# MODELS TO PREDICT SURVIVAL AFTER LIVER TRANSPLANTATION

By

Nathan Rollins Hoot

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biomedical Informatics

December, 2005

Nashville, Tennessee

Approved:

Professor Dominik Aronsky

Professor Irene Feurer

Professor Nancy Lorenzi

Professor C. Wright Pinson

ACKNOWLEDGEMENTS

I am grateful for the support of my thesis advisor, Dr. Dominik Aronsky. Being a teacher involves more than just intelligence, and his patience in working with me was a key contributor to the success of this research. I also want to thank my other thesis committee members. Dr. C. Wright Pinson has been a valuable member of the team through his clinical knowledge and his understanding of policy. Dr. Irene Feurer has taught me a great deal about statistical analysis. Dr. Nancy Lorenzi has served as my committee chair, and her expertise with people and organizational issues have been very helpful.

I would like to thank Dr. Constantin Aliferis for his contributions to this research, as well as the rest of the faculty in the Department of Biomedical Informatics, including Dr. Kevin Johnson, for supporting my work. I appreciate the members of the Vanderbilt Transplant Center who have contributed to my education in transplantation, including Drs. Michael Porayko, Kelly Wright, Ravi Chari, David Raiford, Derek Moore, and Mary Austin. The Medical Scientist Training Program has generously supported my clinical training, and I thank the former director, Dr. David Robertson, and the current director, Dr. Terry Dermody, for allowing me to explore a novel research path.

This work was supported in part by the National Library of Medicine grant LM07450-02. Portions of Chapter IV were reproduced with permission from the American Medical Informatics Association (Hoot N, Aronsky D. Using Bayesian networks to predict survival of liver transplant patients. AMIA Annu Symp Proc. 2005:345-49). Portions of Chapter V were reproduced with permission from Elsevier (Hoot NR, Feurer ID, Pinson CW, Aliferis CF. Modeling liver transplant survival: comparing techniques of deriving predictor sets. J Gastrointestinal Surg. 2005;9(4):563).

The data reported here have been supplied by the United Network for Organ Sharing as the contractor for the Organ Procurement and Transplantation Network. The interpretation and reporting of these data are the responsibility of the author and in no way should be seen as an

official policy of or interpretation by the Organ Procurement and Transplantation Network or the U.S. Government. The data and analyses reported in the 2004 Annual Report of the U.S. Organ Procurement and Transplantation Network and the Scientific Registry of Transplant Recipients have been supplied by the United Network for Organ Sharing and the University Renal Research and Education Association under contract with the U.S. Department of Health and Human Services. The author alone is responsible for reporting and interpreting these data.

My gratitude extends to my other colleagues and friends who are too numerous to mention, in particular Chris Bunick and Justin McCann. Most notably, I am grateful for the love of my family, who have always been unwavering in their support. Even on my down days, they always make me feel as though I deserve the Nobel Prize.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

# LIST OF ABBREVIATIONS

| Abbreviation | Definition |
|---|---|
| APACHE | Acute Physiology and Chronic Health Evaluation |
| AUC | area under the receiver operating characteristic curve |
| CanWAIT | Canadian Waitlisting Algorithm in Transplantation |
| CI | confidence interval |
| CTP | Child-Turcotte-Pugh |
| MELD | Model for End-Stage Liver Disease |
| MeSH | Medical Subject Heading |
| NOTA | National Organ Transplant Act |
| OPTN | Organ Procurement and Transplantation Network |
| ROC | receiver operating characteristic |
| SAPS | Simplified Acute Physiology Score |
| SRTR | Scientific Registry of Transplant Recipients |
| UNOS | United Network for Organ Sharing |

# CHAPTER I

# INTRODUCTION

## Research Motivation

Liver transplantation has become the standard of care for end-stage liver disease in the United States in the past two decades. During this time however, the waiting list of candidates awaiting transplantation has increased dramatically, while the size of the organ donor pool has increased more slowly [1,2]. As shown in figure 1, the national allocation system is currently not in equilibrium, with the quantity of liver grafts demanded exceeding the available supply. As a result of this deficit, more than 2,000 patients die annually while awaiting liver transplantation. Liver transplantation ranks among the most expensive medical services available, costing hundreds of thousands of dollars [3]. In light of the scarcity of both organs and money, careful allocation of liver grafts is critical in order to maximize survival and quality of life for liver transplant patients.

## History of Liver Transplantation

The first successful human liver transplant was performed by Dr. Thomas Starzl in 1967. The surgical foundation necessary for liver transplantation had been established, but underdeveloped technology for immunosuppression made long-term survival after liver transplantation rare during the 1960's and 1970's. Because of this, liver transplantation remained an experimental form of therapy until the 1980's. The advent of cyclosporine represented a breakthrough for liver transplantation, and it became a successful and established form of treatment.

In 1984, the U.S. Congress passed the National Organ Transplant Act (NOTA) [4]. It specified the framework for a national system of organ transplantation. Among its other provisions, it called for the establishment of the Organ Procurement and Transplantation Network

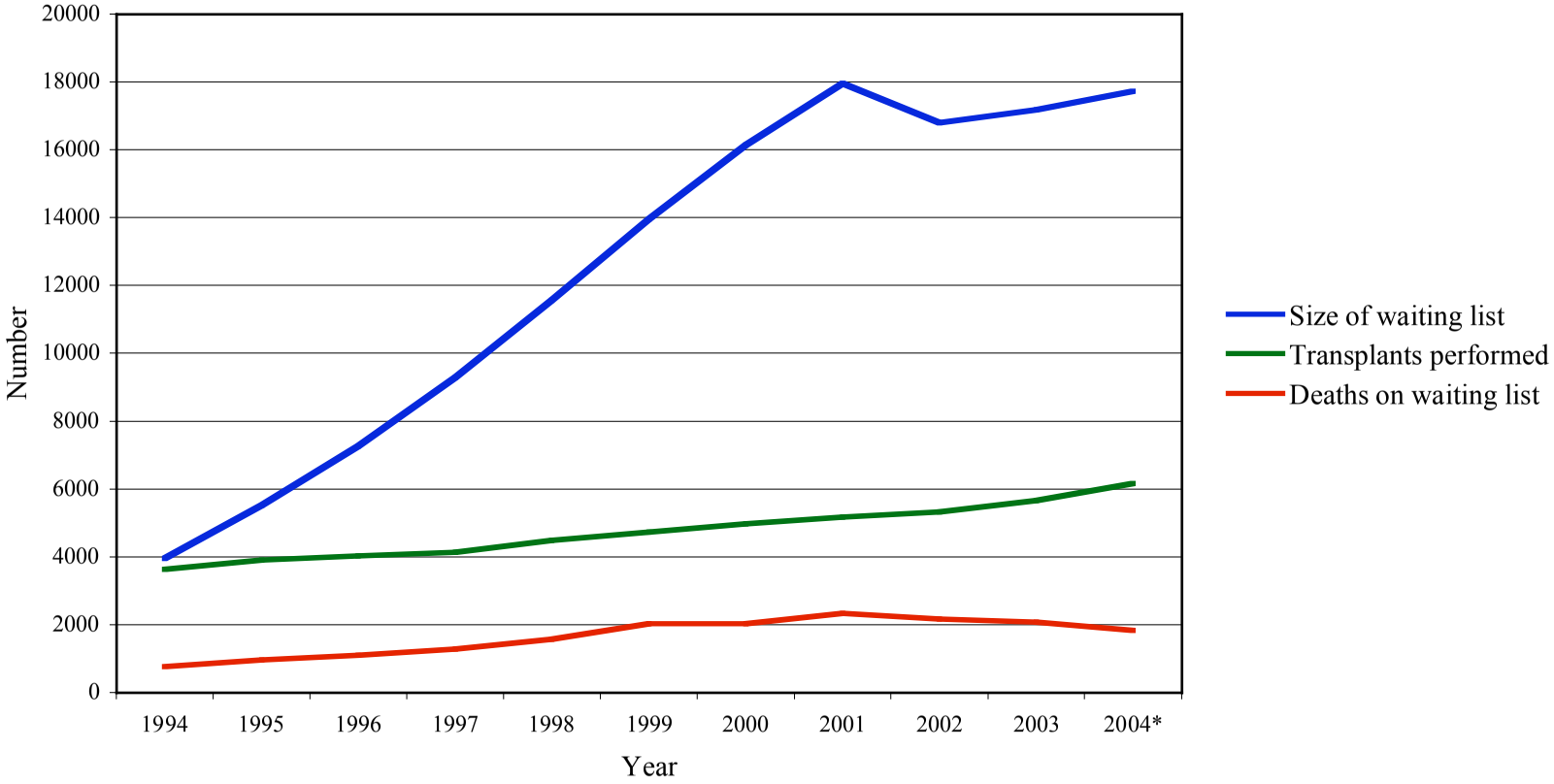



Figure 1. Trends in liver transplant waiting list growth, 1994-2004. Superimposed in the graph are the annual number of candidates awaiting liver transplantation, the total number of liver transplants performed, and the number of deaths among patients awaiting liver transplantation during a 10-year period. Nearly 18,000 candidates are presently on the waiting list, and approximately 6,000 liver transplants were performed in 2004. In recent years, roughly 2,000 patients have been dying annually while awaiting liver transplantation. The decrease in the size of the waiting list in 2002 may be attributable to a decrease in waiting list registrations associated with the implementation of the MELD system in February, 2002. Data from 1994-2003 were obtained from the SRTR Annual Report [1], and data from 2004 were calculated using publicly available information from the OPTN web page [2].

(OPTN), which represents the logistical infrastructure for transplantation in this country. On September 30, 1986, the United Network for Organ Sharing (UNOS) was awarded the initial contract to administer the OPTN, and the contract has been continuously renewed to the present time [5]. The UNOS is responsible to facilitate the matching of organ donors and recipients, to collect data for every transplant that occurs in the country, and to develop organ transplantation policy.

At the present time there are 122 health care institutions in the country currently offering liver transplantation, and a total of 6,164 transplants were performed in 2004 alone [2]. There are currently more than 18,000 candidates awaiting liver transplantation. Approximately 95% of all liver transplants performed use cadaveric liver grafts. While living donor liver transplantation is technically possible, this widespread use of this technique has been limited in part by the risk to the donor. Thus, the procurement of cadaveric livers is the primary limitation on the number of liver transplants that can be performed.

## Allocation Policy

The current mechanism for allocating cadaveric livers is based on the Model for End-Stage Liver Disease (MELD) [6,7]. It was implemented in February 2002, with the goal of creating a "sickest-first" system in which the patients could be ranked by an objective, continuous scale according to the severity of liver disease. Changes in allocation policy are made by the UNOS, with oversight from the U.S. Department of Health and Human Services.

The UNOS provides a framework of principles for making policy decisions about organ allocation. There are two specific and sometimes competing goals that must be considered for decision making in transplantation: Utility and Justice [8]. Utility is defined as "allocating organs to those individuals who will make the 'best' use of them," and Justice is defined as "allocation of organs to those patients in the most immediate need." The UNOS Liver Committee proposes allocation policy in accordance with these goals. Policy changes are proposed based upon a

combination of the clinical expertise of the committee members and experimental inference provided by the Scientific Registry of Transplant Recipients (SRTR).

## Predicting Outcomes

A means to help clinicians and policy-makers predict the outcomes of liver transplantation could prove beneficial both in the clinical and in the research setting. For clinical purposes, a model with good predictive ability may help to distinguish between transplants most likely to result in good outcomes and transplants most likely to result in poor outcomes. Avoiding futility in transplantation was noted in the desiderata for an organ allocation system [9]. If even a small percentage of futile transplants could be avoided each year, the effect on outcomes of patients receiving liver transplants could be substantial both in terms of life-years saved and quality of life gained.

Furthermore, when a liver graft is offered to a transplant center, clinicians may need to select among multiple candidates, all of whom are eligible for the donor liver. A good survival model could be employed to aid matching between donors and recipients. Beyond the issue of the liver shortage, a survival model could also help select the ideal course of therapy for individual patients. Some patients might experience greater benefit from treatments other than transplantation, such as hepatic resection or medical therapy, and a survival model may help clinicians to identify these patients. Thus, individual transplant centers could use a survival model as an adjunct to clinical judgment in selecting candidates for transplantation.

Finally on the research side, the SRTR is responsible for answering inferential requests from the UNOS. An accurate survival model could help policy-makers to assess *a priori* the likely effects of proposed changes in policy. For instance, a model could be employed as a component of a simulation process to model the aggregate effects of changes in allocation policy, before the changes are implemented. Because the utility of a survival model lies in its ability to make predictions at the time of decision-making – that is, just prior to transplantation – this thesis will focus on pre-transplant models.

A variety of measures exist to evaluate the success of a liver transplant. The measures most common to the field are patient survival and graft survival. Patient survival refers to the amount of time that a patient lives following the date of transplantation, while graft survival is defined as the number of days from transplantation to either patient death or retransplantation. Functional performance, as manifested by a patient's ability to perform the activities of daily living or return to work, can be used to represent the quality of life achieved after transplantation [10]. Patients may also report on health-related quality of life using a variety of instruments, such as the SF-36 inventory [11]. Quality-adjusted life years could be used to measure patient survival and quality of life simultaneously.

When dealing with the problem of allocation, the goal is to optimize the utilization of scarce cadaveric organs. Thus, focusing on graft survival allows one to consider both the life of the patient and the use of an organ. The Kaplan-Meier survival curve for graft survival following liver transplantation, shown in figure 2, reveals a steep decline in the number of survivors in the

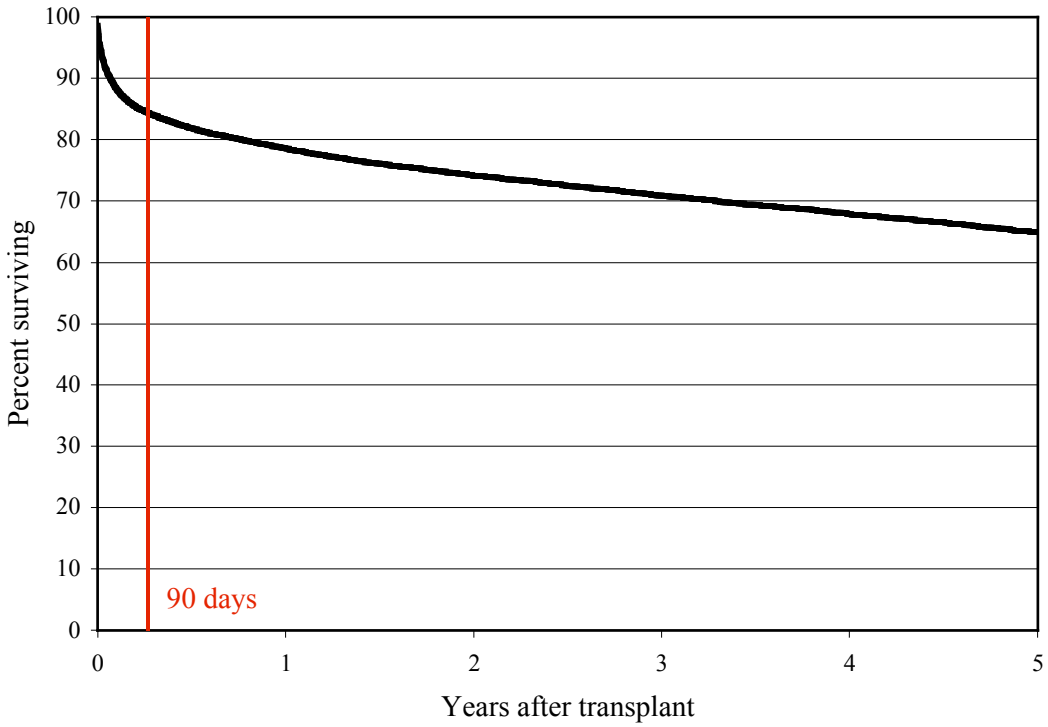



Figure 2. Kaplan-Meier graft survival after liver transplantation, 1995-2000. As illustrated, many of the deaths and retransplantations occur in the early post-operative period, and the survival rate stabilizes after a few months. The red line marks the 90-day anniversary of transplantation.

early post-operative period. As time progresses after transplantation, the slope of the curve flattens and stabilizes. Because of this, 90-day graft survival was adopted as the outcome metric of interest for this thesis.

**Specific Aims**

*Aim 1: To determine whether informatics techniques can improve upon Cox regression in predicting outcomes following liver transplantation.*

*Aim 2: To simplify the machine learning model by identifying the key factors necessary for robust prediction of outcomes.*

*Aim 3: To evaluate the practical significance of a refined model in a clinical setting.*

The motivation and context for the proposed research is described in Chapter I. Chapter II describes a systematic literature review to identify the relevant literature on modeling survival after liver transplantation. Chapter III presents methodology that will be shared by multiple studies in this thesis. The following four chapters detail the experiments that address the Specific Aims. Chapter IV presents a Bayesian network model that was created and validated to address Specific Aim 1. Chapter V describes a variable selection experiment, which will address Specific Aim 2. Chapter VI examines the complexity of the problem in a manner that partially addresses both Specific Aims 1 and 2. Chapter VII describes research on how well clinicians predict survival in comparison to mathematical models. The practical significance of a survival model lies in its ability to improve upon the current standard of care, which is the use of clinical judgment to predict survival. In this way, the study will address Specific Aim 3. The thesis concludes in Chapter VIII with a discussion of the key findings, together with the resulting policy implications.

# CHAPTER II

# SYSTEMATIC LITERATURE REVIEW

## Purpose

The goal of the systematic literature review was to identify articles dealing with the use of statistical models to predict survival after liver transplantation. I will note the strengths and weaknesses of articles written to address this problem. The research described later in this thesis will represent an effort to address the gaps that are identified in the body of past research.

## Methods

Relevant articles were defined to be those "dealing with the creation or validation of a general, pre-transplant, human, adult, statistical model to predict graft or patient survival following liver transplantation." Terms within this definition were defined by the following:

"General" was defined to refer to a model that was intended for use on the whole population of adult liver transplant recipients, rather than a specific sub-population based on disease etiology or retransplantation. When studies listed specific exclusion criteria, such as patients with fulminant hepatic failure, the studies were deemed relevant as long as the exceptions represented a minority of transplants performed.

"Pre-transplant" was defined to refer to a model that could be evaluated at the time of transplantation; in other words, it does not require the availability of intra-operative or post-operative information.

"Statistical model" was defined as a mathematical entity that was created or used for the purpose of making predictions.

Articles that adhere to the above definition should address three general concepts: (1) liver transplantation in humans, (2) an outcome measure, and (3) statistical modeling. I identified all possible medical subject heading (MeSH) terms that fell into any of the three categories using

Table 1. PubMed search strings for systematic literature review on 8/25/2005

| Search String | # of Abstracts |
|---|---|
| 1. Search "Liver Transplantation"[MeSH] | 25494 |
| 2. Search #1 AND "Humans"[MeSH] | 21753 |
| 3. Search #2 AND ("Mortality"[MeSH] OR "Liver Transplantation/mortality"[MeSH] OR "Survival"[MeSH] OR "Prognosis"[MeSH] OR "Host vs Graft Reaction"[MeSH]) | 8094 |
| 4. Search #2 AND ("Statistics"[MeSH] OR "Models, Theoretical"[MeSH] OR "Computing Methodologies"[MeSH]) | 3782 |
| 5. Search #3 AND #4 | 2112 |
| 6. Search #5 AND "English"[la] | 1980 |
| 7. Search #6 AND Limit Publication Date up to 2005/03/31 | 1937 |

Note: Search #7 denotes the final set of abstracts examined by the reviewers.

the MeSH browser [12]. These three concepts, together with the MeSH terms that apply to them are described as follows:

The concept of liver transplantation in humans was described by the MeSH terms "Liver Transplantation" and "Humans".

The concept of an outcome measure was described by the MeSH terms "Treatment Outcome"; “Survival Rate"; "Graft Survival"; "Prognosis"; "Mortality"; "Liver Transplantation/ mortality"; “Host vs Graft Reaction"; "Graft Rejection"; and "Graft Survival".

The concept of statistical modeling was described by the MeSH terms "Proportional Hazards Models"; "Neural Networks (Computer)"; "Models, Statistical"; "Survival Analysis"; "Models, Theoretical"; "Computing Methodologies"; "Artificial Intelligence"; and "Bayes Theorem".

I traced the set of MeSH terms up the hierarchy and identified common parents whenever possible, with the goal of making the search very broad and thorough. Lastly, the search was limited to English articles that were published prior to March 31, 2005. No beginning publication date for the search was specified. The final PubMed search strings are shown in table 1.

Two reviewers independently examined all abstracts returned by the search and identified all potentially relevant articles using the title and abstract. Disagreement was resolved by

consensus opinion. When the correct classification was ambiguous based solely on the abstract, it was included for examination of the full-text article.

The full-text articles were retrieved, and their relevance was examined to obtain the final list of articles that fit the inclusion definition provided above. Aspects of the methodology of each article were abstracted: (1) the data source used, (2) the model validated, (3) the performance measure used; and for articles that created an original model, (4) the modeling technique used, and (5) the experimental design for validation.

## Results

The PubMed search described above returned 1937 abstracts. The two reviewers narrowed the list to 49 abstracts for which the full-text was retrieved. After review of the full-text articles, 22 fit the inclusion criteria [13-34]. The relevant articles are summarized in table 2. Among these articles, 10 focused on validating existing models that had been created for other purposes, and the remaining 12 detailed the creation of original models. Three of the articles were published before 2000, and the remaining 19 articles were published between 2000-2005, as shown in figure 3.

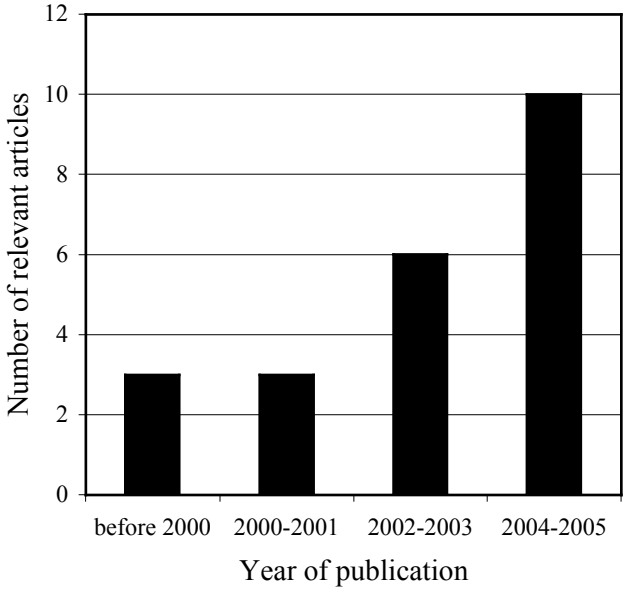



Figure 3. Relevant articles by year of publication. Recent years have seen a growing number of articles published on the use of models to predict survival after liver transplantation.

*Data Sources*

Eight of the relevant articles used a large multi-center data source, and 15 used data from a single transplanting institution. One study created a model using data from a single center and

Table 2. Summary of the literature on liver transplant survival modeling

| Ref # | Author | Year | Data Source | n | Model Validated | Technique | Validation | Performance Measure |
|---|---|---|---|---|---|---|---|---|
| 13 | Shaw | 1985 | Pittsburgh | 118 | original | “trial and error” | same | “separation” |
| 14 | Maggi | 1993 | Milan | 95 | Child-Pugh class | | | simple association |
| 15 | Selberg | 1997 | Hannover | 150 | original | “risk profile” | independent | simple association |
| 16 | Chung | 2000 | Vancouver | 31 | activation, SAPS, APACHE | | | simple association |
| 17 | Adam | 2000 | Europe | 11968 | original | product-limit | same | aggregate mortality |
| 18 | Parmanto | 2001 | Pittsburgh | 293 | original | neural network | independent | operating characteristics |
| 19 | Ghobrial | 2002 | UNOS | 25772 | original | Cox regression | same | discriminatory power |
| 20 | Onaca | 2003 | Baylor Dallas | 669 | MELD | | | simple association |
| 21 | Fernandez-Aguilar | 2003 | Carlos Haya | 185 | MELD, CTP score | | | simple association |
| 22 | Saab | 2003 | UCLA | 404 | MELD | | | simple association |
| 23 | Thuluvath | 2003 | UNOS | 38876 | original | logistic regression | independent | discriminatory power |
| 24 | Bilbao | 2003 | Barcelona | 197 | original | logistic regression | independent | operating characteristics |
| 25 | Desai | 2004 | UNOS | 8983 | original, MELD | Cox regression | independent | discriminatory power |
| 26 | Santori | 2004 | Genoa | 79 | UNOS status, MELD | | | discriminatory power |
| 27 | Austin | 2004 | Oregon HSU | 434 | original | Cox regression | same | discriminatory power |
| 28 | Roberts | 2004 | UNOS | 17044 | original | Cox regression | same | aggregate mortality |
| 29 | Jacob | 2004 | UK/Ireland | 3838 | MELD | | | discriminatory power |
| 30 | Bazarah | 2004 | Halifax | 228 | MELD, CTP score, CanWAIT | | | discriminatory power |
| 31 | Northup | 2004 | UNOS | 1510 | MELD, delta-MELD | | | simple association |
| 32 | Santori | 2005 | Genoa | 69 | UNOS status, MELD | | | discriminatory power |
| 33 | Haydon | 2005 | Queen Elizabeth | 827 | original | neural network | independent | simple association |
| 34 | Moore | 2005 | Vanderbilt | 483 | original | Cox regression | | |

then validated it using data from a multi-center database. The median sample size used, in number of transplant events, was 228 (range: 31-827) for the single-center studies and 11,878 (range: 1,510-38,876) for the multi-center studies. Five of the multi-center studies used data from the UNOS, two used data from the UK and Ireland Liver Transplant Audit, and one used data from the European Liver Transplant Registry.

*Models*

Existing models that were validated by a number of articles included MELD (9 articles), Child-Turcotte-Pugh (CTP) score (2 articles), and UNOS status (2 articles). Other existing indexes were each validated for survival prediction in one article and included Acute Physiology and Chronic Health Evaluation (APACHE) II/III, Canadian Waitlisting Algorithm in Transplantation (CanWAIT), Child-Pugh class, Simplified Acute Physiology Score (SAPS), delta-MELD, and activation status.

Among the articles that dealt with the creation of original models, various techniques of statistical modeling were used, including Cox proportional hazards regression (5 articles), logistic regression (2 articles), neural networks (2 articles), product-limit estimate (1 article), non-specific risk profile (1 article), and "trial-and-error" (1 article).

*Validation*

The relevant studies used various methods of model validation. Among the studies that created original models, five used the same data set for model derivation and validation, six used an independent data set for model validation, and one performed no validation. The validation statistics employed included measures of simple association (8 articles), discriminatory power (8 articles) including area under the receiver operating characteristic curve (AUC) [35], operating characteristics (2 articles), aggregate mortality (2 articles), and qualitative "separation" (1 article).

*Synopses*

Brief synopses of the relevant articles follow, in chronological order of publication. The first model to predict survival after liver transplantation was reported by Shaw et al. in 1985 [13]. They used a cohort of patients transplanted at Pittsburgh to identify key risk factors and develop an empirical scoring system using "trial-and-error". They assessed the prognostic value of their model for 6-month survival by qualitatively assessing the "separation" attained.

Maggi et al. established the predictive value of the traditional Child-Pugh classification with a set of patients transplanted in Milan, Italy [14]. They found no significant difference in survival between patients of the three Child-Pugh classifications A, B, and C at an alpha level of 0.01, and concluded that no definite prognostic value of the Child-Pugh classification could yet be assigned.

Selberg et al. assessed the ability of nutritional and metabolic parameters to predict outcomes of liver transplantation [15]. They studied transplant candidates at the Medical School of Hannover, Germany and noted that resting energy expenditure and body cell mass could be used as a "risk profile" that showed a significant difference in terms of Kaplan-Meier survival curves.

Chung et al. correlated pre-transplant activation status, SAPS, and APACHE II/III with post-transplant survival [16]. Activation status was defined as a four-category ordinal variable that denoted whether a patient was at home, in the hospital, in intensive care, or being mechanically ventilated at the time an organ was offered. They studied 31 patients transplanted at Vancouver Hospital in Canada, and found no significant correlation between any of the risk scores and survival. They concluded that detailed physiological scoring systems were of no greater value in predicting outcome than activation status.

Adam et al. used patients from the European Liver Transplant Registry to develop two original models [17]. Their adult model is described below, and their pediatric model was not relevant to the present study. They selected variables using Cox regression and derived the model using the product-limit estimate. The model included 11 variables and 8 two-way interactions

between variables. They concluded that the "normalized intrinsic mortality risk" predicted by the model was similar to that seen in the whole population.

Parmanto and Doyle used recurrent neural networks trained with the backpropagation-through-time algorithm to predict 90-day graft survival for adults transplanted at Pittsburgh [18]. They validated the model by 6-fold cross validation, and reported model performance in terms of "total performance", defined as (sensitivity + specificity) / 2. They developed a series of models beginning at the time of transplantation, and sequentially including 1, 2, 3, 4, 5, and 6 post-operative days. They showed that model performance improves with each day of post-operative information provided, and nearly perfect classifications, in terms of the total performance metric, may be made with 6 days of post-operative information.

Ghobrial et al. expanded on a model that had been previously developed to predict survival in hepatitis C patients [19]. They used the UNOS database and Cox regression to develop an 8-factor model that included recipient age, donor age, creatinine, total bilirubin, prothrombin time, retransplantation, cold ischemia time, and warm ischemia time. They validated the model with the same data used to derive the model and reported c-statistics of 0.69, 0.68, and 0.67 for 3-month, 6-month, and 1-year patient survival, respectively.

Onaca et al. validated the use of the pre-transplant MELD score to predict mortality within 2 years after transplantation [20]. They studied patients who underwent transplantation at Baylor in Dallas, Texas. The patients were divided into strata of MELD < 15, MELD 15-24, and MELD > 24. They were also categorized by disease: hepatitis C, cholestatic liver disease, and non-cholestatic liver disease. They found significant differences in survival at 3, 6, 12, 18, and 24 months for patients in the high MELD strata.

Fernandez-Aguilar et al. examined the use of the MELD score and CTP score in predicting post-transplant survival [21]. They included adults receiving liver transplants at a single hospital in Spain. Using the chi-square test, a MELD score $\geq 18$ was not significantly associated with 1-year survival, but a CTP score $\geq 10$ was significant at an alpha level of 0.01. They concluded that the MELD score adds no prognostic advantage to the CTP score.

Saab et al. assessed the association of MELD with 1-year patient survival on a cohort of adult liver transplant recipients at the University of California in Los Angeles [22]. They used Cox regression for the purpose of identifying independent risk factors of mortality and demonstrated an independent association between MELD strata and outcome.

Thuluvath et al. developed "user-friendly" models to predict post-transplant survival [23]. They used a very large cohort of patients from the UNOS database and used 2/3 of them, randomly selected, to derive logistic regression models, and the remaining 1/3 to validate the models. They converted the regression coefficients of the resulting models into integers that could be added to obtain a "severity score" ranging from 1 to 30. This calculation was easy to perform by hand, hence their designation as a "user-friendly" model. Their models for 1-month and 1-year patient survival included age, body mass index, UNOS status, diagnosis, total bilirubin, and creatinine. Their model for 5-year patient survival included race in addition to the other predictors. They reported AUC values of 0.7, 0.7, and 0.63 for 1 month, 1 year, and 5 years, respectively.

Bilbao et al. reviewed the literature for predictors of post-transplant survival after liver transplantation [24]. They developed an original model using patients transplanted in Barcelona, Spain. They used logistic regression to predict 3-month patient survival using four variables: Child-Pugh class, renal insufficiency, need for cross-clamping, and malnutrition. The model operating characteristics were 75% sensitivity, 75% specificity, 30% positive predictive value, and 94.4% negative predictive value on the training set; and 80% sensitivity, 88% specificity, 61.5% positive predictive value, and 95.3% negative predictive value on an independent validation set.

Desai et al. used MELD and an original model to predict survival [25]. They used patients from the UNOS database to create a Cox regression model that included age, ventilation, dialysis, and retransplantation. They reported an AUC performance for 3-month survival of 0.54 (95% confidence interval [CI]: 0.50-0.59) for MELD. Their original model achieved AUC

performance of 0.65 (95% CI: 0.61-0.69) on the training set and 0.60 (95% CI: 0.58-0.63 on the validation set.

Santori et al. evaluated the use of UNOS status and MELD score for predicting patient survival at 3 months [26]. They examined patients who were transplanted in Genoa, Italy, and reported c-statistics of 0.524 (95% CI: 0.410-0.636) for UNOS status and 0.677 (95% CI: 0.527-0.762) for MELD. There was no significant difference between the c-statistics for UNOS status and MELD score.

Austin et al. examined whether there was a difference in mortality between veterans and non-veterans after liver transplantation [27]. They compared two cohorts of patients receiving liver transplants at Oregon Health & Science University, including 285 university patients and 149 veterans. Their initial analysis found that veteran status was not a significant predictor of outcome. Next they created a Cox proportional hazards model using the entire study cohort that included gender, donor age, recipient age, and MELD score. They also created an alternate model that avoided the politically charged variables of recipient age and gender, and this model included donor age, alcoholism, and MELD score. They validated the models using the study cohort for predicting 1-year graft survival and reported an AUC of 0.71 for the 4-variable model and 0.66 for the 3-variable model.

Roberts et al. created a library of disease-specific models to predict post-liver transplant survival [28]. They used a cohort of patients from the UNOS database and grouped patients into disease categories determined by the National Clinical Oversight Committee. They built Cox regression models for each of ten disease categories. The models showed that different variables were significant predictors for different diseases. They compared model-predicted Kaplan-Meier survival curves for a standardized patient with actual Kaplan-Meier survival curves and noted similarity between the two. They concluded that disease etiology is a stronger predictor of outcome than MELD score.

Jacob et al. assessed the ability of the MELD score to predict 90-day patient survival [29]. They used data from the UK and Ireland Liver Transplant Audit and evaluated predictive

ability of the model using the c-statistic. They also re-estimated the coefficients of MELD using Cox regression and validated the modified model. The c-statistic for MELD was 0.58 (95% CI: 0.54-0.61), and this value remained less than 0.60 after re-estimating coefficients and considering non-linearity. They concluded that the MELD score might be appropriate to predict pre-transplant outcomes but not post-transplant outcomes.

Bazarah et al. compared the ability of the MELD score, the CTP score, and CanWAIT to predict post-transplant survival [30]. They used patients transplanted at the Queen Elizabeth II Health Sciences Hospital in Canada. They noted significant correlations between all three models' predictions and 90-day graft survival. They calculated c-statistics of 0.67 for MELD, 0.65 for CTP score, and 0.71 for CanWAIT. They concluded that there is no substantiated reason to consider replacing the currently used CanWAIT system for liver transplantation in Canada.

Northup et al. evaluated the utility of pre-transplant delta-MELD score as a predictor of post-transplant survival [31]. The delta-MELD score was defined as the change in a patient's MELD score during the 30 days prior to transplantation. They examined liver transplant recipients from the UNOS database. Using logistic regression, they concluded that absolute MELD score was significantly associated with 90-day survival, but delta-MELD was not.

Santori et al. compared UNOS status with the MELD score for predicting patient survival at 1 month, 3 months, 6 months, and 1 year; and graft survival at 1 week, 1 month, 3 months, 6, months, and 1 year [32]. This was similar in nature to their 2004 study described above, except that their study cohort was reduced and they examined additional endpoints. They assessed performance by AUC, and performance of both models at almost all endpoints was less than 0.6. They reported a trend of better performance for the MELD score, but this was not statistically significant.

Haydon et al. used self-organizing maps, a variant of neural networks, to predict transplant survival [33]. They studied patients transplanted at Queen Elizabeth Hospital in the United Kingdom. Their network contained 37 recipient and 18 donor variables, and it was trained to predict 3-month and 1-year patient survival. The model was validated using a national

multi-center database, and they showed that there was a significant difference by chi-square test in survival for the different output neurons.

Moore et al. simplified previous survival models that had been reported in the literature [34]. They studied patients that underwent liver transplantation at Vanderbilt University Medical Center. They used Cox regression to identify three recipient, donor, and technical factors as independent predictors of graft and patient survival: UNOS status 1/2A versus 2B/3, donor age ≥ 60 years, and cold ischemia time ≥ 12 hours.

## Discussion

The literature review included 22 articles published between 1985 and 2005. Recent years have seen a surge of interest in predicting survival after liver transplantation, possibly due to the increasing scarcity of liver grafts. Great diversity exists among the articles discussed in terms of data sources and methodology employed. Survival modeling is well developed in the literature using Cox regression and logistic regression. A variety of machine learning techniques exist for developing statistical models, yet only two papers have used neural networks [18,33], and no papers have reported the use of other established techniques. Despite the efforts that researchers have made, no well-validated survival model currently exists with discriminatory power greater than 0.7 AUC.

One limitation of the present review is that I did not judge the quality of the studies discussed. This review instead focused on a historical summary of methodology, because a review has already been reported that incorporated qualitative judgment of each article [36]. Jacob et al. performed a systematic literature review similar to ours, except that after identifying the body of relevant literature, they used an assessment tool to judge the quality of the studies. The tool was based on published instruments and was modified for the purpose of their review [37]. Using the tool, they found that five articles fulfilled the instrument's quality criteria to be included in the review [17,19,23-25]. Of the models in these articles, three were based in the

United States and therefore can be easily re-evaluated using the UNOS data set [19,23,25]. These three models will be used later in this research for the purpose of comparative validation.

My research will represent an attempt to improve upon past research in two specific ways. First, all methods of statistical modeling make certain assumptions about the data or the underlying process that generated them, and models perform optimally when those assumptions hold true. The Cox assumptions are not stringent, and the successful use of such techniques is widespread in the medical literature. However, the computational approaches germane to informatics allow for great flexibility and power in selecting model hypotheses, possibly resulting in improved performance. Thus, freedom from the assumptions inherent in Cox regression may allow for the construction of improved models for post-transplant outcomes using alternative techniques. Second, careful validation of a survival model is essential if conclusions are to be made about its generalizable performance. Not all models proposed in the literature were externally validated, and this may lead to overly optimistic conclusions about model performance.

## Conclusion

Various pre-existing and original models have been employed to try to predict survival after liver transplantation. However, there may be room for improvement in the performance of these proposed models [36]. The use of established techniques from machine learning has not been thoroughly explored in this context. Furthermore, the validation techniques employed for some published models may limit the conclusions that may be made about their generalizable performance.

# CHAPTER III

# GENERAL METHODOLOGY

## The UNOS Database

The UNOS operates the national system for organ transplantation. As mandated by policy, all transplanting institutions must report certain information for each transplant performed. The UNOS Liver Committee selects the relevant set of variables to be reported, which are collected on standardized forms made available by the UNOS. The UNOS makes the information publicly available in a de-identified electronic format. The UNOS database was selected for my research, which was approved by the local Institutional Review Board (IRB #040419: "Liver Transplant Outcomes: Bayesian Analysis and Other Artificial Intelligence Techniques").

The provided distribution of the UNOS database contains 372 fields for 62,676 transplants, representing every liver transplant performed in the country since the mid-1980's. The precise beginning date is unknown, because all of the dates contained are consistently shifted by an unknown amount, plus or minus six months. The UNOS database is large and comprehensive; however, certain aspects made it challenging to work with. As common for large-scale databases, the included elements need to be examined for errors and inconsistencies. For one patient, the wrong scale appears to have been used when the height was reported to be 1.85 centimeters, rather than 1.85 meters. Another patient was reported to have a survival time of negative 91 days. Moreover, many fields in the database contain varying proportions of missing values, and the reason for this is unknown.

Moreover, there is a selection bias inherent in the UNOS database, because each patient is selected for transplantation not randomly, but according to a specific allocation policy. Those candidates deemed by clinicians to be too sick for transplantation are removed from the waiting list, and thus are not represented in the set of transplant events. Despite some limitations, the

UNOS database was considered to be the best available information source for my experiments, due to its large sample size compared with single-center databases.

*Database Scrubbing*

All variables were inspected and cleaned prior to analysis using a Perl script. Every data field was verified to contain interpretable information. In other words, each field in the database was verified to contain a number where expected; a string of the form MM/DD/YYYY where a date was expected; or a string of the form HH:MM:SS where a time was expected. Many fields contained meaningless information; for example, many binary yes/no variables were marked 'U' or 'Uncertain'. Similarly, some serological tests were marked as 'Indeterminate' or 'Cannot Disclose'. Ambiguous or meaningless values were treated as if they were missing.

Discrete variables with many categories were abstracted into fewer categories in a clinically justifiable fashion, thus allowing for larger sample sizes within each category. For example, the home state of the donor, which was represented by more than 50 categories, was consolidated into 11 transplant regions defined by the UNOS. Range checking was performed for all continuous variables to verify that they lay within a reasonable range, determined by clinical experts, and outliers were removed and as missing values. As an example, all ages were ensured to be non-negative values. Complete details of the database cleaning process can be found in appendix A.

## Overview of Modeling Techniques

The following is a brief survey of major statistical modeling techniques, each of which has its own strengths and weaknesses. As discussed previously, many studies in the literature proposed Cox regression models to predict liver transplant survival [19,25,27,28,34]. The use of machine learning modeling techniques, some of which have different or more relaxed assumptions than Cox regression, or which do not require *a priori* preference for some functional form, may result in improved performance for the prediction of liver transplant outcomes. A

variety of modeling techniques exist in the field of machine learning, and these have been successful in a variety of tasks ranging from stock market prediction to facial recognition [38-39]. Machine learning methods have also been applied successfully to some difficult problems in medicine, such as classification and prognosis of cancer for example [40].

*Cox Proportional Hazards Regression*

The Cox proportional hazards model has experienced widespread application in predicting survival. Three central assumptions exist in this regression technique [41]:

The linearity assumption states that the relationship between a predictor and the outcome takes a linear functional form. For variables that have a skewed distribution, as laboratory values commonly do, the axis may be transformed by a square root or logarithmic function so that the transformed variable may adhere better to the linearity assumption.

The additivity assumption states that the total effect of different predictors may be estimated simply by summing the individual effects. In cases where a more complex variable interaction is believed to exist, the experimenter may create a composite variable by joining two individual variables together in a single term.

The proportional hazards assumption states that the impact of each predictor on survival does not change over time. Extensions to the Cox model exist, such as the use of time-dependent covariates, to allow for situations where this assumption does not hold [42].

In summary, the Cox assumptions are not rigid, and some methods exist to account for situations when they are violated. These methods tend to require *a priori* preference for some functional form by the experimenter.

*Decision Trees*

A decision tree is a schematic of sequential decisions that branch to arrive at some outcome [43-44]. A decision tree can be easily represented by a flowchart, in which each node involves testing some condition, and the path of questions followed subsequently depends on the

answer to each question before it, until some conclusion node is reached. This series of steps allows a decision tree to make inferences based on data. A variety of algorithms exist for constructing a decision tree from raw data, and for the most part decision trees tend to generate models in accordance with Occam's Razor, preferring simpler explanations over more complex ones.[1]

Decision trees have the advantage that they lead to human-interpretable models. However, they also have a tendency to generate models that over-fit a given data set, although pruning mechanisms may be used to simplify a decision tree to combat this shortcoming. Moreover, decision trees assume that the data set can be appropriately split at every node by considering only one data feature. This assumption does not hold for many complex problems.

*Bayesian Models*

Bayesian reasoning takes a probabilistic approach to model induction. It is founded on the principle that, given a set of attributes together with prior conditional probabilities, Bayes' theorem is used to assign a certain probability to various hypotheses [45]. Bayesian methods tend to operate more efficiently if the input data are discrete. Many techniques exist to convert continuous variables into discrete ones, and the choice of discretization technique may have a large impact on model performance [46]. The Bayesian approach to modeling is tolerant to missing data values. Different flavors of Bayesian modeling exist: the naïve Bayesian classifier, the optimal Bayesian classifier, and Bayesian networks [43].

A naïve Bayesian classifier assumes that various outcomes are mutually exclusive, and that the effects of all predictors are conditionally independent of all other predictors. While these assumptions seem relatively stringent, the naïve Bayesian classifier has been shown to perform very well in many circumstances [47].

---

[1] William of Occam, a 14th century English friar, is generally credited for the principle, "Entia non sunt multiplicanda praeter necessitatem." Loosely interpreted, this means that when facing multiple explanations, all other things being equal, the simpler one should be preferred.

The optimal Bayesian classifier relaxes the assumptions of the naïve Bayesian classifier, but its need for a huge table of conditional probabilities means that it may only achieve optimality with a nearly infinite sample size, and with nearly infinite computational power to estimate the probabilities. Because of this, the optimal Bayesian classifier is impractical to use for many complex problems [48].

A Bayesian network represents a structure of statistical associations between variables, so it creates a human-interpretable directed acyclic graph of relationships. A conditional probability table is populated for each node in the network. A Bayesian network assumes that the data are generated from a faithful distribution, which means that there is a Bayesian network structure that can describe the joint distribution [49,50].

*k-Nearest Neighbors*

The k-nearest neighbors algorithm takes a set of known data points, and it makes a decision about an unknown data point using the *k* data points closest to the unknown point [43,51]. It operates under the assumption that the appropriate prediction for a given data point is denoted by the outcomes of data points "nearest" to it. In this regard, it has some common ground with clustering techniques. A variety of distance measures may be employed. The most common approach is to plot all data points in hyperspace and measure the Euclidean distance between them to determine their similarity.

This assumes that every dimension in the set of features is equally important in making a prediction for novel data. This is known as the "curse of dimensionality" [52], and is considered to be a major weakness in k-nearest neighbors modeling. To compensate for the curse of dimensionality, some extensions have been proposed that involve weighting individual axes according to their impact in predicting the outcome. Moreover, the k-nearest neighbors algorithm assumes that the scale of every axis in the data is identical, and this assumption may be satisfied by normalizing and scaling each variable. Lastly, this technique does not tolerate missing data, so either missing fields must be imputed, or data points with missing variables must be eliminated.

*Neural Networks*

Neural networks are inspired by the biological process of neurons in the brain [43,53]. A given neuron fires a signal when it is stimulated above some threshold. Its output may be sent to other neurons, where it has a stimulatory or inhibitory effect on each of the neurons that it innervates. Collectively, a network of such neurons may exhibit complex reactions to different stimuli. The presentation of various stimuli over time, together with a feedback mechanism, results in those neurons being trained for certain responses. The computational analog of this process can result in the creation of highly complex but powerful models. Artificial neurons may be combined in multiple layers or other structures as desired for some learning task.

The chief assumption in neural networks is that of reinforcement learning – repeated presentation of some stimulus to a neuron, as manifest by patterns in a data set, will result in the pattern becoming recognizable to the network [54]. Otherwise, neural networks are relatively free from assumptions about the functional form of a model. A network with at least two layers and adequate neurons within each layer is capable of learning any arbitrary function.

The power of neural networks carries with it certain disadvantages – first, neural networks are very prone to over-fitting; and second, the model becomes a "black box" that shows predictive power without revealing useful information about the underlying process [55]. Missing data may be either imputed or encoded using dummy input neurons for each variable. Moreover, multiple local optima may be encountered in the training process, potentially preventing the model from achieving a globally optimal state. Finally, training a neural network may be extremely intensive from a computational standpoint.

*Support Vector Machines*

Support vector machines are a relatively novel and very popular development in the machine learning community [56,57]. They have achieved success in many modeling tasks, and they rectify some shortcomings of neural networks. A support vector machine plots data points in multi-dimensional space and draws a separating hyperplane to distinguish between various

outcomes. This space need not be linear, and various functional forms may be modeled through the use of kernel functions. The selection of a kernel function represents an *a priori* preference for some functional form, although these functions are quite versatile and hence not very restrictive. Kernel functions implicitly map each data point into a transformed, higher-dimensional space. This mapping is implicit because, due to the use of mathematical identities, the precise location of each data point is not calculated – only the dot product of two vectors in the transformed space must be known.

The support vector machine formulation has a unique optimum, and training always converges at this optimum. Due to the implicit transformation of input data, this optimum may be solved in a computationally efficient manner. Moreover, support vector machines perform relatively well in the context of noisy data. However, like neural networks they have the disadvantage of being a "black box" – a model may perform very well while revealing little explanatory information about the process that it is modeling. Also similar to neural networks, missing values must be either imputed or encoded using dummy variables.

## Experimental Design

Statistical modeling experiments must be carefully designed in order to make conclusions about the generalizability of resulting models. When a model is derived using one data set and validated using that same data set, valid conclusions may be drawn about how well the model fits that data set; i.e., the experimenters may conclude that "our model fits *these* data with *x* predictive ability." However, no conclusions may be drawn regarding the generalizability of the model; that is, how well the model can make predictions based on unforeseen data points. When the experimenter wishes to assert that "our model fits *other* data with *x* predictive ability" – as is often the case in research involving statistical modeling – different approaches must be used.

Two experimental designs that can be employed to reach conclusions about model generalizability are cross-validation and independent validation [58]. Cross-validation refers to the process of taking a single set of data and randomly dividing it into a number of balanced

splits, with each containing a representative sample of the whole data set. The first split is designated the validation set, and all of the other splits are used as a training set. A model is derived using the training set, and its performance is estimated using the validation set. Next, the second split is designated as the validation set, and all other splits as the training set. The process is repeated once for each of the splits, so that each split serves as a validation set once during the experiment. The process of cross-validation can be further extended to include an inner optimization loop when parameters must be selected for a modeling technique. The inner loop involves using a nested experiment of cross-validation on each of the training sets employed, thus allowing the model parameters to be selected in a manner that is relatively free of bias. The final result of a cross-validated experiment is generally reported as an average of performance for all splits.

Independent validation refers to the process of generating a model using a training set, and then validating the model using an independent source of data. The data may come from a different population, a different time, or a different geographical location. For example, a model to predict survival after liver transplantation could be derived using data from the United States and subsequently validated using data from Europe. This would allow the researcher to conclude that the observed model performance does not depend on a specific process that is unique to the United States. Consequently, there would be reason to believe that the model will perform similarly when presented with other previously unseen data.

# CHAPTER IV

# BAYESIAN NETWORK

## Purpose

I am unaware of previous attempts to use Bayesian networks to model outcomes following liver transplantation. The purpose of the present experiment was to evaluate the feasibility and suitability of Bayesian networks for predicting 90-day graft survival. I hypothesize that a Bayesian network can be created using expert knowledge in the field of liver transplantation to model post-transplant survival, and that this Bayesian network will have better discriminatory power than other survival models in the literature.

## Methods

This study included all liver transplant events in the UNOS database occurring between 2000-2002, and which involved an adult ($\geq$ 18 years) recipient and an adult donor (n = 12,239). Some patients received multiple liver grafts, and thus may be represented in the data by more than one transplant event. The binary outcome variable of interest was 90-day graft survival, where the endpoint of graft survival is defined as death or retransplantation. Transplant events were eliminated if the patient was lost to follow-up prior to 90 days (n = 132). A set of 258 variables in the database were identified as being available by the time of transplantation, so these were used as independent variables in the pool of candidate predictors for survival. These factors included demographic information, such as recipient and donor age, gender, and race; clinical information, such as laboratory values, medical condition, and functional status; and technical information, such as cold and warm ischemia time.

As described briefly in the general methodology (Chapter III), a Bayesian network describes a system of interest by specifying relationships of conditional dependence between its variables [43,49,50]. These relationships are represented by a directed acyclic graph, and this,

together with a joint probability distribution for the nodes it contains, creates a model that can be used for making inferences. Bayesian networks are being used for a variety of medical problems that involve reasoning with uncertainty. For example, Heckerman et al. developed a system to aid in diagnosis of diseases of the lymph nodes [59]. Domain knowledge drawn from clinical expertise and relevant literature was used to select key predictors of survival from the pool of available pre-transplant variables, and subsequently to construct the relationships between nodes in the network [24,60,61].

All data processing and performance analysis was done with Matlab (version 6.5, http://www.mathworks.com), while the Bayesian network was created and simulated using the Netica development environment (version 1.12, http://www.norsys.com). Different network structures were examined to determine the appropriate degree of model complexity. Both continuous and discrete variables were represented in the network, and all continuous variables were discretized into seven equal-width intervals. The Netica software tolerates missing values in data, so no imputation was performed.

I employed a three-fold cross-validated experimental design. All data points from 2000-2001 were randomly split into three stratified folds. Three iterations of model training and validation were conducted, each time estimating model parameters using two-thirds of the data and determining its performance using the remaining third. A final model was created using all data points from 2000 and 2001, and I validated its performance using an independent set consisting of the data from 2002.

Model performance was determined using the receiver operating characteristic (ROC) curve. This curve plots the sensitivity and 1 - specificity levels achieved across all possible thresholds in a binary classification task. The AUC represents the overall discrimination power of a given test, where a value of 0.5 denotes random guessing and 1.0 demonstrates perfect classification [35]. All AUC analyses was performed using ROCKIT (http://www-radiology.uchicago.edu/krl/KRL_ROC/software_index.htm). For the final independent validation

step, I calculated AUC as well as specificity and positive and negative predictive value for fixed sensitivity levels of 95% and 90%.

To compare the performance of my Bayesian network to other models published in the literature, I selected three Cox regression models from the literature [19,23,25]. These studies were used because they were identified as "high-quality" articles in a recent systematic literature review [36]. I ran the independent validation set from 2002 through these three models. Because these models are not tolerant to missing data, I imputed missing values in the independent validation set using close linear surrogates where possible.[2] For other variables no linear surrogate was available, so I used mean imputation to fill in the other missing values. The discriminatory power of these models was estimated in terms of AUC. I performed three pairwise comparisons of AUC between the Bayesian network and each of the other three models. The Bonferroni correction was applied to the significance level as appropriate, resulting in an alpha level of 0.0033.

Table 3. Demographics of the study population

| | 2000-2001 (n = 7,887) | 2002 (n = 4,220) |
|---|---|---|
| 90-Day Graft Survival* | | |
| Yes | 86.6% | 89.2% |
| No | 13.4% | 10.8% |
| Age | | |
| Mean ± SD (in years) | 50.8 ± 10.0 | 51.1 ± 9.7 |
| Gender* | | |
| Female | 34.7% | 31.2% |
| Male | 65.3% | 68.8% |
| Diagnosis* | | |
| Acute Hepatic Necrosis | 8.6% | 6.7% |
| Cholestatic | 11.0% | 10.8% |
| Malignant | 3.6% | 7.8% |
| Metabolic | 3.0% | 2.7% |
| Non-Cholestatic | 68.5% | 66.6% |
| Other | 5.3% | 5.4% |

Note: Data represent transplant events, so unique patients may be represented more than once. Asterisk denotes a significant difference between the populations from 2000-2001 and 2002, $p < 0.01$.

## Results

The descriptive statistics from some key variables in the data set were compared between the study populations from 2000-2001 and 2002, and they are summarized in table 3. The proportion of patients achieving 90-day graft survival improved between 2000-2001 and 2002. The age of patients receiving liver transplants was not significantly different between the two

[2] Specifically, when necessary, prothrombin time was estimated by multiplying the international normalized ratio by a factor of 12.

groups. Gender of liver transplant patients showed a small but statistically significant difference in favor of more males being transplanted in 2002. The diagnostic categories were mostly similar between the two study populations, excepting that the percentage of patients being transplanted for malignancy doubled from 2000-2001 to 2002. In both groups, non-cholestatic liver disease, including alcoholic liver disease and hepatitis C, accounted for a significant majority of liver transplants performed.

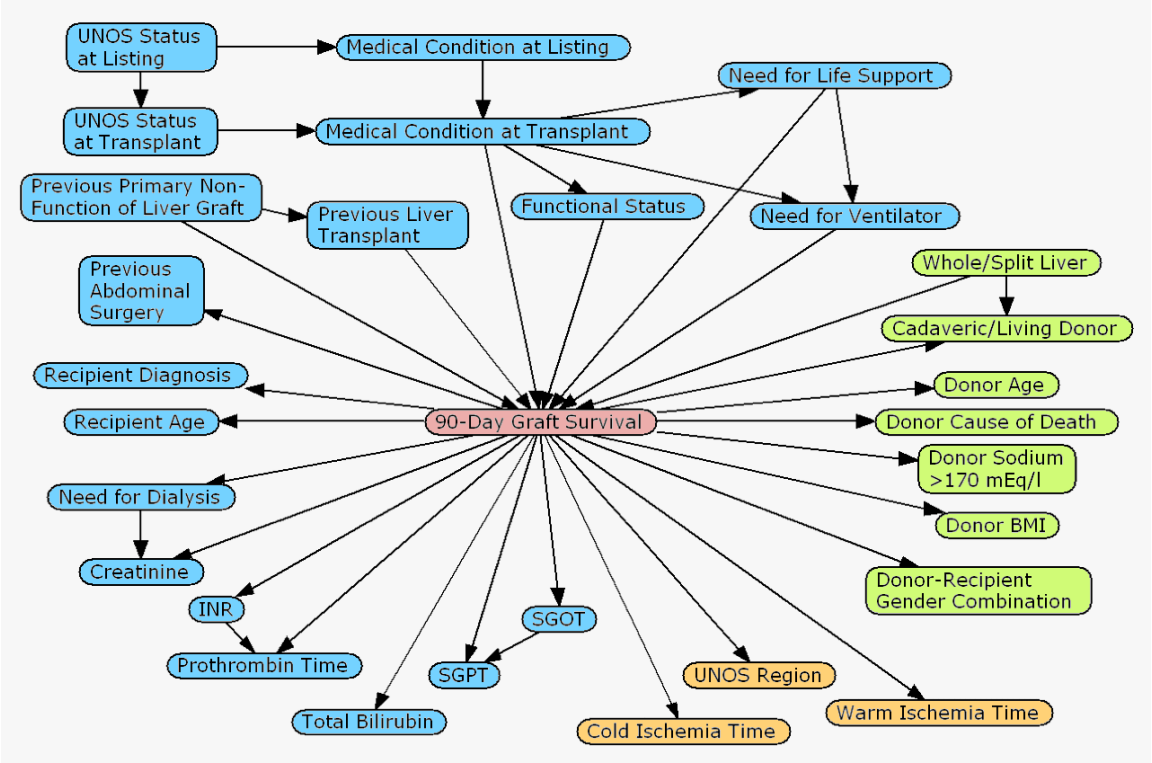
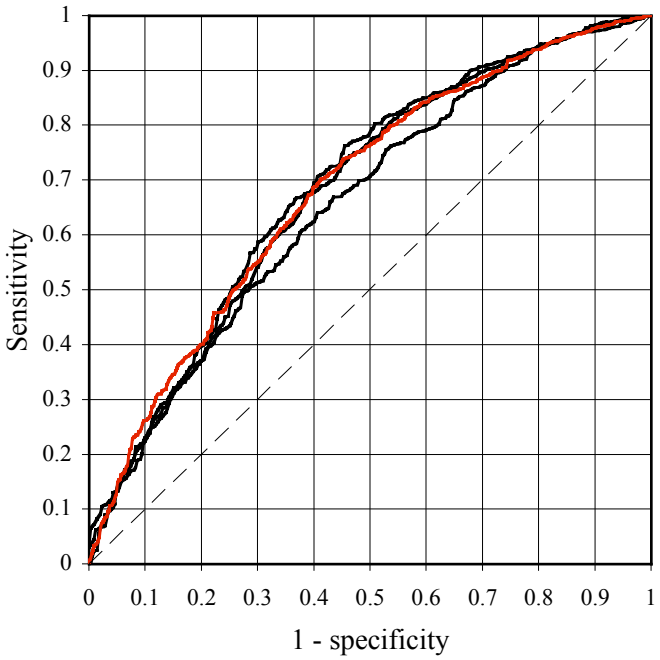



Figure 4. Structure of the final Bayesian network, created using domain knowledge from clinical expertise and literature review. The outcome node is shown in the center and is highlighted with pink. Recipient variables are shown in blue, donor variables are shown in green, and technical variables are shown in orange.

The Bayesian network consisted of 30 nodes, including 29 nodes representing pre-transplant variables and a single dichotomous outcome node for 90-day graft survival. Only seven of the most important predictors of survival were designated as parents of the outcome node, ensuring that its conditional probability table would not grow too large and contain many structural zeros. The other nodes in the network were connected either as grandparents or as children of the outcome node, and the relationships between different predictor nodes were




| Area under curve (95% CI) | |
|---|---|
| Split 1 | 0.659 (0.628-0.689) |
| Split 2 | 0.681 (0.650-0.711) |
| Split 3 | 0.682 (0.651-0.712) |
| Independent | 0.681 (0.654-0.707) |

Figure 5. Receiver operating characteristic curves of Bayesian network performance. Area under curve values are shown for the three cross-validated splits using the training set from 2000-2001, as well as for the independent validation set from 2002, shown in red.

founded on clinical rationale. A total of 39 links connected the network, and the joint probability distribution across all nodes consisted of 2,049 conditional probabilities. The final network structure is shown in figure 4.

The mean performance of the three cross-validated folds as measured by AUC was 0.674, and the AUC for the independent validation set was 0.681 (95% CI: 0.654-0.707). The ROC curves are shown superimposed in figure 5. Using a threshold that fixed the sensitivity level at 95%, the model specificity was 18%, with a positive predictive value of 91% and a negative predictive value of 30%. These operating characteristics are

Table 4. Bayesian network classification at fixed 95% sensitivity

| | | Actual outcome | |
|---|---|---|---|
| | | Survivor | Non-survivor |
| Predicted outcome | Survivor | 3578 | 373 |
| | Non-survivor | 188 | 81 |

| | |
|---|---|
| Sensitivity: | 95% (fixed) |
| Specificity: | 18% |
| Positive Predictive Value: | 91% |
| Negative Predictive Value: | 30% |
| Likelihood Ratio (+): | 1.16 |
| Likelihood Ratio (–): | 0.28 |

summarized in table 4. When the sensitivity level was fixed at 90%, the resulting specificity was 27%, and the positive predictive value remained at 91% while the negative predictive value was 24%.

The AUC performance for predicting 90-day graft survival was 0.617 (95% CI: 0.588-0.646) for the Desai model, 0.616 (95% CI: 0.588-0.643) for the Ghobrial model, and 0.605 (95% CI: 0.576-0.633) for the Thuluvath model. Multiple pairwise comparisons of the alpha level between my model and each of the three others showed that the Bayesian network had significantly greater discriminatory power than the others ($p < 0.0033$). The ROC curves from the four models on the independent validation set are shown in figure 6.

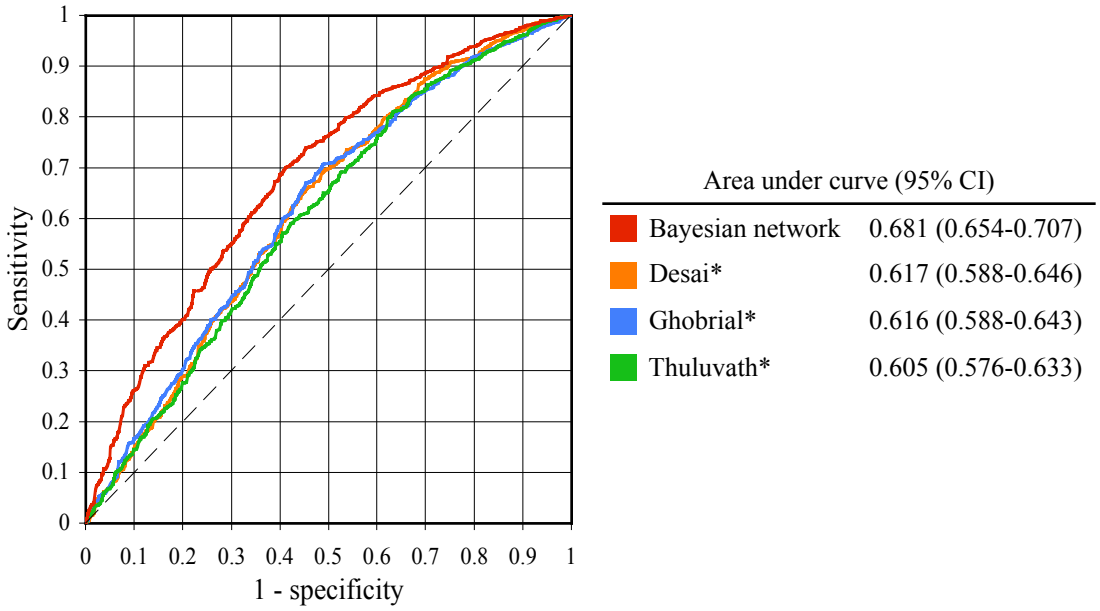



Figure 6. Receiver operating characteristic curves of comparative model performance. The Bayesian network and three models from the literature were validated using the same independent data set from 2002. Asterisks denote statistical significance in a pairwise comparison of AUC versus the Bayesian network, $p < 0.0033$.

## Discussion

A Bayesian network model may be constructed using domain expertise to predict outcomes following liver transplantation. My network for predicting 90-day graft survival performed modestly well, according to the AUC. Performance was found to be highly consistent

between the training and independent validation sets. Model performance was also examined using a fixed threshold that yielded a sensitivity level of 95%, and the resulting specificity level was low. The model showed a very good positive predictive value (91%), whereas the negative predictive value was relatively low. I interpreted this to mean that the model performs very well at the task of identifying good survivors. However, the model does not perform well at recognizing poor survival candidates for liver transplantation. One limitation of the study is that 132 transplant events were eliminated from analysis due to inadequate follow-up information; it is plausible that many of these represented poor outcomes, so a bias may have been introduced to the research.

Of noteworthy mention is the fact that the UNOS instituted a major allocation policy change in February, 2002, by switching from a ranking of four sickness categories – Status 1, 2A, 2B, and 3 – to using a continuous scale called MELD [6,7]. This policy change closely coincided with the division between the training set from 2000-2001 and the independent validation set from 2002. This observation may explain why differences were seen in certain descriptive statistics as noted above. Regardless, since performance was very consistent between both study groups despite the policy change, this suggests that the Bayesian network model may generalize well.

The performance of the Bayesian network measured by AUC was approximately 5-10% higher than previous Cox regression models in the literature, and this difference was statistically significant [19,23,25]. Furthermore, because I used a data set that was independent with respect to all four models, I was able to perform a well-controlled comparative validation to show that the Bayesian network outperforms the others in predicting 90-day graft survival. This conclusion is worded carefully and is not intended to impugn the quality of the other models or the studies that created them.

The present research sheds new light on the field of modeling liver transplant outcomes in a few ways. First, to my knowledge Bayesian networks have not previously been applied to the problem of modeling liver transplant outcomes. The use of this technique may better lend

itself to high model complexity than Cox regression does, as evidenced by the 29 predictors included in the Bayesian network model. Another benefit to this modeling technique, as compared with Cox regression, is that Bayesian analysis is tolerant to missing values such that no imputation is necessary. Moreover, by examining the positive and negative predictive value of the model, I found that the difficult part of the problem lies not in identifying good survivors, but in identifying poor survivors. Lastly, my model incorporates some interactions between important predictors, and while it is possible to explicitly specify variable interactions in a Cox regression model, none of the models used for comparison included them.

Future research will focus on constructing a Bayesian network that incorporates both pre-transplant information as well as some early post-transplant information. The model can be trained with joint probability distribution governing all nodes in the network, even if inferences are made by entering findings only into pre-transplant nodes. The presence of this additional information in the network may help improve predictive power.

## Conclusion

Bayesian networks may be used to model graft survival following liver transplantation with fair performance. This model exhibited greater discriminatory power than three other high-quality models from the literature. The Bayesian network is a candidate for further validation using foreign liver transplant data. Given its high positive predictive value, our model may serve as a useful adjunct to clinical judgment in identifying patients most likely to have good outcomes.

# CHAPTER V

# VARIABLE SELECTION

## Purpose

Selecting variables for predicting liver transplant survival has traditionally relied upon clinical judgment for *a priori* selection of variables, followed by statistical treatment of this subset to empirically identify the most important factors. Computational techniques for feature selection may be used to obtain the minimal set of variables required for optimal prediction, called the "Markov blanket" [62]. The goal of this study was to compare the performance of survival models created from factors chosen using traditional methods versus those selected by an automated technique that approximates the Markov blanket. I evaluated the performance of these predictor sets using several different types of machine learning models.

## Methods

Four different sets of predictors from the UNOS liver transplant database were used to model 90-day graft survival in adult (> 18 years) liver transplant recipients. The variables identified by Desai [25], Ghobrial [19], and Thuluvath [23] were used as baseline predictor sets, derived via well-established methods of univariate and multivariate association.

Among all of the pre-transplant variables in the UNOS database, 148 of them were at least 80% populated for the study period from 1995-1999. Variables with a greater fraction of missing information were deemed to have insufficient data for further analysis. These variables were available at the time of transplantation, and any of them could be potential predictors for post-transplant outcomes.

The automated feature selection technique used repeated tests of conditional independence with forward and backward conditioning to identify key features in the data set. An alpha level of 0.001 was used for all statistical tests in automated feature selection, both for entry

into and exit from the variable pool. This technique was based on the algorithm described by Aliferis et al., and I refer the reader to the original paper for more technical details [63]. My method differed from theirs in that I did not implement the final wrapper step specified in their algorithm. For the feature selection step, continuous variables were discretized by selecting thresholds that maximized the chi-squared association with the outcome variable. Once all variables were in the required discrete form, missing values were encoded as a separate category.

Various types of classifier models were built to predict 90-day graft survival; specifically linear regression, naïve Bayesian classifiers, support vector machines, and neural networks. To ensure that all required values were populated in the data set, I excluded data points that had values missing for any of the required variables in the four predictor sets. All of the feature selection and modeling experiments were conducted using the Matlab environment. (version 6.5, http://www.mathworks.com)

I employed a 3-fold cross-validated experimental design using the data set from 1995-1999 (n = 6,765). The data were randomly divided into balanced splits, and during three alternating iterations two of these splits were used to train a model, while the third was used to validate the model. I also created a final model using the entire data set from 1995-1999, and validated its performance using an independent data set of adults transplanted from 2000-2002 (n = 2,119). The AUC was used to assess model performance [35]. This metric quantifies discriminatory power, where 0.5 denotes random guessing and 1.0 denotes perfect discrimination. I repeated the cross-validation step five times to obtain an average AUC for each predictor set and model type. I used a t-test to evaluate the difference between the means between my predictor set and those published in the literature.

To evaluate the necessity of all predictors, a support vector machine was trained with random fractions of features removed from the predictor set. I repeated the above process of cross-validation five times with all predictors in the set; five times with 25% of the predictors randomly selected for removal; five times with 50% of the predictors randomly selected for

removal; and five times with 75% of the predictors randomly selected for removal. The results were visualized using a scatterplot.

## Results

The automated feature selection technique selected 21 survival predictors, cross-tabulated in table 5. All of these predictors were clinically justifiable, and some of them had not been used in the other models from the literature. The linear regression, naïve Bayes, and support vector machine models tended to show better performance for all predictor sets than the neural network,

Table 5. Comparison of four predictor sets for liver transplant survival

| Predictor | Desai | Ghobrial | Thuluvath | Automated |
|---|---|---|---|---|
| Recipient Variables | | | | |
| Age, recipient | √ | √ | √ | √ |
| Diagnosis of liver disease | | | √ | |
| Serum creatinine | | √ | √ | √ |
| Total bilirubin | | √ | √ | |
| Prothrombin time | | √ | | √ |
| UNOS status, registration | | | | √ |
| UNOS status, transplantation | | | √ | √ |
| Medical condition, registration | | | | √ |
| Medical condition, transplantation | | | | √ |
| Dialysis, registration | | | | √ |
| Dialysis, transplantation | √ | | | √ |
| Life support, transplantation | | | | √ |
| On ventilator, transplantation | √ | | | √ |
| Previous abdominal surgery | | | | √ |
| Previous liver transplant | √ | √ | | √ |
| Previous primary non-function | | | | √ |
| Date of transplant referral | | | | √ |
| Days on waiting list | | | | √ |
| Employment status | | | | √ |
| Body mass index | | | √ | |
| Racial background | | | √ | |
| Donor Variables | | | | |
| Age, donor | | √ | | √ |
| Whole or split liver graft | | | | √ |
| Technical Variables | | | | |
| UNOS region | | | | √ |
| Cold ischemia time | | √ | | √ |
| Warm ischemia time | | √ | | |

as seen in table 6. For these three model types, my predictor set outperformed each of the three others by cross-validation ($p < 0.0001$). Similar trends were seen by prospective validation, except for the Ghobrial predictor set, which showed similar performance to my predictor set.

Table 6. Model performance with four different predictor sets

| Model Type | Desai | Ghobrial | Thuluvath | Automated |
|---|---|---|---|---|
| Logistic regression | 0.63* (0.62) | 0.65* (0.65) | 0.64* (0.63) | 0.67 (0.64) |
| Naïve Bayes | 0.62* (0.60) | 0.64* (0.61) | 0.63* (0.60) | 0.67 (0.63) |
| Neural network | 0.58 (0.56) | 0.64 (0.60) | 0.61 (0.59) | 0.63 (0.62) |
| Support vector | 0.63* (0.62) | 0.65* (0.65) | 0.63* (0.62) | 0.68 (0.65) |

Note: Performance is measured by AUC, with independent validation in parentheses. Asterisk denotes predictor sets that were outperformed by the automated set, $p < 0.0001$.

Moreover, when increasing numbers of variables were randomly removed from my predictor set, a gradual decline in support vector machine performance was seen. The variability of model discriminatory power also became greater as more variables were removed. These trends are shown on a scatterplot in figure 7.

## Discussion

Key predictive factors for liver transplant survival may be identified in a very large database using fully automated methods. My feature set was mostly a superset of the three other sets from the literature; i.e. it included nearly all of the features that the others used, but it also incorporated some additional variables. Validation showed that my technique led to the creation of models with statistically better performance than two out of three baseline data sets. The improved performance may in part be due to the fact that my predictor set contained more information than the other sets; it contained 21 predictors while the others had 8, 4, and 7 predictors. Moreover, this predictor set does not depend on a specific type of model to perform well, as different machine learning techniques all resulted in the creation of similarly performing models. Future experiments may further reduce the size of the predictor set while maintaining predictive power.

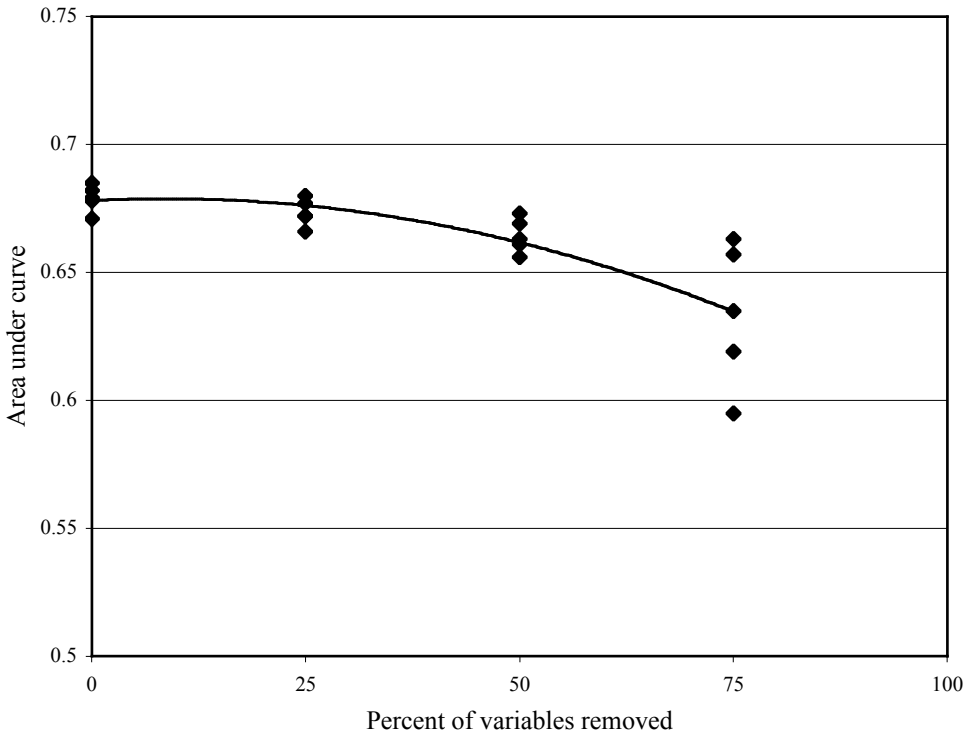



Figure 7. Impact of randomly deleting variables from the automated predictor set. All models in this experiment were support vector machines. Each data point represents the mean performance of one cross-validated experiment with a specified fraction of variables randomly removed. Discriminatory power was measured by the area under the receiver operating characteristic curve. A minor impact on model performance was observed when 25-50% of the variables were removed from the predictor set, and a greater impact on performance was observed when 75% were removed.

## Conclusion

Fully automated feature selection techniques may be used to identify the key predictors of liver transplant survival from a very large database, and this predictor set performed better than two out of three other predictor sets reported in the literature. Furthermore, various machine learning techniques may be used in conjunction with the UNOS database to model survival following liver transplantation.

# CHAPTER VI

# COMPLEXITY ANALYSIS

## Purpose

As described previously, support vector machines have been growing in popularity for machine learning applications due to their mathematical elegance and computational efficiency [57]. They can detect patterns in large data sets consisting of hundreds or thousands of variables. The support vector machine formulation is based upon an implicit mapping of variables to a higher-dimensional space through the use of kernel functions. When a polynomial function is used as the kernel function, all possible variable interactions up to an arbitrary complexity are implicitly considered for the model [56]. For example, a polynomial support vector machine of degree two would implicitly model all main effects and all possible combinations of two-variable interactions.

The goal of this experiment was to evaluate the ability of support vector machines to predict liver transplant outcomes using the entire set of pre-transplant variables available in the UNOS database. To my knowledge, no previously reported model has incorporated all available information in the UNOS database. It is possible that some variables available in the pre-transplant database have previously been overlooked, and they might contain useful information for modeling. A second goal of the experiment was to determine the extent to which complex variable interactions are important for predicting liver transplant survival.

## Methods

The population for this study was all adult (> 18 years) liver transplant recipients in the UNOS database from 1995-1999 (n = 15,747). All of the 258 available pre-transplant variables, described earlier, were included in the data set. Two copies of the data set were created, and they differed in the way missing values were handled. For one data set – called the missing-encoded

data set – all missing values were set to 0. For every variable in the data set, an additional dummy variable was created, and this dummy variable held the value 0 if the corresponding value was missing and 1 if it was populated. For the second data set, the missing-imputed data set, all missing values were filled in using the sample mean of all populated values for that variable. Both data sets were processed according to the technique described earlier in the general methodology chapter. Continuous values were normalized and scaled between 0 and 1, and discrete variables with more than two categories were distributed into multiple binary variables.

The GEMS software package was used for model creation and validation [64]. GEMS automates the process of optimizing and cross-validating a model through a graphical user interface. Using both the missing-encoded and missing-imputed data sets, I conducted a 3-fold cross-validated experiment, using an inner loop to optimize the support vector machine tolerance and kernel parameters. Next, I conducted another experiment using the missing-imputed data set. All support vector machine parameters were fixed, including tolerance, and I varied the degree of the polynomial kernel function. Ten models were created, one for every integer degree from 1 to 10. Discriminatory power for all generated models was measured using the area under the receiver operating characteristic curve, and the relationship between polynomial degree and AUC was visualized using a line graph [35].

## Results

The discriminatory power was similar for the missing-encoded (AUC = 0.666) and the missing-imputed (AUC = 0.661) data sets. When polynomial degree was varied and other parameters remained fixed, the AUC of the support vector machine increased from 0.605 to 0.654 as the degree was increased from 1 to 2. However, using polynomial kernel functions of degree higher than 2, the model discriminatory power remained virtually constant all the way up to degree 10, as seen in figure 8.

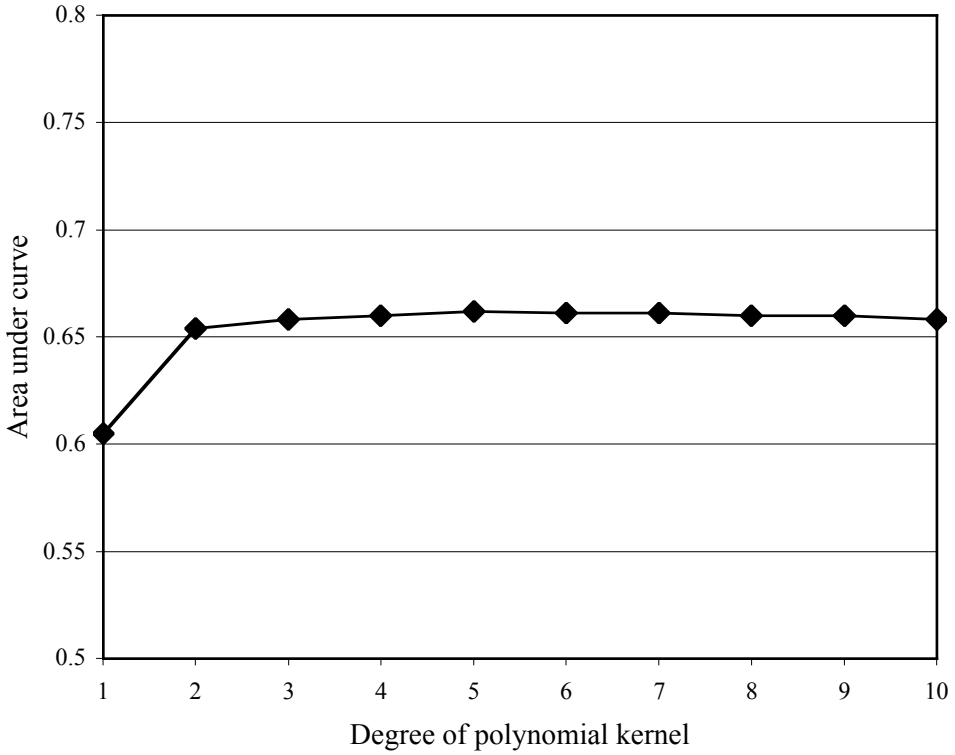



Figure 8. Effect of polynomial kernel degree on support vector machine performance. The degree of the kernel function determines the complexity of the space into which the input is mapped. It may be roughly interpreted as an unknown space in which all possible interactions up to some determined size are considered. Discriminatory power was measured by the area under the receiver operating characteristic curve. Note that the vertical axis does not range from 0 to 1, in order to allow easier visualization of results.

## Discussion

A survival model was constructed with the entire set of available pre-transplant information, using a support vector machine. Given that the missing-encoded and missing-imputed data sets performed similarly, the support vector machine does not seem to be learning patterns of reporting in the data. If the model performance on the missing-encoded data set had significantly exceeded model performance on the missing-imputed data set, questions would have been raised about whether the data alone, or the data combined with the reporting habits of transplant centers, was providing information about post-transplant survival.

The effect of degree of the polynomial kernel function on model performance was striking due to the sharp rise in performance between degrees 1 and 2, and the stable performance at all degrees greater than 2. This observation implies that little or no marginal benefit exists in considering interactions involving more than two variables. A limitation of the present study

exists in that the polynomial kernel function of degree 2 implicitly considers both main effects that are non-linear in shape, as well as two-way interactions between variables. Thus, one cannot conclude which of the two factors accounted for the rise in performance between degrees 1 and 2: whether two-way interactions, or non-linear functions of main effects. Another limitation of the experiment lies in the fact that the support vector machine remains a "black box" in the sense that these results do not provide information about which specific variables or interactions of variables are important for the model. However, the results illustrate that complex interactions involving three or more variables may not need to be considered for post-transplant survival modeling using the UNOS database.

## Conclusion

A support vector machine can be trained to predict post-transplant survival using the entire set of pre-transplant variables available in the UNOS database. Moreover, complex interactions involving more than two variables may not need to be taken into account for predicting liver transplant survival when using this database.

# CHAPTER VII

# CLINICIAN SURVEY

## Purpose

Clinical judgment is the current standard for making decisions about liver transplant allocation. This raises the question, how good are clinicians in predicting outcomes following liver transplantation? In other fields, studies have been published to assess the predictive ability of clinicians. In oncology, a systematic literature review showed that for terminally ill cancer patients, clinicians were found to overestimate survival consistently [65]. Despite this poor calibration, their predictions were highly correlated with actual survival. Thus, clinician predictions did have some discriminatory power. Critical care physicians were found to have difficulty predicting length of stay in ICU patients whose total length of stay was greater than five days [66]. However, for patients with a short length of stay in the ICU, physicians were able to predict the length of stay and outcome fairly accurately, with more experienced clinicians showing better performance than their less experienced colleagues.

I am aware of no previous studies that assessed the prognostic ability of clinicians in the field of liver transplantation. The purpose of the present study is to quantify the ability of expert clinicians to predict outcomes after liver transplantation, and to compare the clinicians' performance to published survival models.

## Methods

*Survey Participants*

The local Institutional Review Board approved this research (IRB #050329, "Assessing Clinical Utility of a Liver Transplant Survival Model"). I invited clinicians at different transplant centers specializing in the care of liver transplant patients to participate in a survey. The eligibility

criteria specified that each participant must be "an attending physician who is up-to-date in the care of liver transplant patients." All clinicians who agreed to participate were sent a packet containing the study instrument, together with a cover letter and an example item.

*Study Instrument*

Case reports were prepared based on 16 actual patients who received liver transplants in 2003-2004 at Vanderbilt University Medical Center. The patients were randomly selected and stratified to achieve a balance of good and poor outcomes. I randomly chose four patients who were healthy at last follow-up, with the original graft intact (range: 8-18 months after transplantation), four patients who experienced a graft failure after 90 days but before 2 years after transplantation, and eight patients who experienced a graft failure before 90 days.

The medical records for the 16 patients were condensed into half-page case reports, fully de-identified. Information was obtained from all available data sources, which consisted of (1) electronic medical records, (2) archived paper charts, (3) the local transplant database, and (4) UNOS donor and recipient information forms. Care was taken to include all pre-transplant information that could be relevant to the post-transplant prognosis. This included (1) basic demographic information, (2) history of the liver disease, (3) medical therapy provided, (4) past medical and surgical history, (5) family and social history, and (6) characteristics of the donor liver.

To ensure that the reports were readable and concise, only noteworthy information was included. Participants were told to assume that any findings not discussed were either within normal limits or otherwise unremarkable, given the patient's health status. For example, the finding of "mild encephalopathy" was not repeated in the case report, as this finding is very commonly seen in patients with liver failure. Severe findings like "stage IV hepatic coma," however, were always included in the reports.

For quality control, each case report was independently reviewed by the medical and surgical directors of the local liver transplant program, and by a surgery resident with experience

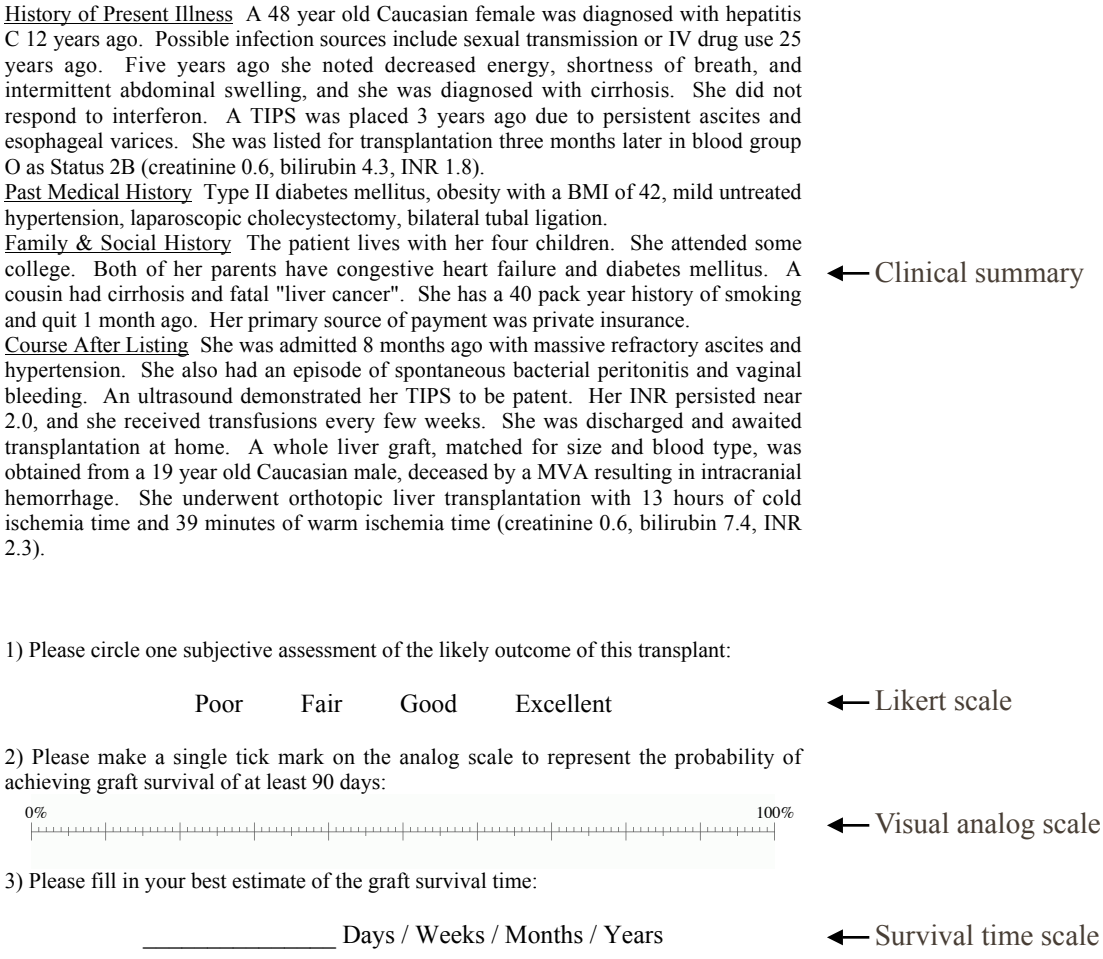



Figure 9. Example of a finished case report. All case reports were presented using the same structured format. Relevant clinical details from the patient's medical record were summarized and followed by three measurements for estimating the prognosis: a 4-point Likert scale, a visual analog scale, and a survival time scale. The survival time scale was not used for any subsequent analysis.

in liver transplant research. The three reviewers were asked to evaluate whether the information provided was (1) adequate for predicting outcomes, (2) consistent in scope among the reports, and (3) objectively described to avoid possible bias by the writer. An example of a case report is shown in figure 9.

Participants were asked to assess the likelihood of 90-day graft survival for each patient using (1) a 4-point Likert scale of 'poor', 'fair', 'good', or 'excellent'; and (2) a visual analog scale representing the probability of survival. The participants were also asked to provide their age, gender, clinical specialty, and years of experience in caring for liver transplant patients after the completion of specialty training.

*Statistical Analysis*

All statistical analysis was performed using the SPSS package (version 13, http:// www.spss.com). The mean and standard deviation were calculated for age and years of experience of the participating clinicians. The frequency was calculated for their gender and clinical specialty. For each clinician, the Pearson correlation was computed between predictions on the Likert scale and the visual analog scale. The distribution of clinician predictions on the visual analog scale for each patient was examined on a box-and-whisker plot and compared with the actual outcome of each case. A t-test was used to determine whether there was a difference in the coefficient of variation in predictions between transplants with good outcomes versus poor outcomes. The ability of each clinician in predicting 90-day graft survival was measured using the AUC, and the 95% confidence intervals were calculated [35]. The AUC is a measure of discriminatory power at all classification thresholds, where 0.5 denotes random guessing and 1.0 denotes perfect discrimination. The non-parametric Mann-Whitney test was used to determine whether there was a difference in AUC between medical and surgical specialties. The bivariate correlation between years of experience and AUC was computed. As a baseline for comparison, four models from the literature were further validated using the data from the cohort of 16 patients. [7,19,23,25] The predictive ability of each model was measured using the AUC.

## Results

Study participation was solicited through a recruitment email to 150 clinicians. Study material was sent out to 30 clinicians who agreed to participate, and 20 clinicians returned completed surveys (67% response rate; 19 males, 1 female; age 47 ± 6 years). These clinicians represented both medical specialties (n = 4) and surgical specialties (n = 16). They had 11 ± 6 years of experience, with a range from 2 years to 21 years. There was no difference in years of experience between the specialty types ($p = 0.920$).

The 4-point Likert scale was used to establish the consistency of predictions for each rater between the different metrics. I examined the correlation between the Likert scale and the

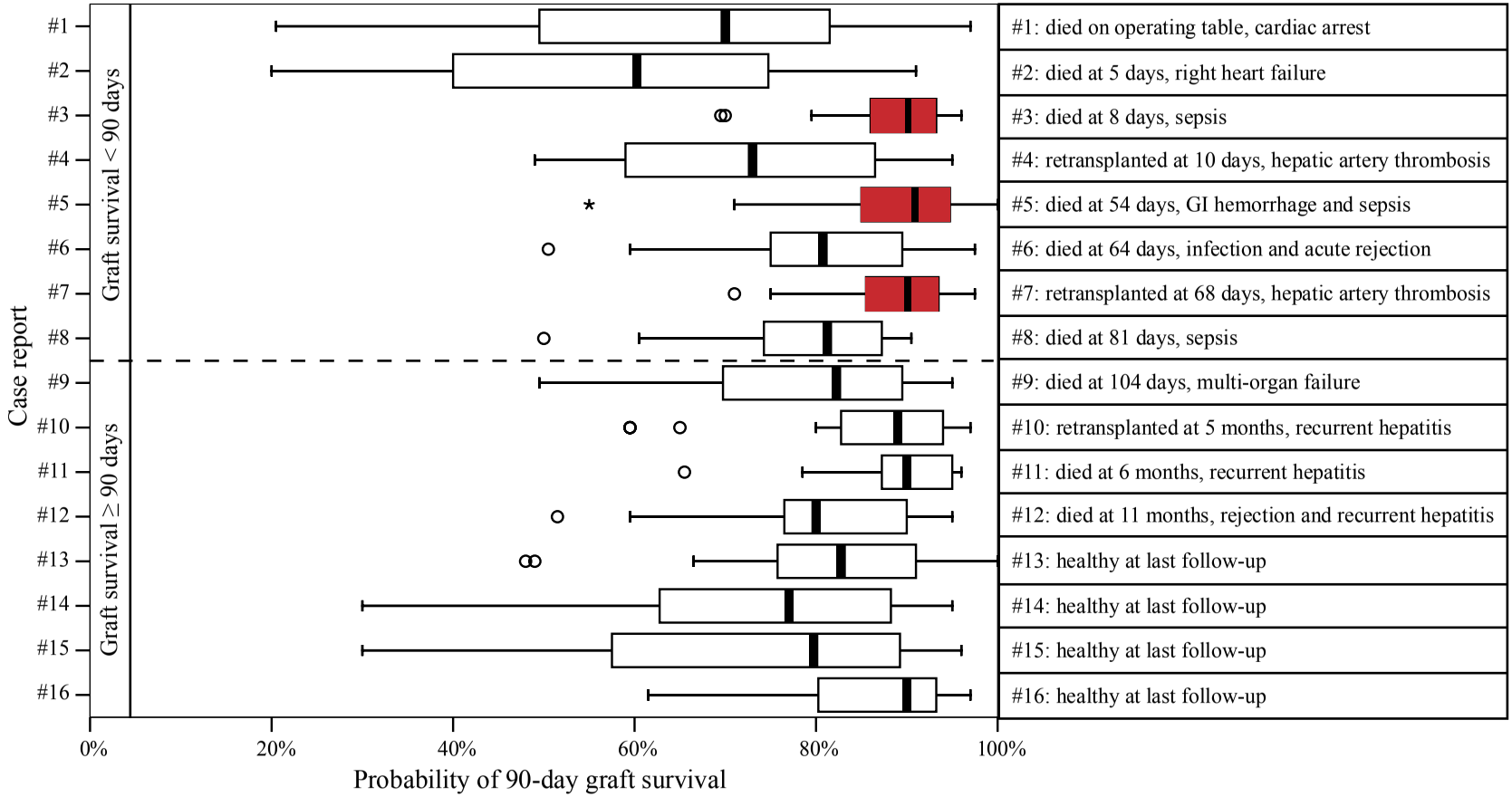



Figure 10. Clinician predictions for 90-day graft survival, shown for each of the 16 case reports. The case reports are sorted by outcome, with poor outcomes at the top and good outcomes at the bottom. The box-and-whisker plots show the median, 25% and 75% values, and extreme values of each distribution. Shown at right are the actual outcomes together with the cause of graft failure, when applicable. Three patients highlighted in red were given favorable estimates of prognosis from the clinicians; however each of the three experienced an early graft failure. All three of these patients experienced post-operative complications of sepsis or hepatic artery thrombosis that contributed to their graft failures.

visual analog scale, and the median Pearson's r was 0.86.

The distributions of the clinicians' prognostic estimates for 90-day graft survival are shown in figure 10. Clinicians tended to agree on the prognosis for some patients and disagree for others, and no relationship was found between the coefficient of variance and outcome ($p = 0.527$). For cases #3, #5, and #7, clinicians estimated the probability of survival to be 88 ± 8%, 88 ± 11%, and 88 ± 7% respectively; however, all three of these patients had early graft failures due to post-operative complications. Two of them had sepsis, leading to death, and one had hepatic artery thrombosis, leading to retransplantation.

One case report (#14) was excluded from AUC performance analysis due to a factual discrepancy in the case report discovered after mailing the summaries. A critical care progress note stated that the patient was not ventilated, as reported in the case summary, but an imaging study prior to transplantation noted life support in place. The remaining 15 cases were used to determine the predictive performance of each clinician, and 13 cases were used for one clinician who left two estimates blank.

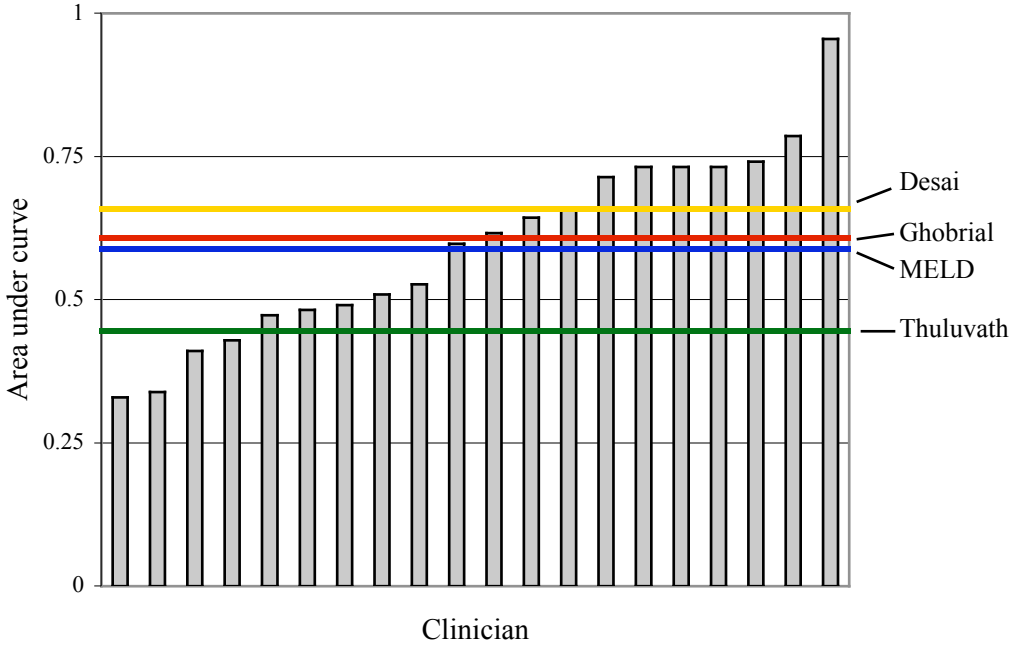



Figure 11. Distribution of individual clinicians' discriminatory power. The area under the receiver operating characteristic curve for each clinician is shown, sorted from lowest to highest. The clinician mean was 0.60 ± 0.16. For comparison, the performance of four baseline models for the cohort of 16 patients is shown on the right: Desai, 0.66; Ghobrial, 0.61; MELD, 0.59; and Thuluvath, 0.45.

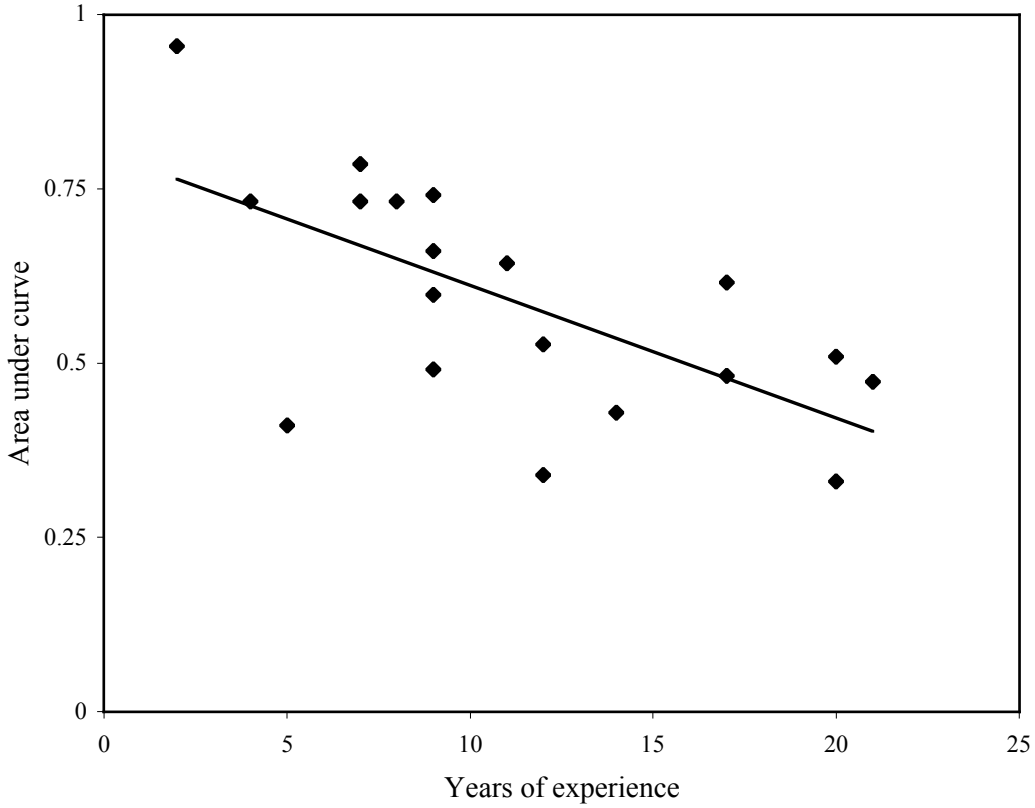



Figure 12. Correlation between experience and discriminatory power. Experience was defined by the number of years a clinician spent caring for liver transplant patients as an attending physician. Discriminatory power was measured by the area under the receiver operating characteristic curve. A significant negative correlation was found, $r = -0.641$, $p = 0.003$.

Individual clinicians' predictive ability for 90-day graft survival was $0.60 \pm 0.16$, as assessed by the AUC. This was comparable to the performance of published models for my cohort of patients (MELD 0.59; Desai 0.66; Ghobrial 0.61; Thuluvath 0.45), as shown in figure 11. The confidence intervals for all AUC calculations were wide, by roughly 50%, due to the small number of patients represented in each receiver operating characteristic curve.

There was no difference in predictive ability between the medical specialties and the surgical specialties ($p = 0.570$). A significant negative correlation ($r = -0.641$, $p = 0.003$) was found to exist between years of experience and predictive power, illustrated by figure 12.

## Discussion

A few observations were noted regarding the distribution of clinician predictions for each case. There were patients about whom clinicians tended to agree, and other patients about whom the clinicians tended to disagree. Furthermore, when clinicians agreed in their predictions for a given patient, the consensus opinion was not always correct, as shown by cases #3, #5, and #7. The patients with good predicted survival and poor actual survival tended to experience complications of generalized sepsis or hepatic artery thrombosis. These complications may be very difficult to predict based on pre-transplant information. Other researchers have postulated that there is a theoretical limit on how well a pre-transplant model of liver transplant survival can perform, because certain complications are affected by peri-operative and early post-operative events [25,36,67]. My observations are consistent with this suggestion.

The results indicate that clinicians predict 90-day graft survival after liver transplantation with a mean AUC performance of 0.60. As a baseline for performance, models in the literature showed similar or worse performance than the clinicians. The statistical significance of this trend was not evaluated, because the confidence intervals were wide for the AUC estimates.

The patient population was selected randomly and stratified to attain a balance of good and poor outcomes. The prevalence of severe complications was higher in my population than in the overall population of patients receiving liver transplants. Considering this, if it holds true that certain complications of liver transplantation are inherently hard to predict, then it logically follows that the predictive ability of a model should be lower with my limited population than with the general population. This is substantiated by the fact that the four published models all showed worse performance than reported in the literature.

The results indicated that less-experienced clinicians have better predictive ability than their more experienced colleagues. At the present time, I can only speculate about potential explanations for this rather surprising finding. It is possible that more experienced clinicians take on more administrative duties as their careers progress, and hence they spend less time with patients. Likewise, it is possible that more experienced clinicians were busier and thus were

unable to devote as much time and attention to the survey as their less experienced colleagues. There could also be a bias in the group of participants or in the study instrument that contributed to this finding. However, this incidental finding requires further examination to distinguish between these possible hypotheses, and future research will focus on this issue.

The study has limitations that merit discussion. First, I have a relatively small sample size of participants. Based on the trends observed in the data analysis, some statistically significant effects were found, and the trends that were not significant were deemed unlikely to become significant if the sample size were increased. Second, I have a small number of cases evaluated by each clinician, resulting in the confidence intervals of AUC performance calculations being relatively wide. To achieve an adequate return rate, I limited the number of cases included, as it took 30 minutes to an hour for the clinicians to read through and make predictions for the 16 included cases. To further increase the number of cases would ask for an impractical time commitment on the part of the attending physicians who completed the survey. Third, the participants in the study never saw firsthand the patients they were assessing; all of the predictions were based on written case reports. The manner of presentation of the case reports may have been an important influence on clinician predictions. However, the preparation of these case reports included a quality control step in anticipation of this limitation, and great efforts were made to include all relevant information in the summaries.

## Conclusion

Based on the survey data, the overall ability of clinicians to predict survival for a set of pre-transplant scenarios was similar to published survival models. In other words, their power to discriminate between likely survivors and non-survivors was relatively modest. Our data are consistent with the postulate that certain post-operative complications of liver transplantation may be inherently hard to predict prior to surgery.

# CHAPTER VIII

## CONCLUSION AND POLICY IMPLICATIONS

The greatest practical value in a liver transplant survival model lies in its ability to make predictions at or before the time of transplantation. Others have shown that it is possible to predict liver transplant survival with good accuracy if post-transplant information is considered [18,67]. However, the utility of a post-transplant model is mitigated by its inability to make predictions at the time of decision-making. Because of this, my thesis has focused on the harder and more clinically-oriented problem of developing a pre-transplant survival model.

To address Specific Aim 1, Chapter IV showed that a Bayesian network can be created to predict liver transplant survival with performance that exceeds published Cox regression models using an independent UNOS data set. Chapter VI demonstrated that a support vector machine can be scaled up to learn a survival model using a very large set of data.

To address Specific Aim 2, Chapter V illustrated that fully automated feature selection techniques can identify the key predictors of liver transplant survival within a very large database, and this predictor set performed better than two out of three published in the literature. Chapter VI showed that a relatively simple model that accounts for main effects and two-way variable interactions may be sufficient for predicting liver transplant survival with the UNOS database.

To address Specific Aim 3, Chapter VII described the ability of clinicians to predict survival after liver transplantation. They perform similarly to mathematical models using our survey cohort, so careful evaluation is necessary if a decision support system is to be used in transplantation.

In addition to addressing the Specific Aims, these collective results highlight the fact that predicting survival after liver transplantation is a challenging problem. Others have suggested that there is a theoretical limit on how well a pre-transplant model can perform [25,36,67]. Two lines of reasoning support this idea. First, model performance may be limited by the content of

the data available, as shown by my extensive search of the UNOS database. Second, the observations from the clinician survey suggest that certain complications of liver transplantation are inherently hard to predict. My thesis, however, does not address where the theoretical limit may be, or how close the current survival models are to reaching that limit.

These findings carry policy implications that pertain both to clinical decision support and to the collection of transplant data. First, my Bayesian network, which was validated using an independent data set and showed generalizable performance, may be useful for clinical decision support in transplantation. I emphasize that it should be used as an adjunct, and not as a replacement, for clinical judgment. Clinicians should be made aware of the strengths and weaknesses of the model, and care should be taken to monitor the impact of its implementation. Second, the collection of additional data pertaining to transplantation may be warranted. For example, these data may include the viral genotype of hepatitis C, a gene expression profile of hepatocellular carcinoma, or mass spectrometry analysis of the donor liver.

In summary, accurate survival prediction after liver transplantation is challenging, and it remains an open problem. However, I have proposed a model for liver transplant survival, and I have carefully validated and compared its performance to models from the literature. I have characterized various aspects of the problem, including which predictors are important, and how complex should a model be in considering them. I have also investigated the ability of clinicians to predict survival, which is the current standard of care for making decisions in transplantation. These elements together pave the way for future refinements in survival modeling, and for a clinical trial of decision support in liver transplantation.

# APPENDIX A

## DETAILS ON CLEANING THE UNOS DATABASE

Several steps of processing occurred in order to convert the UNOS liver transplant database into a cleaner form that would be more useful for survival modeling. These steps are detailed below.

1. *Validation of Data Elements.* The format of data was validated, and all invalid items were removed from the database. All free-text variables were considered to be valid. For all other variables, the following values were deemed to be missing:

    a. Any values which were blank or '**', or which contained a question mark

    b. Any date values which did not have the format MM/DD/YYYY

    c. Any time values which did not have the format HH:MM:SS

    d. Any numeric values which did not have a valid decimal representation

    e. Any values which are connected to text labels in an external file, for which there was no label defined

2. *Ambiguous Information.* All values which represented complete uncertainty, and thus contained no useful information for the present purposes, were removed from the database. The following values were deemed to be missing:

    a. All of the 'U' values for character discrete variables, which often take the form 'Y/N/U' or 'P/N/U'

    b. Any of the values 'Cannot Disclose', 'Indeterminate', 'Not Done', or 'Unknown' for variables which denoted serum status, as noted by SERSTAT labels

    c. For other discrete variables, any of the values 'Unknown', 'UNKNOWN', 'Unknown Duration', 'Not Reported', 'Confirmed Blk.', 'Not Tested', or 'Unknown (for Donor Referral only)'

    d. For the variables ABO and ABO_DON, the value 'UNK'

    e. For the variable PX_STAT, the value 'N'

3. *Range Checking.* For continuous variables, a range was established for values which would be considered reasonable. Any values falling outside of this range were removed from the database, as they resulted either from data entry errors or extreme outliers. The following ranges were established for continuous data:

a. Age: 0 - 100 years

b. Height: 25 - 250 cm

c. Weight: 2.5 - 250 kg

d. Albumin: 0.5 - 7.5 g/dl

e. Alkaline phosphatase: 20 - 2500 U/l

f. Blood urea nitrogen: 5 - 150 mg/dl

g. Creatinine: 0.2 - 10 mg/dl

h. Prothrombin time: 5 - 75 sec

i. Prothrombin control time: 5 - 50 sec

j. International normalized ratio: 1 - 20

k. Transaminases: 5 - 30000 U/l

l. Total bilirubin: 0.2 - 75 mg/dl

m. Cold ischemia: 0 - 48 hours

n. Warm ischemia: 5 - 240 minutes

o. Variables which cannot be non-negative: > 0

p. Dates: > 1/1/1985

q. Year of transplantation: > 1985

4. *Category Abstraction.* Discrete variables which had a large number of possible categories, and thus an insufficient number of values in each category for reliable statistical testing, were condensed into broader categories. The following substitutions were made:

   a. ABO blood type: the values ‘A’, ‘A1’, and ‘A2’ were all marked as ‘A’, and the values ‘AB’, ‘A1B’, and ‘A2B’ were all marked as ‘AB’

   b. Cancer site: all values other than ‘NO’ were marked as ‘YES’

   c. Storage solution: any values aside from ‘VIASPAN (UW/BELZER)’ were marked as ‘OTHER’

   d. Ethnic category: the ‘Non-Hispanic Multiracial’ designation was made part of the ‘Other’ category

   e. Time zone: the values ‘ALASKA’, ‘HAWAII’, and ‘ATLANTIC’ were combined into a category called ‘OTHER’

   f. Payment source: all values aside from ‘HMO/PPO’, ‘MEDICAID’, ‘MEDICARE’, and ‘PRIVATE INSURANCE’ were grouped into a category called ‘OTHER’

g. Employment status: all values designating some reason for not working were marked as 'NOT WORKING', and all values designating some reason for working part time were marked as 'WORKING PART TIME'

h. Patient status: all values other than 'LI: Status 1' were marked as 'OTHER'

i. States: all values were combined into their respective UNOS regions, as defined by the OPTN at http://www.optn.org/latestData/stateData.asp?type=region.

j. Diagnosis: all values were combined into the categories 'NON_CHOLESTATIC', 'CHOLESTATIC', 'ACUTE_HEPATIC_NECROSIS', 'METABOLIC', 'MALIGNANT', 'BILIARY_ATRESIA', and 'OTHER', as defined by the OPTN at http://www.optn.org/organDatasource/about.asp?display=Liver.

5. *Calculated Variables.* A few custom variables were calculated and added into the database. Their values were determined using the following methods:

   a. Patient survival at 30 days, 90 days, 6 months, 1 year, and 5 years: this variable took the value 'Y' if patient observation time was greater than the desired interval, 'N' if patient observation time was less than the desired interval and the patient status was marked as dead, and it was marked as missing otherwise

   b. Graft survival at 30 days, 90 days, 6 months, 1 year, and 5 years: this variable took the value 'Y' if graft observation time was greater than the desired interval, 'N' if graft observation time was less than the desired interval and the patient status was marked as dead or retransplanted, and it was marked as missing otherwise

   c. Body mass index for the donor and recipient: this was determined by the formula (weight in kg / (height in m) squared), and any values less than 5 or greater than 100 were considered absurd and thus marked as missing

   d. Gender match combination: this variable took one of the values 'FF', 'FM', 'MF', or 'MM', representing the concatenation of donor and recipient gender, respectively

6. *Uncleaned Variables.* The following variables were not cleaned and should be excluded from further analysis. The rationale for excluding these items is described below.

   a. DA1, DA2, DB1, DB2, DDR1, DDR2, RA1, RA2, RB1, RB2, RDR1, RDR2: these histocompatibility variables have too many categories and could not be condensed in a systematic manner, and the variables AMAT, AMIS, BMAT, BMIS, DRMAT, and DRMIS act as appropriate surrogates for them

   b. All binary payment form variables (DONATION_*, FREE_*, HMO_PPO_*, MEDICAID_*, MEDICARE_*, OTH_GOVT_*, PRIV_INS_*, SELF_*): these were removed because the PRIMPAY variables act as surrogates for them

   c. DEATH_MECH_DON, DIAL_TY_TCR, HIST_DIABETES_DON, LITYP, RACE, RACE_DON, TX_PROCEDUR_TY: these variables would all need to have their categories condensed for proper analysis, and those condensed variables already exist in the database

d. CLSTR_OLD, CLSTRTYP_OLD, EXTRACRANIAL_CANCER_DON, ECMO, IABP, LI_PUMP, PGE, VAD_TAH: these items lack sufficient variation to be useful for statistical tests

e. CMV_OLD, HBEAB_OLD, HEPD_OLD, HIV_CONF_CAD_DON, HIV_SCRN_CAD_DON, MRCREATG_OLD: collection of these variables stopped during the study interval, and thus they are not populated for prospective validation

7. *Identification of Pre-Transplant Variables.* The following steps were used to eliminate items and reduce the UNOS data to those variables that were available prior to transplant.

   a. Variables collected on the forms TRF or TRF/TRR

   b. Variables marked as TRR/TRF - CALCULATED

   c. Variables collected on the form TRR in the section POST TRANSPLANT CLINICAL INFORMATION

   d. Encrypted identifiers

   e. Items for internal UNOS use (DATASET, DATE_OF_RUN)

   f. Other outcome variables (GTIME, GSTATUS, PTIME, length of stay variables including DISCHARGE_DATE)

   g. Items with OSTXT in the title

REFERENCES


1. 2004 OPTN/SRTR Annual Report 1994-2003. HHS/HRSA/HSB/DOT; UNOS; URREA.

2. Organ procurement and transplantation network [Internet]. Richmond (VA): United Network for Organ Sharing; c2003 [modified 2005 Nov 18; cited 2005 Nov 23]. Available from: http://www.optn.org/latestData/step2.asp?

3. Gilbert JR, Pascual M, Schoenfeld DA, et al. Evolving trends in liver transplantation: an outcome and charge analysis. Transplantation. 1999;67(2):246-53.

4. National Organ Transplant Act: Public Law 98-507. US Statut Large. 1984;98:2339-48.

5. Organ donation and transplantation [Internet]. Richmond (VA): United Network for Organ Sharing; c2005 [cited 2005 Nov 1]. Available from: http://www.unos.org/whoWeAre/history.asp

6. Malinchoc M, Kamath PS, Gordon FD, et al. A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. Hepatology. 2000;31(4):864-71.

7. Kamath PS, Wiesner RH, Malinchoc M, et al. A model to predict survival in patients with end-stage liver disease. Hepatology. 2001;33(2):464-70.

8. Organ donation and transplantation [Internet]. Richmond (VA): United Network for Organ Sharing; c2005 [cited 2005 Nov 24]. Available from: http://www.unos.org/resources/glossary.asp

9. Organ Procurement and Transplantation Network—HRSA. Final rule with comment period. Fed Regist. 1998;63:16296-338.

10. Karnofsky DA, Abelmann WH, Craver LF, et al. The use of nitrogen mustards in the palliative treatment of carcinoma. Cancer. 1948;1:634-56.

11. Ware JE, Kosinski M, Keller SD. SF-36 physical and mental health summary scales: a users' manual. Boston: The Health Institute, New England Medical Center; 1994.

12. MeSH [Internet]. Bethesda (MD): National Library of Medicine; [cited 2005 Aug 25]. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=mesh

13. Shaw BW, Jr., Wood RP, Gordon RD, et al. Influence of selected patient variables and operative blood loss on six-month survival following liver transplantation. Semin Liver Dis. 1985;5(4):385-93.

14. Maggi U, Rossi G, Colledan M, et al. Child-Pugh score and liver transplantation. Transplant Proc. 1993;25(2):1769-70.

15. Selberg O, Bottcher J, Tusch G, et al. Identification of high- and low-risk patients before liver transplantation: a prospective cohort study of nutritional and metabolic parameters in 150 patients. Hepatology. 1997;25(3):652-7.

16. Chung SW, Kirkpatrick AW, Kim HL, et al. Correlation between physiological assessment and outcome after liver transplantation. Am J Surg. 2000;179(5):396-9.

17. Adam R, Cailliez V, Majno P, et al. Normalised intrinsic mortality risk in liver transplantation: European Liver Transplant Registry study. Lancet. 2000;356(9230):621-7.

18. Parmanto B, Doyle HR. Recurrent neural networks for predicting outcomes after liver transplantation: representing temporal sequence of clinical observations. Methods Inf Med. 2001;40(5):386-91.

19. Ghobrial RM, Gornbein J, Steadman R, et al. Pretransplant model to predict posttransplant survival in liver transplant patients. Ann Surg. 2002;236 (3):315-22.

20. Onaca NN, Levy MF, Sanchez EQ, et al. A correlation between the pretransplantation MELD score and mortality in the first two years after liver transplantation. Liver Transpl. 2003;9(2): 117-23.

21. Fernandez-Aguilar JL, Santoyo J, Suarez MA, et al. Is MELD useful in evaluating the surgical risk in liver transplantation candidates? Transplant Proc. 2003;35(2):705-6.

22. Saab S, Wang V, Ibrahim AB, et al. MELD score predicts 1-year patient survival post-orthotopic liver transplantation. Liver Transpl. 2003;9(5):473-6.

23. Thuluvath PJ, Yoo HY, Thompson RE. A model to predict survival at one month, one year, and five years after liver transplantation based on pretransplant clinical characteristics. Liver Transpl. 2003;9(5):527-32.

24. Bilbao I, Armadans L, Lazaro JL, et al. Predictive factors for early mortality following liver transplantation. Clin Transplant. 2003;17(5):401-11.

25. Desai NM, Mange KC, Crawford MD, et al. Predicting outcome after liver transplantation: utility of the model for end-stage liver disease and a newly derived discrimination function. Transplantation. 2004;77(1):99-106.

26. Santori G, Andorno E, Antonucci A, et al. Potential predictive value of the MELD score for short-term mortality after liver transplantation. Transplant Proc. 2004;36(3):533-4.

27. Austin GL, Sasaki AW, Zaman A, et al. Comparative analysis of outcome following liver transplantation in US veterans. Am J Transplant. 2004;4(5):788-95.

28. Roberts MS, Angus DC, Bryce CL, et al. Survival after liver transplantation in the United States: a disease-specific analysis of the UNOS database. Liver Transpl. 2004;10(7):886-97.

29. Jacob M, Copley LP, Lewsey JD, et al. Pretransplant MELD score and post liver transplantation survival in the UK and Ireland. Liver Transpl. 2004;10(7):903-7.

30. Bazarah SM, Peltekian KM, McAlister VC, et al. Utility of MELD and Child-Turcotte-Pugh scores and the Canadian waitlisting algorithm in predicting short-term survival after liver transplant. Clin Invest Med. 2004;27(4):162-7.

31. Northup PG, Berg CL. Preoperative delta-MELD score does not independently predict mortality after liver transplantation. Am J Transplant. 2004;4(10):1643-9.

32. Santori G, Andorno E, Morelli N, et al. MELD score versus conventional UNOS status in predicting short-term mortality after liver transplantation. Transpl Int. 2005;18(1):65-72.

33. Haydon GH, Hiltunen Y, Lucey MR, et al. Self-organizing maps can determine outcome and match recipients and donors at orthotopic liver transplantation. Transplantation. 2005;79(2): 213-8.

34. Moore DE, Feurer ID, Speroff T, et al. Impact of donor, technical, and recipient risk factors on survival and quality of life after liver transplantation. Arch Surg. 2005;140(3):273-7.

35. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36.

36. Jacob M, Lewsey JD, Sharpin C, et al. Systematic review and validation of prognostic models in liver transplantation. Liver Transpl. 2005;11(7):814-25.

37. Counsell C, Dennis M. Systematic review of prognostic models in patients with acute stroke. Cerebrovasc Dis. 2001;12(3):159-70.

38. Walczak S. Gaining competitive advantage for trading in emerging capital markets with neural networks. J Management Information Systems. 1999;16(2):177-92.

39. Fernandez R, Viennet E. Face identification using support vector machines. Proc European Symposium on Artificial Neural Networks. 1999:195-200.

40. Aliferis CF, Hardin D, Massion PP. Machine learning models for lung cancer classification using array comparative genomic hybridization. Proc AMIA Symp. 2002:7-11.

41. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.

42. Hosmer DW, Lemeshow S. Applied survival analysis: regression modeling of time to event data. New York: Wiley; 1999.

43. Mitchell TM. Machine Learning. New York: McGraw-Hill; 1997.

44. Breiman L, Friedman JH, Olshen RA, et al. Classification and regression trees. Belmont, CA: Wadsworth International Group; 1984.

45. Bayes T. An essay towards solving a problem in the doctrine of chances. Philosophical Trans Royal Society of London. 1763;53:370-418.

46. Dougherty J, Kohavi R, Sahami M. Supervised and unsupervised discretization of continuous features. Machine Learning: Proc Twelfth International Conference. 1995:194-202.

47. Domingos P, Pazzani M. Beyond independence: conditions for the optimality of the simple Bayesian classifier. Proc Thirteenth International Conference on Machine Learning. 1996:105-12.

48. Dempster AP. A generalization of Bayesian inference. J Royal Statistical Society, Series B. 1968;30:205-47.

49. Jensen FV. An introduction to Bayesian networks. New York: Springer Verlag; 1996.

50. van der Gaag LC, Abu-Hanna A. Bayesian networks in biomedicine and health care. Artificial Intelligence in Medicine. 2004;30:201-14.

51. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967;13:21-7.

52. Duda RO, Hart PE. Pattern classification and scene analysis. New York: Wiley; 1973.

53. Wasserman PD. Neural computing: theory and practice. New York: Van Nostrand Reinhold; 1989.

54. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review. 1959;65:386-408.

55. Bishop CM. Neural networks for pattern recognition. Oxford: Oxford University Press; 1995.

56. Burges CJC. A tutorial on support vector machines for pattern recognition. Knowledge Discovery and Data Mining. 1998;2(2):1-43.

57. Schölkopf B, Burges CJC, Smola AJ. Advances in kernel methods: support vector learning. Cambridge, MA: MIT Press; 1999.

58. Russell SJ, Norvig P. Artificial intelligence: a modern approach. 2nd ed. Upper Saddle River, NJ: Prentice Hall/Pearson Education; 2003.

59. Heckerman DE, Horvitz EJ, Nathwani BN. Toward normative expert systems: Part I. The Pathfinder project. Methods Inf Med. 1992;31(2):90-105.

60. Markmann JF, Markmann JW, Markmann DA, et al. Preoperative factors associated with outcome and their impact on resource use in 1148 consecutive primary liver transplants. Transplantation. 2001;72(6):1113-22.

61. Moore DE, Feurer ID, Rodgers S, Jr., et al. Is there racial disparity in outcomes after solid organ transplantation? Am J Surg. 2004;188(5):571-4.

62. Tsamardinos I, Aliferis CF. Towards principled feature selection: relevancy, filters, and wrappers. Proc Ninth International Workshop on Artificial Intelligence and Statistics. 2003.

63. Aliferis CF, Tsamardinos I, Statnikov A. HITON: a novel Markov Blanket algorithm for optimal variable selection. AMIA Annu Symp Proc. 2003:21-5.

64. Statnikov A, Tsamardinos I, Dosbayev Y, et al. GEMS: a system for automated cancer diagnosis and biomarker discovery from microarray gene expression data. Int J Med Inform. 2005;74(7-8):491-503.

65. Glare P, Virik K, Jones M, et al. A systematic review of physicians' survival predictions in terminally ill cancer patients. BMJ. 2003;327(7408):195.

66. Gusmao Vicente F, Polito Lomar F, Melot C, et al. Can the experienced ICU physician predict ICU length of stay and outcome better than less experienced colleagues? Intensive Care Med. 2004;30(4):655-9.

67. Markmann JF, Markmann JW, Desai NM, et al. Operative parameters that predict the outcomes of hepatic transplantation. J Am Coll Surg. 2003;196(4):566-72.