_p are the eigenvalues of the Hessian matrix of the error function in Eq. (6.16) and α_{LI-1} is the hyperparameter α from the previous learning iteration.

6.5 Case Study and Discussion

To confirm the measurement model and determine the composite scores of the latent and composite variables, PLS was preferred in this study because it does not place much importance on the sample size and data distribution assumptions [24]. Additionally, it can handle the formative measurement models which are required in our approach. The results for the reflective part of the model, namely the part of the model pertaining to *transplant success* (Model 1 in Figure 6.1) are presented in Table 6.2.

Table 6.2 Internal consistency, reliability, and convergent validity measures

Latent Variable	CR	Cronbach's Alpha	AVE	Items and their loadings
Transplant Success	0.736	0.728	0.699	GTIME → 0.841
				FUNC_STAT_TRF → 0.994

Composite reliability (CR) is a criterion of scale reliability. It can assess the internal consistency of the item and is given by Eq. (6.19) following the same parameters from Eq. (6.19) [25].

$$CR_{\xi h} = \frac{(\sum \pi_h)^2}{(\sum \pi_h)^2 + \sum \xi_h} \tag{6.19}$$

On the other hand, Cronbach's alpha measures the extent to which the observable variables can explain their corresponding latent variable and is also supportive reliability measurement criterion [26]. For this latent variable (*transplant success*) CR measure was found to be 0.736 and Cronbach's alpha was 0.728, both of which pass the widely accepted threshold value of 0.7 [27]. These two measures ensure that this latent variable is internally consistent i.e. reliable and stable. On the other hand, to check the convergent validity of the latent variables the average variance extracted (AVE) should be calculated as in Eq. (6.20).

$$AVE = \frac{\sum {\pi_h}^2}{\sum {\pi_h}^2 + \sum 1 - {\pi_h}^2} \tag{6.20}$$

AVE should exceed the 0.5 threshold value as a rule-of-thumb, which was 0.699 in our results for this latent variable. Also, all item loadings should be at least 0.70 and were observed as 0.841 for GTIME and 0.994 for FUNC_STAT_TRF in our analysis.

The measurement models pertaining to the 3 composite variables, i.e. recipient's profile, donor's profile, and match level, (Model 2 in Figure 6.1) are constructed by formative models because all item measures are independent of one another and are viewed as items that constitute their corresponding composite variables. In formative model cases, abovementioned internal consistency, reliability, and convergent validity criteria (i.e. Cronbach's alpha, CR, and AVE) are not deemed appropriate [24], [28]. In assessing formative models, Petter *et al.* [29] place great importance on prior data collection phase and rather propose to assess *content validity* essentially by "evaluating if the set of indicators under-specify the domain of the construct based on explicated facets in the *theory base*". For formative models, PLS weights represent a comparable effect of indicators on composite variables [31]. *Construct validity* can be assessed by eliminating the non-significant items in expense of losing the content validity to some extent or alternatively non-significant items can be kept to preserve the content validity [29]. Formative model results of our model are presented in Table 6.3 in detail.

Note that considering the t-statistics in Table 3, all of the indicators were found to be significant at the 0.05 significance level, and therefore kept in the model. Negative PLS weights indicate the fact that the individual variable affects in a negative direction in its corresponding composite variable. In other words, for example the total days a patient waited for a transplant on the waiting list (DAYSWAITCHORN) has a negative impact on the recipient's profile quantified by the PLS weight of -0.527, which in turn negatively affects his/her strength to undergo a successful transplant. All the negative and positive signs of the other indicators are to be interpreted in the same fashion.

Table 6.3 Formative model results for composite variables

Indicators	Composite Variables	PLS Weights	t-statistic
AGE	Recipient's Profile	-0.301	6.529
DAYSWAITCHORN		-0.527	2.396
FUNC_STAT_TRR		0.113	8.739
HGT_CM_TRR		-0.182	4.269
MED_COND_TRR		0.624	6.954
STERNOTOMY_TRR		0.018	5.343
WGT_KG_TRR		-0.475	7.885
ABO_MAT	Match Level	0.413	6.957
AMAT		0.082	5.691
BMAT		0.276	3.098
DRMAT		0.037	7.738
EINT		0.593	2.131
GINT		0.727	4.799
HLAMAT		0.073	5.325
AGE_DON	Donor's Profile	-0.929	7.694
HGT_CM_DON		-0.228	6.746
HIST_ALCOHOL_OLD_DON		-0.032	7.348
HIST_CANCER_DON		-0.015	9.718
HIST_CIG_DON		-0.106	8.047
HIST_COCAINE_DON		-0.034	2.106
HIST_DIABETES_DON		-0.029	5.541
HIST_HYPERTENS_DON		-0.073	3.698
HIST_IV_DRUG_DON		-0.049	7.379
HIST_MI_DON		-0.051	8.432
WGT_KG_DON		-0.088	6.454

As for the structural portion of our model, we used normalized composite scores of the latent/composite variables received from PLS path model as inputs to the regression tree models which were implemented with CHAID, CART, and M5. Based on the favorable results provided by CART we present the results of PLS-based CART model as in Table 6.4.

The results in Table 6.4 are based on the testing dataset. In this study, to estimate the performance of the prediction models a 10-fold cross-validation approach was used and hence the results presented in Table 6.4 are the 10-fold cross-validated results for each model.

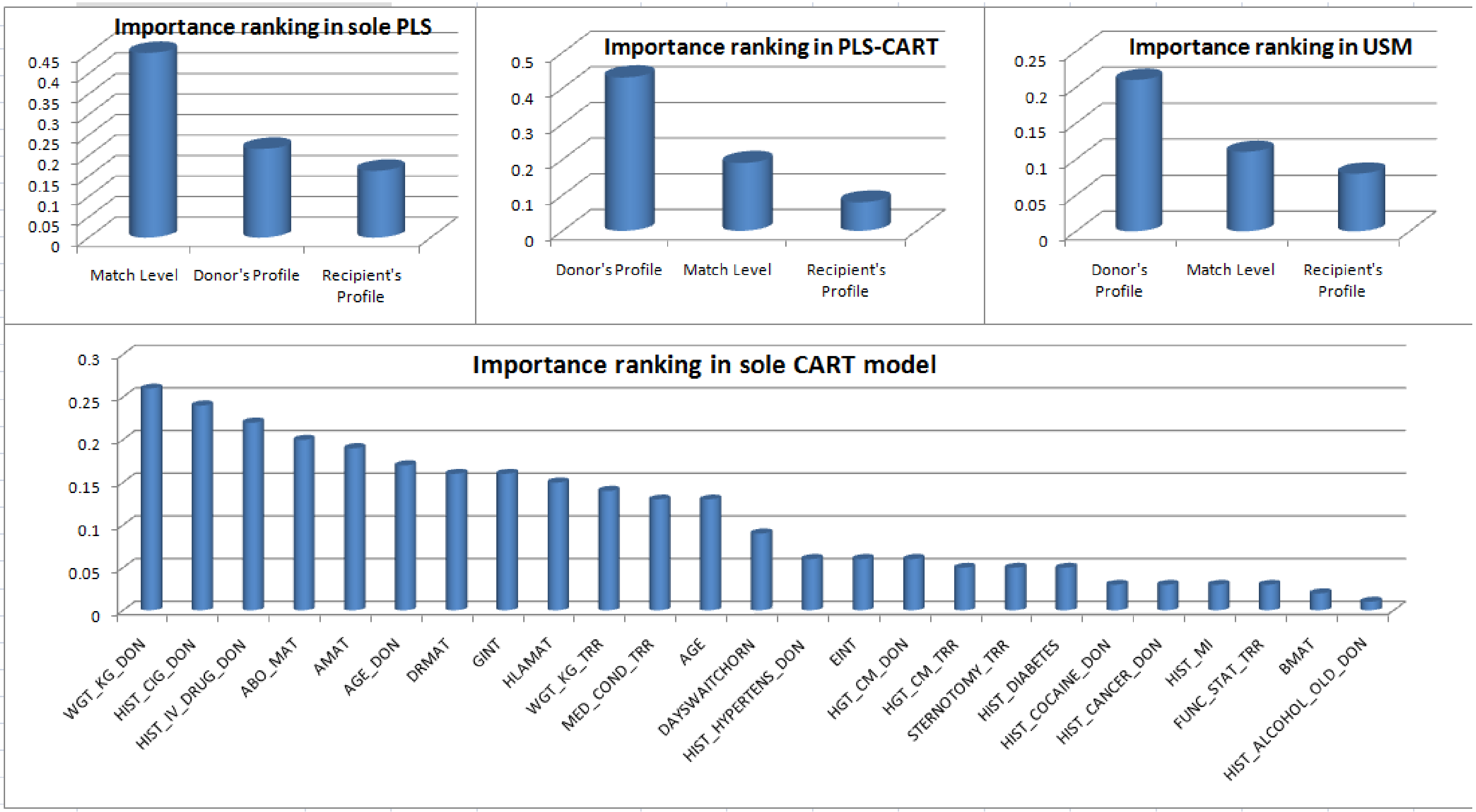
In a 2 GHz Intel Core 2 Duo® PC, USM model required 28 hours to complete 50-sample bootstrapping whereas the analysis using our proposed structural equation modeling-based CART model was completed within a few minutes (~3-4 min).

Table 6.4 Comparison of R^2 values from each model

Prediction Models Performance Measure	Sole PLS	Sole CART	PLS-based CART	USM
R^2 value	0.34	0.56	0.68	0.73

In this study, variable importance measures were also investigated to judge the relative importance of each composite variable. Variable importance ranking in decision trees uses surrogate splitting to produce a scale (a relative importance measure) for each predictor variable included in the analysis. The computational details regarding these measures can be found in Breiman *et al.* [23]. 10-fold cross-validated variable importance ranking for sole PLS, sole CART, and PLS-based CART model results are illustrated in Figure 6.3. Regarding the sole PLS variable ranking, path coefficients are reported all of which were found to be significant at 0.05 level. Note that in Figure 6.3, our proposed structural equation model-based decision tree model and the universal structure modeling ranked the variables exactly in the same order. This consistency between the two models could be attributed to the fact that both models are capable of discovering nonlinear relationships among the predictors, which was not possible to capture with sole partial least squares-based path modeling. These two models agree that in predicting the *transplant success* the ascending rank order of composite variables is as follows: *donor's profile, match level,* and *recipient's profile*. Based on this consistency in Figure 6.3 and high prediction accuracy provided by the two models as in Table 6.4, we can conclude that the most important predictors of *transplant success* would be ranked as such.

Figure 6.3 Variable importance ranking by different models



Likewise, when all predictor variables are tapped into the CART model, nine topmost important variables (WGT_KG_DON, HIST_CIG_DON, HIST_IV_DRUG_DON, ABO_MAT, AMAT, AGE_DON, DRMAT, GINT, HLAMAT) also belong to the top-ranked two composite variables, namely donor's profile and match level with a 10-fold cross-validated variable importance ranking approach. In the USM model, nonlinear relations were sought, and at 0.05 significance level transplant success was revealed to have a significant nonlinear relationship with the donor's profile.

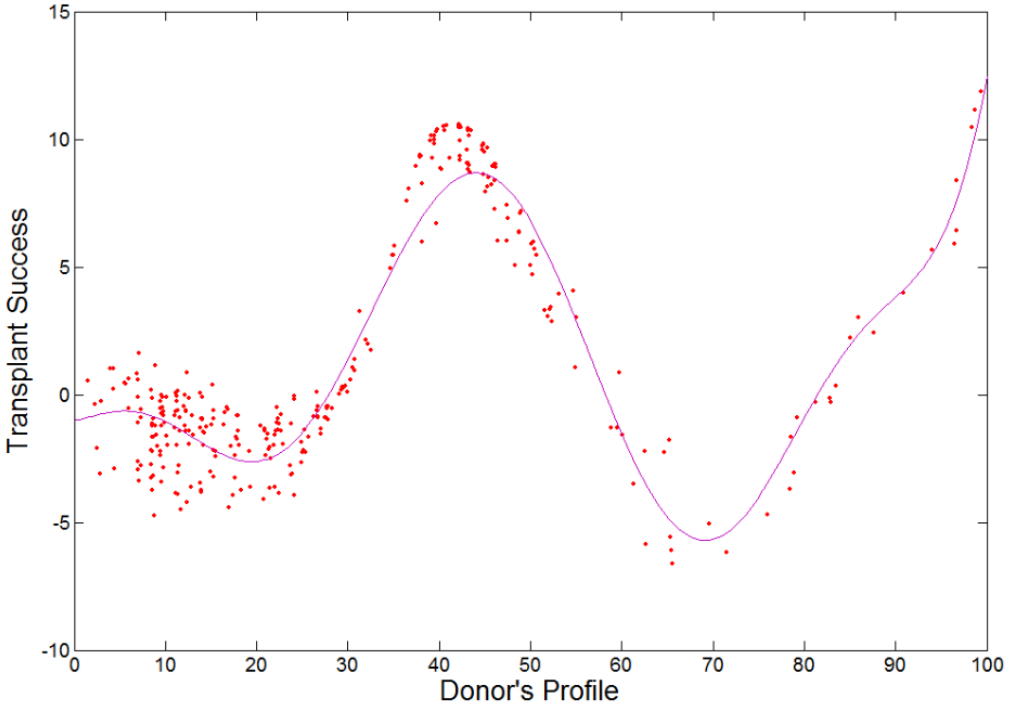


Figure 6.4 Nonlinearity revealed by the USM model

In Figure 6.4 the causing composite variable, donor's profile, is represented against the affected latent variable, transplant success. The line getting through the observed cases is the additive function explaining the nonlinear cause of donor's profile on transplant success. Note that the 0-100 bandwidth of x-axis is the scale of the normalized latent variable of donor's profile. The y-axis represents the variation in transplant success caused by the causing variable, i.e. donor's profile. High nonlinearity observed here explains why sole PLS model could not reveal the high impact of the composite variable donor's profile while ranking the composite variables in terms of their importance.

Interaction effect (IE) of two independent latent/composite variables (ξ^j and ξ^k) on ξ^i (shortly IE^i_{jk}) is expressed as the portion of variable ξ^i's explained variance that can be attributed to the interaction between ξ^j and ξ^k and is given by Eq.(6.21) [23].

$$IE^i_{jk} = \frac{\sum_{n=1}^{N} \left| \frac{\widehat{z^i_{jk}} - \widehat{a_j} - \widehat{a_k}}{\hat{\xi} - \bar{\xi}} \right|}{N} \tag{6.21}$$

where $\hat{a}$ is the additive score of a polynomial regression of ξ on a and $\hat{z}$ is the outcome of a universal regression with the two latent variables j and k as regressors on z^i_{jk}. Here a and z can be given by Eqs.(6.22) and (6.23), respectively.

$$a^i_j = f^i(\xi^1, \dots, \xi^j, \dots, \xi^n) - f^i(\xi^1, \dots, \overline{\xi^j}, \dots, \xi^l) \tag{6.22}$$

where a_j^i is the change in ξ^i caused by the additive effect of ξ^j, f is the neural network function, and ξ^1 and ξ^n are the latent variables. By setting the value of ξ^j to its mean value ($\overline{\xi^j}$), the change in ξ^i which is provided by ξ^j can be captured.

Similarly, Eq. (6.25) represents the change in ξ^icaused by the interactive effect of ξ^j and ξ^k.

$$z_{jk}^i = f^i(\xi^1, \dots, \xi^j, \dots, \xi^k, \dots, \xi^l) - f^i\left(\xi^1, \dots, \overline{\xi^j}, \dots, \overline{\xi^k}, \dots, \xi^l\right) \qquad (6.23)$$

In our analysis, only one such an interaction effect on transplant success was observed at 0.05 significance, which was caused by the interaction of recipient's profile and match level as shown in Figure 6.5. The IE value of this effect through Eq. (6.21) was 0.82. This is translated into that 82 % of the explained variance of the latent variable transplant success has been explained by the interaction effect caused by recipient's profile and match level. In other words, referring to Table 6.4, 82 % of the explained variance by USM with 73 % can be attributed to the interaction effect and the rest is explained by *individual* effects of all composite variables, i.e. recipient's profile, donor's profile, and match level.

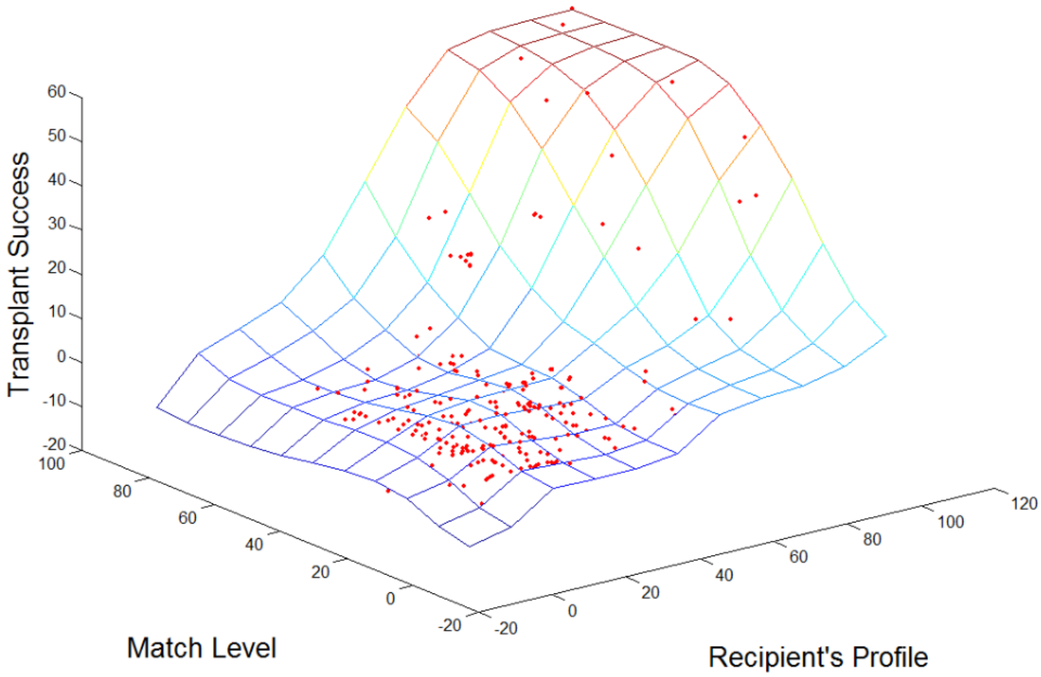


Figure 6.5 Interaction effect of treatment and match level on transplant success

The final structural equation modeling-based CART model is also pictorially presented in Figure 6.6. One of the most straightforward sample rules extracted by this final model is as follows: if the donor’s profile score is higher than 0.766 and the match level score is higher than 0.841, then transplant success would be 94.6 %. The rest of the rules can also be visualized in Figure 6.6.

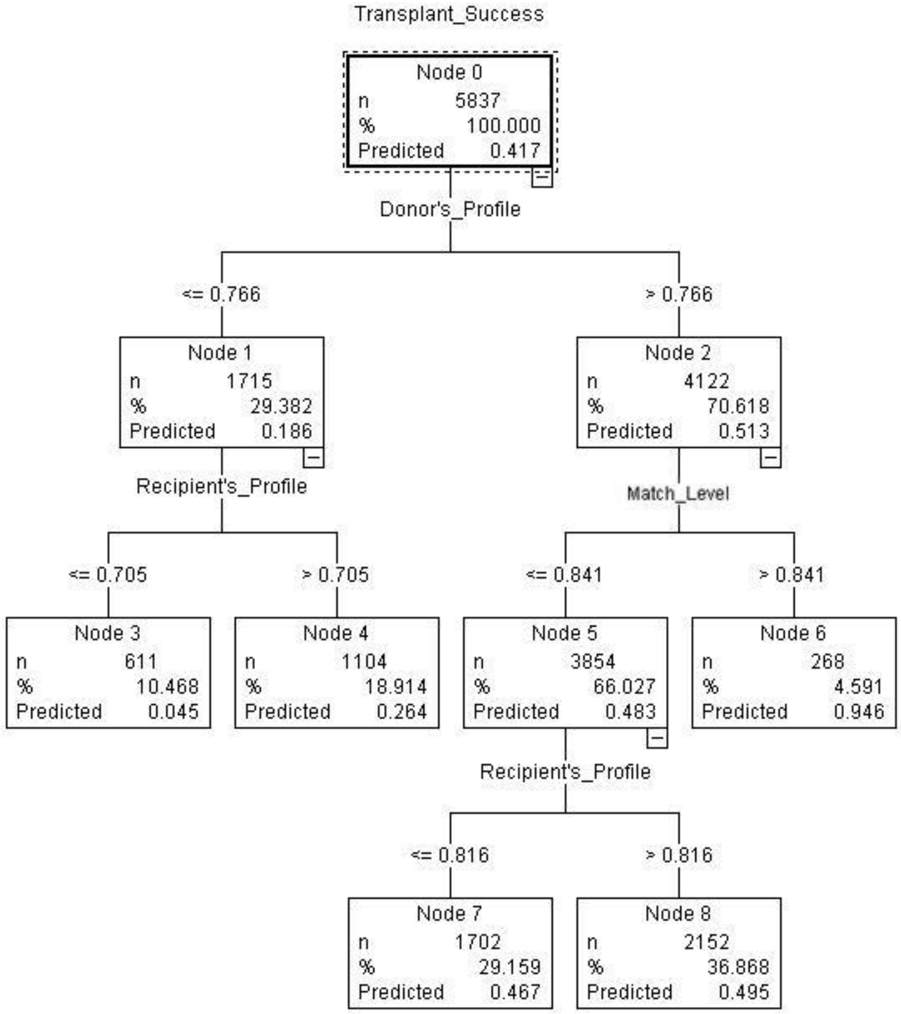


Figure 6.6 The final PLS-based CART model

6.6 Conclusions

Medical experts are trained to reason “medically” whereas data miners place more importance on model’s performance, e.g. prediction accuracy. Since research designs vary in both areas, such differences grow even more later on [8]. In addition to this conflict of interests, some machine learning methods (e.g. neural networks) are powerful in terms of predictive ability, yet they are black boxes. Namely, they give no (or very limited) explanation of the “reasoning” used behind the scene to achieve high predictive accuracy. Therefore, their acceptance by the medical experts is limited [8]. Our proposed method balances this trade-off and overcomes aforementioned issues that have been faced in collaboration between medical experts and data miners. Our integrated method with structural equation modeling and decision trees is proven to be fairly capable in terms of predictive accuracy with an R^2 value of 0.68 as well as interpretability with a much lower computational time requirement compared to Bayesian neural networks-based USM technique. Proposed method not only covers nonlinear relations among various variables but also brings more explanation into the scene to make the lung transplant procedure more understandable and transparent in terms of variables used for modeling and prediction. It provides concise rules which can be visualized in the final decision tree.

A main future research stream of this study might be to create a decision support system equipped with a user-friendly frontend and a near-transparent backend application which would help medical professionals to deal with voluminous data more effectively and efficiently (e.g. providing reliable results in a very short time period). Having entered hundreds of predictive variables into the system, a medical professional can then visualize the summarized information through our proposed method and make decisions for the upcoming transplants. In other words, having a potential donor organ on hand a medical expert could make a rapid decision as to which potential recipient patient to allocate the donor organ. As explained in this study, this could be achieved by utilizing the most critical variables which are related to recipient's and donor's profiles and their match level information as opposed to using a couple of hundreds of variables.

REFERENCES

[1] S. Daar, D.R. Salomon, R.M. Ferguson, J.H. Helderman, P. Macchiarini, New directions for organ transplantation Nature 392 (1998) 11-12.

[2] J.M.A. Smith, J. Vanhaecke, A. Haverich, E. De Veries, L. Roels, G. Persijn, G. Laufer, Waiting for a Thoracic Transplant in Eurotransplant, Transplant International 19 (2006).

[3] R.J. Ruth, L. Wysezewianski, G. Herline, Kidney Transplantation: A Simulation Model for Examining Demand and Supply. Management Science 31 (1985) 515-526.

[4] D. Sheppard, D. McPhee, C. Darke, B. Shretha, R. Moore, A. Jurewitz, A. Gray, Predicting Cytomegalovirus Disease After Renal Transplantation: An Artificial Neural Network Approach, International Journal of Medical Informatics 54 (1999) 55-76.

[5] R.S. Lin, S.D. Horn, J.F. Hurdle, S. Goldfarb-Rumyantzev, Single and Multiple Time-Point Prediction Models in Kidney Transplant Outcomes, Journal of Biomedical Informatics 41 (2008) 944-952.

[6] R.N. Pierson, M.L. Barr, K.P. McCulluough, T. Egan, E. Garrity, M. Jessup, S. Murray, Thoracic Organ Transplantation, American Journal of Transplantation 4 (2004) 93-105.

[7] F.L. Grover, M.L. Barr, L.B. Edwards, F.J. Martinez, R.N. Pierson, B.R. Rosengard, S. Murray, Thoracic Transplantation, American Journal of Transplantation 3 (2003) 91-102.

[8] Schmitt M., Teodorescu H.N., Jain A., Jain S., Jain L.C., Computational Intelligence Processing in Medical Processing, Studies in Fuzziness and Soft Computing, Springer-Verlag, 2002.

[9] A. Oztekin, D. Delen, Z.J. Kong, Predicting the graft survival for heart–lung transplantation patients: An integrated data mining methodology, International Journal of Medical Informatics 78 (12) (2009).

[10] D. Delen, A. Oztekin, Z.J. Kong, A Machine Learning-based Approach to Prognostic Analysis of Thoracic Transplantations, Artificial Intelligence in Medicine 49 (1) (2009).

[11] A.E. Molhazn, H.C. Northcott, L. Hayduk, Quality of Life of Patients with End Stage Renal Disease: A Structural Equation Model, Quality of Life Research 5 (4) (1996).

[12] M. Biewen, Item nonresponse and inequality measurement: Evidence from the German earnings distribution, Allgemeines Statistisches Archiv 85 (2001).

[13] S.H. Hsu, W.H. Chen, M.J. Hsieh, Robustness testing of PLS, LISREL, EQS and ANN-based SEM for measuring customer satisfaction, Total Quality Management 17 (3) 2006.

[14] W.W. Chin, P.R. Newsted, Structural equation modeling analysis with small samples using partial least squares, in: Hoyle, R., Editor, Statistical Strategies for Small Sample Research, Sage Publications (Beverly Hills, CA, 1999).

[15] M. Tenenhaus, V.E. Vinzi, Y.M. Chatelin, C. Lauro, PLS path modeling, Computational Statistics and Data Analysis 48 (1) (2005) 159-206.

[16] J. Quinlan, Induction of decision trees, Machine Learning 1 (1986) 81–106.

[17] J. Quinlan, C4.5: programs for machine learning, Morgan Kaufmann, San Mateo, CA, 1993.

[18] J. R. Quinlan, Learning with continuous classes, In: *Proceedings of 5th Australian Joint Conference on Artificial Intelligence*, Adams & Sterling, eds., World Scientific, Singapore (1992), pp. 343–348.

[19] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, Classification and regression trees, Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA, 1984.

[20] G. V. Kass, An Exploratory Technique for Investigating Large Quantities of Categorical Data, Applied Statistics 29 (1980) 119-127.

[21] S. Dreiseitl, L. Ohno-Machado, Logistic Regression and Artificial Neural Network Classification models: A Methodology Review, Journal of Biomedical Informatics 35 (2002) 352-359.

[22] B. Efron, R. Tibshirani, Statistical Data Analysis in the Computer Age, Science 253 (1991) 390-395.

[23] F. Buckler, T. Hennig-Thurau, Identifying Hidden Structures in Marketing's Structural Models through Universal Structure Modeling: An Explorative Bayesian Neural Network Complement to LISREL and PLS, Marketing, Journal of Research and Management 4 (2008) 47-66.

[24] W.W. Chin, The partial least squares approach to structural equation modeling, in: G.A. Marcoulides (Ed.), Modern Methods for Business Research, Lawrence Erlbaum, Mahway, NJ, 1998, pp. 295–336.

[25] C. Fornell, D.F. Larcker, Evaluating structural equation models with unobservable variables and measurement error, Journal of Marketing Research 18 (1981) 39-50.

[26] E.J. Pedhazur, L.P. Schmelkin, Measurement, design and analysis: an integrated approach, Lawrence Erlbaum, Hillsdale, 1991.

[27] M.Y. Yi, F.D. Davis, Developing and validating an observational learning model of computer software training and skill acquisition, Information Systems Research 14 (2) (2003) 146–169.

[28] D. Gefen, D. Straub, and M. Boudreau, "Structural equation modeling and regression: Guidelines for research practice," Commun. AIS, vol. 7, pp. 1–78, 2000.

[29] S. Petter, D. Straub and A. Rai, Specifying formative constructs in information systems research, MIS Quarterly **31** (4) (2007), pp. 623–656.

[31] K. Bollen and R. Lennox, "Conventional wisdom on measurement: A structural equation perspective," Psych. Bull., vol. 110, pp. 305–314, 1991.

CHAPTER VII

SIMULATION MODELING TO VALIDATE THE MATCHING INDEX AND TO FINE-TUNE ITS WEIGHTS

To validate the matching index for organs through the composite score, a simulation model would be developed in this chapter. At the first step, the simulation model is to be validated against the current organ allocation scheme. In the next step, through the response surface methodology the weights of the matching index will be fine-tuning.

7.1 Background

Since the first successful lung transplantation in 1983, the lung allocation policy has gradually evolved from allocating the organs purely based on waiting list time to the current day lung allocation policy called Lung Allocation Score (LAS), which is an intricate process involving different kinds of people, resources and organizations [1]. There are numerous reasons for the current complex state of lung allocation policy, but the most noteworthy is due to the fact that the lung transplantation became an accepted treatment option for patients, resulting in rapid increase in patients registering their names for lung transplantation. However, the relative scarcity of suitable lung donors among the pool of conventional brain dead organ donors has resulted in increasing waiting time for the patients on the list [2].

Prior to 2005, organs were allocated to patients purely based on the amount of time that candidates had accumulated on the waiting list and ABO match (donor-recipient blood group match level). Offers were first made to candidates within the OPO (organ procurement organization) donor service area where the donor was located, and then within expanding 500-nautical-mile zones around the donor hospital. The Organ Procurement and Transplantation Network (OPTN) focused on the use of objective medical criteria and medical urgency. To achieve this, the effect of waiting time should be minimized and broader geographic sharing of donor organs should be encouraged. It was decided to implement Lung Allocation Score (LAS), which is a multivariate model designed to predict the risk of death during the following year on the waiting list and the likelihood of survival during the first year after the transplantation [1]-[4]. The primary objective of LAS is to decrease waiting list mortality, prioritize candidates based on medical urgency, and decrease the relevance of waiting list time on prioritization of donor lung [2]. This algorithm was focused on minimizing deaths on the waiting list and maximizing the benefit of transplant by incorporating post-transplant survival into the algorithm.

7.2 The Current Lung Allocation and Transplantation Process

This process has three categories which can be listed as 1) pre-transplant process, 2) LAS calculation, and 3) organ matching process.

7.2.1 Pre-transplantation Process

In the pre-transplantation process, the corresponding physician recommends lung transplantation to the patient based on the patient's medical condition. Once the patient is willing, s/he can approach a lung transplant center. Here the patient needs to complete a transplant work-up. During the transplant work-up the patient will participate in a series of medical tests and consult a transplant physician, social worker and financial coordinator. Based on the test results and the review from all the consulted people, the transplant center will register the patient as a candidate for lung transplantation. This same information will also be used to calculate the patients Lung Allocation Score. The entire patients registration process shall be carried out by the corresponding transplant center on UNet^{SM}, which is a web-based electronic utility used by UNOS (which is an OPTN contractor). Once the registration process is completed, the patients name is added to the waiting list [5].

7.2.2 Lung Allocation Score (LAS) Calculation Process

Before explaining LAS calculation, some other terminology (such as transplant benefit, urgency, and post-transplant survival) needs to be introduced.

1. Transplant Benefit

The concept of survival, with or without a transplant, constitutes the central theme in transplant benefit. To predict that benefit, the area under the predicted survival curve is used, which represents the total days of predicted survival within one year on the waiting list and one year following the transplant. In Figure 7.1, the shaded area under the waiting list curve is the measure of predicted number of days of survival without a transplant during an additional year on the list, which is a measure of urgency [6]. This is named as waiting list urgency measure (WL_i).

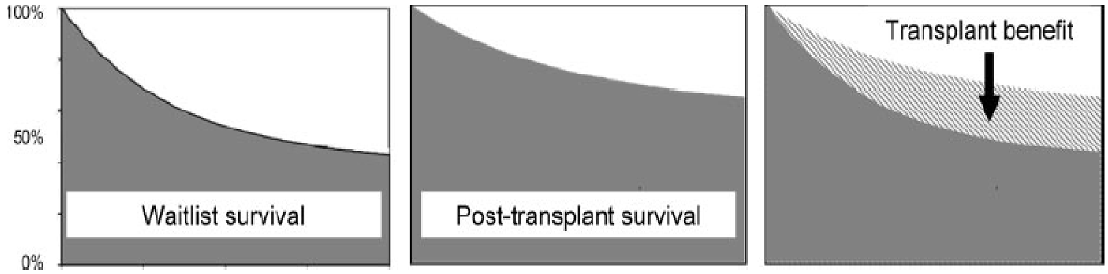


Figure 7.1 Transplant benefit calculation through survival curves [1]

On the other hand, the area under the post-transplant survival curve shows the number of days survived after the transplant, which is named as post-transplant survival measure and shown by PT_i. The difference between these measures is a measure of "transplant benefit" ($Benefit_i$). This is translated into the number of expected additional days of life over the next year if a particular patient receives a transplant, rather than remaining on waiting list. The calculation of $Benefit_i$ is summarized in Eq. (7.1) [7].

$$\text{Benefit}_i = \text{PT}_i - \text{WL}_i \tag{7.1}$$

2. Urgency and Post-transplant Survivability

The OPTN committee evaluated the options to select the relative importance of urgency and transplant benefit. Weighing in favor of benefit alone would offer organs to patients with a high chance of survival on the waiting list over the short term. In contrast, weighing in favor of urgency alone would allocate organs to patients with poor post-transplant outcomes over equivalently urgent patients who could have a better outcome. These two are represented in Figure 7.2 (a) and (b). After mathematical modeling, OPTN observed that a 45^o bar is the optimal approach to balance both measures, as shown in Figure 7.2 (c). Therefore, the raw allocation score was written as in Eq. (7.2) [6]-[7].

$$\text{Raw score}_i = \text{PT}_i – 2^* \text{WL}_i \tag{7.2}$$

where "2" in the equation refers to the 45^o angle in the graph of Figure 7.2 (c). Changing this angle to 60^o or 90^o caused more number of predicted deaths. The possible values of Raw score_i range between +365 and -730, which represent the two extremes of 100% survival post-transplant but dying today without a transplant to a 100% chance of living for one year on the waiting list but a 100% probability of dying before the first day just after the transplant. To eliminate the negative scores the raw score was decided to be normalized to a continuous scale of 0-100 as shown in Eq. (7.3) [6]-[7].

$$LAS = [100*(Raw\ Score+2*365)]/3*365 \quad (7.3)$$

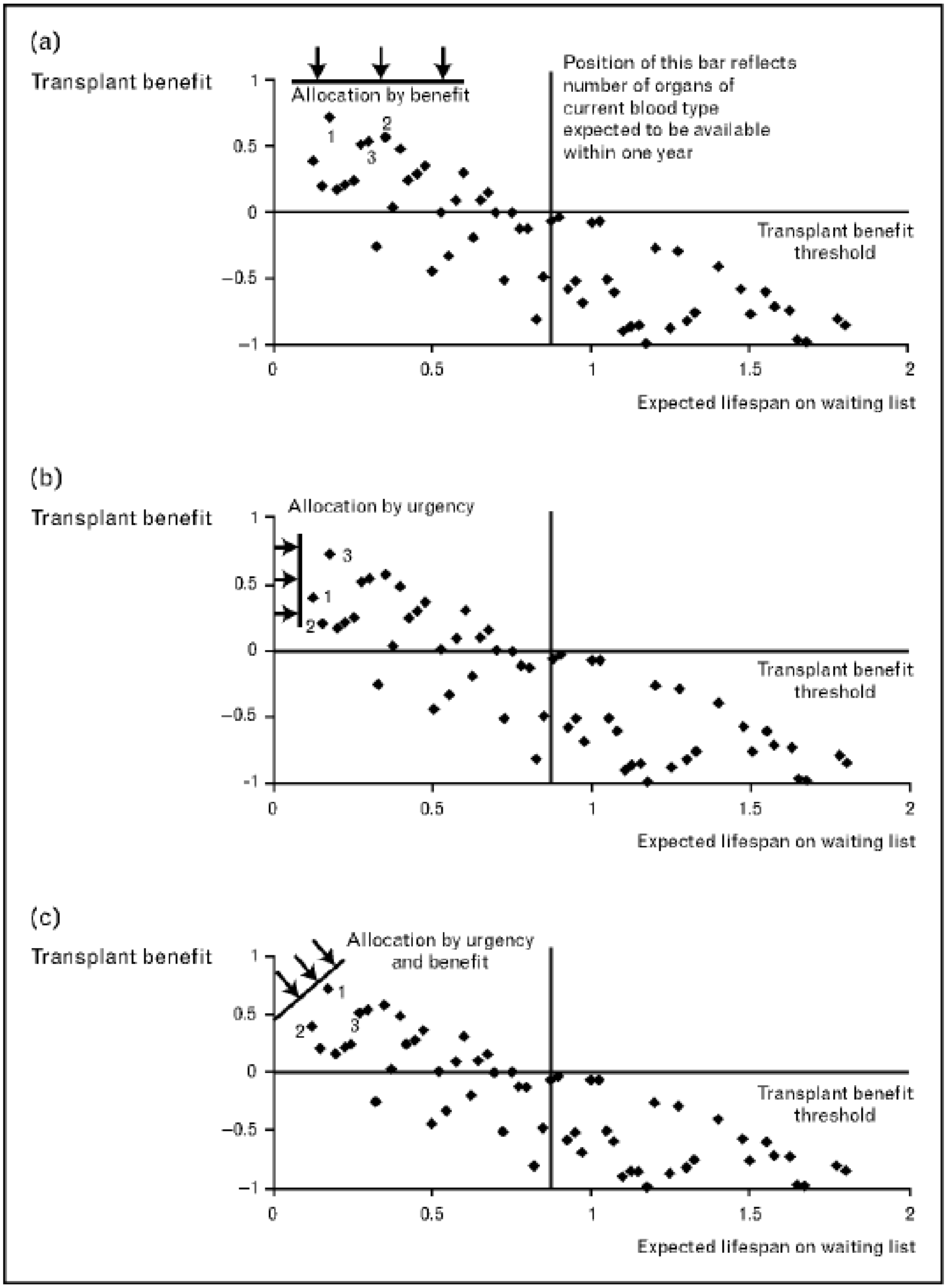


Figure 7.2 Urgency vs. post-transplant survival [7]

7.2.3 Lung Matching Process

The final and arguably the most important step in lung allocation process is the matching process. Whenever a donor lung becomes available, a match list is created to match the lungs with a suitable candidate based on the distance from the donor hospital to their transplant center, ABO type, and age group.

The donor location is one of the most important factors in the lung allocation process since the organs are perishable items and cannot survive more than a specific length of time. Once the lung becomes available, it is first offered locally to the candidates within the OPO's limits. If a suitable recipient cannot be identified, the zonal allocation process starts. The zones are delineated by concentric circles of 500 (Zone A), 1000 (Zone B), 1500 (Zone C) and 2500 (Zone D) nautical mile radii with the donor hospital at the center. Zone A will extend to all transplant centers which are within 500 miles from the donor hospital but not in the local area of the donor hospital. Zone B will extend to all transplant centers between 500 and 1000 miles. Similarly, Zone C and D will follow the same 500-mile radii increments. On the other hand, Zone E will extend to all transplant centers beyond 2500 miles. Figure 7.3 represents the geographic sequence of lung allocation process. Since there is considerable difference between pediatric and adult patients and their potential lung sizes, the matching process takes age groups of donor and recipients into account. The prioritization matrix is summarized in Table 7.1. The candidate with the highest LAS score in a particular age group will have priority over others [5]. If no appropriate recipient is found among local candidates in any of the three age groups, then the potential compatible recipients in Zone A will be offered the donor lungs. If still an appropriate recipient is not found in any of the three age groups, then the potential compatible recipients in Zone B will be offered the donor lungs. Similarly this process is repeated in successive zones until a suitable recipient is found [5], [8].

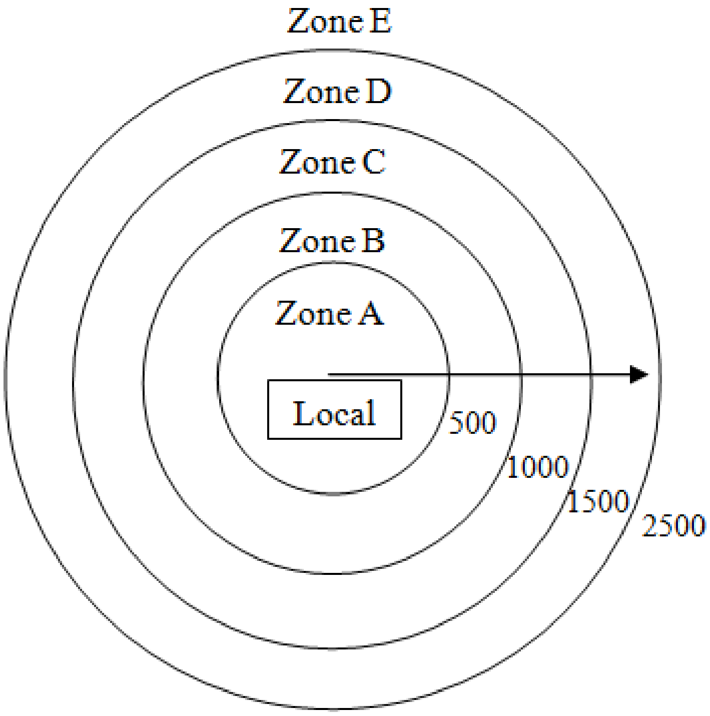


Figure 7.3 Geographic sequence of lung allocation process [5]

Table 7.1 Age group prioritization matrix [8]

	Donor Age <12	**Donor Age 12−17**	**Donor Age 18+**
1st Priority Candidate	Recipient Age <12	Recipient Age 12−17	Recipient Age 12+
2nd Priority Candidate	Recipient Age 12−17	Recipient Age <12	Recipient Age <12
3rd Priority Candidate	Recipient Age 18+	Recipient Age 18+	

The ABO types of the donor and the recipient plays a major role in deciding because it critically affects the success of the lung transplantation. There are two levels of ABO match level: identical and compatible. The first preference is given to the candidates who have identical match with the donor, and then the compatible ABO match.

The recipient candidates are categorized into two classes: adults and pediatric candidates. The ones older than 12 years are adults and ranked based on aforementioned LAS score whereas younger ones (less than 12 years) are pediatric candidates and are ranked based on the waiting time on the UNOS waiting list [5], [8]. Figure 7.4 summarizes the current lung allocation method in a schema.

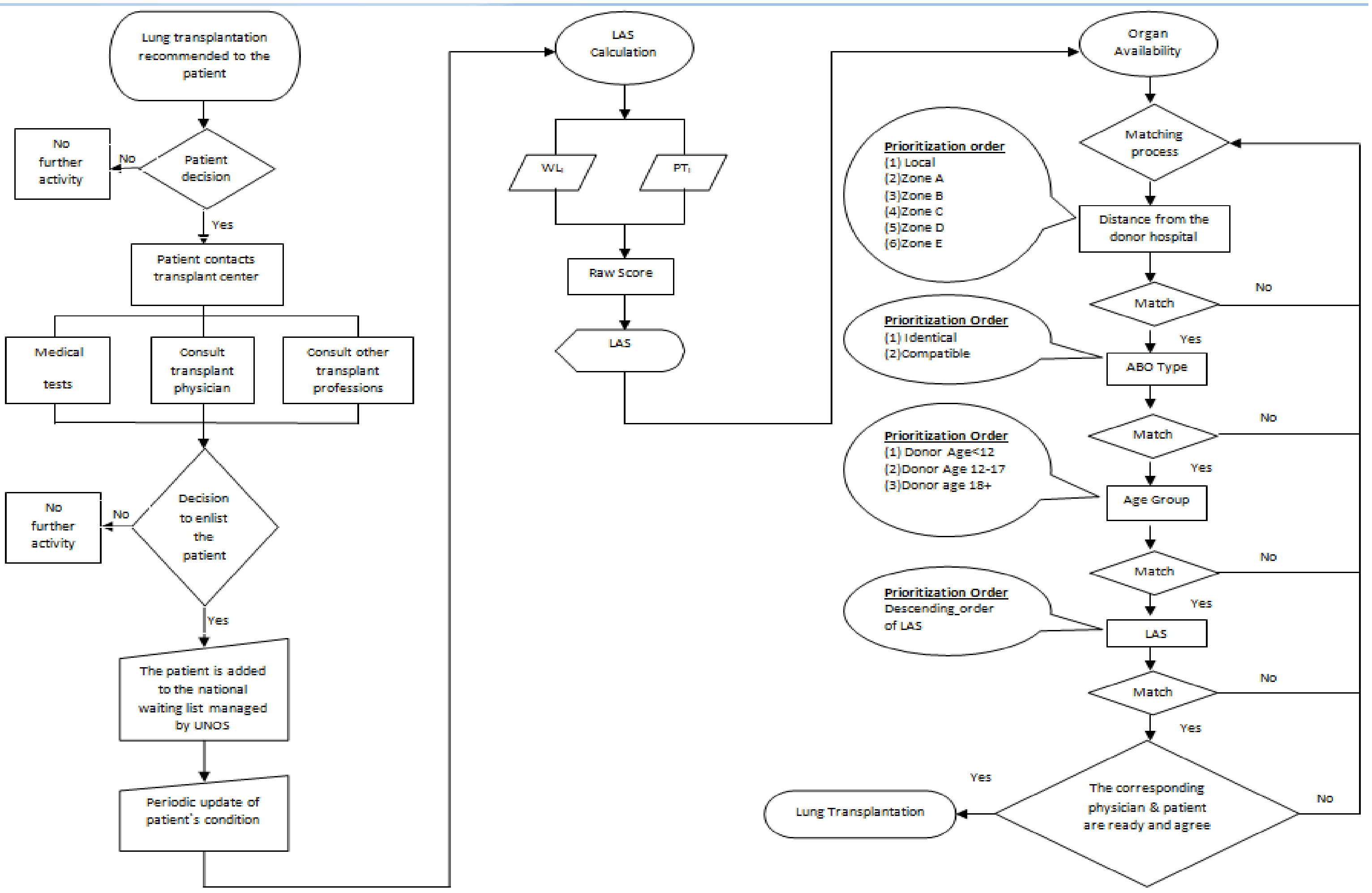


Figure 7.4 Schematic representation of the current lung allocation process

7.3 Validation of the Composite Score of Matching Index for Lung Transplants

A composite score for lung matching index was created in Chapter VI. In this current chapter, we will search for the potential improvements if such an index is used for the lung allocation procedure. Since the matching and allocation are quite complex step-by-step procedures with various performance measures to be considered simultaneously, we use simulation modeling in this chapter to validate the proposed matching index. The potential modules in this simulation model would be patient arrival, donor arrival, waiting for allocation, prioritization and matching, and transplantation modules. After constructing the simulation model based on the current allocation scheme as summarized in Section 7.2, the verification of it should be ensured. If the simulation model runs correctly without errors, then the validation of it should also be certified. Validation can be tested by checking the model output results in comparison with the actual outputs presented by UNOS allocation scheme. If the validation of the simulation fails, the model formulation and construction would be revised and corrected. If it produces statistically similar enough results compared to the baseline model of UNOS, then we can proceed to incorporate our proposed matching scheme. What-if scenario analyses would be conducted at this stage to observe how the output measures change based on potential changes in the logic of matching and allocation.

Since our proposed matching index is a combination of various input factors to satisfy 3 output measures (as listed previously), the determination of weights for each factor would be analyzed through response surface methodology (RSM). RSM can be used as a post-simulation analysis to significantly reduce the number of simulation runs. It gives an idea of how the response surface changes over various regions of input-factor space to find the optimal settings for them. By utilizing the RSM technique, we can determine the suboptimal weights for our composite score matching index. The flowchart for the simulation modeling approach in our proposed method is illustrated in Figure 7.5.

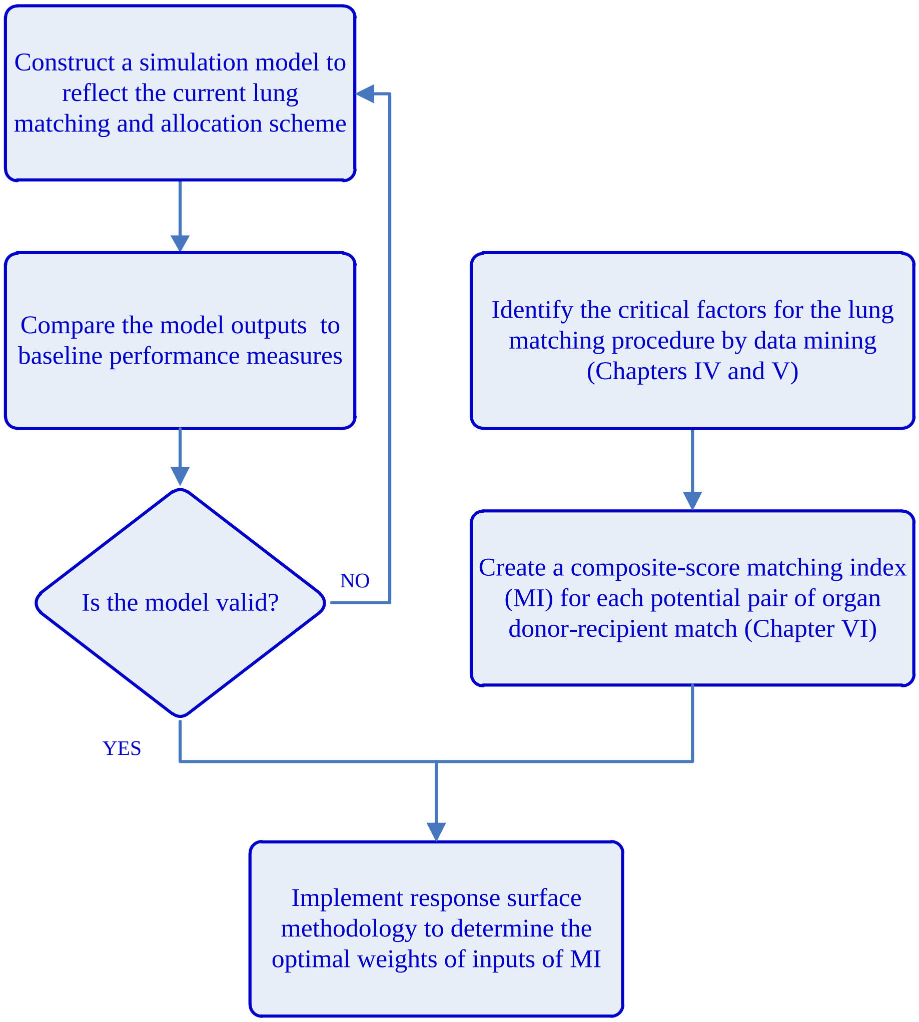

Construct a simulation model to reflect the current lung matching and allocation scheme
Compare the model outputs to baseline performance measures
Is the model valid?
NO
YES
Identify the critical factors for the lung matching procedure by data mining (Chapters IV and V)
Create a composite-score matching index (MI) for each potential pair of organ donor-recipient match (Chapter VI)
Implement response surface methodology to determine the optimal weights of inputs of MI

Figure 7.5 Proposed simulation modeling framework

7.4 Simulation Modeling

In this study, Simio$^{®}$ simulation software package was used to develop a simulation model. Simio$^{®}$ has changed the modeling basis from process orientation to object orientation and has taken its name from the notion of **sim**ulation modeling framework based on **i**ntelligent **o**bjects [9]. The modelers can construct intelligent objects to be utilized in multiple modeling projects, which makes the object orientation very simple to utilize and in turn effective to run [9].

There are two main input streams for our simulation model, namely *donor arrival module* and *patient arrival module*. The donor arrival module provides donor organ arrivals and assigns the related attributes to be used in further modeling. In the same vein, the patient arrival module creates patient arrivals and assigns their related attributes to each and every patient. Since the purpose of this study is to compare and contrast lung allocation score (LAS) system against our proposed composite score matching index, we use the UNOS lung transplant data starting from the year 2005 up to 2008 (the LAS system was developed in 2005). Another critical module of the simulation model is *the matching and allocation module*, which takes into account the steps as represented in Figure 7.4. As a donor organ enters the system, this module determines which patients match with the organ, how to prioritize them based on the criteria of *distance, ABO type, age group,* and *LAS*. Our contribution to the matching and allocation module is in the determination of which patient should be given the priority based on the composite score matching index derived in Chapter VI.

A sample screenshot of Simio is presented in Figure 7.6 to show how the decision logic works in the matching and allocation module in the simulation model. This process flow basically models the prioritization of the candidate patients based on their age category, as summarized in Table 7.1.

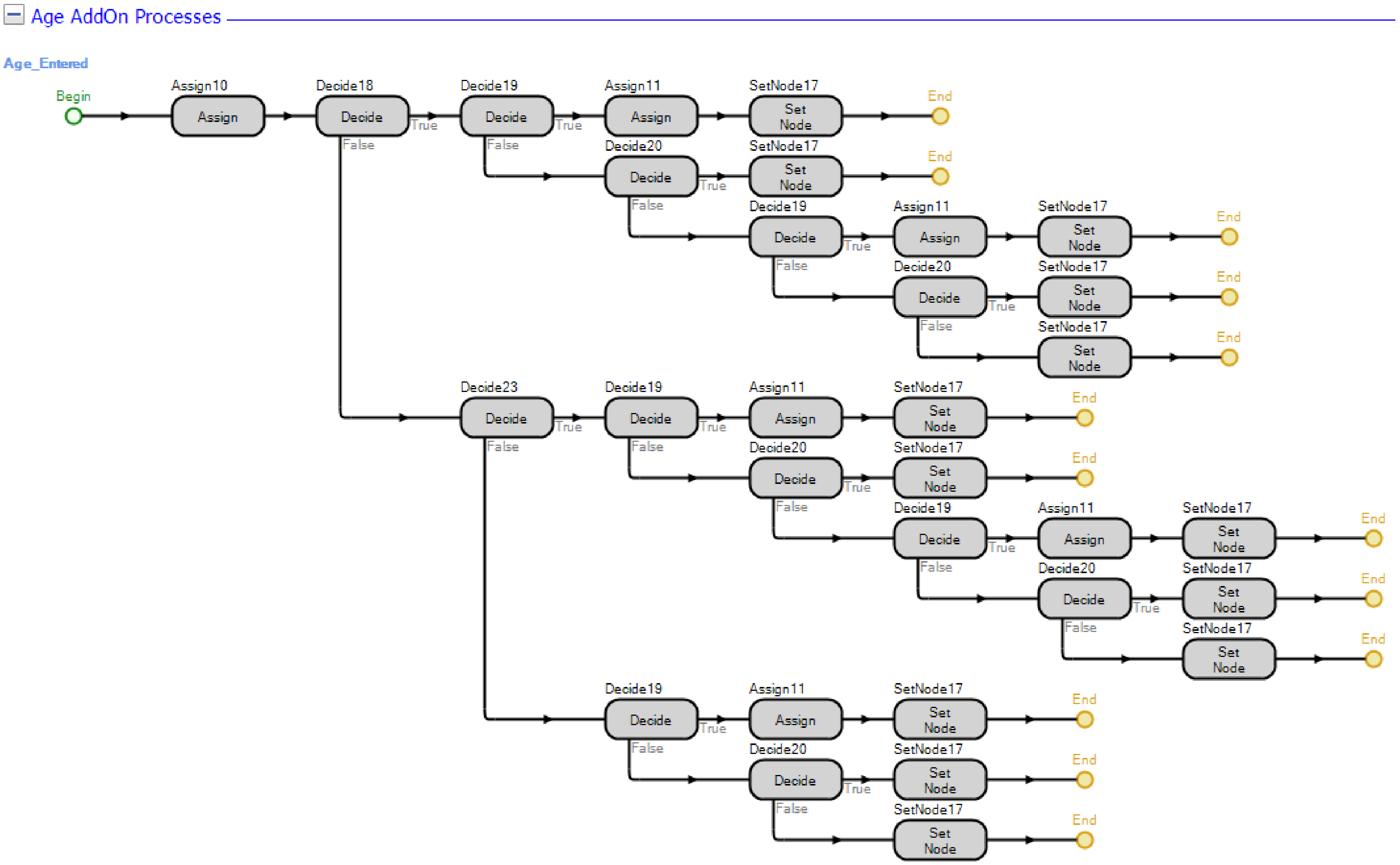


Figure 7.6 The Simio screenshot for matching and allocation module with regard to *age match*

7.5. Response Surface Methodology-based Simulation Optimization

Response surface methodology (RSM) is widely used for simulation-based optimization, which drastically minimizes the number of required experimental runs [10]. RSM is conducted by the integration of polynomial equation using regression analysis and a functional relationship between the output (dependent/response) variable y and the set of input (independent) variables x_i [11].

As the first step, a first-order polynomial function of input variables along with two-way interactions is fit as given in Eq. (7.4)

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{j=1}^{k} \sum_{i=1}^{k} \beta_{ij} x_i x_j + \varepsilon \tag{7.4}$$

where x_i refers to the input variables and y refers to the response variable to model. β’s refer to the unknown regression coefficients to be determined by the method of least squares so that the random model error ε would be minimized. If the model does not perform well and has some significant curvature, then a second-order model is fit via Eq. (7.5)

$$y = \beta_0 + \sum_{i=1}^{k} (\beta_i x_i + \beta_{ii} x_{ii}^2) + \sum_{j=1}^{k} \sum_{i=1}^{k} \beta_{ij} x_i x_j + \varepsilon \tag{7.5}$$

The output measures (responses) are computed based on each experiment-based simulation. The model adequacy is measured via coefficient of determination (i.e. R^2) as given in Eq. (7.6)

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(F_t - Y_t)^2}{\sum_{t=1}^{n}(Y_t - \bar{Y}_t)^2} \tag{7.6}$$

where F_t , Y_t , and $\overline{Y_t}$ refer to predicted, actual, and mean values of the response variable, respectively. Since adding more input variables would increase the value of R^2, the adjusted R^2 is usually also considered for model fit as given in Eq. (7.7)

$$R_{adj}^2 = 1 - \frac{k-1}{k-p}(1 - R^2) \tag{7.7}$$

where k is the total number of observations and p is the number of regression coefficients. These two measures, R^2 and R_{adj}^2 are supposed to be close to each other and both close to 1 for a good fitted model [11].

Central composite design (CCD) is the main technique used in modeling second-order response surface models. If the response surface is modeled dependent on three input variables (which is the case for our study), the surface is approximated to be a hypercube or a sphere as shown in Figure 7.7 (with a radius of $\sqrt{k}$ where k is the number of input variables) [11].

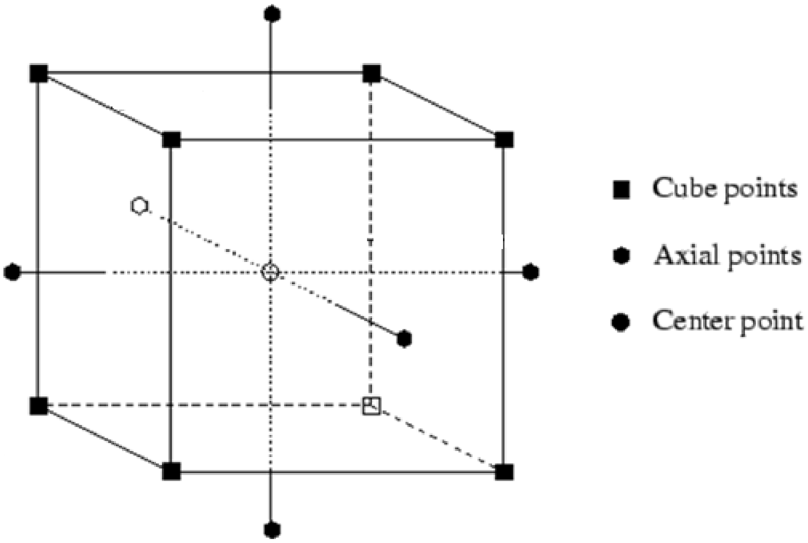


Figure 7.7 Central composite design for three input variables [12]

7.6 Simulation Modeling Results

As explained in Figure 7.5, the first step for the simulation is to verify and validate that the model mimics the real world. In order to check that, the performance measures taken into account based on the LAS system are compared and contrasted against our simulation outputs. This part of the study is focused on the data between 2005 and 2007 due to the data existence at the time of the study. The output measures, i.e. survival rate and average waiting time came out to be in alliance with the actual system outputs within the time frame of 2005-2007. Table 7.2 summarizes the outputs of the simulation and presents the actual system values.

Table 7.2 The comparison of performance measures of the simulation vs. actual system

	Survival Rate	Average Waiting Time
Simulation Outputs	83%	283 days
Actual System Outputs	81%	267 days

As seen in Table 7.2, the values determined by the simulation model are a good representation of the real-world. In statistical terms, the estimated survival rate at 83% has a standard error of 0.0099 and therefore the 95% confidence interval (i.e. [0.811, 0.849]) includes the real-world output of 0.81. In the same fashion, the average waiting time on the waiting list has a standard error of 8.3 and hence the 95% confidence interval for the simulated output, [266.732, 299.268], consists of the actual system output value of 267. Having showed that the simulation model replicates the real-world output measures and hence mimics the real-world lung matching and allocation scheme correctly, the next step is to discover if our composite score matching index helps improve this system in terms these output measures.

Table 7.3 Performance outputs received via the matching index

	Survival Rate	Average Waiting Time
New Value via Matching Index	86%	271 days
SE	0.0012	5.8
95% CI	(85.76%, 86.24%)	(259.632, 282.368)

Table 7.3 summarizes the results received by implementing our proposed composite score matching index instead of the currently implemented LAS scoring system. Since

the fundamental contribution of the matching index is essentially focused on a more efficient matching between the donor organ and the recipient by making use of voluminous UNOS dataset more effectively, it provides an improved survival rate (from 81% to 86%) while the average waiting time of the patients on the waiting list does not improve. This makes sense since in the queuing systems the waiting time could only be decreased by either decreasing the service time and/or by decreasing the service demands. Neither of these is the case for our study on hand. Therefore, the implementation of the composite score matching index (as derived in Chapter VI) is justified via this simulation output.

7.7 Fine-Tuning for the Weights of the Matching Index Components

In Section 7.6, *the matching index* was shown to be a better way of matching candidate recipients with donor organs. In this current section, we develop a response surface method (particularly a central composite design with *k=3* variables) to optimize the weights of the components of the matching index as explained in Chapter VI in detail, namely *recipient's profile, donor's profile,* and *match level*. In our setting, the input variables are the weights (coefficients) of these three latent variables where the weights refer to the importance of each latent variable. In doing so, it can be determined how to weight each of these latent variables so as to receive a satisficing matching in terms of survival rate and waiting time on the list. Since the problem refers to a multi-response (multi-criteria) optimization problem, we implement the concept of *desirability approach* developed by Derringer and Suich [13] and refined by Castillo *et al.* [14]. In this approach, each response *y_i* is mapped to a desirability function, *$d_i(y_i)$*, which take values between 0 and 1. If the response *y_i* is at its desired level (target/goal value), *$d_i(y_i)$* would be 1. On the other hand, *$d_i(y_i)$* would take 0 if *y_i* is out of its desirable range. Via a geometric mean calculation as given in Eq. (7.8), these individual desirabilities are maximized to calculate the *overall desirability* (D).

$$D = [(y_1) * d_1(y_1) * \ldots * d_n(y_n)]^{1/n} \tag{7.8}$$

where *n* is the total number of responses [11]. If the response is to be maximized, the corresponding individual desirability function is given by Eq. (7.9).

$$d_i(y_i) = \begin{cases} 0 \; if \; y_i(x) < L_i \\ \left(\frac{y_i(x)-L_i}{T_i-L_i}\right)^s \; if \; L_i \leq y_i(x) \leq T_i \\ 1 \; if \; y_i(x) > T_i \end{cases} \tag{7.9}$$

where *y_i* is the response to be maximized, *L_i* is the lower value, and *T_i* is the target value for the the response *y_i*. The exponent *s* determines the importance of hitting (being closer to) the target value *T_i* [11]. Similarly following the same notation, if the response is to be minimized, the desirability function would be written as in Eq. (7.10).

$$d_i(y_i) = \begin{cases} 1 \; if \; y_i(x) < T_i \\ \left(\dfrac{y_i(x) - U_i}{T_i - U_i}\right)^s \; if \; T_i \leq y_i(x) \leq U_i \\ 0 \; if \; y_i(x) > U_i \end{cases} \tag{7.10}$$

where U_i is an upper value for the response.

7.8 Results and Discussion

The generic central composite design (CCD) matrix with *k=3* (x_1*=recipient's profile*, x_2*= match level*, and x_3*=donor's profile*) as given in Table 7.4 was used to conduct the response surface methodology in the simulation model of this study. Since the CCD is utilized to determine the optimal weights/coefficients of the three latent variables (x_1, x_2, x_3), the coding-uncoding these weights was realized by the scale transformation as summarized in Eq. (7.11).

$$x_{h(new)} = \frac{x_{h(old)} - x_{\min(old)}}{x_{\max(old)} - x_{\min(old)}}\left(x_{\max(new)} - x_{\min(new)}\right) + x_{\min(new)} \tag{7.11}$$

The fitted simulation meta-models based on the coded x_i units for the two response values are summarized as follows in Eq. (7.12) and Eq. (7.13).

Table 7.4 Central composite design matrix

x_1	x_2	x_3
-1	-1	-1
1	-1	-1
-1	1	-1
1	1	-1
-1	-1	1
1	-1	1
-1	1	1
1	1	1
-1.682	0	0
1.682	0	0
0	-1.682	0
0	1.682	0
0	0	-1.682
0	0	1.682
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0

$$\hat{y}_1 = 0.809 - 0.028x_1 + 0.075x_2 + 0.042x_1^2 \tag{7.12}$$

$$\hat{y}_2 = 275.643 - 9.405x_1 + 28.661x_2 + 16.774x_1^2 - 18.389x_1x_2 \tag{7.13}$$

The R^2 and R_{adj}^2 values for each of the response values were found to be as follows showing the model adequacy for both of the response surface models: $R^2(\hat{y}_1) = 83.26\%$; $R_{adj}^2(\hat{y}_1) = 68.19\%$ and $R^2(\hat{y}_2) = 86.48\%$; $R_{adj}^2(\hat{y}_1) = 74.31\%$. Since the CCD-based metamodeling performs satisfactorily based on the R^2 and R_{adj}^2 values, the next step is to reveal the combined (overall) desirability of these two responses, y_1 and y_2.

Assigning different levels to *s* values in Eqs. (7.9) and (7.10) makes the overall desirability function, D, either convex (via s<1) or concave (via s>1). If s is assigned to be 1, it is approximated to be linear towards T_i [13]. Since the shape of the function is not known apriori, it is suggested in the literature to adopt a trial-and-error approach with various settings at 0.1, 1, and 10 [14]. Hence, in this study all potential combinations were searched and the best one was received by s=0.9 for y_1 and s=0.8 for y_2, which revealed the highest overall desirability value at D=0.82. The survival time was targeted at 0.90 and was restricted to be bigger than 0.85. The waiting time was targeted to be 270 days and was restricted to be smaller than 365 days, namely one year. The optimal solution for the weights of the latent variables, x^**(recipient's profile, match level, donor's profile)*, the responses at the optimal solution, the corresponding individual desirability function results, and the overall desirability are tabulated in Table 7.5. Using Eq. (7.11), uncoding of the weights was also conducted and is shown in Table 7.5.

Table 7.5 The CCD-based optimal weighting scheme and the desirability functions

Coded optimal solution (x*)	**Uncoded optimal solution**	**$y_1(x^*)$**	**$y_2(x^*)$**	**$d_1(y_1)$**	**$d_2(y_2)$**	**$D(y_1,y_2)$**
[-0.9664, 0.9344, 0.5267]	[0.2127, 0.778, 0.657]	0.90	308.02	1.00	0.66	0.82

Note that the response surface method-based simulation optimization placed the most importance on the *match level* between the donor and the recipient which is followed by the *donor's profile.* This may be attributed to the fact that once the recipient is prioritized and determined with regard to our newly derived matching index, the survival rate of the recipient would be dependent on the extent how much the donor and the recipient are fitting to each other in terms of medical matching. Also, the survival of the recipient is strongly affected by the medical history of the donor such as the donor's history regarding the usage of alcohol, cigarette, and cocaine. While this approach did not cause a dramatic increase in the average waiting time of patients on the list (a slight change from 267 days to 308 days), it helped improve the survival rate from 81% to 90%.

7.9 Conclusions

This study is primarily focused on the validation of the composite score matching index for lung transplant patients which was derived by using a structural equation model-based decision tree model. The simulation model results showed that the matching indexing of the recipients in terms of prioritization and then allocation of the donor organ accordingly provided an improved survival rate (from 81% to 86%) with a slight deterioration in the average waiting time (from 267 days to 271 days). Sticking to the matching index formulation, a response surface method-based simulation was deployed to develop meta-models to fine-tune the weights of the matching index components and to optimize the lung allocation system. This was realized by jointly optimizing the lung transplant measures, namely justice principle (in terms of waiting time) and utility principle (in terms of survival rate) via the desirability approach along with a central composite design method. The results showed that without making a policy change in the current UNOS-based lung allocation system the survival rate can still be increased up to 90% through the suggested comprehensive data analysis-based matching index derivation. Such a methodology not only provides an improved utility of the donor organs but also presents a means for the medical experts to gain in depth control of the voluminous data while making their decisions.

The integration of response surface method-based simulation helps determine the importance of the components of the transplant decision process and also provides a low cost tool for medical professionals to conduct what-if analyses effectively and efficiently. Additionally, this integration provides a generic model which can later be evaluated based on the potential changes and improvements in the transplant systems of other organ types i.e. liver, kidney, and etc.

REFERENCES

[1] S.Q. Davis, E.R.J. Garrity, Organ allocation in lung transplant, CHEST 132 (2007) 1646-1651.

[2] A. Iribarne, M.J. Russo, R.R. Davies *et al.*, Despite the decreased wait-list times for lung transplantation, lung allocation scores continue to increase, CHEST 135 (2009) 923-928.

[3] R.R. Hachem, E.P. Trulock, The new lung allocation system and its impact on waitlist characteristics and post- transplant outcomes, in: Seminars in Thoracic and Cardiovascular Surgery 20 (2008) pp. 139-142.

[4] T.M. Egan, S. Murray, R.T. Bustami *et al.*, Development of new lung allocation system in the United States, American Journal of Transplantation 6 (2006), 1212-1227.

[5] United Network for Organ. Organ Distribution: Allocation of Thoracic Organs Policies: http://www.unos.org/PoliciesandBylaws2/policies/pdfs/policy_9.pdf Accessed on July 28, 2009.

[6] United Network for Organ. Organ Distribution: A guide to calculating the lung allocation score: http://www.unos.org/SharedContentDocuments/lung_allocation_score_updated_01072009.pdf Accessed on July 28, 2009.

[7] T.M. Egan, The new lung allocation system in the United States, Current Opinion in Organ Transplantation 11 (2006), 490-495.

[8] United Network for Organ. Organ Distribution: LAS Brochure for Medical Professionals:http://www.unos.org/SharedContentDocuments/Lung_Professional(1).pdf Accessed on July 28, 2009.

[9] C.D. Pegden, Introduction to Simio, in: Proceedings of the Winter Simulation Conference (2008) 229-235.

[10] G.E.P. Box, K.B. Wilson, On the Experimental Attainment of Optimum Conditions, Journal of Royal Statistical Society 13 (1951) 1-38.

[11] D.C. Montgomery, *Design and Analysis of Experiments*, (John Wiley and Sons, 2001).

[12] Adapted from: www.iue.tuwien.ac.at/ phd/plasun/img17.gif (Accessed on November 1, 2010)

[13] G.C. Derringer, R. Suich, Simultaneous Optimization of Several Response Variables, Journal of Quality Technology 12 (1980) 214-219.

[14] E.D. Castillo, D.C. Montgomery, D.R. McCarvile, Modified Desirability Functions for Multiple Response Optimization, Journal of Quality Technology 28 (1996) 334-345.

CHAPTER VIII

CONCLUSIONS AND FUTURE WORK

Healthcare has recently become one of the most important research domains within the industrial engineering studies. Within this domain, resource allocation, particularly matching and allocation of scarce number of donor organs with a long list of candidate patients, has attracted researchers' attention more than the others. This is mainly attributed to the fact that the voluminous data collected for system modeling have not been efficiently utilized. Therefore, this dissertation has targeted at modeling of lung transplant procedures through a methodological data analysis-based strategy. The major contribution of the study and future plans for research are summarized in the follow subsections.

8.1 Conclusions

UNOS lung transplantation dataset is used within this study to reveal the unknown patterns lying under the data. Although voluminous data has been recorded for the above purpose, a small subset of it has been explored in the literature based on the intuitions and experiences of medical experts. This study provides an effective way for selection of critical variables for survival analysis and prediction of lung transplant outcomes (e.g. survival time after transplant) by utilizing data mining techniques.

By deploying an integrated method with clustering algorithms and medical domain knowledge, potential organ recipients (candidate patients) are categorized in terms of their risk severity after the transplant. A prognostic index is derived for this purpose to group patients in terms of risk groups, e.g. low, medium, and high risk. Such a grouping would lead to a wise approach for medical experts to plan on an appropriate means of treating organ recipients and suggesting a more proper follow-up and clinical visit scheduling.

A sophisticated structural equation model-based decision tree is developed to simultaneously predict the performance outcomes of the lung transplant, namely, survival time and functional status of the patient after the transplant. The UNOS-based large dataset is grouped into three major representative higher-level components in light of the discussions with medical experts in the transplant surgery. This integration of structural equation modeling with decision trees not only provides a satisfactory level of accuracy, but it also presents more interpretability of the massive dataset and the related lung procedure. Moreover, a single composite-score matching index has been derived for each potential match of candidate recipient and the donor organ.

The matching index derived via the structural equation model-based decision tree is validated to be a more effective way of matching lungs to the patients since it achieves an increased survivability with an ignorable amount of deterioration in the average waiting time on the list. This is realized by a simulation study using the findings of the structural equation model-based decision tree model. The simulation-based optimization using the response surface methodology provides a cost-effective tool to determine the weights of the matching index factors. The developed simulation model can be further utilized to evaluate potential future changes in the lung matching system before deploying them in the real life.

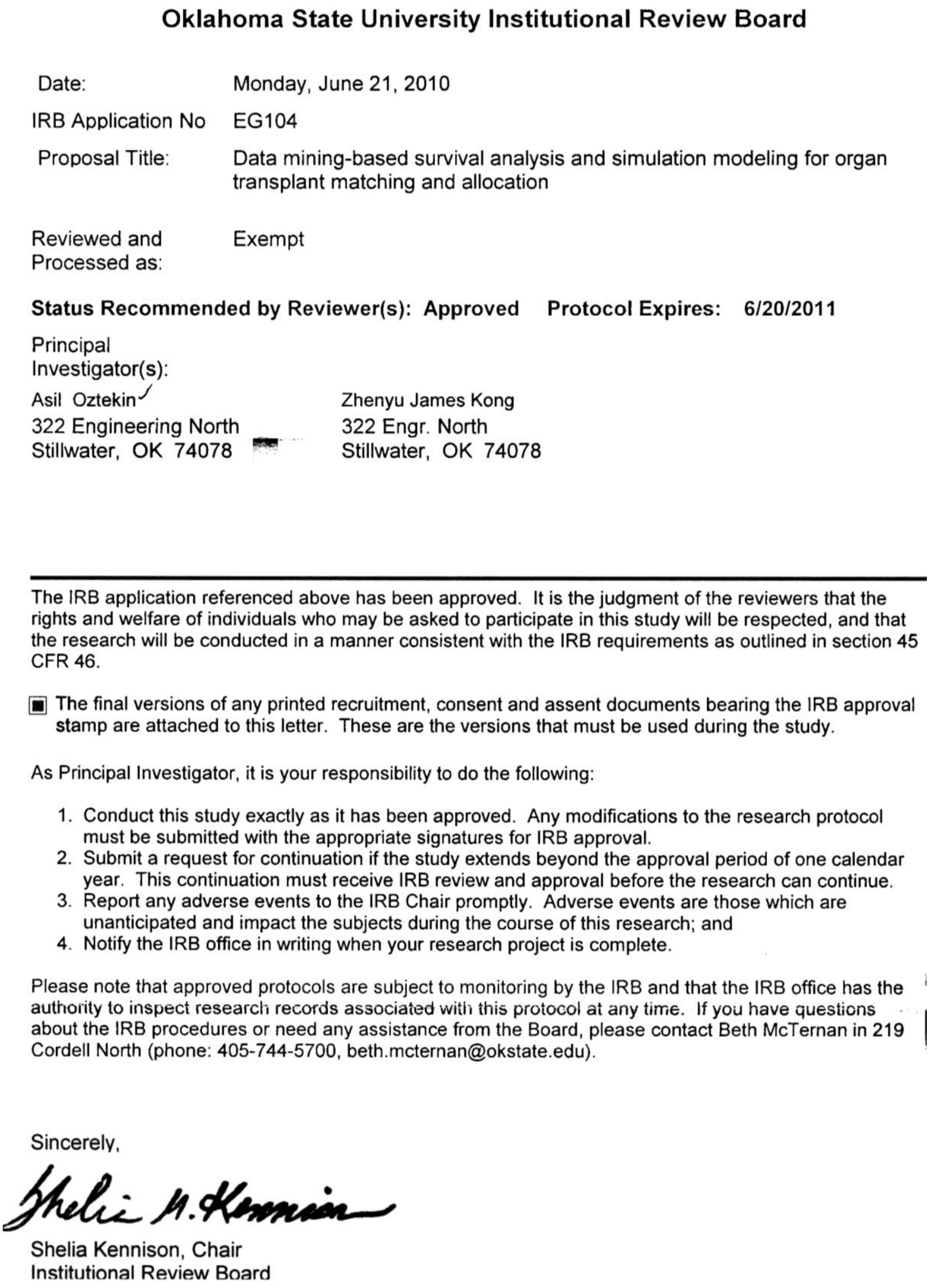
8.2 Future Work

Targeted at improving the organ matching and allocation system in the US, the prospected future work could be focused on data analysis and modeling of other organ matching and allocation systems such as liver, kidney, etc.

First of all, a decision support system (DSS) equipped with a user-friendly frontend and a backend application would be developed in order to make the proposed modeling approach usable for the medical professionals in this domain. Such a DSS would enable medical experts to deal with voluminous data more effectively and efficiently by providing reliable and accurate organ matching and allocation results in a very short period of time.

Secondly, in this research the risk groups of patients (low, medium, and high risk) have been modeled in the subsequent analysis as a whole. A future research extension would be devising a separate matching index and hence suggesting a matching and allocation scheme based on the risk group which patients belong to. Such a scheme may be more appropriate and convincing for various stakeholders of the lung allocation system.

Thirdly, the current lung allocation scheme is severely affected and limited by the ischemia time of the donor organ. Transferring the extensive use of the Radio Frequency Identification (RFID) technology from manufacturing supply chains into the supply chain management of the donor organs would potentially provide an extended ischemia time and hence an improved lung allocation system. This RFID implementation could be realized via sensor-based modeling and hopefully lead to a fairer system diminishing the effect of distance of potential organ recipient from the donor organ.



Oklahoma State University Institutional Review Board

Date: Monday, June 21, 2010

IRB Application No EG104

Proposal Title: Data mining-based survival analysis and simulation modeling for organ transplant matching and allocation

Reviewed and Processed as: Exempt

Status Recommended by Reviewer(s): Approved Protocol Expires: 6/20/2011

Principal Investigator(s):

Asil Oztekin 322 Engineering North Stillwater, OK 74078	Zhenyu James Kong 322 Engr. North Stillwater, OK 74078

The IRB application referenced above has been approved. It is the judgment of the reviewers that the rights and welfare of individuals who may be asked to participate in this study will be respected, and that the research will be conducted in a manner consistent with the IRB requirements as outlined in section 45 CFR 46.

☒ The final versions of any printed recruitment, consent and assent documents bearing the IRB approval stamp are attached to this letter. These are the versions that must be used during the study.

As Principal Investigator, it is your responsibility to do the following:

1. Conduct this study exactly as it has been approved. Any modifications to the research protocol must be submitted with the appropriate signatures for IRB approval.
2. Submit a request for continuation if the study extends beyond the approval period of one calendar year. This continuation must receive IRB review and approval before the research can continue.
3. Report any adverse events to the IRB Chair promptly. Adverse events are those which are unanticipated and impact the subjects during the course of this research; and
4. Notify the IRB office in writing when your research project is complete.

Please note that approved protocols are subject to monitoring by the IRB and that the IRB office has the authority to inspect research records associated with this protocol at any time. If you have questions about the IRB procedures or need any assistance from the Board, please contact Beth McTernan in 219 Cordell North (phone: 405-744-5700, beth.mcternan@okstate.edu).

Sincerely,

Shelia M. Kennison

Shelia Kennison, Chair
Institutional Review Board

VITA

Asil Oztekin

Candidate for the Degree of

Doctor of Philosophy

Dissertation: DATA MINING-BASED SURVIVAL ANALYSIS AND SIMUALTION MODELING FOR LUNG TRANSPLANT

Major Field: Industrial Engineering and Management

Biographical:

Personal Data: Born in Manisa, Turkey on June 22nd, 1982.

Education:

Completed the requirements for the Doctor of Philosophy degree with a major in Industrial Engineering and Management at Oklahoma State University, Stillwater, Oklahoma in December, 2010.

Received the Master of Science degree with a major in Industrial Engineering at Fatih University, Istanbul, Turkey in July, 2006.

Received the Bachelor of Science degree with a major in Industrial Engineering at Yildiz Technical University, Istanbul, Turkey in July, 2004.

Experience: Graduate Teaching and Research Assistant, School of Industrial Engineering and Management, Oklahoma State University from January 2007 to present; Graduate Teaching and Research Assistant, Department of Industrial Engineering, Fatih University from January 2005 to July 2006.

Awards: Recipient of Outstanding Research Assistant Award, presented by Alpha Pi Mu Industrial Engineering Honor Society, Oklahoma State University Chapter, Fall 2009; Recipient of US nation-wide Hanel Storage Systems Honor Scholarship, presented by Material Handling Education Foundation Inc., Summer 2010.

Name: Asil Oztekin Date of Degree: December, 2010

Institution: Oklahoma State University Location: Stillwater, Oklahoma

Title of Study: DATA MINING-BASED SURVIVAL ANALYSIS AND SIMULATION MODELING FOR LUNG TRANSPLANT

Pages in Study: 163 Candidate for the Degree of Doctor of Philosophy

Major Field: Industrial Engineering and Management

Scope and Method of Study: The objective of this research is to develop a decision support methodology for the lung transplant procedure by investigating the UNOS nation-wide dataset via data mining-based survival analysis and simulation-based optimization. Traditional statistical techniques have various limitations which hinder the exploration of the information hidden under the voluminous data. The deployment of the structural equation modeling integrated with decision trees provides a more effective matching between the donor organ and the recipient. Such an integration preceded by powerful data mining models to determine which variables to include for survival analysis is validated via the simulation-based optimization.

Findings and Conclusions: The suggested data mining-based survival analysis was superior to the conventional statistical methods in predicting the lung graft survivability and in determining the critical variables to include in organ matching and allocation. The proposed matching index derived via structural equation model-based decision trees was validated to be a more effective priority-ranking mechanism than the current lung allocation scoring system. This validation was established by a simulation-based optimization model. It was demonstrated that with this novel matching index, a substantial improvement was achieved in the survival rate while only a short delay was caused in the average waiting time of candidate patients on the list. Furthermore, via the response surface methodology-based simulation optimization the optimal weighting scheme for the components of the novel matching index was determined by jointly optimizing the lung transplant performance measures, namely, the justice principle in terms of the waiting time and the utility principle in terms of the survival rate. The study presents uniqueness in that it provides a means to integrate the data mining modeling as well as simulation optimization with the survival analysis so that more useful information hidden in the large amount of data can be discovered. The developed methodology improves the modeling of matching and allocation system in terms of both interpretability and predictability. This will be beneficial to medical professionals at a great deal.

ADVISER'S APPROVAL: Dr. Zhenyu (James) Kong