UNIVERSITY OF CALGARY

Does the Model for End-stage Liver Disease Accurately Predict the Occurrence of Ninety-Day Wait-List Mortality in Atlantic Canadian Liver Transplant Candidates? A Validation Study.

by

Paul Douglas Renfrew

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMMUNITY HEALTH SCIENCES

CALGARY, ALBERTA

JANUARY, 2011

Abstract

The Model for End-stage Liver Disease (MELD) is a prognostic model for End-Stage Liver Disease (ESLD) that has been shown to accurately predict the risk of mortality in liver transplant wait-list cohorts. In consideration of its adoption as a component of a national organ allocation system, it is important that the predictive performance of the MELD be characterized in Canadian liver transplant candidates. The primary objective of this study was to determine if the MELD accurately predicts the occurence of 90-day mortality in Atlantic Canadian adults with ESLD awaiting liver transplantation. The MELD was found to have excellent accuracy of discrimination and analysis of calibration produced no evidence that MELD-generated 90-day mortality estimates differed significantly from the observed incidence. The results of this study indicate that the MELD is generalizable to adult liver transplant candidates with ESLD residing in Atlantic Canada and consequently would be suitable for the prioritization of these patients for allograft allocation.

Preface

As stated in its title, the primary objective of this study was to externally validate the Model for End-stage Liver Disease using data from a five year cohort of Atlantic Canadian liver transplant candidates with end-stage liver disease. The report on my study, which begins in Chapter Three, is preceded by two background chapters. The intent of these two chapters was to summarize all that I have had to learn, regarding both content and methodology, to enable me to design and conduct this research project. The first chapter endeavours to comprehensively cover issues related to liver transplant organ allocation and the Model for End-stage Liver Disease. The second background chapter deals with the principles and methodologies of prognostic model validation and discusses, in depth, performance measures for the evaluation of predictive models for binary outcomes. The objective of these background chapters is to provide the reader with sufficient information that they will easily understand the subsequent chapters describing this study and its results. Chapters Three and Four detail the objectives and methodology of the study, and Chapters Five through Nine report its results. Each results chapter addresses a specific study objective and includes a discussion and conclusions pertaining to that objective. My intent with this format was that each results chapter could be consumed as it were a stand alone study. The final chapter, Chapter Ten, provides a summary of all study conclusions. Material and analyses that I felt were interesting, but not directly contributory to the study objectives, were assigned to Appendices A through E. I am very proud of this work and I hope it can contribute in some way to improving the care of patients in Atlantic Canada, and maybe beyond.

Acknowledgements

I have many people that I need to thank! First, I would like to thank my supervisor, Dr. Elijah Dixon, and the other members of my committee; Dr. Christopher Doig, Dr. Hude Quan and Dr. Michele Molinari, for their patience, and provision of guidance in the design and execution of this research project. In particular, I must acknowledge my appreciation of Dr. Molinari's encouragement and support, which enabled me to complete my studies. I also wish to thank Dr. Marilynne Hebert, who facilitated my continuation of graduate studies with the University of Calgary, despite my relocation to Nova Scotia. Dr. Oliver Bathe has to take the credit for starting me on this academic path, and so I want to acknowledge and thank him too.

Special thanks go out to Catherine Guimont and Mary Jane MacNeil for showing a genuine interest in my project and providing guidance and assistance that greatly facilitated my data acquisition. Also, thanks to Terry Reardon for his prompt responses to my data requests.

I also need to express my gratitude to Joanne Benson, Ray Kim and Terry Therneau of the Mayo Clinic College of Medicine in Rochester Minnesota for their taking the time to respond to my inquiries pertaining to the MELD and MELDNa. I also want to thank Michael Pencina of Boston University for his assistance with questions pertaining to the Net Reclassification Improvement and Integrated Discrimination Improvement.

Above all others, I want express my thanks, and affection, to Morgan Brooke Slater, whom I suspect is second only to me in happiness on completion of this endeavour.

Dedication

This work is dedicated in memory of Michael John Kestle, a gifted surgeon, a mentor, and the best colleague you could ever have.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BcA	Bias-corrected and Accelerated
BMI	Body Mass Index
CanWAIT	Canadian Wait-listing Algorithm in Transplantation
CBS	Canadian Blood Services
CDHA	Capital District Health Authority
CORR	Canadian Organ Replacement Registry
c-statistic	Concordance Statistic
CTP	Child-Turcotte-Pugh
ESLD	End-Stage Liver Disease
HBV	Hepatitis B Virus
HCC	Hepatocellular Cancer
HCV	Hepatitis C Virus
IDI	Integrated Discrimination Improvement
INR	International Normalized Ratio of prothrombin time
IQR	Inter-Quartile Range
MELD	Model for End-stage Liver Disease
MELD+D	Model for End-stage Liver Disease with activated disease aetiology variable
MELDNa	Model for End-stage Liver Disease combined with serum sodium concentration

MELD-UNOS	Model for End-stage Liver Disease, United Network for Organ Sharing modification
MOTP	Multi-Organ Transplant Programme
NRI	Net Reclassification Improvement
OPTN	Organ Procurement and Transplant Network
PBC	Primary Biliary Cirrhosis
PHGIH	Portal Hypertensive Gastrointestinal Haemorrhage
PSC	Primary Sclerosing Cholangitis
PSE	Porto-Systemic Encephalopathy
RED	Recognized Exceptional Diagnosis
ROC	Receiver Operating Characteristic
SRTR	Scientific Registry of Transplant Recipients
TIPS	Transjugular Intrahepatic Porto-systemic Shunt
UNOS	United Network for Organ Sharing
USDHHS	United States Department of Health and Human Services

Chapter One: Background – Liver Transplant Allograft Allocation and the Model for End-Stage Liver Disease

1.1 Liver Transplantation and End-Stage Liver Disease

Liver transplantation is accepted as the optimum therapy for irreversible acute liver failure, decompensated end-stage liver disease (ESLD), early-stage hepatocellular cancer (HCC), and selected genetic metabolic diseases.[1] Of these four broad categorical indications, decompensated ESLD is the predominant primary indication for liver replacement, accounting for between two thirds to three quarters of all individuals placed on the waiting list.[2] ESLD, also known as cirrhosis, represents the final common end point of all chronic liver diseases and can be categorized clinically as either "compensated" or "decompensated". This categorization is based on the absence or presence of symptoms related to loss of hepatic metabolic function and/or stigmata of portal hypertension. These symptoms include; jaundice, coagulopathy, muscular atrophy, ascites, portal hypertensive gastrointestinal haemorrhage (PHGIH), and porto-systemic encephalopathy (PSE). Individuals with compensated ESLD have been found to develop symptoms, and therefore transition to decompensated status, at an annual rate of five to seven percent (Figure 1.1).[3] This occurrence heralds a dramatic change in prognosis, as the median survival for individuals with compensated ESLD exceeds 12 years, while the median survival for individuals with decompensated ESLD is approximately two years. The corresponding one-year and five-year survival estimates for compensated vs. decompensated ESLD are 95 vs. 66 percent and 80 vs. 25 percent (Figure 1.2).[3] Specific manifestations of decompensation are associated with diminished survival. The presence of ascites is associated with a one-year mortality risk of 20 percent and a five-

Figure 1.1: Cumulative proportion of patients, from a cohort of 806, transitioning from compensated to decompensated ESLD[3] (reproduced with permission from Elsevier Limited)

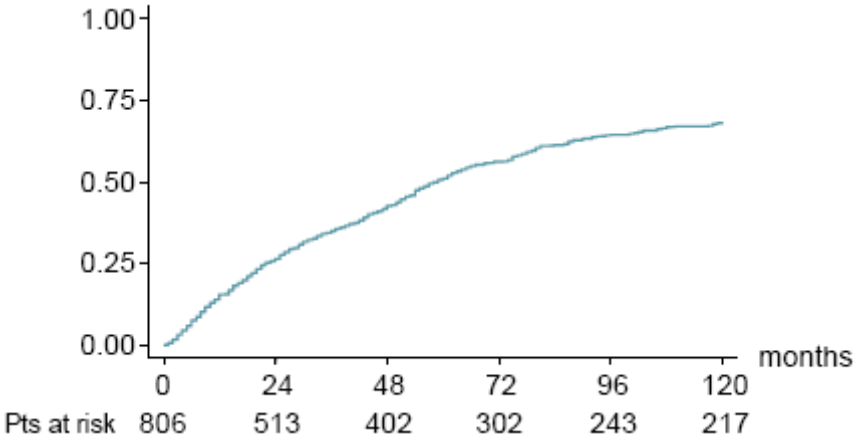
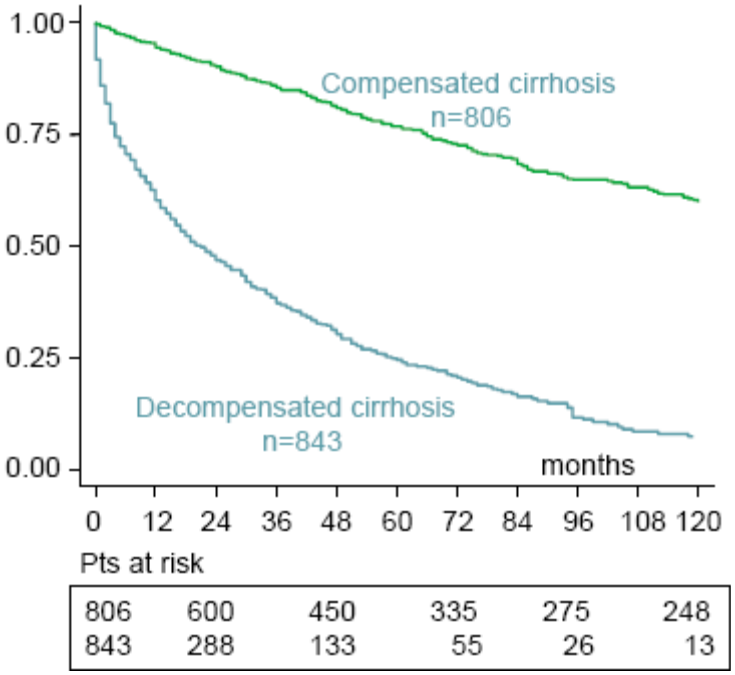


Figure 1.2: Survival of a cohort of 1,649 individuals with ESLD according to compensated vs. decompensated status at time of diagnosis[3] (reproduced with permission from Elsevier Limited)

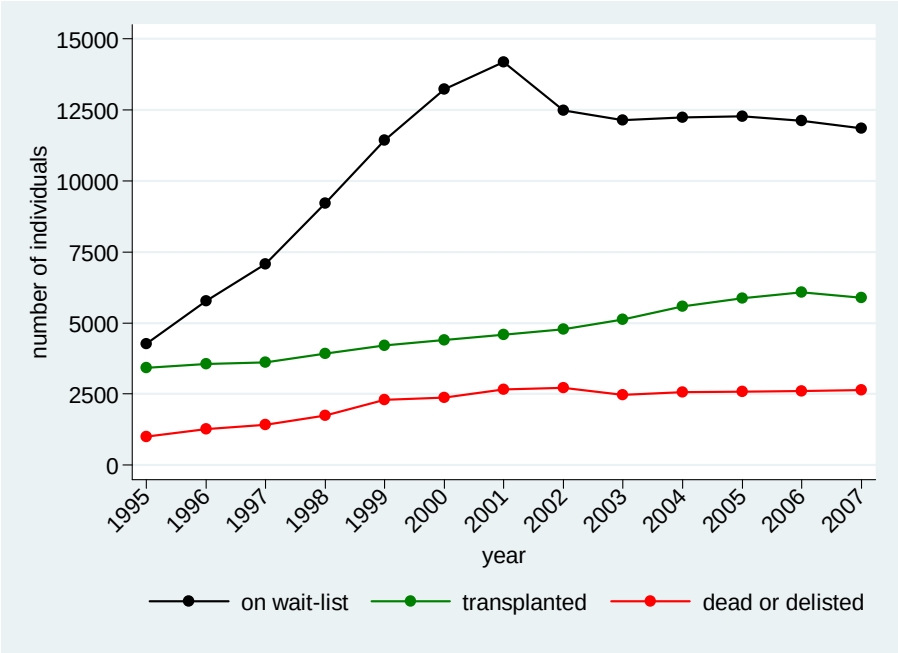


year mortality of approximately 65 percent. Ascites which is refractory to medical management is associated with a mean survival of approximately one year. Individuals experiencing their first episode of bleeding from oesophageal varices have a one-year mortality risk of approximately 40 percent. The occurrence of portosystemic encephalopathy has been determined to be associated with a 60 to 70 percent increase in the relative risk of mortality at one year.[3, 4] In contrast to these rather dismal outcomes facing individuals with decompensated ESLD, liver transplantation results in one, five and ten-year survivals of approximately 85, 70 and 55 percent.[5, 6]

1.2 Demand, Supply and Trends in Waiting List Outcomes

As liver transplantation evolved from an experimental to widely accepted form of treatment, the demand for liver replacement has grown. Unlike most medical treatments in the developed world, the ability to provide liver transplantation is clearly constrained by resource limitations, specifically the availability of deceased donor allografts. In Canada the number of individuals 18 years of age and older on the waiting list for liver transplant at years end increased from 126 in 1995 to 570 in 2008.[7, 8] This constitutes a 4.5-fold relative increase in the demand for liver replacement over this 14 year period. During this time the actual number of transplants performed increased only 1.5-fold, from 268 in 1995 to 409 in 2008. This inequity between demand and supply is even more concerning given that over the 14 year interval there was only a 1.3-fold increase in the number of deceased donor allografts available for transplantation, from 268 to 351 (14 percent of liver transplants performed in 2008 to individuals 18-years and older utilized living donor allografts).[7, 8] Similar trends have been described in the United States of America, and in Europe.[9, 10] As a consequence of this imbalance between demand for,

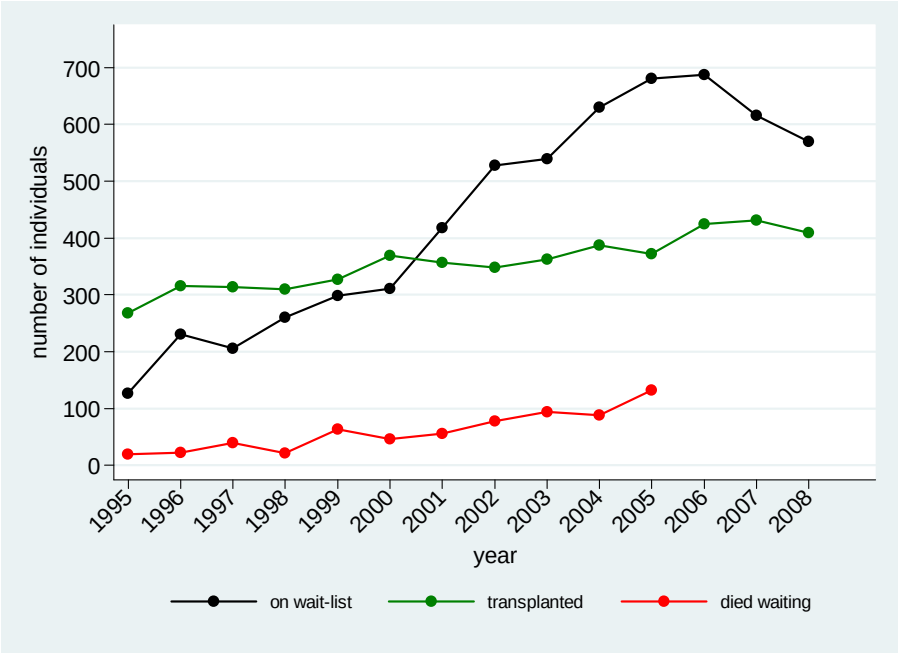
Figure 1.3: U.S. trends in liver transplant waiting list outcomes 1995 to 2007[11-13]



and supply of, liver allografts, there has been a progressive rise in the absolute number of individuals dying while on the liver transplant waiting list. In the United States, from 1995 to 2007, there was a 2.8-fold increase in the number of active adult candidates on the liver transplant waiting list at year-end (from 4,261 to 11,854), yet only a 1.7-fold increase in the number of transplants performed, utilizing both deceased and live donors (from 3,428 to 5,887). From 1995 to 2000, paralleling the increase in the waiting list size, wait-list deaths increased from 843 to plateau at approximately 2,000 annually. Unfortunately, the number of adults losing candidacy for being too sick to undergo transplantation continued to rise, so that overall the number of individuals experiencing wait-list failure continued to increase. From 1995 to 2007 the absolute number of individuals dying or becoming too ill to be transplanted, while on the waiting list,

increased 2.6 times, from 1,002 to 2,633 (Figure 1.3).[11-13] In Canada, as a consequence of the reporting of liver transplant wait-list information not being mandatory, the accurate determination of wait-list outcomes is less certain. Never-the-less, reports from the Canadian Organ Replacement Registry (CORR) indicate that from 1995 to 2005 the absolute number of adult candidates dying while awaiting liver transplant increased from 19 to 132.[7, 14] This information, which indicates a 6.9-fold increase in wait-list mortality over the time period, is particularly sobering since it does not include individuals that lost candidacy due to becoming too ill to undergo transplantation (Figure 1.4).

Figure 1.4: Canadian trends in liver transplant waiting list outcomes from 1995 to 2008[7, 8, 14]



1.3 Wait-List Mortality

Waiting list mortality can be defined as the occurrence of death while a liver transplant candidate awaits the availability of a suitable allograft. Comprehensively it should also include the occurrence of a candidate being removed from the waiting list as they became too ill, due to progression of their liver disease or associated complications, to survive the transplant procedure, as all these delisted individuals ultimately die. Both of these outcomes constitute a wait-list failure and, unless otherwise specified, will be referred to interchangeably as "wait-list mortality" or "wait-list failure" throughout the remainder of this monograph.

In the United States of America, from 1998 to 2007, the annual incidence proportion of individuals on the national waiting list that died (delisted not included) ranged from 6.7 percent (in 2007) to 9.8 percent (in 1999), and the corresponding annual incidence rate for wait-list death ranged from 113 to 166 per 1,000 person-years at risk.[12] A commonly utilized reference point for the measurement of wait-list mortality is status at 90 days after listing. In a large American wait-list cohort study by Wiesner et al the incidence of 90-day wait-list death was 12.0 percent (412 of 3,437 candidates) and, when a further 95 patients that were delisted as they became too sick to transplant are included, the overall incidence of 90-day wait-list mortality was 14.8 percent.[15] It must be pointed out that the cohort was comprised of individuals with higher illness acuity (status 2A and 2B, explained in section 1.6.1) and therefore the risk of 90-day failure is likely inflated. A very large, and more inclusive, study by Kim et al found the 90-day incidence of wait-list death in a cohort of 13,940 candidates to be 6.4 percent.[16] The specific number of candidates delisted for being too sick to tolerate transplantation was

not described in this study, although overall only a small number, 114 candidates, were withdrawn for all causes. In an analysis of 320 individuals on the waiting list of the liver transplant programme based in Edmonton, Alberta the incidence of death or delisting at 90 days was 9.7 percent and 14.7 percent at one year.[17]

The determinants of wait-list mortality can be broadly categorized as liver disease-related and waiting time-related. The severity of the liver disease at the time of waiting list placement and the rate of disease progression is predominantly a function of the underlying liver disease(s). Disease severity at the time of listing is however influenced by the timeliness of referral and the transplant assessment process. Individuals that are referred late in their disease process or are subjected to a protracted evaluation period will enter the waiting list later in their disease process compared to those identified as potential liver transplant candidates early on their disease timeline and assessed efficiently. Once on the waiting list, disease severity and tempo of progression, although, again, primarily a function of the underlying liver disease(s), can be influenced by access to, and the expertise of, medical care while awaiting transplantation. Waiting-time determinants can be sub classified into those that influence allograft availability, those that influence allograft eligibility, and the allograft allocation process. As indicated earlier, the incidence of deceased donorship is the predominant constraint on allograft availability and is influenced by sociocultural and legislative factors. Once a donor liver becomes available candidate blood group and physical size influences their eligibility for a specific allograft. Finally, allograft allocation policy can play a pivotal role in determining the duration of the waiting time for an individual candidate.

1.4 General Principles of Liver Transplant Allocation

In the earliest days of liver transplantation it was a somewhat formidable experimental procedure, performed in only a few centres worldwide, and consequently there were only a small number of candidates. The clinician-investigators involved in the transplantation process made the allocation decisions without needing to consider any broader implications. As progress was made on many fronts the transplant procedure and perioperative management were refined, and by the early 1980s liver transplantation was recognized as an accepted therapy for many previously terminal liver diseases. Paralleling this progress, the donor pool, which had originally been restricted to individuals declared dead by cardiopulmonary criteria, was expanded by the development and acceptance of criteria for the declaration of neurological death. The deceased donor resource was also enhanced by improvements in donor management and organ preservation. Never-the-less, as has been discussed previously, the growth of the candidate population has outstripped the growth of the allograft supply, emphasizing the need for an equitable and efficient strategy to distribute this precious resource.[18]

Effective allocation of scarce medical resources, such as donor allografts, requires balancing two competing forces, the needs of the individual and the needs of the population. Early liver allocation strategies were modelled after renal allograft allocation and utilized time on waiting list to prioritize allocation. Although this is fair in the setting of end-stage renal failure, where patients enter the waiting list with a similar severity of disease and renal replacement therapy can maintain the candidate until it is their turn to receive an allograft, individuals entering the liver transplant waiting list have heterogeneous disease severity and rate of progression, with no means of providing long-

term liver replacement therapy. Use of time on wait-list to prioritize liver transplant candidates favours individuals entered early with less severe and slowly progressing liver disease, as individuals with more severe and rapidly progressive liver disease succumb before they can amass sufficient waiting time to ascend the wait-list.[19] Subsequently it has been recognized, and in some cases mandated, that for liver transplantation a candidate's medical need should be employed as the measure by which allocation is prioritized.[20] Although individual medical need is the dominant determinant of liver allograft allocation strategy, this must be tempered by considering the needs of the entire population of individuals requiring liver replacement. Following the battlefield triage model, the medical acuity of individuals selected to receive an allograft must not be such that they have no chance of surviving transplantation. Thus, to balance individual medical justice with medical utility, allocation must give priority to the sickest individuals that have a realistic probability of having a positive outcome post-transplantation.[18]

For this philosophy to be optimally executed, the allocation system must be efficient, objective and reproducible. This requires clear definitions of need and outcome. In the setting of liver transplantation medical need is defined by the severity of liver disease, and has been measured in several ways. One of the earliest measures utilized the setting in which the candidate was receiving care, e.g. in an acute care unit, in hospital, or outpatient, as a surrogate for severity of liver disease. Such a system is still employed as the basis of the rudimentary allocation system in Canada. This measure has been acknowledged as being subject to manipulation by clinicians and does not permit stratification of need for the vast majority of candidates who are outpatients.[21] In an effort to achieve more objective and reproducible measurement of liver disease severity,

disease severity indices have been employed. These endeavour to stratify a candidate's disease severity based on the status of factors that have been determined to be associated with mortality in individuals with liver disease. The most prominent are the Child-Turcotte-Pugh (CTP) classification and score, and the Model for End-stage Liver Disease (MELD) both of which will be discussed in detail in subsequent sections.[22-24]

Given that, with very few exceptions, liver transplantation is undertaken to resolve otherwise lethal illness, survival is the outcome of interest by which the allocation process is measured. While this is a concrete and easily quantifiable outcome, the optimal timing of its evaluation is debatable and can dramatically influence the allocation strategy. Survival of liver transplant candidates can be divided into survival on the waiting list and survival following transplantation. Allocation strategies which endeavour to minimize wait-list mortality recognize foremost the needs of the individual. However, unchecked, this type of allocation scheme will result in transplantation being undertaken for individuals with poor prospect for survival post-transplant while depriving others with less severe liver disease, but more favourable long-term outcome, the allograft resource. Conversely, an allocation strategy which aims to maximize post-transplant survival will favour individuals with lower illness acuity (and absence of comorbid conditions such as malignancy) as they have the greatest likelihood of surviving the wait-list and peri-operative periods, and achieving long-term survival. As a consequence, individuals with more severe liver disease will be denied an allograft even though they might have a reasonable prospect for survival if transplanted expeditiously. The ideal allocation system therefore, should aim to optimize both pre-transplant and post-transplant survival. To this

point in time such an allocation strategy has yet to be devised, and most formal systems focus primarily on wait-list outcome.[18, 25]

Details of allocation methods differ internationally but share the principles that have been discussed thus far. Common to all allocation systems is an emergent priority status category for individuals with acute liver failure or allograft non-function post-transplantation (primary or secondary to hepatic arterial thrombosis). These high status individuals have a very high risk of short-term mortality and are eligible for the first available, blood group appropriate, allograft from an expanded (frequently national) geographic area. This is an example of patient-oriented allocation, in that the allograft is directed to an individual based on their medical need. Some nations employ patient-oriented allocation for liver transplant candidates of all levels of illness acuity. This requires a central registry on to which all candidates are enrolled and the application of a standard prioritization method. The United States of America and Eurotransplant (Austria, Belgium, Germany, Luxembourg, the Netherlands and Slovenia) employ a patient-oriented allocation model. The alternative is centre-oriented allocation, whereby, in the absence of emergent candidates, donor allografts are allocated to the transplant centre, which in turn allocates the liver to one of the candidates on its waiting list using the centre-specific prioritization process. Centre-oriented allocation is a hold-over from the earliest days of liver transplantation, when there were relatively few programs competing for allografts. With this type of system there is no centralized administration of a waiting list, however nations that report satisfactory function of this type of system typically have some form of central administrative structure to actively monitor

transplant activity. Centre-oriented systems are employed by Spain, the United Kingdom, Scandiatransplant (Denmark, Finland, Iceland, Norway and Sweden) and Canada.[18, 25]

1.5 Origins of the Model for End-Stage Liver Disease (MELD)

The MELD score is a prognostic model for end-stage liver disease that is based on a Cox regression model that was developed, by Malinchoc et al of the Mayo Clinic in Rochester, Minnesota, to estimate post-procedural mortality in cirrhotic patients undergoing transjugular intrahepatic porto-systemic shunts (TIPS).[26] This original Mayo End-Stage Liver Disease model was derived from factors, found by multivariable Cox proportional hazards regression, to be independently associated with survival in a multicentre cohort of 231 individuals with decompensated cirrhosis that had non-emergent placement of a TIPS. The modelling process was quite restrictive, requiring a univariable threshold p-value of less-than 0.05 for inclusion in the multivariable model, and setting the level of significance for multivariable coefficients at less-than 0.01. From this process four independent predictors of survival were identified: serum creatinine, serum total bilirubin, the International Normalized Ratio (INR) of prothrombin time, and aetiology of primary liver disease (alcoholic or cholestatic, coded 0, vs. all others, coded 1). The resulting model was:

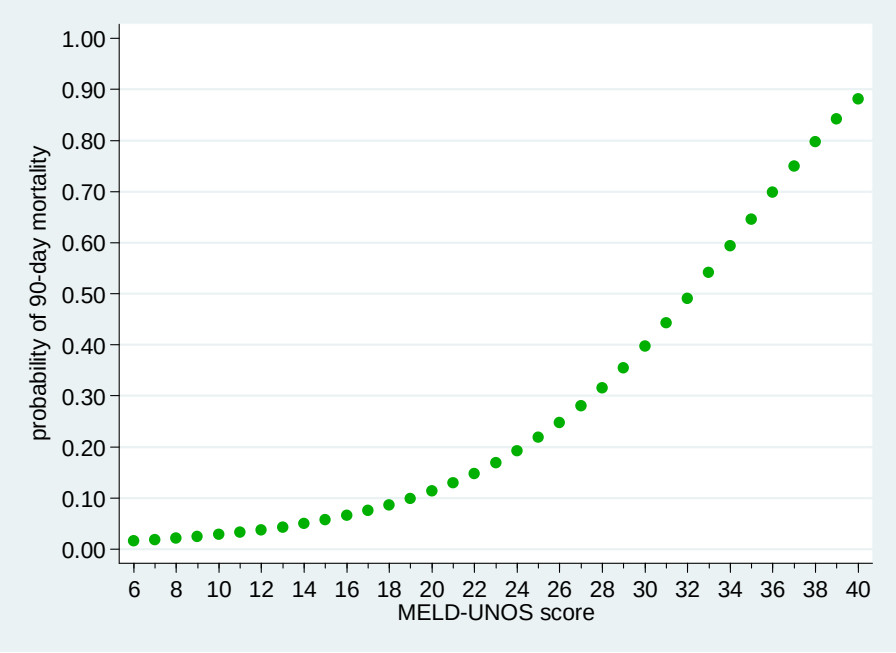
$$\text{Risk Score (R)} = (0.957 \times \ln(\text{creatinine mg/dl})) + (0.378 \times \ln(\text{total bilirubin mg/dl}) + (1.120 \times \ln(\text{INR})) + (0.643 \times \text{liver disease aetiology})$$

Subsequent to the calculation of an individual's risk score an estimate of their probability of survival at a specific point in time post-TIPS could be generated by inserting their risk score into the underlying survival function:

$$P_{survival}(t) = S(t)^{\exp(R - Ro)},$$

where S(t) is the estimated probability of survival corresponding to the mean risk score (R_0) of the cohort at a specific time (*t*) post-TIPS. Figure 1.5 demonstrates the MELD derived probability estimates for 90-day ESLD-related mortality.

Figure 1.5: MELD derived probability estimate for ESLD-related for 90-day mortality



Malinchoc et al evaluated their new model's performance by measuring its ability to predict 90-day survival in the validation cohort using Receiver Operating Characteristic (ROC) analysis. The authors justified the 90-day period as this was the average time to transplantation at the time of the study and also it was felt that longer term survival might be confounded by late complications of the TIPS, such as stenosis. This analysis found the area under the ROC curve to be 0.85 (95%CI: 0.78-0.91), which implies good discrimination of the model.[27] In addition to the ROC analysis, quantile-of-risk assessment was performed which demonstrated that the predicted survival tended

to reflect observed survival of the validation cohort. To further document the performance of the model it was applied to a cohort of 71 TIPS patients from the Netherlands. Comparison of observed and estimated survival curves for these patients detected no statistically significant difference.

Following this derivation study the Mayo group next reported on the ability of a slightly modified form of their original model, which they christened the Model for End-Stage Liver Disease, to estimate short-term mortality in four cohorts of patients with ESLD of diverse aetiology and severity.[24] The MELD score for each patient was determined by calculating their risk score using the original Mayo model formula and then, for ease-of-use, multiplying the result by ten and rounding it to the nearest integer. As with the derivation study, the primary outcome was survival status after 90 days of observation. The concordance statistic (c-statistic), which is analogous to the area under the ROC curve and therefore quantifies the discriminative performance of a prognostic model, was the summary, and sole, performance measure for this validation analysis. The first validation cohort consisted of 282 inpatients with ESLD admitted to the Mayo Clinic over the five year period, from January 1994 to January 1999. The causes of ESLD in this group included: alcohol-related cirrhosis 30 percent; chronic viral hepatitis 21 percent; and cholestatic liver disease in 15 percent. The median MELD score was nine (IQR: 4-14) and 90-day mortality, measured from day of hospitalization, was 20.9 percent in this group. The second validation cohort was comprised of 491 newly diagnosed cirrhotics followed by an ambulatory clinic in Palermo, Italy. Chronic viral hepatitis was the primary cause of ESLD in the majority of this cohort (89%). The median MELD score for this group was ten (IQR: 8-13) and the 90-day mortality was 6.9 percent. The third

validation cohort was composed of 326 outpatients with primary biliary cirrhosis (PBC) that were registered by the Mayo Clinic between 1973 and 1984. MELD scores were low in this group, with a median of 1 (IQR: -2-5), and only 1.5 percent of this cohort died within 90-days of registration. The fourth validation cohort consisted of 1,179 Mayo Clinic patients that were diagnosed with ESLD between 1984 and 1988, a time when liver transplantation was not as widely applied. The median MELD score in this group was seven (IQR: 3-13) and their 90-day mortality was 18.7 percent.

The c-statistics for the four cohorts ranged from 0.78 to 0.87 and are summarized in Table 1.1. Values in this range are reflective of acceptable to excellent discrimination.[27] The authors also analysed the contribution of the disease aetiology variable to the discriminative performance of the MELD score, citing several concerns about its inclusion in the model, the most compelling of which was that it introduced difficulties in applying the model when there was coexistent causes of chronic liver disease (e.g. chronic hepatitis C virus infection and alcohol-related cirrhosis). Results showed that the exclusion of the disease aetiology variable from the score did not produce any important difference in its discrimination (decreased c-statistic by 0.01 in cohort 1, increased by 0.02 in cohort 2, no change in cohorts 3 and 4). Analysis to explore whether the addition of variables to represent specific complications of portal hypertension, ascites, spontaneous bacterial peritonitis, variceal bleeding, and encephalopathy, augmented score discrimination was also performed. This disclosed minimal, if any, improvement in discrimination across the four cohorts (greatest c-statistic improvement 0.03).

Table 1.1: Discriminatory performance of the MELD in first validation study [24]

cohort	c-statistic (95%CI)
1. hospitalized ESLD, Mayo 1994-1999 (n = 282)	0.87 (0.82 – 0.92)
2. ambulatory non-cholestatic ESLD, Palermo 1981-1984 (n = 491)	0.80 (0.69 – 0.90)
3. ambulatory PBC, Mayo 1973-1984 (n = 326)	0.87 (0.71 – 1.00)
4. historical ESLD, Mayo 1984-1988 (n = 1,179)	0.78 (0.74 – 0.81)

In their summary of this first report on the MELD, Kamath et al stated that the score was "...a reliable measure of short-term mortality risk in individuals with ESLD of diverse aetiology and severity." and proposed that it be adopted in the United States of America as the disease severity index by which liver transplant candidates with ESLD be prioritized for allograft allocation.

1.6 Liver Allograft Allocation in the United States of America

Since the first liver transplants were performed in 1963 in Denver by Starzl et al the United States of America has been a leader in the development and application of liver transplantation.[28] In the late 1970s as liver transplantation emerged from its experimental phase and demand for allografts grew, individual transplant programmes banded together to form regional organ procurement organizations to increase recruitment of deceased donors, yet there were no national guidelines for organ allocation. Early regional allocation practices were adopted from renal transplantation and employed first-come, first-served prioritization. As has been discussed earlier, it quickly became recognized that although reasonable in renal transplantation, waiting time

was not appropriate for prioritization in liver transplantation. The passage of the National Organ Transplantation Act in 1984 was the first step in the development of a national organ procurement and allocation system, referred to as the Organ Procurement and Transplant Network (OPTN). In 1987 the United Network for Organ Sharing (UNOS) was mandated by the United States Department of Health and Human Services (USDHHS) to administer the OPTN and thus became responsible for all aspects of organ donation and transplantation in the United States of America.[18]

Initially UNOS continued with waiting time-based prioritization for candidates without an emergent indication. Subsequently however, in an effort to introduce a measure of medical need into the allocation process, the physical location in which the candidate was receiving care (e.g. intensive care unit, hospitalized, outpatient) was employed as a surrogate measure of medical acuity. While this measure permitted prioritization of hospitalized individuals with more severe ESLD it could not provide stratification for the majority of candidates with chronic liver disease that waited as outpatients. In this group waiting time was, once again, employed for prioritization. Additionally, potentially as a consequence of frustration with the emphasis on waiting time, it was observed that setting-of-care could be manipulated by clinicians to hasten a candidate's receipt of an allograft.

1.6.1 Application of the Child-Turcotte-Pugh ESLD Severity Index for Liver Allocation

In an effort to address the limitations of location-based prioritization on January 19th, 1998 the Child-Trucotte-Pugh (CTP) classification was adopted by UNOS as the disease severity index by which transplant candidates with pre-existing liver disease were prioritized (non-acute liver failure).[29-31] This classification system was originally

described by Child and Turcotte in 1964 as a means of estimating risk of post-operative mortality following porto-systemic shunt surgery.[22] Pugh et al made a minor modification to the index by altering one of the five risk index variables in 1973.[23] The CTP classification has become the most widely used disease severity index for patients with chronic liver disease. The classification system is based upon five variables: three objective laboratory values, serum total bilirubin, INR and serum albumin; and two subjective clinical findings, presence and severity of ascites and porto-systemic encephalopathy. Each of the five variables has three ordinal levels of assignment, scored one through three points, hence possible scores range from five to 15. Categorical CTP classification is based on total score, with scores five to six assigned to class "A", seven to nine to class "B" and individuals with scores of 10 to 15 assigned to the highest class of disease severity, "C" (Table 1.2).

Table 1.2: Child – Turcotte – Pugh score and classification

variable	1 point	2 points	3 points
total bilirubin (μmol/L)	< 35	35 – 50	> 50
albumin (g/L)	> 35	28 – 35	< 28
INR	< 1.7	1.7 – 2.3	> 2.3
ascites	none	controlled	refractory
encephalopathy*	0	I – II	III - IV

* Cited grading systems for encephalopathy vary. Pugh et al refer to grading system of Trey et al, which was based upon system of Adams and Foley.[23, 32-34] In recent practice West Haven criteria are employed, which are based on system of Parsons-Smith et al. [35, 36] (see Appendix A)

Although the CTP classification was a more objective measure of severity of ESLD than setting-of-care, its adoption failed to resolve the perceived limitations of the liver allocation system. Primarily this was due to the method in which the disease severity index was applied; however characteristics of the score were also implicated. Although the score was capable of generating an 11 level ordinal classification of disease severity it was applied using only three categories, mirroring the three levels of the CTP classification. A CTP score of seven was designated as the minimal listing criterion score and individuals with scores from seven to nine (CTP "B") were assigned a "status 3" designation, while individuals with scores of 10 to 15 (CTP "C") were given a "status 2" designation, further subdivided into "2A" if experiencing specified manifestations of acute decompensation or "2B" if stable. Although application of the CTP in this fashion permitted objective assignment into these three categories of disease severity it did not permit prioritization within categories. Once again, it was left to time on waiting list to assign order to individuals within each specific status category, and therefore waiting time continued to be a strong determinant of organ allocation.[21, 24] In addition to this method of application which constrained its discriminative ability, the CTP index had several characteristics which were criticised. The ordinal nature of the severity scoring system limited the potential for discrimination of severity by the three objective, laboratory-based, risk index variables. Conversion of these continuous measurements to three ordered categories results in a ceiling effect with loss of potentially important prognostic information by ignoring the implication of extreme values in the highest severity score category (e.g. total bilirubin 51 μmol/L scored the same as 500 μmol/L). Additionally, all five variables in the CTP score have arbitrarily been given equivalent

weight in the risk index even though the prognostic influence of the variables is very likely to differ (e.g. grade IV PSE is given a similar score as a serum albumin of 27 g/L although the implications of these two observations is likely quite different). Finally, and the most prevalent criticism of the CTP score, is its inclusion of the two clinical variables, the presence and severity of ascites and PSE, which, due to their subjective nature, are susceptible to variability, and even embellishment, in their assignment.[15, 24] In the end, although the incorporation of the CTP disease severity index into the UNOS allocation system was a step in the right direction, it failed to address many problems; in particular it failed to diminish the dominance of waiting time duration as a determinant of allocation priority.

1.6.2 Adoption of the MELD for Liver Allocation in the United States of America

By the late 1990s several developments firmly established that an allocation system based upon objectively defined medical need was required for liver transplantation. The acceptance that waiting time was an inappropriate determinant of prioritization was given an evidence-based foundation with the publication of two large cohort studies that documented poor correlation between length of time on waiting list and the risk of mortality while awaiting liver transplantation. [19, 37] Concurrently, to address the deficiencies of the liver allocation process in April of 1998 the USDHHS proposed the “Final Rule” stating that the role of waiting time in organ allocation strategy should be minimized and instead, organs should be allocated to transplant candidates based on medical urgency.[38] By late 1999 the “Final Rule” was enacted and thus established “...performance goals for bringing about: standardized criteria for placing patients on transplant waiting lists, standardized criteria for defining a patient’s medical

status, and allocation policies the make the most effective use of organs, especially by making them available whenever feasible to the most medically urgent patients who are appropriate candidates for transplantation."[20] In response to this directive UNOS created a Liver Disease Severity Score Committee that was tasked with identifying an alternative to the CTP score. In consideration of adopting the MELD for this role, UNOS performed a validation study of the MELD to evaluate its ability to predict 90-day mortality in a cohort of individuals with ESLD.[15] Using the data from 3,437 status 2A and 2B candidates assigned to the OPTN wait-list from November 1999 to December 2001, the discrimination of the MELD and CTP score was measured by the c-statistic. This group had ESLD of diverse aetiology, with chronic hepatitis C infection (36.4%) and alcohol-related cirrhosis (27.6%) accounting for the majority of the cases. The MELD score was calculated using the method described by Kamath et al and, as per those author's suggestion, the influence of the disease aetiology variable was omitted although the coefficient retained as a constant so that results could be compared to previous studies.[24] Additionally, to avoid negative values of the score, the lowest limit of the laboratory variables was set at one. With these two conditions in place the lowest possible value of the score would be six, while the upper limit remained unbounded. The mean MELD score for the cohort was 19.6. Within 90 days of wait-listing 1,040 individuals were transplanted, and therefore censored from the 90-day outcome analysis, while 412 died (12.0%) and 1,859 candidates were still alive and waiting at the end of the 90 day observation period. The c-statistic for the MELD score was 0.83 (95%CI: 0.81-0.84), reflective of good discriminative ability, and significantly superior to that of the CTP score, c-statistic 0.76 (95%CI: 0.74-0.79)($p < 0.001$). A compelling argument was

made by Heuman and Mihas, that the CTP was handicapped in this comparison by virtue of the validation cohort being limited to status two patients and therefore a limited range of CTP scores, from 10 to 15.[39] These correspondents produced results of their own validation analysis based on OPTN wait-list data from May 10 to November 30, 2001 which included 10,554 candidates of status 2A, 2B and 3 designations. In this more inclusive cohort the discriminative performance of both the MELD and CTP scores was found to be acceptable (c-statistics: 0.76 and 0.77 respectively) with no significant difference found between the two (p-value not provided). It is important to note that in both of these studies it was the performance of the CTP score that was evaluated, as opposed to the three level classification employed by UNOS for liver allocation, thus even in the restricted cohort it reflects the performance of the CTP in its more favourable application. In the end Wiesner et al clarified that although it could be debated as to whether the MELD had superior discrimination to the CTP, at worst it offered similar performance while eliminating subjective variables from the risk index.[39] Based on the favourable results of the study conducted by the UNOS Liver Disease Severity Score Committee, the OPTN Board of Directors approved the adoption of the MELD as the disease severity index by which all non-status one liver transplant candidates would be prioritized for allograft allocation on February 27th, 2002.[15] When employed for this purpose the methodology of score calculation utilized in the UNOS validation study was adopted with two further modifications. The first was the empirical establishment of the upper limit of serum creatinine at 4.0 mg/dl (353.6 mmol/L) for candidates with laboratory values exceeding that limit or those on haemodialysis. This was done to address concerns that application of the MELD might unduly favour transplantation of

individuals with renal failure, leading to more futile transplants as renal dysfunction is known to be associated with inferior post-transplant survival. Conversely, as a consequence of effective haemodialysis serum creatinine in those patients might be artificially low leading to underscoring, were a correction factor not in place. The second modification was put in place to address concerns regarding the potential for an allocation system based on the MELD disease severity index to increase the number of futile transplants. As was discussed in the section on basic principles of organ allocation, unchecked, an allocation system based solely on medical need will result in transplantation of very sick individuals that have marginal chance of long-term survival, thus wasting precious allografts. To counter this tendency, by consensus the MELD score was capped at 40. It was felt that this limit would not discourage the transplantation of critically ill individuals with reasonable hope of salvage yet not promote futile transplantation.[21, 40]

The end result of the modifications by Kamath et al and the UNOS conditions is a continuous disease severity scale for ESLD with 35 potential assignments from six to 40.[15, 24] With the introduction of the MELD the previous categorical status 2A, 2B and 3 designations became redundant and were amalgamated as “status 2”, while the “status 1” designation remained for individuals with de novo or post-transplant acute liver failure. Under the UNOS allocation system for adults (candidates ≥ 18 years of age) a deceased donor allograft was first offered to local then regional, blood group appropriate, status 1 candidates. If there was no status 1 candidate the organ is next offered to local then regional status 2 candidates ranked by descending order of MELD score. In the event that there were two local or regional candidates with the same score,

the organ was offered to the individual with the longest waiting time at that score or greater. Unlike the earlier allocation systems total waiting time was not used, and every time a candidate's score increased waiting time was reset to zero. If a candidate's score declined they retained the waiting time accrued at higher score levels. Finally, if the organ was not accepted at a local or regional level it was offered to national status 1 and subsequently status 2 candidates. With minor modifications that will be discussed subsequently, the UNOS allocation policy continues to follow this algorithm.[18, 41]

In the first year following the adoption of the MELD-based allocation policy by UNOS there was a significant reduction in the number of individuals placed on the waiting list for liver transplantation and listed candidates had much higher MELD scores at listing compared to the year prior to implementation.[42] These changes were felt to reflect the elimination of the influence of time on wait-list in allocation prioritization which had previously been a motivator to list patients earlier in their disease process. Despite the higher disease acuity of candidates there was a 3.5 percent absolute reduction in the number of patients removed from the waiting list due to death or becoming too ill to undergo transplantation ($p = 0.076$). Analysis at 18 months further quantified a relative reduction in wait-list mortality of 18 percent compared to the pre-MELD era.[43] Median time to transplantation was also found to decrease from 656 to 416 days.[44] An analysis of the UNOS liver transplant database, comparing the year leading up to the introduction of the MELD-based allocation system to the year following its introduction, found that despite MELD scores reflecting significantly higher pre-transplant disease severity post implementation (mean 20.5 vs. 17.0, $p < 0.001$) there was no significant difference in one-year post transplant patient or allograft survival.[45] The implication of this finding

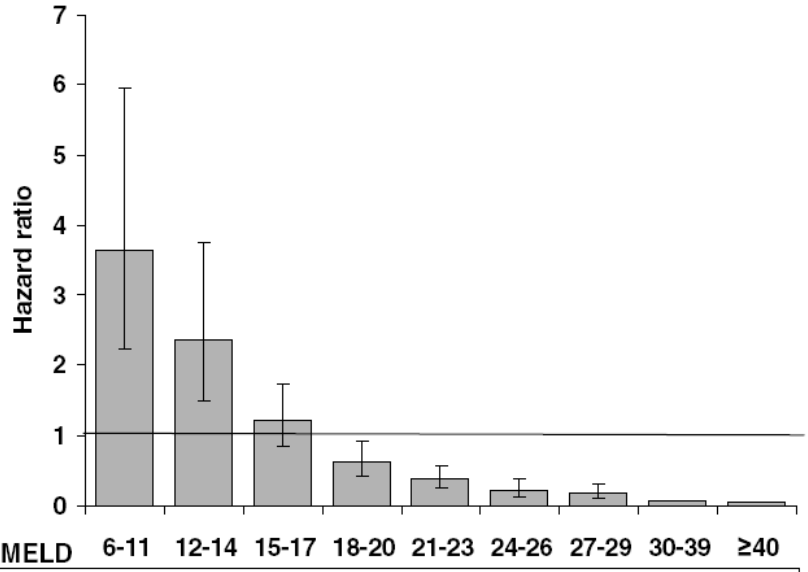
was that in serving to facilitate organ allocation to individuals with greater medical need, the new allocation system did not appear to lead to more futile transplants.

Although ESLD is the predominant primary indication for liver replacement there are several minority indications in which prognosis is determined by factors other than liver disease severity. Early developers of liver allocation policy in the United States of America recognized that ESLD severity indices such as the CTP and MELD would not be accurate metrics of medical need for these indications and established policies to permit exceptions to ESLD risk index-based prioritization. Some of these diseases have formal strategies under UNOS policy and are referred to as "Recognized Exceptional Diagnoses" (REDs), while others are dealt with on a case-by-case basis by regional review boards. The leading RED is hepatocellular carcinoma (HCC), which in the five years preceding the implementation of MELD-based allocation accounted for approximately five percent of the liver transplants performed annually in the United States of America.[46] Although 80 to 90 percent of individuals with HCC have underlying chronic liver disease, many have preserved liver function and consequently, have low MELD scores.[47] In these individuals it is the risk of progression and subsequent death from their primary hepatic malignancy that defines their medical need, not the severity of their underlying ESLD. In recognition of this, under the CTP-based allocation system patients with cirrhosis and HCC meeting criteria for transplantation (solitary tumour ≤ 5 cm in maximum diameter, or up to 3 tumours ≤ 3 cm in maximum diameter, with no evidence of vascular invasion or metastases [48]) were assigned a 2B status.[15] When the MELD-based allocation system was implemented, UNOS directed that candidates with stage I tumours (solitary tumour < 2 cm) would be assigned a MELD

score equivalent to a 15 percent risk of 90-day mortality, and those with stage II tumours (solitary tumour with a maximum dimension of > 2 cm but ≤ 5 cm, or up to 3 tumours with maximum dimension ≤ 3 cm) would be assigned a MELD score equivalent to a 30 percent risk of 90-day mortality from ESLD. Initially under these conditions the assigned scores for stage I and stage II candidates was 24 and 29 respectively. UNOS closely followed up on the impact of this policy and it was disclosed that these conditions overestimated the risk of progression and hence gave an unfair wait-list advantage to candidates with HCC. This lead to a marked increase in the number of transplants performed for this indication.[46] As a consequence of this observation the assigned MELD score for HCC was progressively titrated downwards so that eventually, by March of 2005, only candidates with stage II HCC were assigned a MELD score of 22, equivalent to a 6-month risk of mortality from ESLD of 15 percent.[43, 46, 49]

Another example of evidence-based adjustment of liver allocation policy by the UNOS is the "Share MELD 15" policy. This was implemented in 2005 and prompted by research by Merion et al that demonstrated, using data from the Scientific Registry of Transplant Recipients (SRTR) from September 2001 to June 2003, that the one-year survival benefit of liver transplantation was limited to individuals with MELD scores of 15 and greater (Figure 1.6).[50] This implied that for status 2 candidates with lower chronic liver disease severity (MELD < 15) that transplantation was more hazardous than remaining on the waiting list. Also of interest was the observation that patients with MELD scores less than 15 showed only gradual progression, if any, of their disease severity score over time, with only five percent progressing to a score associated with a significant survival benefit (MELD ≥ 18) in the one-year observation period of the study.

Figure 1.6: Hazard ratio for 1-year mortality by MELD-UNOS score for recipients of liver transplant vs. candidates on the liver transplant waiting list [50] (reproduced with permission from John Wiley & Sons)



MELD	6-11	12-14	15-17	18-20	21-23	24-26	27-29	30-39	≥40
Hazard Ratio	3.64	2.35	1.21	0.62	0.38	0.22	0.18	0.07	0.04
p-values	<0.001	<0.001	0.41	<0.01	<0.001	<0.001	<0.001	<0.001	<0.001

The authors suggested that, to improve the distribution of allografts to those of greater medical need, allografts should be offered to candidates with higher MELD scores over a larger geographic area before they were offered to individuals with lower MELD scores. Subsequent to this recommendation, the Share MELD 15 policy adjusted the geographic sequence in which deceased donor allografts were offered, so that livers were offered to local then regional candidates with MELD scores equal to or greater than 15 before they were offered to individuals with scores less than 15.[49]

At the present time the UNOS algorithm for allograft allocation of deceased donor livers offers the organ first to local then regional status 1 candidates; then to local then regional status 2 candidates with MELD scores equal to or greater than 15, in descending

score order; then to local or regional candidates with MELD scores less than 15, in descending score order; and finally to national status 1 then status 2 candidates, in descending MELD score order.[41]

1.7 Origins of the Sodium Augmented Model for End-Stage Liver Disease: The MELDNa

From the time of its first adoption by UNOS as a key component in liver allograft allocation strategy it was recognized that the MELD would not be static.[15] To that end several modifications to the MELD have been proposed over the years in an effort to refine its predicative performance. Augmentation of the MELD by incorporating the influence of low serum sodium concentration into the model has been the most studied modification. Hyponatremia had long been recognized as a negative prognostic indictor in individuals with cirrhosis. In the closing paragraph of a 1956 publication by the late Dame Sheila Sherlock it was stated that "In patients with liver disease, serum sodium levels below 130 mmol/L must be regarded as serious and, if below 125 mmol/L, ominous."[51] In 2004 Heuman et al were the first to report that hyponatremia was associated with short-term mortality in liver transplant wait-list patients, independent of MELD score.[52] Using prospective liver transplant waiting list data from a cohort of 507, predominantly male, patients enrolled in the Department of Veterans Affairs Health System, the investigators found that MELD score, serum sodium concentration less than 135 mmol/L and refractory ascites were independent predictors of 180-day wait-list mortality. Of particular interest, they found that the influence of hyponatremia and/or refractory ascites was limited to individuals with MELD scores less than 21 (the median MELD score for individuals that experienced 180-day mortality). In their conclusions the

authors emphasized that individuals with these factors formed a unique subset of patients at increased risk for mortality despite their lower MELD scores, and should receive expedited consideration for transplantation. Further studies from North and South America, and Europe confirmed the independent association of low serum sodium concentration with wait-list mortality.[53-56] Studies by Ruf et al and Biggins et al endeavoured in their analysis to assess for an interaction between low serum sodium concentration and MELD score, however failed to confirm the phenomenon which had been observed by Heuman et al.[54, 55] Each of the four studies also compared the discrimination of a risk model combining MELD score and hyponatremia to the MELD score alone. Three of the studies demonstrated small (absolute) improvements in the c-statistic ranging from 0.02 to 0.04, which were not statistically significant.[53, 55, 56] In the smallest study (n = 262) by Ruf et al a very small difference of 0.014 was found between the c-statistic for the MELD and that of a sodium augmented MELD model (respective c-statistics 0.894 and 0.908), which was reported to have statistical significance (p = 0.006), although the methodology of the comparison is unclear.[54]

The benchmark study investigating the impact of hyponatremia on liver transplant wait-list mortality was reported by Kim et al in 2008.[16] This study was based on analysis of data provided by the OPTN on all adults listed for first liver transplant for ESLD indication in the United States of America in 2005 and 2006. In those two years 14,130 individuals meeting the inclusion criteria were placed on the waiting list, of which 13,940 had complete data for analysis. The median MELD score at wait list enrolment was 15 (range 6 to 40) and 31 percent of the patients had hyponatremia (defined in this study as a serum sodium concentrations < 135 mmol/L). Cox regression analysis was

applied to the data from the 2005 cohort, consisting of 6,769 candidates, to analyse the association between the MELD score, hyponatremia and the occurrence of 90-day mortality. This confirmed MELD score and hyponatremia as independent predictors of 90-day mortality. An inverse linear relationship was disclosed between declining serum sodium concentration and the occurrence of 90-day mortality. With each one mmol/L decrease in the serum sodium concentration, over a range from 140 to 125 mmol/L, the relative hazard for mortality increased by five percent (Hazard Ratio: 1.05; 95%CI: 1.03-1.08; $p < 0.001$). Interaction analysis also confirmed the findings of Heuman et al, that the influence of hyponatremia was most pronounced in lower MELD scores. Subsequently, a new prediction model was derived incorporating MELD score and serum sodium concentration as independent variables, which had the acronym MELDNa assigned to it. The formula for this new model is:

$$\text{MELDNa} = \text{MELD} + (140 - \text{serum sodium mmol/L}) - [0.025 \times \text{MELD} \times (140 - \text{serum sodium mmol/L})],$$

where the serum sodium concentration is bound between 125 and 140 mmol/L. The MELDNa was then validated with the data of the 2006 calendar year cohort (n = 7,171) and compared to the performance of the MELD. The calibration performance of the MELDNa was reported as superior to that of the MELD, however both models showed statistically significant discrepancy between the observed incidence and estimated probability of mortality within 90 days of waiting list placement (MELD $\chi^2 = 51.3$; $p < 0.001$ and MELDNa $\chi^2 = 17.8$; $p = 0.023$). Discrimination of both the MELDNa and MELD was very good, with c-statistics of 0.883 and 0.868 respectively (confidence intervals not reported). It was reported that this 0.015 difference was statistically

significant ($p < 0.001$). Although it could be argued that this small performance difference is more of a statistical phenomenon than a clinically important one, the authors produced a provocative analysis that estimated that, had the MELDNa been used to guide liver allocation in 2006, 32 of the 477 wait-list deaths that year (7%) might have been prevented by more accurate prioritization and timely transplantation. Similar to the authors of the original MELD report, Kim et al proposed in their conclusions that the MELDNa should be trialed for liver transplant allograft allocation.[24]

1.8 Liver Allograft Allocation in Canada

Unlike the United States of America, in Canada there is no national system of organ donation and allocation. Provincial transplant organizations have the responsibility for donor organ procurement and, with the exception of high status candidates, organs are allocated to the regional centre providing transplant services to that province (i.e. centre-based allocation). A categorical classification system based on physical location and medical status of the candidate has been developed by the Canadian Liver Transplant Study Group and is referred to as the Canadian Wait-listing Algorithm in Transplantation (CanWAIT)(Table 1.3).[57] There is minimal peer-reviewed literature on this instrument and no published evidence to document that it has ever been validated in wait-list candidates. With the exception of an informal agreement between the seven liver transplant centres which facilitates nationwide allograft recruitment for critically ill patients (CanWAIT status 3F, 4, 4F), liver allografts procured in a transplant programme's catchment are allocated to candidates on that programme's waiting list. Each transplant centre manages its own waiting list and prioritization of candidates is at the discretion of the programme. The lack of a national system for organ donation and

Table 1.3: Canadian Wait-list Algorithm in Transplantation (CanWAIT)[57]

status	criteria
1	at home
1T	at home with a liver tumour
2	in hospital with stable condition
3	in intensive or equivalent care facility, but not requiring mechanical ventilation support, with either: • creatinine > 200 μmol/L or rising by > 50 μmol/L per day • grade 3 or grade 4 encephalopathy
3F	in intensive or equivalent care facility for fulminant liver failure, but not requiring mechanical ventilation support, who fulfils the King's College criteria* for high risk of mortality without liver transplantation
4	in intensive care requiring mechanical ventilation support; without transplantation, death is considered imminent
4F	in intensive care requiring mechanical ventilation support for fulminant liver failure, including non-function of a primary graft; without liver transplantation, death is considered imminent
0	on hold

*see Appendix B for King's College criteria for liver transplantation in acute liver failure [58]

transplantation in Canada has been implicated as one of the causes for a low rate of donorship (e.g. 12.8 per million persons vs. 21.5 per million persons in the USA in 2005)

and potentially the inefficient utilization of donor organs.[59, 60] To remedy this situation on August 12th, 2008 the Canadian government entered into an agreement with Canadian Blood Services (CBS) to develop a national organ donation and allocation system.[61] Although the details of this system will be a work in progress for the next several years, it has been proposed by members of the Canadian liver transplant community that an allocation system similar to that employed by UNOS be adopted.

1.8.1 Adoption of a MELD-Based Liver Allocation System in Canada?

A Canadian liver allocation system based on objectively defined medical need is highly desirable. As discussed in previous sections, a liver disease severity index such as the MELD, facilitates allocation prioritization by objectively defining the medical need of candidates. If it is to be utilized as a component of deceased donor liver allocation strategy in Canada, it is imperative that the MELD be validated in the population of Canadian patients with ESLD awaiting liver transplantation. Optimally, this validation would be done using a cohort drawn from a national wait-list database. Currently however, there is no national registry for liver transplant candidates to permit this analysis. As an initial solution to this deficiency, regional validation studies conducted by individual transplant centres using their own wait-list data, have an important role to play. In addition to contributing to the greater goal of nationwide validation, regional studies may permit detection of regional differences in model predictive performance and identify additional mortality risk factors. Differences in MELD score adjusted wait-list withdrawal have been documented between donor service areas in the United States of America and developers of the MELD concede that accuracy of the mortality estimates generated by the MELD may vary due to differences between populations to which the

score is applied.[40, 62] These differences may be attributable to geographically unique risk factors for wait-list mortality not accounted for by the MELD. Recalling the determinants of waiting list mortality discussed in section 1.3, examples of potential causes of regional differences in wait-list mortality include: a unique pattern of aetiology of ESLD; the increased prevalence of a specific comorbidity in candidates; or diminished access to specialist care or tertiary medical services. It has been recognized that if an allocation policy involves transfer of allografts between procurement regions, regional differences in the performance of a common disease severity stratification tool, such as the MELD, could disadvantage individuals residing in regions where the model underestimates wait-list mortality risk.[21] In this situation candidates suffer the double affront of having their risk of wait-list mortality understated by the prioritization tool while having diminished access to an allograft as there would be a net efflux of donor livers from their region. It is therefore imperative that the predictive accuracy of the MELD be characterized in the Canadian liver transplant wait-list population prior to it being adopted as a component of the national organ allocation system.

To this point in time, the only validation study of MELD for Canadian liver transplant wait-list patients was performed by Burak in 2005.[17] As a component of this study, the discriminative performance of the MELD to estimate 90-day mortality in a cohort of 320 patients wait-listed for liver transplantation in Alberta between January 1st, 1998 and December 31st, 2002, was evaluated. Unlike the major MELD validation studies, this cohort included individuals whose primary indication for liver replacement was acute liver failure and HCC in its principle analysis.[15, 16, 24] Additionally it needs to be acknowledged that 105 of the 320 candidates (32.8%) were CanWAIT status “0”

(inactive) at the time of wait list enrolment. This group of individuals, which experienced 15 of the 49 over-all mortality events (30.6%), although placed on the waiting list, may not have been capable of actually receiving a liver. The influence of this group on the final results was not clarified in the analysis. The results disclosed adequate discriminative ability of the MELD (area under ROC curve 0.79; 95%CI: 0.70-0.88), which was similar to the performance of the CTP score (area under ROC curve 0.78; 95%CI: 0.69-0.86), but significantly superior to the performance of CanWAIT status (area under ROC curve 0.59; 95%CI: 0.48-0.71). The calibration of the MELD was not evaluated in Burak's study. We performed a pilot validation study which suggested that the MELD and MELDNa significantly underestimated 90-day wait-list mortality in Atlantic Canadian patients awaiting liver transplantation (unpublished data). Despite this deficiency of calibration, discrimination was acceptable and survival analysis disclosed that the MELD and MELDNa scores were the most significant independent predictors of over-all wait-list mortality. The current study was devised to conduct a definitive independent external validation study of the MELD applied to the Atlantic Canadian liver transplant wait-list population.

Chapter Two: Background – Validating Prognostic Models

2.1 Validation

Validity is defined as a state of relative absence of bias or systematic error.[63] Validation of a prognostic model is the process by which its ability to generate unbiased estimates of the occurrence of interest in its target population is characterized. Validity, and the validation process, is commonly subdivided into "internal" and "external" components.[63-65]

2.1.1 Internal Validation

Internal validation addresses the soundness of the methodology used to derive the predictive model. This is framed by the intended purpose of the model and dependent on the process used to select the derivation cohort; the methods used to select and measure the exposure and outcome variables; and the integrity of the modeling process.[63] The cohort used to evaluate risk factors must be representative of the population to which the model is to be applied, and large enough to permit sufficiently powered multivariable analysis of all important independent variables. As a rule of thumb, a minimum of at least ten outcome events are necessary for each independent variable included in a multivariable regression model.[64, 66, 67] When selecting variables, it is important that all previously recognized and clinically important determinants are considered in the modeling process. Balance must be struck in determining the number of variables to be included in the model, as large models risk being over-fitted and too specific to the characteristics of the derivation cohort, while small models risk being under-fitted if important predictors are left out.[64, 65] The objective in the end should be to produce a model with a relatively small number of strong predictors of the outcome of interest. On

the other side of the model equation, it is important that the outcome, dependent, variable is clearly defined, justified (or widely accepted) and clinically relevant. Once the model is derived data splitting, cross-validation, or bootstrapping techniques can be used to analyze its performance. Data splitting involves dividing the study cohort into two groups prior to analysis. The first group, referred to as the "training" or derivation cohort, is used to develop the model and then the performance of the model is evaluated using the data of the second group, referred to as the "test" or validation cohort. The optimum method of data splitting is subject to debate. Historically, random splitting of the study cohort has been the most commonly used method and results in a training and test data sets that are similar, other than for chance variation. For this reason random validation is considered to be a weak method of internal validation. An alternative is to split the cohort in a non-random way, such as by different time periods, which results in greater variability between the derivation and validation cohorts and therefore offers a more rigorous evaluation of model performance. One of the obvious disadvantages of the data splitting method is that is reduces the sample size and power of both the derivation and validation cohort analyses. Statistical methods have been developed which permit utilization of the entire cohort in model derivation and internal validation and therefore offer a superior alternative to the data splitting method. Cross-validation involves performing multiple repetitions of the data splitting technique with cohort data to generate estimates of model overoptimism which can subsequently be used to adjust model performance measures. Although superior to simple data splitting cross-validation is computationally intensive and yields inferior results to the third statistical method for internal validation, bootstrapping. Bootstrapping involves multiple re-samplings with replacement from the

study sample and, similar to cross-validation, is used to generate measures of overoptimism, known as “shrinkage” factors. Shrinkage factors quantify over-fitting and can be used to adjust model coefficients and/or measures of model performance. This process counters the overoptimism that arises when a model’s fit is too specific to the characteristics of the derivation cohort.[64, 68] In the end, although requisite, internal validation provides only a (first) impression of a model’s performance and does not guarantee that a prognostic model will be acceptable for the purpose that it was intended. The most stringent test of a prognostic model is external validation.

2.1.2 External Validation

External validation of a prognostic model involves the quantification of its predictive accuracy in a new sample that differs in some way from the derivation cohort. The objective is to determine the degree to which the outcome estimates generated by the model are generalizable to broader populations. Generalizability is therefore the fundamental concept of external validation. The population that investigators wish to generalize a prognostic model to may be similar to that which the model was originally derived to target, or quite different (i.e. predicting mortality with the MELD in patients undergoing TIPS for decompensated ESLD [69] versus predicting mortality with the MELD in individuals undergoing liver resection [70]). Justice et al have defined two components of generalizability; reproducibility and transportability.[65] Reproducibility is a reflection of the degree to which the prognostic model is fit to true determinants of the outcome of interest rather than to random variability (noise) unique to the derivation cohort. Models generated in settings with low event rates per variable or that have not utilized more rigorous methods of internal validation, such as bootstrapping, may be

over-fit and hence fail to reproduce their originally reported levels of predictive accuracy when evaluated in subsequent cohorts. To an extent reproducibility can be evaluated by internal validation but external validation provides a more exacting assessment of this component of generalizability. Transportability, the second determinant of generalizability, is a reflection of the extent to which all relevant risk factors for the outcome of interest are included in the prediction model. Failure to include important independent predictors leads to an under-fit model. This may be due to simple omission from the modelling process or may be due to a sampling bias which results in the derivation cohort being homogeneous, or lacking, for an important risk factor, hence it goes undetected. Five types of transportability are discussed in detail by Justice et al; historical, geographical, methodological, spectrum, and (follow-up) interval.[65] Historical transportability is characterized by the maintenance of model predictive performance when evaluated in cohorts from different temporal periods. Geographical transportability is defined by the maintenance of prognostic accuracy when evaluated in cohorts from different locations. Methodological transportability requires that model estimates remain accurate when different definitions or measurement methods for component variables are used. Spectrum transportability reflects the model's ability to maintain its accuracy when evaluated in cohorts with different degrees of disease severity or patterns of disease aetiology. Finally, follow-up interval transportability refers to the maintenance of model prognostic performance despite outcome assessment after different intervals of follow-up.

As with any scientific hypothesis, establishing the external validity of a prognostic model is an ongoing process involving assessment of performance accuracy in

Table 2.1: Hierarchy of external validation for predictive systems[65]

level of validation	components of generalizability evaluated
1. prospective validation • new cohort • same investigators/site as derivation	• reproducibility • historical transportability
2. independent validation • new cohort • different investigators/site	• reproducibility • historical transportability • methodological transportability • spectrum transportability
3. multisite validation • aggregated cohort from multiple sites	• reproducibility • historical transportability • methodological transportability • spectrum transportability
4. multiple independent validations • multiple cohorts • multiple investigators/sites	• reproducibility • historical transportability • methodological transportability • spectrum transportability
5. multiple independent validations with life-table analysis • multiple cohorts • multiple investigators/sites • different outcome intervals	• reproducibility • historical transportability • methodological transportability • spectrum transportability • follow-up interval transportability

increasingly diverse applications. Justice et al has proposed a five-level hierarchy of external validation for prognostic models with successive levels imposing greater variability of the previously discussed determinants of generalizability to test the model (Table 2.1).[65] Although much of the focus of the validation process is on establishing the statistical soundness of the model, the ascent through the validation process also affords the opportunity to assess the clinical validity of the model. Clinical validity implies that the prognostic model produces useful estimates that facilitate making management decisions with patients in the real world. Establishing this property requires evaluation of the model performance from the point of view of the clinical objective that drove its original derivation.[64]

2.2 Performance Measures for Prognostic Models

Prognostic models, such as the MELD, generate an estimate of the probability that an individual will experience the outcome of interest within a specified period of time based on the status of their risk factor variables at the beginning of the observation period. Both logistic and Cox regression models can be utilized for this purpose. In the case of a logistic regression model the outcome probability can be calculated from the logit link function, while for a Cox regression the survival function for the specific time-point is used to determine the probability estimate.[64] In both cases the probability estimate for the outcome is continuous, with a range from zero to one, while the observed outcome is dichotomous, being either zero or one.[71] In the following section common methods for evaluating the accuracy of prognostic models for dichotomous outcomes will be discussed.

Accuracy describes the closeness of a measurement, or prediction, to the true value of the quantity, or outcome, of interest.[71] In clinical practice the estimates generated by a prognostic model are utilized for two purposes; to estimate risk of an outcome at an individual level (e.g. to guide the need for adjuvant therapy in oncology), or for classifying individuals by degree of risk (e.g. for risk stratification in allograft allocation). Mirroring these two clinical applications, the accuracy of prognostic models is typically characterized by two components, calibration and discrimination.[72] These two principle measures will subsequently be discussed in depth, followed by a brief discussion of accuracy scores, which have not been widely applied to prognostic models.

2.2.1 Calibration

Calibration quantifies how closely the probability estimates generated by the model are to the subsequently observed incidence of the outcome. As the observed incidence of the outcome for an individual is either zero or one-hundred percent, measures of calibration require the comparison of the observed outcome incidence in a group of individuals to the expected incidence, which is based on the aggregated probability estimates of the group. Comparison of observed and estimated values for an entire cohort can provide a crude measure of calibration of the model, while division of the cohort into multiple risk strata can permit statistical testing of overall goodness-of-fit and identify specific risk quantiles where model calibration is suboptimal.[72] The most commonly employed summary measure of calibration is the Hosmer-Lemeshow goodness-of-fit test. This test divides the cohort into subgroups based on deciles of estimated risk and uses the observed and estimated incidences from each subgroup to calculate the test statistic. The Hosmer-Lemeshow test statistic is distributed as χ^2 and

when applied for the external validation of a (fixed) model the degrees of freedom are equivalent to the number of subgroups.[27] The null hypothesis of the statistical test is that the model fits. As with any statistical test of hypothesis, the result of the Hosmer-Lemeshow test cannot definitely inform the investigator that the model is well calibrated; it just tells them that the degree of any lack-of-fit is not large enough to reject this null hypothesis. This caveat is particularly relevant given that the Hosmer-Lemeshow (and other) global goodness-of-fit tests have been found to be underpowered when sample sizes are small to moderate ($n < 400$).[27, 73] Although sub-grouping is commonly done by deciles-of-risk, other quantiles-or-risk can be employed for clinical or statistical reasons. For outcomes with recognized risk thresholds the use of clinically dictated risk quantiles based on these thresholds has been proposed as being more informative.[72, 74] The number of risk quantiles may also be dictated by the number of events in the cohort as it is generally accepted that a minimum of five outcomes in each subgroup is necessary.[27] Cohorts with low event rates may therefore require reduction in the number of quantiles to achieve subgroups with greater than five outcome events. An undesirable consequence of the grouping methodology required for Hosmer-Lemeshow goodness-of-fit testing is that the manner in which risk quantiles are formed has been found to have the potential to influence the results of this goodness-of-fit testing method.[73] Consequently, it may be advisable to conduct goodness-of-fit testing using several grouping strategies (i.e. by deciles and clinically relevant quantiles) before making conclusions regarding the adequacy of model calibration.

Although the Hosmer-Lemeshow goodness-of-fit test provides a method of summary statistical characterization of global model calibration it does not reveal focal

deficiencies in calibration across the range of model probability estimates. Useful diagnostic tools for this purpose include risk stratification tables and graphical methods, such as calibration curves and quantile-of-risk histograms.[75] Similar to the Hosmer-Lemeshow goodness-of-fit methodology, these tools involve comparison of observed to expected outcome frequencies in groups defined by model probability estimates (or risk scores) and permit the identification of risk estimate ranges where the model is under or over calibrated. Although the interpretation of these tabular and graphical methods is subjective, they serve as important adjuncts to statistical goodness-of-fit testing.

2.2.2 Discrimination

The other component of prognostic model accuracy that is frequently evaluated is discrimination. Measures of discrimination quantify the ability of the model to assign individuals to the appropriate outcome group. A model with favourable discrimination will tend to estimate higher probabilities for the outcome in individuals that experience the outcome than in those that do not. As such, it is sorting ability, rather than the absolute values of the risk estimates generated by the model, that determines discrimination accuracy. For a very basic prognostic model that estimates future outcome occurrence based on the status of a, single, binary predictor variable, discrimination accuracy can be evaluated by determining sensitivity, specificity, and total proportion correctly assigned from a simple two-by-two table. More complex prognostic models that attempt to separate individuals into two outcome groups based on an ordinal or continuous dependent (classification) variable require alternate methods to quantify their discrimination. Two analogous measures for this purpose are the area under the Receiver Operating Characteristic (ROC) curve and the Concordance statistic (c-statistic, c-index).

ROC analysis permits characterization of model sensitivity and specificity across the range of ordinal or continuous risk scores generated by a prognostic model. This is done by calculating serial sensitivities and specificities using each classification variable value as the threshold for assignment to the binary outcome status. The ROC curve is subsequently generated by plotting the sensitivity (proportion of true positives) versus one minus specificity (proportion of false positives) associated with each classification cut-point on the y and x axes, respectively. There are several methods for plotting the curve itself. A non-parametric ROC curve simply connects the plotted data points sequentially with straight lines, producing a jagged curve, while smoothed curves may be generated using kernel density functions or parametric methods.[72, 76] Non-parametric and parametric ROC curves will be discussed further.

The summary measure for ROC analysis is the area under the ROC curve. For the assessment of prognostic models the full area under the curve is the most common metric. There are some instances, predominantly in studies of diagnostic test accuracy, where partial areas, within clinically relevant sensitivity or specificity ranges, are of interest.[76] The area under the ROC curve can be interpreted in several ways. The most theoretically correct definition is that the area under the ROC curve is an estimate of the probability that if a pair of observations, one positive for the outcome of interest and the other negative, is randomly selected from a sample, the risk ranking derived from the prognostic model will be greater for the individual experiencing the outcome than for the individual not experiencing the outcome.[77] It can also be defined as the average sensitivity for all levels of specificity or the average specificity for all levels of sensitivity.[76] The area under the ROC curve can range from 0.0 to 1.0. An area of 1.0

indicates perfect discrimination (model assigns observations into appropriate outcome group 100% of the time), while an area of 0.5 indicates the instrument possesses no intrinsic discriminative ability (model assigns observations into appropriate outcome group 50% of time, same as random chance). Areas under 0.5 indicate sorting into outcome groups opposite of what is expected, implying that outcomes have been incorrectly coded. Discrimination can be empirically qualified as: "unacceptable", for areas under the ROC curve from 0.50 to 0.69; "acceptable", for areas of 0.70 to 0.7.9; and "excellent" for models generating predictions associated with areas under the ROC curve of 0.80 and greater (as will be discussed in the subsequent section, it is rare for prognostic models to have ROC areas greater than 0.9).[27]

The simplest method of estimating the area under the ROC curve is to connect the data points serially with straight lines and calculate the area using the algebraic rule for the area of a trapezoid. This type of curve, and estimation method, is referred to as the empirical or non-parametric method. Non-parametric ROC estimates impose no assumptions regarding the structure of the data and are therefore widely applicable, although they do have a tendency to underestimate ROC area when the number of threshold data points for the classification variable is less than five.[76, 78] The non-parametric area under the ROC curve is equivalent to the c-statistic, which is a two-sample Wilcoxon rank sum statistic (a.k.a. Mann-Whitney U statistic). This is calculated by ranking observations by their outcome predictions and then comparing ranks between individuals with and without the outcome. This equivalency of the empirical ROC area to these non-parametric statistics affords methods to estimate confidence intervals, most commonly by the method described by DeLong et al.[78] With the exception of a few

situations, these non-parametric methods for calculation of confidence intervals for the area under the ROC curve have been found to be robust. Obuchowski and Lieber performed a simulation study which found that the DeLong method for calculation of confidence intervals was acceptable when both outcome groups contained 30 or more observations and model ROC areas were less than 0.90.[79] As the area under the ROC curve increased over 0.90 greater sample sizes were required ($n \geq 150$). Bootstrap methods produced more appropriate confidence interval estimates when outcome groups were smaller than 30, although no single bootstrap methodology (percentile, *t*, or bias-corrected and accelerated interval methods) was consistently optimal across all simulation scenarios. Of note, the same study found that when used to compare two ROC curve areas the non-parametric method for calculation of the confidence interval of the difference was acceptable even when sample size was small ($n = 20$). A simulation study by Faraggi et al, which compared several non-parametric and parametric methods for estimation of ROC area, noted that although the non-parametric estimate was not always the best in specific situations, it was consistently a close second regardless of sample size or shape of distribution.[80]

Overall, the non-parametric, empirical, method is acknowledged as a robust and generalizable method for the estimation of the area under the ROC curve. However, as has been mentioned, there are situations where alternative methods may be more appropriate. For example, in cases where there are less than five threshold data points of the classification variable empirical ROC curve methodology tends to underestimate ROC curve area.[76, 78, 80] In these situations, parametric, smoothed curve, methods may be more appropriate. In general these methods are based on the assumption that the

distributions of the classification variable for the two outcome groups can be approximated by a known statistical distribution. The binormal method, which assumes a common transformation of the classification variable distribution for both outcome groups to a normal distribution, is the most frequently employed parametric method. The assumption of normality is somewhat lenient and the binormal method has been shown to be appropriate for most applications with the exception of poorly separated and/or complex classification distributions.[76, 80, 81] As such, the binormal parametric method would appear to be the most reasonable alternative when non-parametric methods are not felt to be appropriate.[80, 81]

2.2.3 Calibration versus Discrimination?

Although, intuitively, one would assume that a model with good calibration would also possess good discrimination, and vice versa, this is not necessarily the case. In general, a well calibrated model, by virtue of accurately estimating the risk of the outcome of interest, will tend to sort individuals into the appropriate outcome risk group. There is, however, a limit on model discrimination imposed by the distribution of risk in the population. In a provocative commentary, Diamond demonstrated by simulation that when risk is uniformly distributed within the population, a perfectly calibrated model can only achieve a area under the ROC curve of 0.83.[82] As the distribution of risk in the population becomes skewed higher ROC curve areas are attainable, but for a perfectly calibrated prognostic model are unlikely to exceed a maximum of 0.90 (unless the risk distribution is very strongly skewed, or bimodal).[83] On the other hand, methods used to measure discrimination, such as the area under the ROC curve and c-statistic, do not take the actual values of the probability predictions generated by the model into account

when estimating these measures. As a result, these measures of discrimination are completely insensitive to the calibration accuracy of the model. For example, if the outcome incidence in a validation cohort is ten percent, a model which generates an estimated probability of event occurrence of 90 percent for all individuals that experience the outcome and 89 percent for all individuals that do not, will have excellent (perfect) discrimination, despite being very poorly calibrated. These phenomena impose a trade off between calibration and discrimination which must be recognized when evaluating model performance.[82]

Given this situation, the intended purpose of the prognostic model should dictate the relative importance of calibration or discrimination accuracy when evaluating the model's performance. For models which are used solely to rank individuals by risk and do not use the risk estimates for further applications (e.g. as an intervention indication threshold), discrimination is the sole performance measure of relevance. As an example of this scenario, prior to the implementation of the "Share MELD 15" directive, the MELD score was employed purely for determining the ranking order of the individual for deceased donor liver allograft allocation, and the score associated probability estimates for 90 day mortality had no other management implication. This may explain why some of the early MELD validation studies did not report an evaluation of calibration as part of model assessment.[15, 24] Emphasis on discriminatory performance, sometimes at the expense of calibration, may also be due to the greater implications of deficiencies in discrimination during model derivation. Models that are found to have suboptimal calibration can be recalibrated; however for models that discriminate poorly there is no remediation.[68] It should also be pointed out that in some cases, due to the nature of the

instrument, discrimination is the only evaluable performance measure. The CTP prognostic tool is not based on regression methods and therefore renders a risk score that is not transformable to a unique risk estimate. Subsequently, assessment of the calibration performance of the CTP is not possible.

The prevalence of ROC analysis of discrimination in the assessment of prognostic models is in part as a result of the extrapolation of this technique from studies of diagnostic accuracy. It must be kept in mind that the objective of a diagnostic test is intrinsically different than that of a prognostic instrument. With a diagnostic test the outcome of interest is already present, or absent, and as such the ability of the tool to discriminate between observations with, and without, the outcome is of paramount interest. In the case of a prognostic model the outcome of interest has yet to occur, if at all, and subsequently the accuracy of the predictions of the future event (i.e. calibration) and the instrument's sorting ability (i.e. discrimination) are both relevant to the characterization of model performance.

In any scenario where a model's prediction is utilized to facilitate management decisions based on the relationship of the individual's risk estimate to an established risk threshold, calibration is an important aspect of model performance. An example of this would be in oncology, where an individual's model estimated probability of cancer occurrence, or recurrence, would prompt more, or less, intensive surveillance or treatment, based on its relationship to a "high", or "low" risk threshold. Similarly, a MELD score of 15 was suggested as an evidence-based threshold, at and above which the risk of death on the waiting list exceeds the risk of death associated with liver transplantation.[50] This threshold is now utilized by UNOS in its allocation policy and

therefore, in addition to ranking accuracy, the accuracy of the absolute risk estimates generated by the MELD are of importance (i.e. individuals with MELD scores ≥ 15 are eligible for allografts procured from an expanded geographic area, and therefore have a greater likelihood of receiving a deceased donor liver, compared to individuals with scores < 15).[49] Therefore, although discrimination is considered to be a requisite model performance component, for prognostic models accuracy of calibration is no less important.[72] In fact, in the clinical setting, where patients and their clinicians are primarily concerned about risk at the individual's level, not rank relative to others with the condition, model calibration is the most important component of performance.[83] Subsequently, a comprehensive evaluation of the accuracy of a prognostic model should include assessment of both calibration and discrimination.

2.2.4 Accuracy Scores

Initially developed to assess the accuracy of meteorological predictions, the Brier score has been proposed, and occasionally used, to characterize the accuracy of prognostic models.[64, 75, 84, 85] The score is the mean of the squared errors for the cohort (i.e. error calculated by subtracting the observed outcome, 1 = event, 0 = non-event, from the probability estimate) and therefore ranges from 0.0 to 1.0. Other than a score of 0.0 indicating perfectly correct prediction, and 1.0 perfectly incorrect prediction, the score lacks a user friendly interpretation (the square root of the score is the average error). The score can be decomposed into calibration and discrimination components, but again interpretation of these measures is not intuitive.[75, 86] Statistical tests have been developed to characterize the predictions of a single model, akin to goodness-of-fit, and compare between Brier scores from different models.[87, 88] The Brier score has yet to

gain popularity for the evaluation of prognostic models, potentially due to the obscure nature of the measure.

2.3 Comparing the Performance of Prognostic Models

The most prevalent method utilized in performance comparisons between prognostic models is comparison of the areas under the ROC curve (or c-statistic) for the respective models. From a statistical perspective, Obuchowski and Lieber have shown that non-parametric methods are appropriate, and preferred, for this purpose.[79] Area under the ROC curve analysis has the advantage that it can be applied to compare ranking instruments that differ markedly in terms of; predictor (independent) variables, classification (dependent) variable form (e.g. ordinal vs. continuous), and even the basic form of the instrument (e.g. categorical risk index vs. regression model). For example, ROC analysis can be used to compare the discriminatory accuracy of the MELD and the CTP scores despite the fact that one is a regression model that generates a continuous classification variable based on continuous independent variables and the other is a categorical risk index that generates an ordinal classification variable based on ordinal independent variables. Additionally, unlike many of the other methods which have been discussed, software for conducting this analysis is readily available, and user friendly. ROC analysis affords comparison of the discriminative performance between two models but, as has been discussed in the previous section, fails to address the other important component of model accuracy, calibration. Additionally it must be recognized that the area under the ROC curve can be relatively insensitive to efforts to improve model accuracy by the addition of new predictor variables that are associated with the outcome, unless the effect measure is very large.[83] Pepe et al found that adding an independent

variable with an odds ratio of three, indicating a strong association with the outcome, to a model had little impact on the area under the ROC curve.[89] Furthermore, a model which is already composed of strong predictors is unlikely to have its discriminatory performance significantly improved (i.e. increase in area under ROC curve) by the addition of a new, strongly associated, independent variable, particularly if this new variable is correlated with predictors already in the model.[72] Thus, if relying solely on improvements in area under the ROC curve to evaluate whether a prognostic model is augmented by the addition of new predictor variables, important improvements in model performance (i.e. calibration) may be missed.

Unfortunately there are few other commonly available methods for performance comparisons between fixed prognostic models (i.e. outside derivation and internal validation settings). Comparison of calibration between two models may be done by examining the Hosmer-Lemeshow goodness-of-fit χ^2 statistics, and related p-values. In this comparison the model with the larger p-value would empirically be deemed to have superior calibration, although statistical characterization of the significance of this comparison is not possible.[16, 85] As with the one model setting, subjective evaluation of calibration curves or quantile-of-risk histograms can augment this comparative analysis.

Risk stratification and reclassification tables are recent developments that endeavour to facilitate the comparison of prognostic models without and with a new predictor (i.e. nested models).[74, 83, 85] A comparison of the performance between the MELD and its recently derived serum sodium augmented modification, the MELDNa, is an example of a situation where this tool would be applicable. Proponents of this analysis

method caution that it is not appropriate for the comparison of non-nested models.[74] Methodology is not yet standardized, but in general involves cross tabulating observed and estimated outcome occurrence within clinically relevant risk strata for the standard model and the "new" model which contains the additional predictor variable. From these tables the calibration of both old and new models can be ascertained and patterns of risk reclassification examined. By documenting how the new model reclassifies risk for individuals that experience the outcome of interest, and those that do not, it can be determined if the model is bringing about potentially beneficial or harmful reassignment. A new model will only be beneficial if it systematically redistributes observations to more representative risk strata. To use an example from organ allocation, if a new model consistently reassigns individuals that experienced wait-list mortality to a higher risk score, and vice versa for those that survive, then it would be implied that the new model was superior to the old. Pencina et al have proposed two summary measures that are derived from reclassification tables and are based on this concept.[90] The Net Reclassification Improvement (NRI) and the Integrated Discrimination Improvement (IDI) are measures of discrimination and characterize model performance by evaluating the net or average direction of reclassification for individuals that experience the outcome and those that do not. The NRI reflects the net proportion of individuals benefiting (or harmed if negative) from the risk reclassification brought about by the new model and is applicable to scenarios where there are *a priori* meaningful risk categories. The IDI jointly quantifies the overall improvement (or deterioration if negative) in sensitivity and specificity associated with the new form of the model and does not require the existence of recognized risk thresholds (although its interpretation is abstract, being the change in

the difference between the average true positive proportion and the average false positive proportion brought about by the new model). Both measures are simple to calculate and allow tests of hypothesis that quantify statistically whether the change in classification associated with the new prognostic model is significant (H_0: NRI or IDI = 0).

Similar to the performance evaluation of a single prognostic model, the intended application of the prediction should dictate whether calibration, discrimination, or both, need to be contrasted when comparing the accuracy of prognostic models. There are several methods that permit the comparison of discrimination however options for comparisons of calibration accuracy are few. Hopefully the rapid proliferation of prognostic models in clinical medicine will be accompanied by the development of new methods to evaluate and compare the performance of these tools.

Chapter Three: Research Rationale, Questions and Objectives

3.1 Rationale

The Model for End-stage Liver Disease (MELD) is a prognostic model for End-Stage Liver Disease (ESLD) which generates a continuous disease severity score that reflects an individual's ESLD-related mortality risk based on the status of three objective laboratory variables. The MELD score has been shown to accurately predict the risk of 90-day mortality in patients awaiting liver transplantation in the United States of America and it has subsequently been utilized there to prioritize deceased donor liver allograft allocation since 2002. Presently in Canada, there is no formal national organ allocation system. This situation has been recognized as suboptimal and there is ongoing discussion as to whether a MELD-based allocation system for liver transplantation should be adopted in Canada. For this deliberation to be evidence-based it is important that the predictive performance of the MELD be characterized for the Canadian liver transplant wait-list population. In the absence of a national wait-list data base to facilitate a national validation study, regional external validation studies of the MELD have an important role to play in the evaluation of the MELD. The sole study thus far evaluated the discriminative performance of the MELD in a cohort of Western Canadians. The present study will expand the evidence base by comprehensively assessing the generalizability of the MELD prognostic model to the population of Atlantic Canadian adults with ESLD awaiting liver transplantation.

3.2 Research Questions

3.2.1 Primary Research Question

Does the MELD prognostic model accurately estimate the risk of 90-day wait-list mortality in Atlantic Canadian adults with ESLD awaiting liver transplantation? (i.e. Is the MELD generalizable to the population of adult liver transplant candidates with ESLD residing in Atlantic Canada?)

3.2.2 Secondary Research Questions

1. Does the recently derived modification of the MELD which incorporates serum sodium as an independent variable into the model, the MELDNa, accurately estimate the risk of 90-day wait-list mortality in Atlantic Canadian adults with ESLD awaiting liver transplantation?
2. Compared to the MELD, does the MELDNa more accurately estimate the risk of 90-day wait-list mortality in Atlantic Canadian adults with ESLD awaiting liver transplantation?
3. Does the re-introduction of the disease aetiology variable to the MELD, as originally described by the Mayo group, result in more accurate estimation of 90-day wait-list mortality in Atlantic Canadian adults with ESLD waiting for liver transplantation, compared to the MELD?
4. Compared to the Child-Turcotte-Pugh (CTP) disease severity index score, does the MELD more accurately estimate the risk of 90-day wait-list mortality in Atlantic Canadian adults with ESLD awaiting liver transplantation?

3.3 Research Objectives

3.3.1 Primary Research Objective

The primary objective of this study is to measure how accurately the MELD prognostic model estimates 90-day wait-list mortality in a consecutive cohort of adult liver transplant candidates with ESLD placed on the waiting list of the Atlantic Multi-Organ Transplant Programme (MOTP).

3.3.2 Secondary Research Objectives

1. To measure how accurately the MELDNa prognostic model estimates 90-day wait-list mortality in a consecutive cohort of adult liver transplant candidates with ESLD placed on the waiting list of the Atlantic MOTP.
2. To compare how accurately the MELD and MELDNa prognostic models estimate 90-day wait-list mortality in a consecutive cohort of adult liver transplant candidates with ESLD placed on the waiting list of the Atlantic MOTP.
3. To compare how accurately the MELD prognostic models without and with the disease aetiology variable estimate 90-day wait-list mortality in a consecutive cohort of adult liver transplant candidates with ESLD placed on the waiting list of the Atlantic MOTP.
4. To compare how accurately the MELD prognostic model and CTP disease severity index estimate 90-day wait-list mortality in a consecutive cohort of adult liver transplant candidates with ESLD placed on the waiting list of the Atlantic MOTP.

Chapter Four: Methods

4.1 Design

This is an independent external validation study of the MELD prognostic model which employed data from a consecutive cohort of adults with ESLD placed on the waiting list for first liver transplantation by the Atlantic MOTP over a five year period.

4.2 Setting

The Atlantic Multi-Organ Transplant Programme (MOTP) liver transplant programme is affiliated with the Capital District Health Authority (CDHA) and is based at the Victoria General Site of the Queen Elizabeth II Health Sciences Centre in Halifax, Nova Scotia. The programme serves individuals that reside in the four provinces of Atlantic Canada; Nova Scotia, New Brunswick, Newfoundland and Labrador, and Prince Edward Island. This catchment area has a total population of approximately 2.3 million, 46 percent of which reside in a census rural setting (2006 census data).[91] The Atlantic MOTP started carrying out liver transplantation on July 3rd, 1985. In May of 2001 the liver transplant service suspended its surgical activity due to the lack of a sufficient number of liver transplant surgeons. As a consequence of this closure, patients from the Atlantic Provinces being considered for liver transplantation had to travel to other centres, primarily London, Ontario, for transplant assessment and performance. With the arrival of a new liver transplant surgeon in July of 2004 the MOTP made preparations to reactivate the liver transplant service, and subsequently, on December 1st, 2004, the service reassumed responsibility for all aspects of adult liver transplantation for Atlantic Canada.

4.3 Data Sources

The validation cohort was identified from a prospective database maintained by the CDHA and Atlantic MOTP. This clinical database facilitates assessment and follow-up of all patients referred to the MOTP for consideration of liver transplantation. Demographic and health-related information is prospectively entered into the database by the administrative and clinical staff of the MOTP. Additionally, information on referral outcome, waiting list placement occurrence and outcome, and occurrence of mortality (wait-listed and transplanted patients) are recorded in the database. At the time of wait-list assignment individuals are informed and invited by MOTP clinical personnel (who are not involved research projects) to give written consent for their health information to be utilized for research purposes. Although the prospective database is the primary data source for the study sample its accuracy was cross checked and information augmented by retrospective review of the CDHA medical record and the records of weekly Atlantic MOTP multidisciplinary liver transplant service meetings.

4.4 Sample

This study utilized a convenience sample. As previously discussed, after a three and one-half year hiatus the Atlantic MOTP liver transplant service reassumed responsibility for all aspects of adult liver transplantation for Atlantic Canada on December 1st, 2004. Within the limitation imposed by this temporal constraint, the objective was to achieve a comprehensive, consecutive, validation cohort of adult liver transplant candidates representative of current clinical practice in Atlantic Canada.

4.4.1 Source Population

The study sample was drawn from the population of all individuals recorded in the prospective database as referred to the Atlantic MOTP for consideration of first liver transplant from December 1st, 2004 to November 30th, 2009. This study therefore reflects the first five-year experience of the current iteration of the Atlantic MOTP liver transplant service.

*4.4.2 Sample Inclusion Criteria**

The study sample was composed of all individuals meeting the following criteria:

1. Official documentation of active assignment to the waiting list for liver transplantation (i.e. the candidate had to actually be eligible for an allograft, if one was available).
2. Candidate for first (a.k.a. primary) liver transplant.
3. Age 18 years and older at the time of wait-list placement.
4. ESLD as primary indication for liver replacement. The MELD prognostic model estimates survival related to severity of ESLD. Wait-list survival of individuals with other indications, such as hepatocellular cancer, may be dictated by processes other than ESLD.
5. Informed consent to have data utilized for research purposes provided by individual, or designate.

*4.4.3 Sample Exclusion Criteria**

Individuals with the following criteria were excluded from the study sample:

1. Lack of formal documentation of active assignment to liver transplant waiting list.

2. Recipient of at least one prior liver transplant.
3. Age less than 18 years at the time of wait-list placement.
4. Indication for liver replacement other than ESLD (e.g. acute liver failure, hepatocellular cancer or other malignancy)
5. Lack of informed consent to utilize personal health data for research purposes.

*These inclusion and exclusion criteria are in keeping with previously published major validation studies of the MELD.[15, 16]

4.5 Data Collection and Contents

4.5.1 Measured Variables

Over 80 variables were collected for each sample observation from the previously mentioned data sources. The prospective Atlantic MOTP database provided the bulk of the demographic information and health information pertaining to the primary disease indicating liver replacement, as well as other liver diseases. Additionally the database was the principle source of information on timing and status of referral, wait-list, and mortality outcomes. All laboratory investigation results, and their date of assay, were extracted by retrospective review of the CDHA medical record, which also provided information on referring physician characteristics, medical comorbidities, manifestations of liver disease, and corroborating information pertaining to outcomes (referral, wait-list, and mortality). Regarding the laboratory investigations, the most complete set of results pertinent to the calculation of the MELD and CTP scores, with a sampling date closest to the date of active assignment to the liver transplant waiting list, was recorded. The records of weekly multidisciplinary liver transplant service meetings for the five year

period were also retrospectively reviewed to provide corroborating information regarding outcomes (referral, wait-list, and mortality).

4.5.1.1 Follow-up and Outcome Definitions

Wait-list observation began on the date of active assignment to the liver transplant waiting list. Wait-list outcome was followed up until March 1st, 2010 to permit a minimum 90-day period of waiting list observation for all patients in the cohort. Individuals on the Atlantic MOTP liver transplant waiting list are very actively followed by dedicated pre-liver transplant nurses that monitor the pre-transplant care of candidates and update the prospective database when any outcome events occur. As a result of this surveillance, waiting list outcome, and date of its occurrence, was known with absolute certainty for all individuals in the cohort.

The outcome of interest for this study was liver transplant waiting list mortality. Wait-list mortality, also referred to as wait-list failure, was defined as:

1. Death of the candidate while on the waiting list.
2. Withdrawal of the candidate from the waiting list as they were deemed to have become too ill to survive the transplantation process (i.e. terminally ill), due to progression or complications of their ESLD. This event will also be referred to as “delisting”.

The following defined non-failure outcomes:

1. Removal from the waiting list due to the occurrence of liver transplantation.
2. Candidate alive and still awaiting transplantation at the end of follow-up (March 1st, 2010).

3. Candidate withdrawn from waiting list as condition improved to the point that liver transplantation was no longer deemed to be necessary.
4. Candidate transferred to the wait-list of another liver transplant programme as they had permanently relocated to that region.
5. Individual's candidacy permanently revoked due to detection of psycho-social or comorbid medical condition that rendered them unsuitable for liver transplantation (e.g. substance abuse, lack of sufficient social support, non-compliance, refractory cardiac or pulmonary pathology, malignancy). Also, individuals that requested to be withdrawn from the waiting list as they no longer wished to undergo transplantation are included in this group.

For the candidates whose waiting list outcome met the criteria for wait-list failure the event time was measured at the date of death, for those that died while waiting, or the date of withdrawal from the waiting list, for those that were delisted (i.e. too ill to undergo transplantation). All other candidates had their observation time censored on the date of waiting list withdrawal or, for those still awaiting transplantation at the end of the follow-up period, March 1st, 2010.

The principle validation analyses carried out in this study are based on mortality status at 90 days post placement on the Atlantic MOTP liver transplant waiting list. Although the MELD has been shown to be predictive of ESLD-related mortality over periods of observation from one week up to one year, 90-day mortality is the metric utilized most commonly in liver transplant wait-list cohorts.[15-17, 24] One of the rationales cited for the selection of the 90-day reference point was that at the time of the first report on the Mayo model for ESLD 90 days was the average waiting time for liver

transplantation.[26] As the MELD and related models estimate probability of ESLD-related mortality, analysis of their predictive accuracy must be limited to candidates that have already experienced that failure outcome and those still at risk for it, after 90 days of wait-list observation. This means only observations that experienced wait-list mortality within 90 days and those still awaiting liver transplantation after 90 days (i.e. those with non-censored 90-day outcomes) were included in the validation cohort. Individuals that were transplanted within 90 days cease to be at risk of ESLD-related mortality once transplanted, and therefore cannot be included in the cohort used for the assessment of prognostic model performance. Additionally, observations withdrawn from the waiting list within 90 days of listing for reasons of improvement, transfer, or revocation, had to be excluded from the validation cohort as they cease to be systematically followed and therefore their outcome status at 90-days might not be know with absolute certainty for all patients (only 2 candidates were withdrawn, due to unsuitability, within 90 days).

4.5.2 Calculated Variables

The laboratory investigations required for the calculation of the risk scores for the three variations of the MELD models are; serum total bilirubin, serum creatinine, the International Normalized Ratio for prothrombin time (INR), and serum sodium.[16, 24, 26] As the MELD is based on the measurement of concentration in mg/dL units (as opposed to the molar unit of concentration measurement used in Canada), serum total bilirubin and serum creatinine were converted to mg/dL by dividing the molar concentration (μmol/L) by 17.1 and 88.4 respectively. In addition to three laboratory-based variables the MELD+D score includes disease aetiology as an independent variable. The status of this variable was defined, as per Malinchoc et al ("0" for alcohol-

related or cholestatic liver disease, i.e. PSC or PBC, “1” for all others), by the diagnosis recorded as the primary indication for transplantation.[26]

The laboratory variables required for the determination of the CTP score are serum total bilirubin, the INR, and serum albumin.[22, 23] The two clinical component variables of the CTP risk index are absence or presence, and severity of, ascites and portosystemic encephalopathy (PSE). The status of these two variables was determined by the retrospective review of the clinical record up until the date of wait-list placement. To be defined as present ascites had to have been detectable by physical examination. Encephalopathy was deemed as present when detectable by mental status exam at the time of assessment or when there was a clear cut history consistent with the diagnosis. Due to the retrospective and subjective nature of the measurement of these two clinical variables, other than “always absent” or “ever present”, grading of the severity of ascites and encephalopathy was not feasible. The implications of this limitation will be discussed further in the section on the calculation of the CTP score and classification.

As indicated previously, the most complete set of laboratory results with the closest temporal relationship to the date of active wait-list assignment, coupled with non-laboratory-based variables measured up until the time of wait-listing when applicable, was employed for the calculation of risk index scores. This study defines these scores as the measures of risk for future ESLD-related mortality at the time of waiting list placement and they are therefore the central focus of the analysis of this study.

4.5.2.1 Calculation of the MELD-UNOS score and associated probability estimate for 90-day mortality.

The MELD score was calculated using the formula defined by Kamath et al and the variable constraints implemented by the UNOS, which have been previously discussed.[21, 24] To reiterate, the UNOS modifications are: the omission of the disease aetiology variable, with the retention of the coefficient as a intercept; setting the minimum value for all variables at 1.0; setting the maximum value for serum creatinine at 4.0 for observations with values in excess of 4.0 mg/dL or individuals receiving haemodialysis; and capping the risk score at a maximum of 40.[21] After applying the constraints to the independent variables the MELD-UNOS risk score is calculated with the formula:

$$\text{MELD score} = (9.57 \times \ln(\text{creatinine mg/dl})) + (3.78 \times \ln(\text{total bilirubin mg/dl}) + (11.20 \times \ln(\text{INR})) + 6.43$$

The MELD-UNOS score is expressed as an integer and, as indicated, capped at a maximum value of 40. The score therefore ranges from a minimum of six to a maximum of 40.

The probability estimate for 90-day mortality related to each individual observation's MELD-UNOS score was calculated by entering the score into the survival function for 90-day mortality which was kindly provided by investigators at the Mayo Clinic College of Medicine in Rochester, Minnesota, that were involved in the development of the MELD and MELDNa.[92] The formula utilized for the MELD-UNOS-related estimate for 90-day mortality was:

$$P(\text{90-day mortality}) = 1 - 0.97191^{\exp(0.143842 \times (\text{MELD score} - 10))}$$

Thus, for each patient in the cohort a MELD-UNOS score and associated probability estimate for 90-day mortality was calculated.

4.5.2.2 Calculation of the MELDNa score and associated probability estimate for 90-day mortality.

The MELDNa score was calculated using the formula and method defined by Kim et al.[16] The two variables required for the calculation of the MELDNa score are the MELD-UNOS score and the laboratory value for serum sodium concentration. The value of serum sodium is bounded between a lower limit of 125 mmol/L and an upper limit of 140 mmol/L (values < 125 set to 125, values > 140 set to 140). After applying these constraints, the MELDNa is calculated with the formula:

$$\text{MELDNa score} = \text{MELD} + (140 - \text{serum Na}) - [0.025 \times \text{MELD} \times (140 - \text{serum Na})]$$

As with the MELD-UNOS score, the result is rounded to the nearest integer and ranges from six to 40.

The probability estimate for 90-day mortality related to each individual's MELDNa score was calculated by entering the score into the survival function for 90-day mortality which was kindly provided by investigators at the Mayo Clinic College of Medicine in Rochester, Minnesota.[92] The formula utilized for the MELDNa-related estimate for 90-day mortality was:

$$P(\text{90-day mortality}) = 1 - 0.99262^{\exp(0.209401 \times (\text{MELD score} - 10))}$$

4.5.2.3 Calculation of the MELD+D score and associated probability estimate for 90-day mortality.

To evaluate whether the re-insertion of the disease aetiology variable into the MELD enhanced the accuracy of the model, a MELD+D score was calculated. To do this the primary liver disease responsible for liver replacement was used to assign the status

of the disease aetiology variable which was coded as per the description of the investigators at the Mayo Clinic, Malichoc et al and Kamath et al.[24, 26] For individuals whose primary indication was alcohol-related cirrhosis or cholestatic liver disease (i.e. PSC or PBC) the disease aetiology variable was coded as "0" (unexposed), while for all other indications the variable was coded "1" (exposed). In order that the comparison reflect only the influence of the addition of this variable, the same constraints applied to the three laboratory-based independent variables used in the calculation of the MELD-UNOS score were applied when calculating the MELD+D score with the following equation:

$$\text{MELD+D score} = (9.57 \times \ln(\text{creatinine mg/dl})) + (3.78 \times \ln(\text{total bilirubin mg/dl}) + (11.20 \times \ln(\text{INR})) + (6.43 \times \text{disease aetiology})$$

As with the MELD-UNOS score the result is rounded to the nearest integer. Again, to maintain consistency with the methodology used in the calculation of the MELD-UNOS score, the MELD+D score was capped at a maximum of 40. The MELD+D score therefore has a potential range of zero to 40.

The probability estimate for 90-day mortality associated with the MELD+D score was calculated using the same survival function that was used to generate the MELD-UNOS probability estimate, defined above (section 4.5.3.1).

4.5.2.4 Calculation of the CTP score and classification.

As has been previously discussed in section 1.6.1, the CTP disease severity index is comprised of three laboratory-based variables, serum total bilirubin, serum albumin and the INR, and two clinical variables, ascites and portosystemic encephalopathy.[22, 23] Each of these variables is assigned an ordinal severity score, from one to three, based

on the observed value's relationship to its respective index threshold-based ranges. The CTP score is the sum of the scores for each of the five variables, and can therefore range from five to 15. Based on the CTP score, the CTP classification in turn assigns observations to one of three ordered severity categories: "A" for scores five to six; "B" for scores seven to nine; and "C" for CTP scores of ten and greater.

For the laboratory-based variables CTP severity score assignment was straightforward and based on values from the most complete set of laboratory results with the closest temporal relationship to the date of active wait-list assignment. As alluded to earlier, score assignment for the clinical variables was hampered by their subjective nature, inconsistent reporting, and the retrospective method of their measurement. Due to these limitations, measurement of ascites and PSE was restricted to "always absent" or "ever present", up until the time of waiting list assignment. Although this permitted assignment of "always absent" observations to the lowest severity score rating for ascites (none) and PSE (West Haven grade 0), it did not permit differentiation of "ever present" observations into the middle or highest severity categories. As a compromise to this limitation, in order to permit the calculation of a CTP score and classification, "ever present" observations for ascites or PSE were assigned a score of 2.5 (instead of 2 or 3). It is recognized that this methodology may diminish the discrimination of the CTP score, but it also highlights the difficulty inherent with subjective variables.

Unlike the regression-derived MELD prognostic models, the CTP is a disease severity index and a method to derive survival estimates from its risk scores has not been developed. The evaluation of the calibration accuracy of the CTP is therefore not applicable.

4.6 Statistical Analysis

All statistical analyses were performed using Stata/IC 11 software (StataCorp LP, College Station, Texas). To supplement ROC analytic capabilities, package st0154 was installed from http://www.stata-journal.com/software/sj9-1.[93]

4.6.1 Descriptive Statistics

For continuous variables, unless otherwise specified, the median is used as the measure of central location, and the range and/or interquartile range (IQR) as the corresponding measure of variation. Categorical variables were expressed as proportions.

The Kaplan-Meier estimator of the survival function (a.k.a. the product limit estimator) was employed to describe censored time-to-event data.[94]

4.6.2 Inferential Statistics

For comparisons between continuous variables the non-parametric Wilcoxon rank sum test was used (a.k.a. Mann-Whitney U test). An exception to this was for the comparison between the mean model estimated probability of 90-day wait-list mortality and the observed incidence of 90-day wait-list mortality, for which a one-sample *t* test was used. For comparisons between proportions Pearson's χ^2 test was utilized, unless a cell contained less than five observations, in which case Fisher's exact test was used.

All tests of hypothesis were two-sided and a less than five percent probability of type I error was the threshold for significance.

4.6.3 Validation Analysis

The general principles of validation and accuracy measures for prognostic instruments were discussed in detail in Chapter Two. The two principle components by

which the accuracy of a prognostic model is characterized are calibration and discrimination.

4.6.3.1 Calibration

Three methods were used to evaluate the calibration of the MELD models. The first involved comparing the mean of the probability estimates for 90-day mortality of the entire cohort to the observed incidence. As indicated above, a one sample *t* test was used for this comparison with the null hypothesis that the mean estimate equalled the observed incidence of 90-day mortality in the validation cohort. The Hosmer-Lemeshow goodness-of-fit test was the second method utilized, and provides statistical quantification of global model calibration. Typically, this analysis entails dividing the cohort into ten subgroups (i.e. deciles-of-risk) however due to the small size of the validation cohort this was not felt to be appropriate. Instead, the cohort was divided into four quartiles-of-risk risk based on MELD score. When the Hosmer-Lemeshow goodness-of-fit test is applied for the testing of fixed models the degrees of freedom is equivalent to the number of risk quantiles used in the analysis.[27] Therefore, in this application, the degrees of freedom for the test would be equal to four. Risk stratification tables and quantile-of-risk histograms were the third methodologies employed to provide characterization of model calibration across the full range of observed risk predictions. Two grouping strategies were used for these tabular and graphical methods of calibration analysis; by quartile-of-risk, and, as suggested by Cook and Janes et al, by clinically relevant risk quantiles.[72, 74] For the former, the MELD-based quartiles created for Hosmer-Lemeshow goodness-of-fit testing were used, and for the later, the cohort was divided into three subgroups based on model probability estimates for 90-day mortality. As the reported incidence of

90-day mortality in large MELD validation studies ranges from six to 12 percent, five and ten percent thresholds were used to divide the cohort into three risk strata.[15, 16] Although somewhat arbitrary, the intent of these cut-offs was to create “low”, “intermediate” and “high” risk groups for 90-day waiting list failure. As will be discussed subsequently, these clinical risk strata also facilitate comparisons between models via the creation of reclassification tables.

Measurement of calibration accuracy quantifies the degree to which the model generated predictions approximate the observed outcome, Criteria for acceptable model calibration were pre-specified as:

1. No evidence of a significant difference between the mean model-estimated probability of 90-day wait-list mortality and observed incidence of 90-day waiting list mortality for the cohort.
2. No evidence of significant lack of fit by Hosmer-Lemeshow goodness-of-fit testing (i.e. failure to reject the null hypothesis, no lack-of-fit).
3. Favourable model estimate approximation of observed outcome incidence across all risk quantiles, when subjectively analyzed by tabular and graphical methods (subjective).

4.6.3.2 Discrimination

Model discrimination was characterized by measurement of the area under the non-parametric ROC curve. This assumption-free method for the estimation of the area under the ROC curve is acknowledged as robust and generalizable when the classification variable has five or more potential values.[76, 78, 80] In light of small sample size and low event occurrence (bias-corrected and accelerated (BcA)) bootstrap confidence

intervals for the area under the ROC curve were calculated using 1,000 replications. For illustrative purposes, non-parametric confidence intervals were also calculated using the method of DeLong et al.[78]

Measurement of discrimination accuracy quantifies the degree to which the model, via its predictions, assigns observations to their appropriate outcome group. Based on recommendations by Hosmer and Lemeshow, the criteria for acceptable model discrimination were pre-specified as:

1. Area under the ROC curve equal to 0.80 or more, defined as excellent model discrimination accuracy.
2. Area under the ROC curve from 0.70 to 0.79, defined as acceptable model discriminative accuracy.
3. Area under the ROC curve of 0.50 to 0.69, defined as unacceptable model discrimination accuracy.[27]

4.6.4 Comparisons of Accuracy between Predictive Models

Secondary objectives involve comparisons between the MELD-UNOS, the MELDNa, the MELD+D, and the CTP score. As discussed in section 2.3, calculation of the difference between the areas under the ROC curve, the Net Reclassification Improvement (NRI) and the Integrated Discrimination Improvement (IDI) are all metrics that permit the comparison of discrimination between two instruments (the later two for nested models only). Methods for comparisons of calibration accuracy are less prevalent.

Comparisons of discrimination accuracy will be performed by comparing the area under the ROC curve for the respective MELD models and the CTP disease severity index. Obuchowski and Lieber have shown that non-parametric methods are appropriate,

and the preferred method, for calculating the confidence interval for the difference between areas, even when sample size and event incidence is small. Never-the-less, 1,000 replication bootstrap confidence intervals will also be calculated to provide supporting evidence for the analysis.[79]

The MELDNa and MELD+D both contain the baseline MELD-UNOS variables and are augmented by an additional variable (serum sodium for MELDNa, disease aetiology variable for the MELD+D). They are therefore considered to be "nested" models. Consequently, comparison of the χ^2 statistics derived from the Hosmer-Lemeshow goodness-of-fit tests for the respective models can give an impression of the fit of one model relative to the other (i.e. the model with the lower χ^2 value has relatively better fit). However, unless one model is found to be significantly lacking fit while the other is not, the statistical significance of a difference cannot be determined.

As with one model analysis, subjective evaluations of quantile-of-risk histograms were used to explore differences in the calibration profile between models.

Risk stratification and reclassification tables were employed to compare the nested MELDNa and MELD+D models to the MELD-UNOS. Accompanying these tabular tools the NRI and the IDI statistics, and their associated hypothesis tests, were calculated per Pencina et al.[90]

As previously indicated, all tests of hypothesis were two-sided and a less than five percent probability of type I error was the threshold for significance.

4.7 Ethical Considerations

Prior to initiation of this study ethical review and approval was obtained from the Capital Health Research Ethics Board of the Capital District Health Authority, Halifax,

Nova Scotia (REB#: CDHA-RS/2009-075) and the Conjoint Health Research Ethics Board of the University of Calgary, Calgary, Alberta (Ethics ID: E-23065).

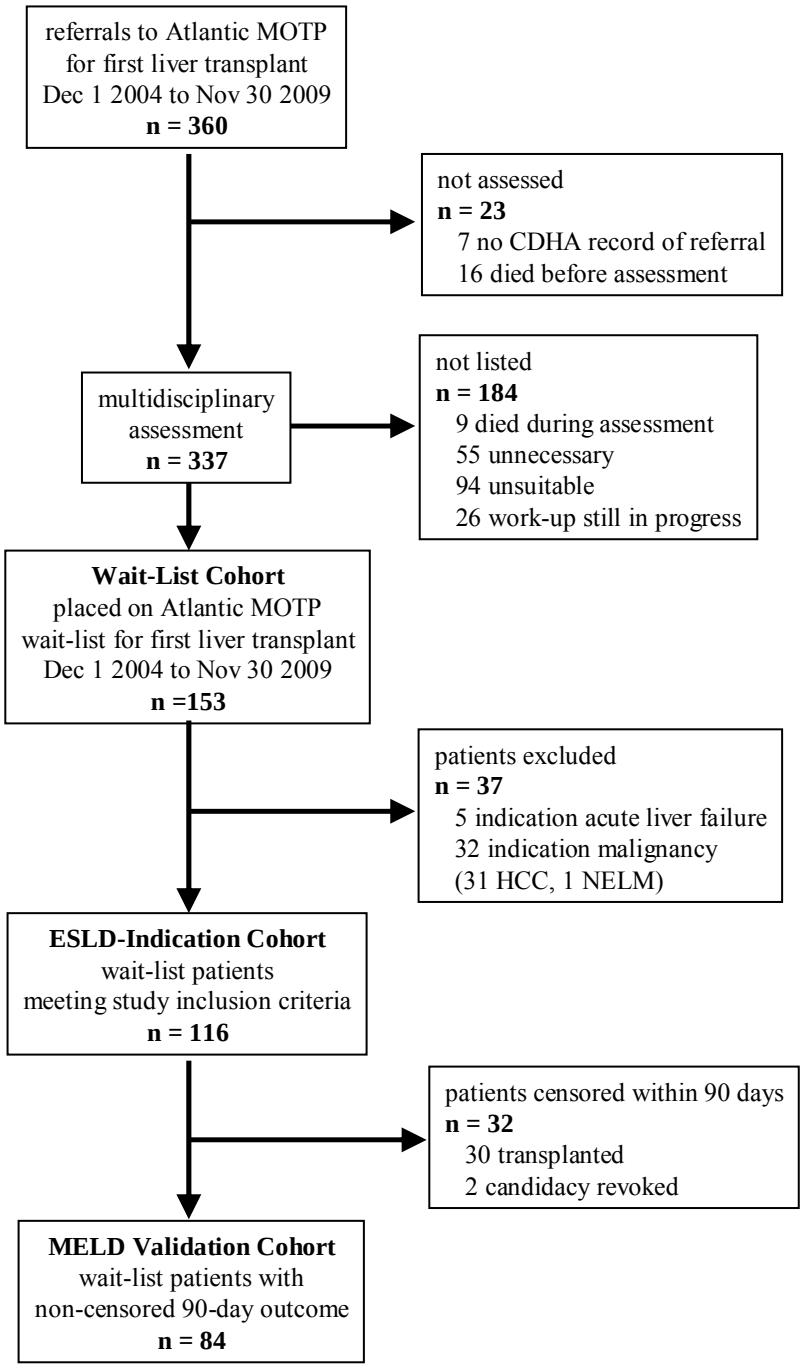
Chapter Five: Results – Description of the Study Sample

5.1 The Atlantic MOTP Liver Transplant Referral Population

After a three and one-half year hiatus the Atlantic MOTP liver transplant service was reactivated on December 1st, 2004. Between that date and December 1st, 2009 360 individuals were recorded in the prospective Atlantic MOTP database as having been referred for consideration of first liver transplant (see Figure 5.1). For seven of these entries there was only rudimentary information in the database and there was no evidence in the CDHA medical record to indicate that a formal referral was ever made. Additionally, 16 individuals died before liver transplant candidacy assessment was initiated (see Appendix C for a discussion of these individuals). Consequently, 337 individuals entered into the multidisciplinary liver transplant candidacy assessment process. 59.1 percent of the individuals assessed were male and 40.9 percent were female. The median age at referral was 56.1 years (range: 16.3 – 72.3). Four patients were under 18 years of age (ages 16.3, 16.4, 17.5 and 17.7 years). Roughly paralleling the relative populations of the four Atlantic Provinces, the province of residence was recorded as: Nova Scotia for 53.8 percent of the referrals, New Brunswick, for 28.2 percent, Newfoundland and Labrador for 13.8 percent, and Prince Edward Island for 4.2 percent (2006 census proportions: NS 40.0%, NB 32.0%, NL 22.1%, PE 5.9%).[91] The three leading primary liver diseases were: alcohol-related ESLD, 20.5 percent; hepatocellular cancer (HCC), 16.6 percent; and ESLD due to chronic hepatitis C virus (HCV) infection. 14.0 percent. The cholestatic liver diseases, primary biliary cirrhosis (PBC) and primary sclerosing cholangitis (PSC), accounted for 18.1 percent of the

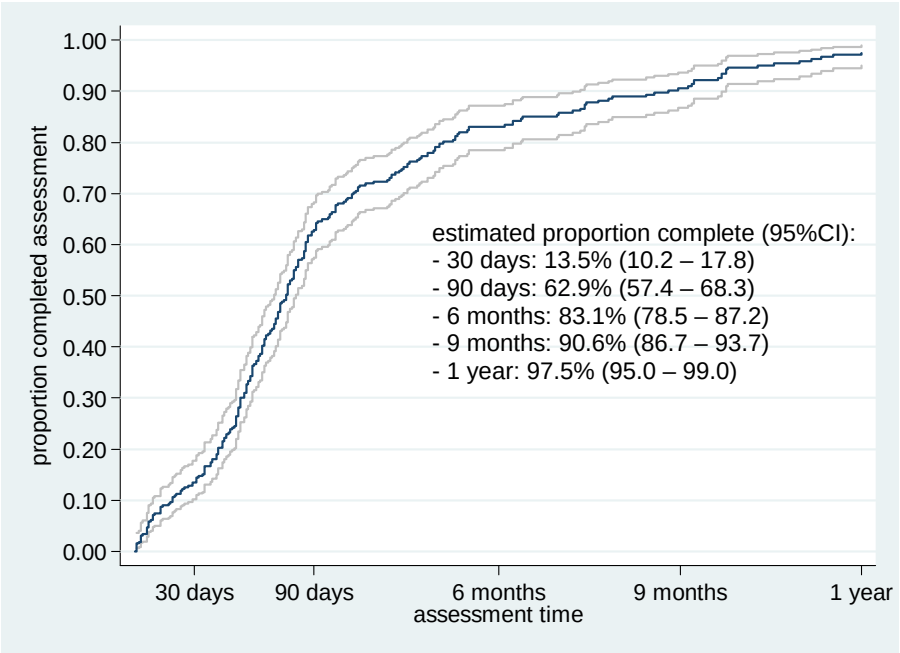
assessed referrals (10.1% and 8.0% respectively). Eleven individuals were referred with acute liver failure (3.3%).

Figure 5.1: Study cohort derivation



Of the 337 individuals that began multidisciplinary transplant assessment, 302 completed the process, nine patients died during assessment (see Appendix C), and for 26 assessment was still ongoing as of November 30th, 2009. The Kaplan-Meier estimated median assessment time was 76 days (95%CI: 70 – 81))(see Figure 5.2). After assessment, liver replacement was not felt to be necessary for 55 individuals (18.2%) and 94 individuals (31.1%) were deemed not to be suitable transplantation candidates due to psychosocial or medical reasons. Ultimately, of the 302 individuals that were referred and subsequently completed the multidisciplinary transplant assessment process of the Atlantic MOTP between December 1st, 2004 and November 30th, 2009, 153 individuals (50.7%) were deemed to be suitable candidates and were therefore assigned to the waiting list for liver transplantation.

Figure 5.2: Kaplan-Meier estimate of time to completion of multidisciplinary liver transplant candidacy assessment by Atlantic MOTP



5.2 The Liver Transplant Wait-List Cohort

One-hundred and fifty-three individuals referred between December 1st, 2004 and November 30th, 2009 had been placed on the waiting list for first liver replacement by November 30th, 2009 (see Figure 5.1). The first date of wait-list assignment for this cohort was February 16th, 2005. There were six additional patients placed on the waiting list for primary liver transplant during this time frame that were excluded from this study as they were referred prior to December 1st, 2004 (referred 1 day, 16 days, 27 days, 5 months, 6 months, and 14 months prior to December 1st, 2004).

5.2.1 Wait-List Cohort Characteristics

Characteristics of the 153 wait-list candidates are summarized in Table 5.1. Slightly over one-half of the cohort was male (54.9%). While exactly half of the ESLD indication subgroup were male, there were proportionately more males (78.1%) in the group of individuals whose primary indication was hepatic malignancy and more females (80.0%) in the group whose primary indication was acute liver failure ($p = 0.001$). The greater proportion of males in the malignant indication subgroup was principally accounted for by the large number of males with HCC as a sequelae of chronic HCV infection-related cirrhosis, which alone accounted for half of the 32 patients in this primary indication group. The median age of the cohort at the time of wait-list placement was 56.0 years (range: 21.6 – 69.0). Age at wait-listing did not differ between males and females ($p = 0.302$). With a median age of 34.5 years (range: 30.5 – 44.4), the five candidates whose primary indication was acute liver failure were significantly younger than the remainder of the wait-list cohort (median: 56.2 years (range: 21.6 – 69.0)) ($p = 0.001$). The majority of candidates resided in Nova Scotia (53.6%). Approximately one-

Table 5.1: Wait-list cohort characteristics

variable – measure	observation (total n = 153)
sex – frequency (proportion):	
male	84 (54.9%)
female	69 (45.1%)
age in years at wait-listing – median (range)	56.0 (21.6 – 69.0)
male	56.1 (26.9 – 68.3)
female	53.4 (21.6 – 69.0)
province of residence – frequency (proportion):	
Nova Scotia	82 (53.6%)
New Brunswick	38 (24.8%)
Newfoundland & Labrador	27 (17.7%)
Prince Edward Island	6 (3.9%)
rural postal code – frequency (proportion)	38 (24.8%)
referring physician specialty – frequency (proportion):	
hepatology	59 (38.6%)
gastroenterology	54 (35.3%)
transplant surgery	20 (13.1%)
general internal medicine	10 (6.5%)
general practitioner/family medicine	6 (3.9%)
other	4 (2.6%)

continues on next page

variable – measure	observation (total n = 153)
indication for liver transplantation – frequency (proportion):	
decompensated ESLD	116 (75.8%)
primary or secondary hepatic malignancy	32 (20.9%)
acute liver failure	5 (3.3%)
primary liver disease – frequency (proportion):	
hepatocellular cancer	31 (20.3%)
alcohol-related ESLD	23 (15.0%)
HCV-related ESLD	22 (14.4%)
primary biliary cirrhosis	19 (12.4%)
primary sclerosing cholangitis	17 (11.1%)
cryptogenic ESLD	10 (6.5%)
autoimmune hepatitis	9 (5.9%)
non-alcohol-related steatohepatitis	7 (4.6%)
polycystic liver disease	4 (2.6%)
acetaminophen-related acute liver failure	2
other ESLD (α-1 antitrypsin deficiency, congenital hepatic fibrosis, cystic fibrosis, HBV-related ESLD, Wilson's disease)	1 each
other acute liver failure (HBV, green tea, aetiology unknown)	1 each
other hepatic malignancy (neuroendocrine liver metastases)	1

continues on next page

variable – measure	observation (total n = 153)
symptoms of decompensated ESLD – frequency (proportion):	
at least one of any of the below	114 (74.5%)
ascites	104 (68.0%)
spontaneous bacterial peritonitis	18 (11.8%)
hepatorenal syndrome	5 (3.3%)
portosystemic encephalopathy	66 (43.1%)
portal hypertensive gastrointestinal haemorrhage	43 (28.1%)
at least one medical comorbidity – frequency (proportion)	82 (53.6%)
cardiovascular comorbidity – frequency (proportion):	44 (28.8%)
hypertension	34 (22.2%)
atherosclerotic heart disease	6 (3.9%)
pulmonary comorbidity – frequency (proportion):	18 (11.8%)
asthma	7 (4.6%)
chronic obstructive pulmonary disease	6 (3.9%)
renal comorbidity – frequency (proportion):	9 (5.9%)
chronic kidney disease	5 (3.3%)
diabetes mellitus – frequency (proportion)	41 (26.8%)
active cigarette smoking	45 (29.4%)
Body Mass Index – median (range)	27 (18 – 42)
obesity (i.e. BMI ≥ 30) – frequency (proportion)	52 (34.0%)

continues on next page

variable – measure	observation (total n = 153)
ABO blood group – frequency (proportion):	
O	64 (41.8%)
A	64 (41.8%)
B	18 (11.8%)
AB	7 (4.6%)

quarter (24.8%) of the cohort resided in an area that had a rural postal code designation. Approximately one-half of the patients (51.7%) were referred by a hepatologist or a transplant surgeon and slightly over one-third (35.3%) were referred by a gastroenterologist. Very few patients had been referred directly from a primary care physician.

Decompensated ESLD was leading indication for liver replacement and accounted for three-quarters of the cohort (75.8%). Hepatic malignancy, almost exclusively HCC, was the second commonest indication (20.9%), while acute liver failure accounted for only five of the 153 candidates (3.3%). The top three primary liver disease indications were: hepatocellular cancer (20.3%), alcohol-related ESLD (15.0%), and ESLD due to chronic HCV infection (14.4%). The overall prevalence of HCV-related liver disease, as a primary or secondary diagnosis, was 26.8 percent. Primary biliary cirrhosis and primary sclerosing cholangitis were the fourth and fifth leading primary disease indications, and when considered as a group, cholestatic liver disease accounted

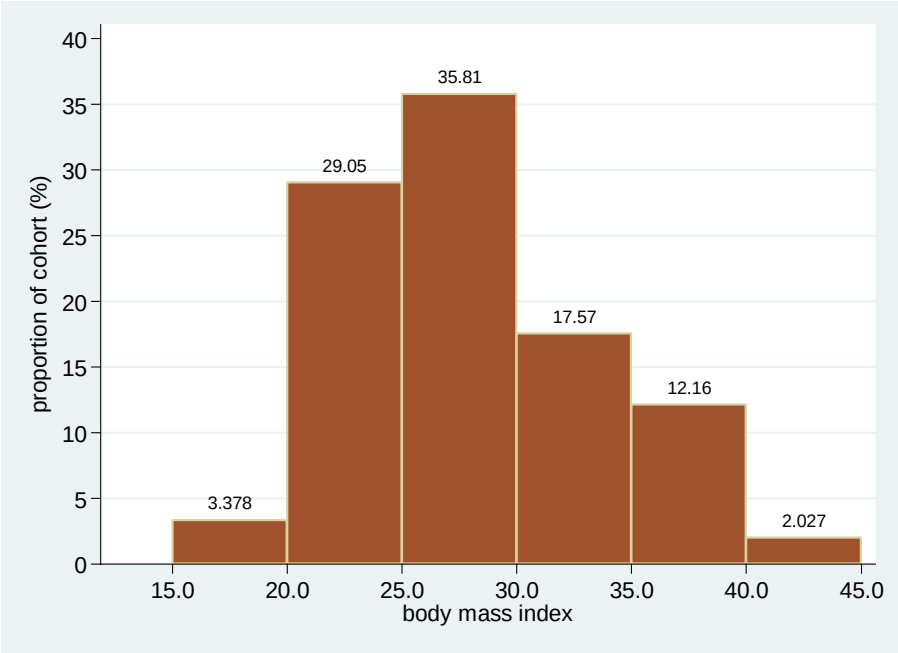
for 23.5 percent of the total wait-list cohort. Thirty-one of the 32 patients in the malignant indication subgroup had HCC (96.9%). In 19 of these individuals (61.3%) HCV infection was the cause of the underlying chronic liver disease. Alcohol-related cirrhosis was the, distant, second commonest aetiology of underlying chronic liver disease in patients requiring liver transplantation for HCC, accounting for 12.9 percent.

Almost three-quarters of the cohort (74.5%) were documented to have experienced symptoms of decompensated ESLD by the date of their waiting list assignment. In 21 of the 39 individuals that were not reported to have symptoms related to decompensated ESLD, HCC was the primary disease indication for transplantation (PSC accounted for eight others and acute liver failure for five). Ascites was the commonest symptom of decompensated ESLD, and was present to some extent in 91.2 percent of symptomatic individuals. Eighteen of the 104 patients with ascites (17.3%) had experienced at least one episode of spontaneous bacterial peritonitis, five had hepatorenal syndrome (4.8%), and four individuals had an associated hepatic hydrothorax. Portosystemic encephalopathy was the second most common symptom of decompensated ESLD, present in 57.9 percent of symptomatic individuals. Over one-third of the 114 symptomatic individuals (37.7%) had experienced at least one episode of portal hypertensive gastrointestinal haemorrhage prior to their listing for transplantation. Six candidates had a history of previous portosystemic shunting (two surgical, four TIPS). The indications for shunt placement were: ascites in three patients, portal hypertensive gastrointestinal haemorrhage in two, and undefined for one individual.

Just over one-half of the cohort (53.6%) had at least one medical comorbidity (excluding active cigarette smoking). The commonest single comorbidity was diabetes

mellitus, which was present in 26.8 percent of candidates. With a prevalence of 22.2 percent, hypertension was the second commonest specific comorbidity. Symptomatic atherosclerotic heart disease was rare (3.9%). Five candidates had mild to moderate pulmonary hypertension. In general, pulmonary disease was infrequent (11.8%) with asthma being the commonest reported pulmonary comorbidity (4.6%). It should be noted however that 29.4 percent of the cohort were active cigarette smokers at the time of their transplant assessment. Five individuals (3.6%) had chronic kidney disease of which three were on haemodialysis. The median body mass index of the candidates was 27 (range: 18 – 42). Slightly over one-third of the cohort was obese (34.0%), defined by a body mass index (BMI) of 30 and greater (see Figure 5.3).[95]

Figure 5.3: Body mass index (BMI) of Atlantic MOTP liver transplant wait-list cohort (n = 153)



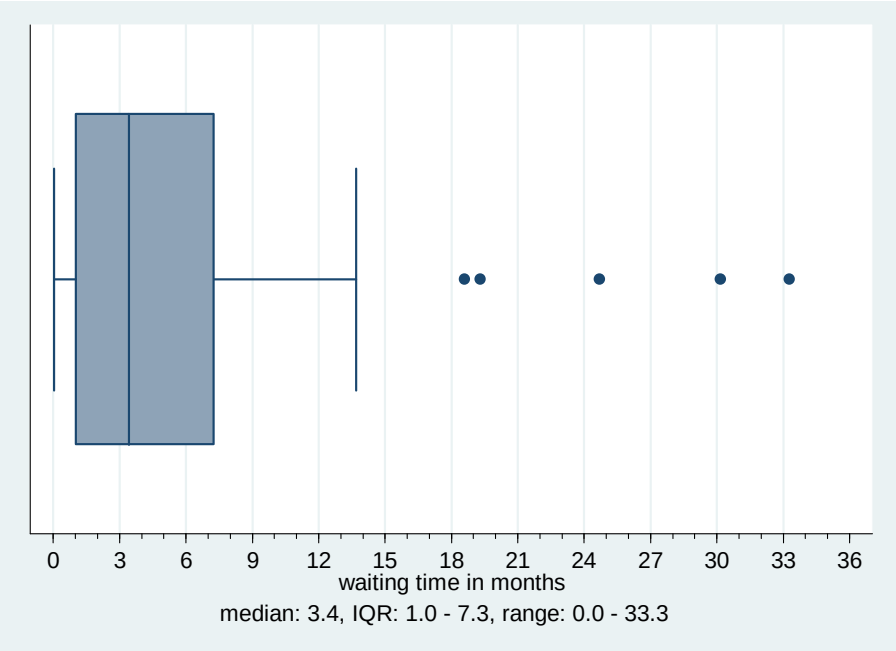
5.2.2 Wait-List Cohort Outcomes

The wait-list cohort was followed until March 1st, 2010. Information pertaining to wait-list outcome was complete for all 153 candidates, and is summarized in Table 5.2. By the end of follow-up only nine of these 153 patients (5.9%) were still awaiting transplantation. The median waiting time for this group of nine was 752 days (range: 340 – 1839). Two other candidates (1.3%) moved away from Atlantic Canada, and were therefore transferred to the care of other liver transplant programmes, after 144 and 235 days of waiting time. The remaining 142 individuals all experienced a definitive wait-list outcome. The condition of 13 patients (8.5%) improved to the point that liver transplantation was no longer deemed to be necessary and they were therefore withdrawn from the waiting list. In all of these patients the primary indication for liver replacement had been decompensated ESLD. Another 11 individuals (7.2%) had their candidacy revoked. Seven of these patients were found during their wait-list follow-up to have non-hepatic medical comorbidities that rendered them unsuitable for liver transplantation; two withdrew consent for transplantation; one had candidacy revoked due to non-compliance and ongoing substance abuse; and one was found 34 days after listing to have an HCC which was more extensive than pre-listing investigations had reported (i.e. under-staged). Happily, the majority of the cohort, 103 candidates (67.3%), received a liver transplant after a median waiting time of 104 days (range: 1 – 1,012)(see Figure 5.4). The date the first individual from the cohort received a liver transplant was May 31st, 2005. By primary indication, the incidence of liver replacement was: 62.9 percent (73 of 116 candidates) for ESLD, 80.0 percent (4 of 5) for acute liver failure, and 81.3 percent (26 of 32) for individuals with hepatic malignancy.

Table 5.2: Overall outcomes for total wait-list cohort (n = 153)

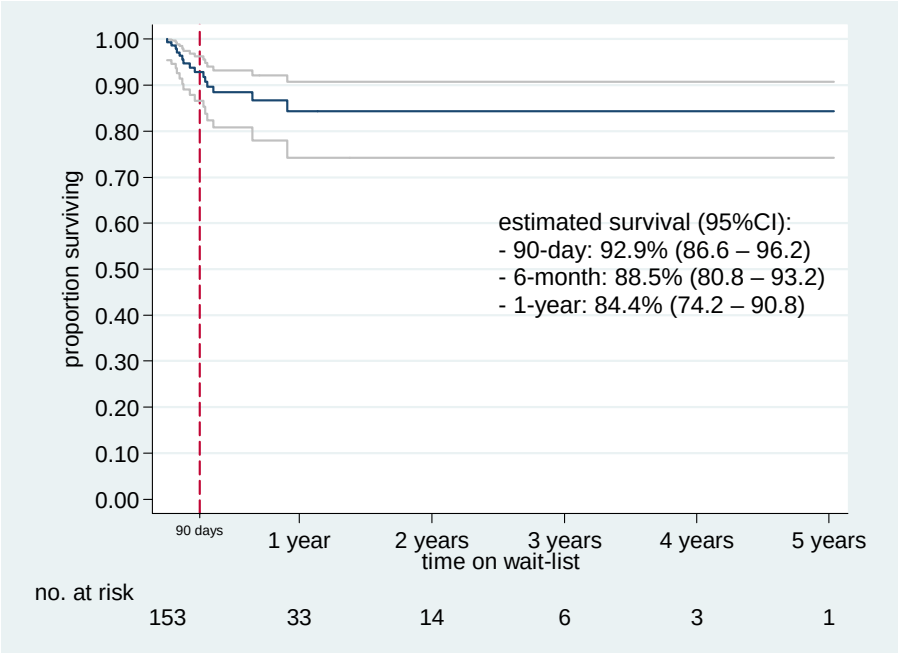
outcome	incidence frequency (proportion)	time-to-event in days median (range)
still awaiting transplant	9 (5.9%)	752 (340 – 1,839)
transferred	2 (1.3%)	144 and 235
withdrawn – unnecessary	13 (8.5%)	554 (122 – 1,510)
withdrawn – unsuitable	11 (7.2%)	236 (29 – 1,112)
transplanted	103 (67.3%)	104 (1 – 1,012)
died	7 (4.6%)	34 (1 – 128)
withdrawn - terminal	8 (5.2%)	70 (13 – 332)
total	153 (100.0%)	128 (1 – 1,839)

Figure 5.4: Duration of waiting time for the 103 individuals, from the wait-list cohort of 153, that received a liver transplant



Of the 153 liver transplant candidates, seven (4.6%) died while waiting for an allograft and eight (5.2%) were withdrawn from the waiting list as they became too ill, due to progression or complications of their liver disease, to survive transplantation. The overall occurrence of wait-list failure for this cohort was therefore 9.8 percent (95%CI: 5.0 – 14.6). The corresponding incidence rate for overall wait-list failure is 14.5 per 100 person-years at risk (95%CI: 8.7 – 24.0). Figure 5.5 describes the estimated wait-list survival experience for the cohort. The Kaplan-Meier estimate for 90-day and one-year wait-list failure was 7.2 percent (95%CI: 3.8 – 13.4) and 15.7 percent (95%CI: 9.3 – 25.8).

Figure 5.5: Kaplan-Meier estimate of liver transplant waiting list survival for wait-list cohort (n = 153)



Within 90 days of placement on the waiting list 49 of the 153 candidates (32.0%) had received a liver transplant, three (2.0%) had their candidacy revoked (due to

detection of severe pulmonary hypertension in one, dementia in another, and extensive HCC in a third candidate), and nine patients (5.9%) died (n = 4) or were delisted (n = 5) as they became too ill to tolerate transplantation. The remaining 92 patients (60.1%) were still awaiting a donor allograft after 90 days of waiting. The incidence of 90-day wait-list mortality for the cohort of 153 candidates was therefore 5.9 percent (95%CI: 2.1 – 9.7).

5.2.2.1 Patterns of overall mortality in the total wait-list cohort

Patterns of overall wait-list mortality were explored via non-parametric univariable comparative analysis of variables between the group of candidates that experienced wait-list failure and those that did not. The results of these analyses are summarized in tabular form in Appendix D.

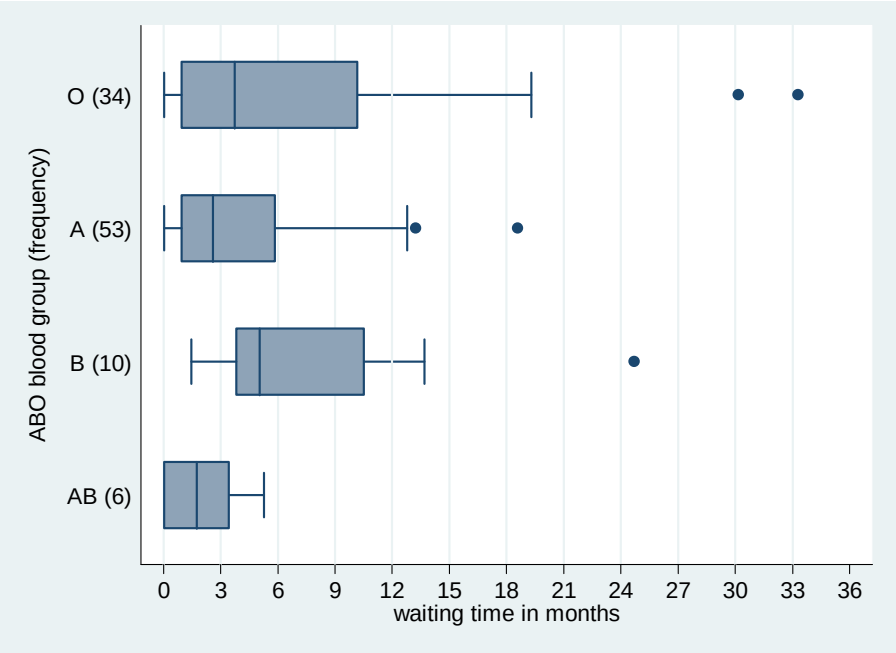
The risk for wait-list failure did not appear to differ between males and females (10.7% vs. 8.7%). Age at the time of wait-list registration did not significantly differ between individuals that survived, median 56.0 years (range: 21.6 – 69.0), and those that did not, median 54.8 years (range: 26.4 – 68.3). Although the smallest of the four provinces, candidates residing in Prince Edward Island appeared to be relatively overrepresented in the occurrence of wait-list failure. Residence in an area served by a rural postal designation did not appear to be associated with wait-list failure (risk ratio: 1.1; 95%CI: 0.4 – 3.3, p = 1.00).

No significant associations were found between the presence of general or specific medical comorbidities and waiting list mortality. The body mass index of survivors and failures were not found to differ, and the presence of obesity was not significantly associated with wait-list failure (risk ratio: 1.7; 95%CI: 0.7 – 4.4, p = 0.275). Cigarette smoking was quite prevalent in the cohort but did not appear to be

associated with mortality while awaiting transplantation (risk ratio: 1.6; 95%CI: 0.6 – 4.2. p = 0.343).

ABO blood group showed a significant association with waiting list mortality. Blood group O individuals constituted 41.8 percent of the cohort but accounted for 12 (80.0%) of the 15 cohort failures. The observed risk of wait-list failure by respective blood group was: O 18.8 percent; A 3.1 percent; B 5.6 percent; and AB 0.0 percent. Relative to the three other blood groups, candidates with blood group O were 5.6 times more likely to experience mortality while awaiting transplant (95%CI: 1.6 – 18.9, p = 0.002). Potentially influential in this phenomenon, blood group O individuals had the lowest incidence of transplantation in the cohort at 53.1 percent (A 82.8%; B 55.6%; AB 85.7%), and for those that were transplanted, the second longest median waiting time (see Figure 5.6).

Figure 5.6: Waiting time for 103 individuals from cohort of 153 that received a liver transplant by ABO blood group



The relative incidence of wait-list failure was greatest in the group of five individuals whose primary indication was acute liver failure, where there was one mortality (20.0%). The decompensated ESLD and malignant indication subgroups experienced similar relative incidences of mortality of 9.5 percent (11 out of 116) and 9.4 percent (three out of 32), respectively. Although nine individuals with autoimmune hepatitis accounted for only a small proportion of the entire cohort (5.9%), with three wait-list failures they accounted for 20.0 percent of the mortality events. With two and three failures respectively, individuals with cryptogenic cirrhosis and HCV-related ESLD were also slightly overrepresented in relative mortality. Conversely, individuals whose primary liver pathologies were alcohol-related ESLD or cholestatic liver disease (i.e. primary biliary cirrhosis or primary sclerosing cholangitis) were significantly underrepresented with respect to wait-list mortality. This observation is interesting given that it concurs with the first report on the MELD, where chronic liver disease not related to alcohol or cholestatic liver disease was a predictor of mortality in patients with ESLD, prompting its inclusion as a variable in the original Mayo ESLD risk model.[26]

The presence of symptoms of decompensated ESLD did not appear to be consistently associated with wait-list mortality. A notable exception was the occurrence of spontaneous bacterial peritonitis, which was significantly associated with wait-list failure. Five of the 18 candidates that had spontaneous bacterial peritonitis died, and they accounted for one-third of the 15 wait-list failures (risk ratio: 3.8; 95%CI: 1.4 – 9.7, $p = 0.006$).

5.2.3 Wait-List Cohort Risk Scores

Complete laboratory data for calculation of the MELD prognostic scores was available for all 153 candidates (Table 5.3). A serum albumin result was missing, and therefore precluded determination of the CTP score and classification, for ten patients. Median time from the acquisition of the laboratory investigations to date of waiting list assignment was two days prior (range: 102 days prior to 35 days after). 40.5 percent of patients had their risk score laboratory tests obtained within seven days of listing; 71.9 percent within 14 days, and 91.5 percent within 30 days.

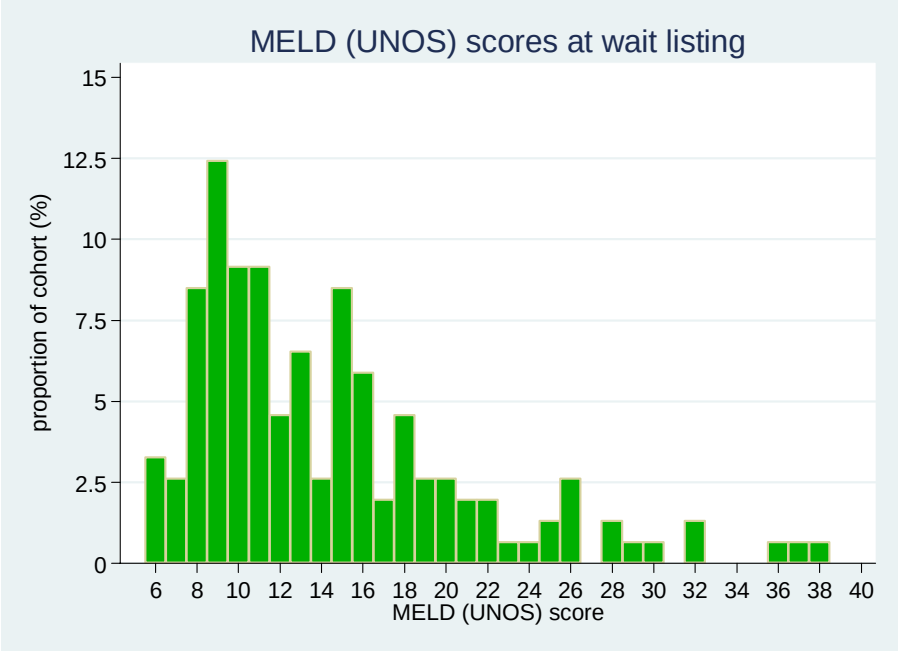
Table 5.3: Wait-list cohort listing laboratory results (n = 153)

variable	median (range)	mean (standard deviation)
total bilirubin (μmol/L)	32 (4 – 700)	69.8 (101.2)
creatinine (μmol/L)	80 (16 – 680)	95.6 (73.7)
prothrombin time (INR)	1.3 (0.9 – 7.2)	1.47 (0.72)
sodium (mmol/L)	137 (116 – 153)	136.5 (5.3)
albumin (g/L)*	29 (13 – 50)	29.4 (6.8)

*serum albumin result missing for ten candidates, therefore n = 143

The median wait-list MELD-UNOS and MELDNa scores for the wait-list cohort were 13 (range: 6 – 38) and 15 (range: 6 – 38) respectively (see Figure 5.7). Of the 143 individuals for which a CTP score could be determined, 27 (18.9%) were CTP “A”, 50 (35.0%) were CTP “B”, and 66 (46.1%) were CTP class “C” (see Figure 5.8). The median CTP score was 9 (range: 5 – 14). Wait-list MELD-UNOS, MELDNa and CTP scores differed significantly between individuals that experienced wait-list failure (both

Figure 5.7: MELD-UNOS and MELDNa scores at the time of waiting list assignment for the entire wait-list cohort (N = 153)



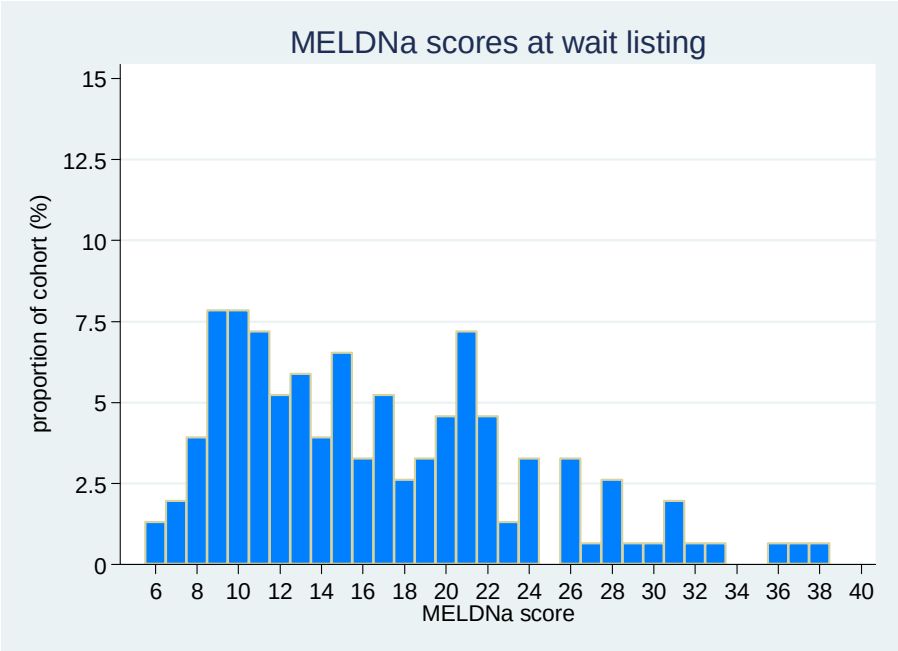
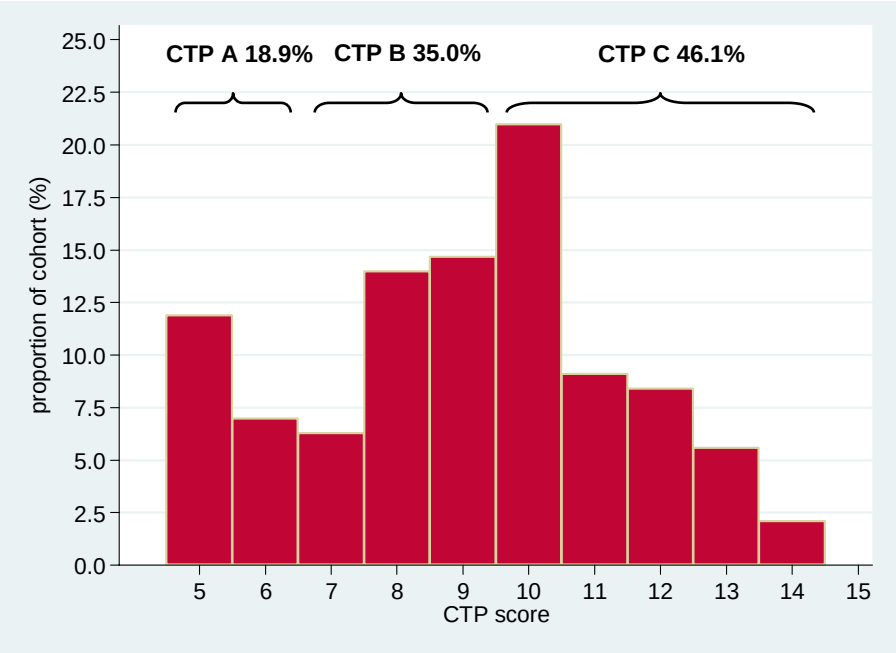


Figure 5.8: CTP score and classification at the time of waiting list assignment for the entire wait-list cohort (n = 153)



overall and 90-day) and survivors (overall survival median score failures vs. survivors: MELD-UNOS 18 vs. 12, $p = 0.005$; MELDNa 21 vs. 15, $p = 0.025$; CTP 10 vs. 9, $p = 0.048$; 90-day survival median score failures vs. survivors: MELD-UNOS 20 vs. 11, $p < 0.001$; MELDNa 22 vs. 13, $p < 0.001$; CTP 11 vs. 8.5, $p = 0.015$).

5.3 The ESLD-Indication Cohort

As has been discussed, the MELD prognostic models estimate risk of ESLD-related mortality in patients with chronic liver disease, including individuals with ESLD awaiting liver transplantation. The study inclusion and exclusion criteria, specified in section 4.4.2, were applied to the wait-list cohort of 153 individuals referred and subsequently placed on the Atlantic MOTP liver transplant waiting list between December 1st, 2004 and November 30th, 2009 (see Figure 5.1). Consequently, five

patients, whose primary indication for liver replacement was acute liver failure, and 32 patients, whose primary indication was hepatic malignancy, were excluded from the study cohort. This resulted in a cohort of 116 individuals, age 18 years and older, whose primary indication for liver replacement was ESLD, which were placed on the waiting list for first liver transplant between December 1st, 2004 and November 30th, 2009, and subsequently observed until March 1st 2010. Many of the features of this group have been covered in the discussion of the general wait-list cohort, therefore the basic characteristics of this ESLD-indication cohort are summarized in Table 5.4.

Table 5.4 ESLD-indication cohort characteristics

variable	observation (n = 116)
sex – frequency (proportion):	
male	58 (50.0%)
female	58 (50.0%)
age in years at wait-listing – median (range)	54.6 (21.6 – 69.0)
primary liver disease – frequency (proportion):	
alcohol-related ESLD	23 (19.8%)
HCV-related ESLD	22 (19.0%)
primary biliary cirrhosis	19 (16.4%)
primary sclerosing cholangitis	17 (14.7%)
cryptogenic	10 (8.6%)
other	25 (16.3%)

5.3.1 ESLD-Indication Cohort Outcomes

No patients from the cohort were lost to follow-up. Outcomes at the end of observation, March 1st, 2010, are summarized in Table 5.5. Two patients were transferred to the care of other liver transplant programmes, after 144 and 235 days of waiting list observation, as they moved out of the Atlantic MOTP's service region. The condition of 13 patients (11.2%) improved to the point that transplantation was not felt to be necessary, and they were therefore withdrawn from the waiting list. The candidacy of nine patients (7.8%) was revoked due to detection of conditions which contraindicated transplantation (medical comorbidities in seven, behavioural issues in one, and consent withdrawal in one). Approaching two-thirds of the patients in the cohort, 62.9 percent, underwent liver transplantation. The waiting time-to-transplant of these 73 patients is displayed in Figure 5.9. For the entire cohort of 116 candidates, there were a total of 11 wait-list failures (five deaths and six delistings) resulting in an overall occurrence of wait-list mortality of 9.5 percent (95%CI: 4.1 -14.9). All mortality events occurred within eight months of waiting list assignment. The incidence rate for wait-list failure was 11.6 per 100 person-years of observation (95%CI: 6.5 – 21.0). Figure 5.10 depicts the Kaplan-Meier estimate of wait-list survival for this ESLD-indication cohort. The Kaplan-Meier estimated 90-day and one-year wait-list mortality was 7.0 percent (95%CI: 3.4 – 14.1) and 12.7 percent (95%CI: 7.1 – 22.2), respectively.

After 90 days of wait-list observation 77 of the 116 candidates (66.4%) were alive and still awaiting a donor allograft, while the remaining 39 had already experienced a wait-list outcome. Thirty candidates (25.9%) had received a liver transplant and two were removed from the waiting list due to the detection of non-hepatic comorbidities that

Table 5.5: ESLD-indication cohort overall wait-list outcomes as of March 1st, 2010

outcome	incidence frequency (proportion)	time-to-event in days median (range)
still awaiting transplant	8 (6.9%)	777 (382 – 1839)
transferred	2 (1.7%)	144 and 235
withdrawn – unnecessary	13 (11.2%)	554 (122 – 1510)
withdrawn – unsuitable	9 (7.8%)	367 (29 – 1112)
transplanted	73 (62.9%)	122 (1 – 1012)
died	5 (4.3%)	34 (24 – 128)
withdrawn - terminal	6 (5.2%)	54 (13 – 236)
total	116 (100.0%)	148 (1 – 1839)

Figure 5.9: Waiting time-to-transplant for the 73 individuals that were transplanted from the ESLD primary indication cohort (n = 116)

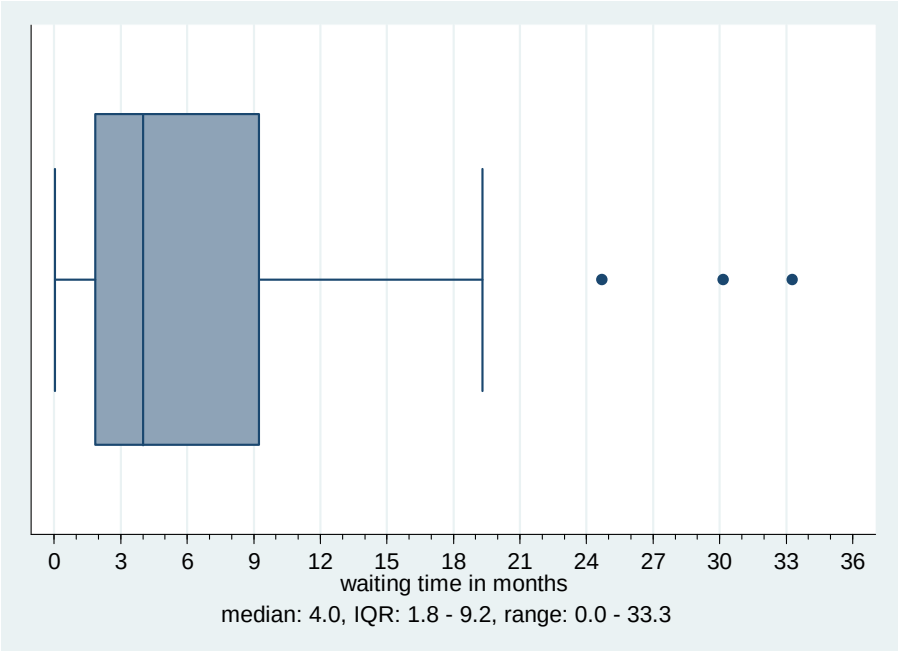
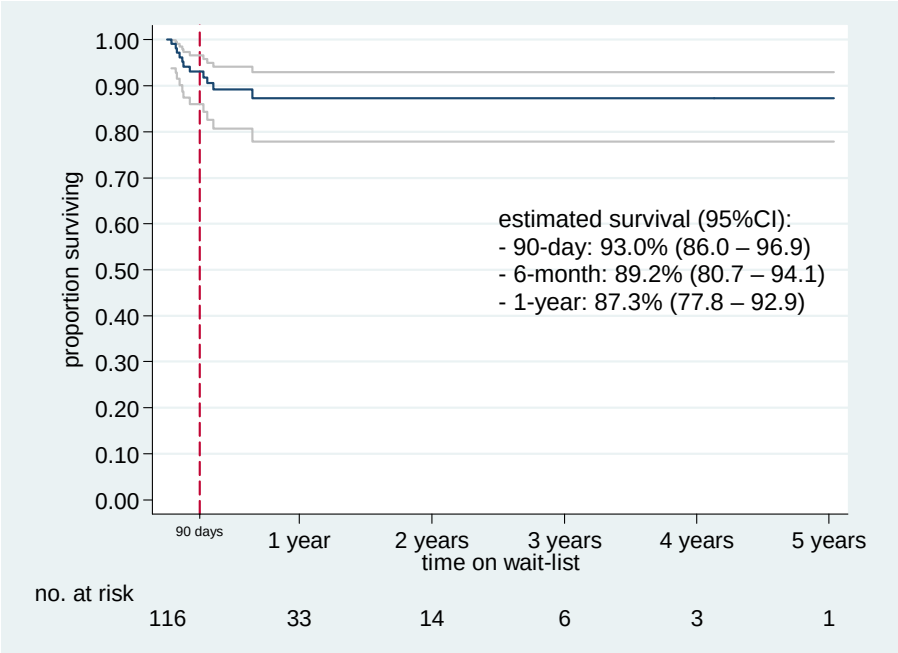


Figure 5.10: Kaplan-Meier estimate of liver transplant waiting list survival for the ESLD-indication cohort (n = 116)



rendered them unsuitable for transplantation (dementia and severe pulmonary hypertension). One of the patients that was withdrawn is known to have died of complications of ESLD 92 days after the date their of wait-list assignment, while the outcome of the other patient is unknown. A total of seven patients experienced wait-list failure within 90 days of registration (three died, four delisted), therefore the incidence of 90-day wait-list mortality in this cohort of 116 patients, whose primary indication for liver replacement was ESLD, was 6.0 percent (95%CI: 1.6 – 10.4).

5.3.2 ESLD-Indication Cohort Risk Scores

The laboratory investigation data necessary for the calculation of the MELD scores was complete for all 116 patients (Table 5.6). Ten patients were missing serum albumin results, precluding determination of the CTP score and classification for these

Table 5.6: ESLD-indication cohort listing laboratory results (n = 116)

variable	median (range)	mean (standard deviation)
total bilirubin (μmol/L)	38 (7 – 700)	74.0 (99.0)
creatinine (μmol/L)	83.5 (16 – 680)	100.2 (80.7)
prothrombin time (INR)	1.3 (1.0 – 4.2)	1.45 (0.51)
sodium (mmol/L)	136 (116 – 148)	135.7 (5.3)
albumin (g/L)*	28 (13 – 46)	28.4 (6.7)

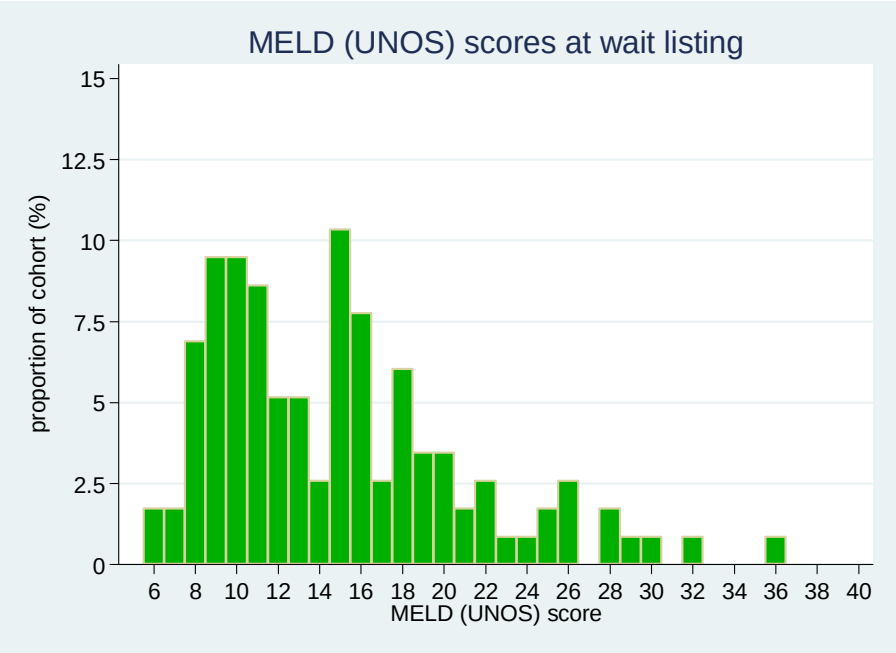
*serum albumin missing for ten candidates, therefore n = 106

observations. The median time from acquisition of laboratory investigations to wait-list assignment was three days prior (range: 57 days prior to 35 days after). 36.2 percent of the cohort had their listing risk score investigations obtained within seven days of assignment and 90.5 percent within 30 days.

The median MELD-UNOS and MELDNa scores for the ESLD-indication cohort were 14 (range: 6 – 36) and 17 (range: 6 – 36) respectively (see Figure 5.11). For the 106 candidates for which a CTP score could be determined, the median CTP score was 9.5 (range: 5 – 14)(see Figure 5.12). Seven of this group of 106 (6.6%) were CTP class “A”, 43 (40.6%) were CTP “B”, and 56 (52.8%) were CTP class “C”. MELD-UNOS, MELDNa, and CTP scores all differed significantly between the group of individuals that experienced wait-list failure (both overall and 90-day) and those that did not (overall mortality median score failures vs. survivors: MELD-UNOS 19 vs. 13, p = 0.010; MELDNa 22 vs. 17, p = 0.028; CTP 12 vs. 9.5, p = 0.027; 90-day mortality median score

failures vs. survivors: MELD-UNOS 20 vs. 11, $p < 0.001$; MELDNa 22 vs. 14, $p = 0.002$; CTP 12 vs. 9, $p = 0.007$).

Figure 5.11: ESLD-indication cohort wait-listing MELD-UNOS and MELDNa scores (n = 116)



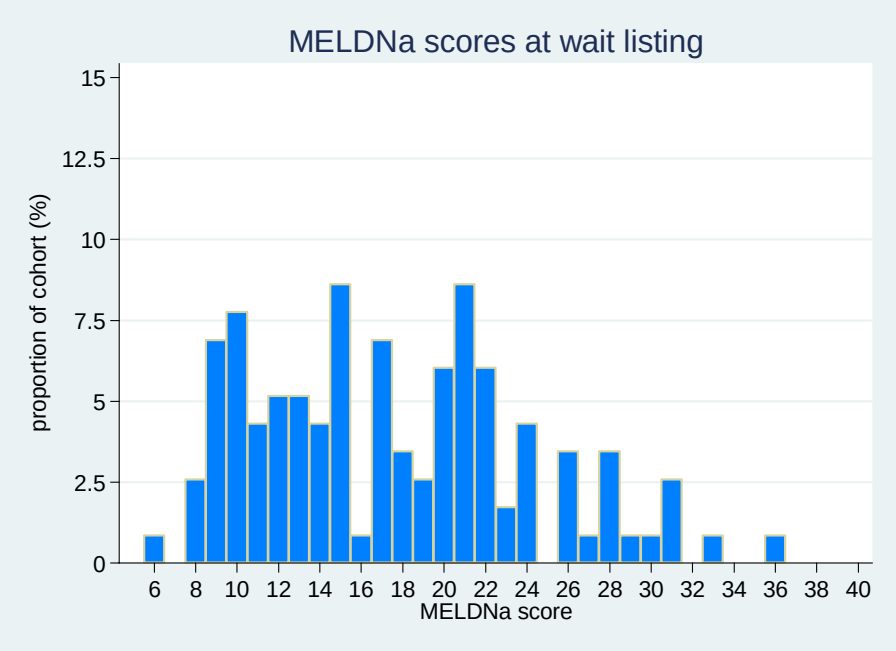
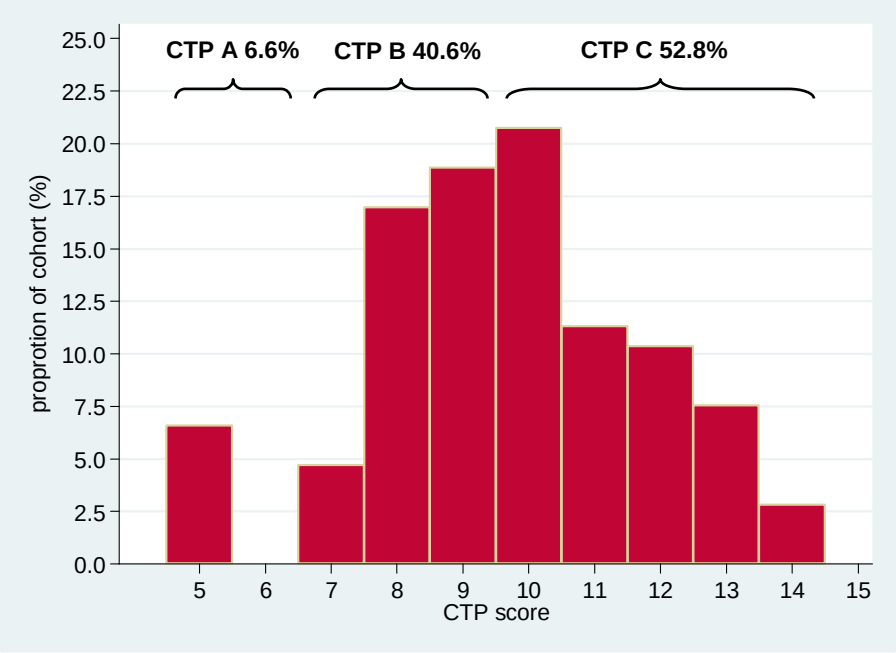


Figure 5.12: CTP score and classification at the time of wait-list assignment for the 106 candidates of the ESLD-indication cohort (n = 116) with complete risk index variables (serum albumin missing in 10 candidates)



5.4 The MELD Validation Cohort

As discussed in depth in the methodology chapter, the principle analysis carried out in this study evaluated the ability of the MELD to predict ESLD-related mortality within 90 days of liver transplant wait-list assignment. To be included in the cohort used for this validation purpose requires that the individual be at risk for ESLD-related mortality (death or delisting) until either this failure outcome occurs or 90 days of observation has elapsed. Candidates that receive a liver transplant within 90 days of wait-list assignment cannot be included in the MELD validation cohort as they have the natural history of their ESLD interrupted, and cease to be at risk for ESLD-related mortality, before 90 days of observation. Additionally, the observations of individuals

withdrawn within 90 days of registration due to improvement in condition, revocation of candidacy, or transfer were censored from the validation cohort as these individuals cease to be systematically followed and therefore their outcome could not be consistently ascertained. These conditions were thus applied to the ESLD-indication cohort to define the validation cohort (see Figure 5.1).

Of the 116 individuals with ESLD as their primary indication for liver replacement, 30 received a liver transplant and two had their candidacy revoked within 90 days of wait-list placement (described above in section 5.3.1). This left 84 liver transplant candidates with non-censored 90-day outcomes to form the validation cohort. Seven of these 84 individuals experienced wait-list failure within 90 days of registration, while the remaining 77 were alive and still awaiting transplantation after 90 days. The incidence of wait-list mortality in this MELD validation cohort was therefore 8.3 percent (95%CI: 2.3 – 14.4).

The laboratory investigation data necessary for calculation for the MELD-UNOS and MELDNa scores was available for all 84 candidates and these scores are reported in Chapters Six and Seven, respectively. A serum albumin value was missing for seven patients and therefore the CTP score and classification could be determined for 77 of the candidates in the validation cohort, the results of which are reported in Chapter Nine.

5.5 Discusion

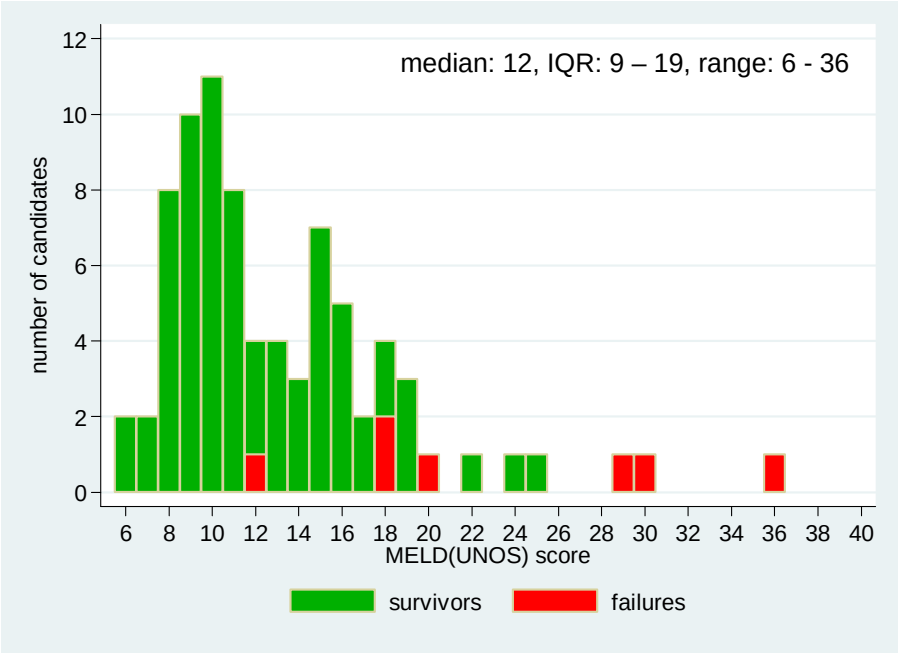
This chapter provides a characterization of the liver transplant referral population serviced by the Atlantic MOTP, the liver transplant assessment process, and the population of individuals assigned to the liver transplant waiting list over the first five years of operation after reactivation of the Atlantic MOTP liver transplant service on

December 1st, 2004. This analysis disclosed that a substantial number of referred patients succumbed before the assessment process could even be initiated which may point to a need to implement methods to increase physician accessibility to, and knowledge of, the availability and indications for liver transplantation. The estimated median assessment time for potential candidates was less than two and one-half months and following assessment approximately one-half of referrals were deemed to be appropriate candidates for liver replacement. The top three primary indications for liver replacement were hepatocellular cancer, alcohol-related ESLD, and cirrhosis due to chronic HCV infection. Approximately two-thirds of all individuals registered for liver transplantation over the five year observation period received a liver after a median waiting time of approximately three and one-half months, while approximately ten percent died or were removed from the waiting list as they became too ill to tolerate transplantation. O blood group, non-alcohol-related, non-cholestatic primary liver disease, and the occurrence of spontaneous bacterial peritonitis prior to wait-list assignment were all found to be significantly associated with wait-list failure. For individuals in whom ESLD was the primary indication for liver replacement, the observed incidence of 90-day wait-list failure was 6.0 percent (95%CI: 1.6 – 10.4). Candidates that experienced wait-list mortality, both overall and 90-day, had significantly greater MELD-UNOS, MELDNa, and CTP scores than individuals that survived.

Chapter Six: Results – Validation of the MELD Prognostic Model

The primary objective of this study was to determine if the MELD prognostic model accurately estimates the risk of 90-day wait-list mortality in Atlantic Canadian adults with ESLD awaiting liver transplantation. Of the 84 patients with ESLD that composed the validation cohort (i.e. those with non-censored 90-day outcomes), seven patients were withdrawn from the waiting list within 90 days of registration due to death (n = 3) or progression of their ESLD to the point that they were deemed too ill to survive transplantation (n = 4), while the remaining 77 candidates survived beyond 90 days. Data from this group of 84 candidates, whose wait-listing risk scores are summarized in Figure 6.1, was used to evaluate the calibration and discrimination of the MELD-UNOS.

Figure 6.1: MELD-UNOS scores from time of wait-list registration for the 84 individuals with non-censored 90-day outcomes ($n_{survivors}$ = 77, $n_{failures}$ = 7) that formed the validation cohort



6.1 Calibration of the MELD-UNOS

Calibration accuracy was assessed by comparing the observed to model estimated incidence of 90-day wait-list mortality. The observed 90-day incidence of wait-list failure in the validation cohort was 8.3 percent. The mean MELD-UNOS estimated probability of 90-day wait-list mortality for the cohort was 7.0 percent (95%CI: 4.8 – 9.1). There was no evidence that the mean MELD-UNOS derived estimate for 90-day mortality in the wait-list cohort differed from the observed incidence ($p = 0.215$).

As specified, two grouping strategies were used to assess global goodness-of-fit. For the first, the cohort was divided into four groups based on MELD-UNOS score quartiles in the manner described by Hosmer and Lemeshow.[27] Table 6.1 and Figure 6.2 provide, respectively, a tabular and graphical summary of the observed and expected event occurrences across the four observed risk score quartiles. The Hosmer-Lemeshow goodness-of-fit statistic was 4.159. The p-value corresponding to this test statistic value on a χ^2 distribution with four degrees of freedom is 0.385. While this result implies no evidence of significant lack-of-fit, the validity of this result must be viewed with caution as it is generally viewed that cells must have a minimum of five observations for the distributional assumptions upon which the Hosmer-Lemeshow statistic is based to be met.[27] Subjective evaluation of calibration within specific MELD-UNOS score quartiles is hampered by the low event rate in the lowest three quartiles. While it appears that the MELD-UNOS score may underestimate mortality in patients with scores of 17 and greater, it is difficult to gain insight on its performance in the lower risk groups defined by this grouping strategy. It would appear, however, that the model does not underestimate mortality in these lower three score quartiles. The scarcity of outcomes in

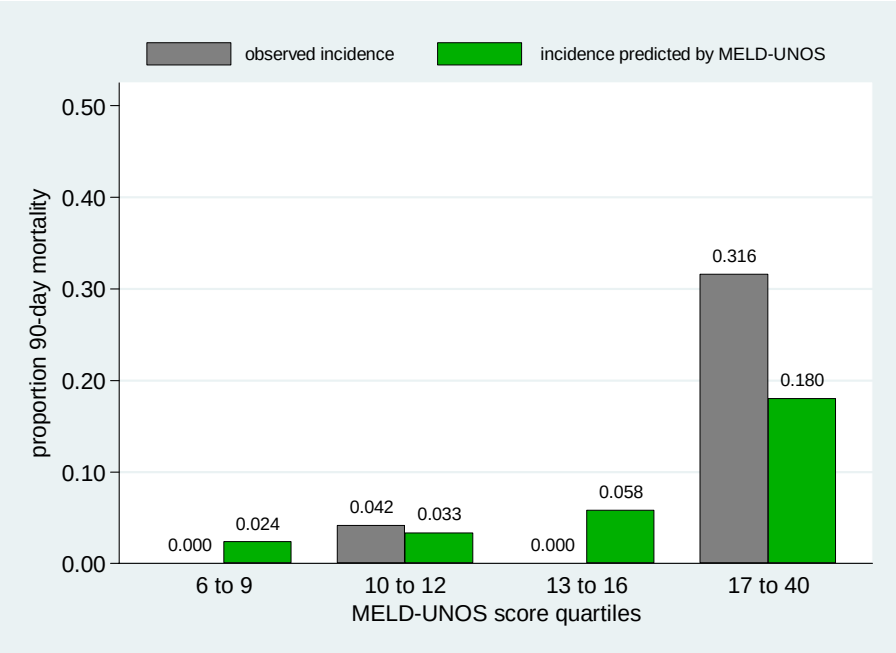
Table 6.1: MELD-UNOS calibration analysis by risk score quartiles

MELD score quartile	Observed failure incidence	Estimated failure incidence	events Observed Expected	non-events Observed Expected	H-L G-o-F χ^2
6 to 9 n = 22	0.000	0.024	0 0.5	22 21.5	0.531
10 to 12 n = 24	0.042	0.033	1 0.8	23 23.2	0.052
13 to 16 n = 19	0.000	0.058	0 1.1	19 17.9	1.166
17 to 40 n = 19	0.316	0.180	6 3.4	13 15.6	2.410
total cohort n = 84	0.083	0.070	7 5.9	77 78.1	4.159*

*$\chi^2(4)$, p = 0.385

the lower three score quartiles is not surprising given that the probability estimates for 90-day mortality corresponding to this range of MELD-UNOS score ranges from 1.6 to a maximum of 7.1 percent and, based on the sum of the probability estimates for the candidates with MELD-UNOS scores from six to 16, only 2.4 failures would have been expected in this subgroup of 65 candidates.

Figure 6.2: Observed versus MELD-UNOS estimated incidence of 90-day wait-list mortality in validation cohort (n = 84) by risk score quartiles



The second grouping strategy entailed subdividing the cohort into three risk strata based on each individual's MELD-UNOS derived probability estimated for 90-day mortality. As described in section 4.6.3.1, thresholds of five percent and ten percent were utilized to assign observations to "low" (< 5.0%), "intermediate" (5.0 to 9.9%), and "high" (≥ 10.0%), risk of failure within 90 days of wait-list assignment. Table 6.2 and Figure 6.3 provide tabular and graphical summarization of the results of analysis of calibration via this grouping strategy. The Hosmer-Lemeshow goodness-of-fit statistic was 1.216, which corresponds to a p-value, computed from a χ^2 distribution with three degrees of freedom, of 0.749. As indicated above, the validity of this statistical quantification of fit must be viewed with caution due to the low event rate. The absolute difference between observed and MELD-UNOS predicted mortality in the low and

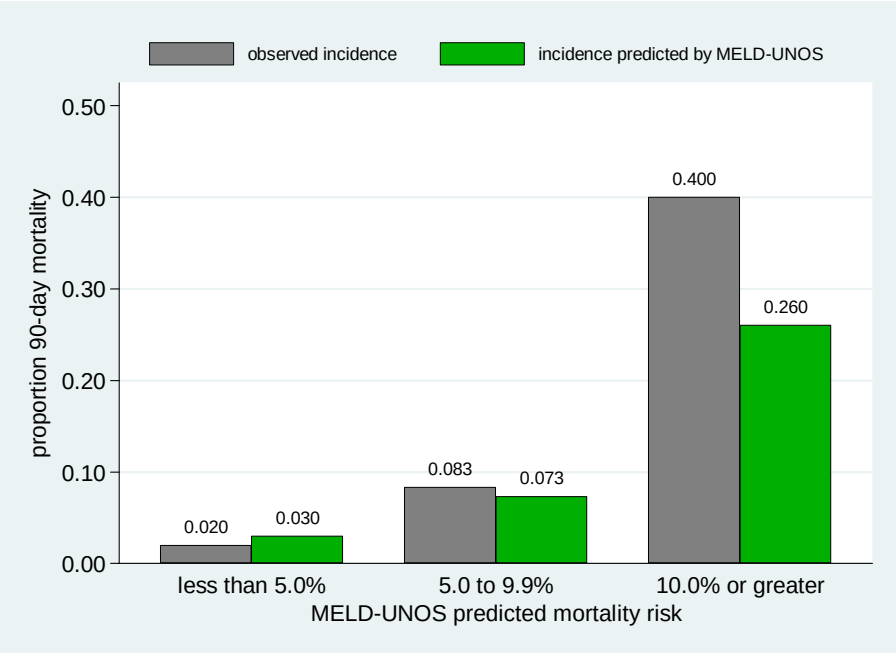
Table 6.2: MELD-UNOS calibration analysis by three risk strata

90-day mortality risk strata	Observed failure incidence	Estimated failure incidence	events Observed Expected	non-events Observed Expected	H-L G-o-F χ^2
< 5.0% "low" n = 50	0.020	0.030	1 1.5	49 48.5	0.172
5.0 – 9.9% "intermediate" n = 24	0.083	0.073	2 1.8	22 22.2	0.025
≥ 10.0% "high" n = 10	0.400	0.260	4 2.6	6 7.4	1.019
total cohort n = 84	0.083	0.070	7 5.9	77 77.1	1.216*

*$\chi^2(3)$, p = 0.749

intermediate risk strata was quite small, one percent for both. For the highest risk group the absolute difference between observed and predicted failure was 14 percent. This disparity may not be surprising considering that the stratum contained only ten patients with risk estimates ranging from 10.8 percent to 73.4 percent.

Figure 6.3: Observed versus MELD-UNOS estimated incidence of 90-day wait-list mortality in validation cohort (n = 84) by three risk strata (low, intermediate, high)



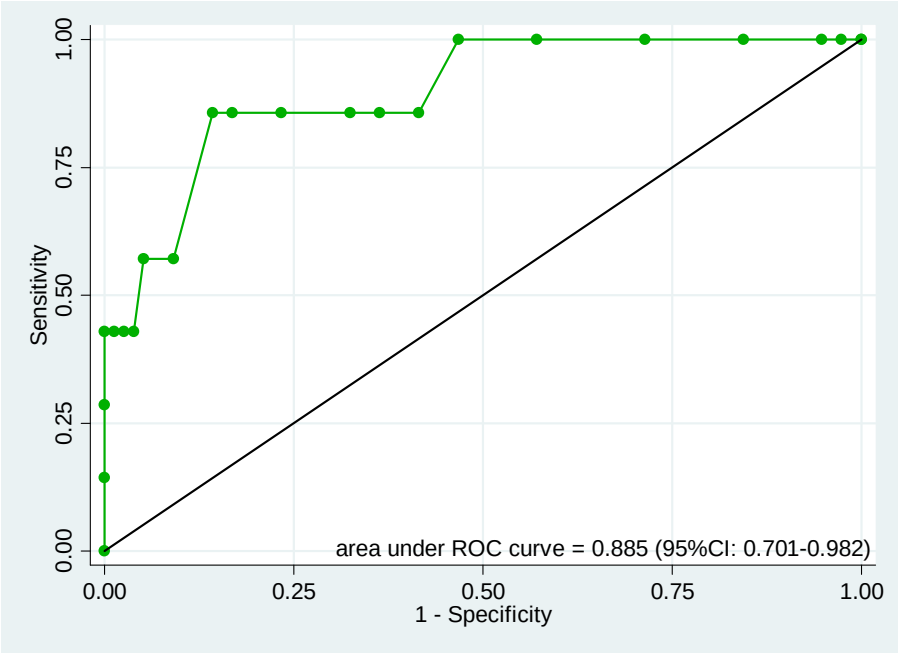
The results of a second clinically based grouping strategy in presented in Appendix E. In this analysis the cohort was divided into two groups using a MELD score of 15 as the threshold. The intent of using this cut-point was to assess calibration accuracy in the context of the Share MELD 15 policy with view to exploring the potential impact of the adoption of this policy in conjunction with a MELD-UNOS-based allocation system.

6.2 Discrimination of the MELD-UNOS

The measurement of discrimination quantifies the ability of the model, via its probability estimates, to assign observations to the appropriate outcome group (e.g. survival or mortality at 90 days). A model with favourable discrimination will tend to estimate higher probabilities (or risk scores) for the outcome in individuals that

experience the outcome than in those that do not. In Figure 6.1 the wait-listing MELD-UNOS scores for observations that survived 90 days and those that did not, are contrasted. From this it can be appreciated that the MELD-UNOS tended to assign higher scores to individuals that experienced mortality by 90 days than to 90-day survivors (the lowest MELD-UNOS score for failures was 12 which was greater than the score of over half of the survivors). Figure 6.4 displays the non-parametric ROC curve for the MELD-UNOS prediction of 90-day ESLD-related mortality in the validation cohort. The area under the ROC curve, and BcA bootstrap confidence interval for the estimate, was 0.885 (95%CI: 0.701 – 0.982) (the non-parametric confidence interval, calculated by the method of DeLong [78], was 0.752 – 1.000). This result means that if a randomly chosen pair of observations were drawn from the validation cohort, one having experienced the outcome and one not, 88.5 percent of the time the MELD-UNOS will have assigned a higher risk score to the observation that experienced the outcome than to the observation that did not. Alternatively, 88.5 percent is the average sensitivity, or specificity, for all levels of specificity, or sensitivity, associated with the range of classification variable thresholds. An area under the ROC curve, or c-statistic, of this value is considered to reflect excellent accuracy of discrimination. Even if, in a “worst case” scenario, the most inferior limit of the confidence interval was the true population value for the area under the curve, this value would still be generally considered to be reflective of acceptable discrimination accuracy.[27]

Figure 6.4: Non-parametric receiver operating characteristic curve for the MELD-UNOS prediction of 90-day wait-list mortality in the validation cohort (n = 84)



6.3 Discussion

Overall, in the context of the pre-specified study criteria for calibration accuracy, this analysis found no evidence that the MELD-UNOS generated estimates for 90-day wait-list mortality differed substantially from the observed incidence of mortality in this cohort of candidates with ESLD awaiting liver transplantation. Admittedly, to an extent, the analysis of global goodness-of-fit is hampered by the low event rate. It is also recognized that the Hosmer-Lemeshow goodness-of-fit test has been shown to be underpowered when sample sizes are small.[27, 73] The combined implication of the low event rate in the cohort and moderate validation cohort size is that even if a significant lack-of-fit existed, it could go undetected by this analysis. Through the evaluation of calibration within specific score/risk strata created by the various grouping strategies it

did appear that the MELD-UNOS had a tendency to underestimate failure incidence in individuals with risk scores in the upper teens and greater. This trend was apparent with all grouping strategies and will be discussed subsequently. In the lower ranges of the score (approximately < 15) however, calibration accuracy appeared satisfactory, with estimated mortality risk very closely approximating observations.

The only other study to include an analysis of the calibration performance of the MELD-UNOS was the MELDNa validation study reported by Kim et al.[16] This study, discussed in section 1.7, utilized a large cohort of individuals enrolled on the UNOS wait-list for first liver transplant in the years 2005 and 2006. Calibration of the MELD-UNOS was characterized using the data from the 7,171 candidates registered in 2006. Global calibration goodness-of-fit was summarized by the Hosmer-Lemeshow statistic, which found evidence of significant lack-of-fit ($p > 0.001$). The authors have indicated that due to the large number of individuals in the cohort this analysis is sensitive to even small departures in fit.[96] Despite this statistical finding of lack-of-fit, evaluation of the deciles-of-risk histogram disclosed very close approximation (within three percent) of observed to expected incidences of mortality in all but the two most high risk deciles. Unfortunately in this analysis the deciles were based on the candidates MELDNa scores, rather than the MELD-UNOS scores, and therefore specific score range comparisons of that study's results with the current study is not possible (MELD-UNOS and MELDNa scores have different survival functions and therefore equivalent scores do not have equivalent risk estimates associated with them). It is however reasonably safe to say that both studies suggest a tendency for the MELD-UNOS to underestimate mortality in groups with higher MELD-estimated ESLD acuity. It must be noted that to an extent this

apparent relative lack of calibration seen in higher risk quantiles may be due in part to the observation that scores in this range correspond to the steepest part of the exponential probability curve (see Figure 1.5). As a result, the variability of probability estimates in higher score quantiles is likely to be greater, potentially rendering the mean probability estimate used in the analysis to be less representative of the expected mortality for the quantile. In this study the low number of events and small cohort size may exacerbate this imprecision. When contemplating the above discussion it also should be pointed out that the clinical implications of model calibration deficiencies at higher risk scores are less than if they were present in the lower risk ranges. Candidates with high risk scores are already at the front of the line, so to speak, and more accurate estimation of their actual mortality risk is unlikely to further enhance their chance of receiving a liver.

From the perspective of the pre-specified study criteria, this study found the MELD-UNOS to have excellent discrimination accuracy. These results are comparable to the c-statistics from previous large studies by Wiesner et al (c-statistic: 0.83; 95%CI: 0.81 – 0.84) and Kim et al (c-statistic: 0.868; no confidence interval provided), and are possibly superior to the findings of the only previous Canadian study, by Burak (area under ROC curve: 0.79; 95%CI: 0.70 – 0.88). This finding of excellent discrimination accuracy of the MELD-UNOS when applied to this Atlantic Canadian liver transplant wait-list cohort is encouraging as, as was discussed in section 2.2.3, accurate discrimination is the requisite performance characteristic of a predictive model.

6.4 Conclusions

In summary, the most important finding from this analysis was that the MELD-UNOS was found to have excellent accuracy of discrimination when applied for the

prediction of 90-day wait-list failure in this cohort. If the predictions generated by the MELD-UNOS are used solely for ranking of liver transplant candidates this is the only relevant performance measure for the model. Regarding the calibration of the MELD-UNOS, this study produced no conclusive evidence to indicate that the predictive model produced inaccurately calibrated estimates of the 90-day probability of wait-list failure in this cohort. As indicated, the validity of the statistical analysis of global goodness-of-fit was hampered by the low event rate. For risk scores up to the mid-teens the estimated probabilities closely approximated the observed incidences of mortality, but with higher risk scores it appeared that the MELD-UNOS tended to underestimate 90-day wait-list failure. Overall, based on these findings, the MELD-UNOS predictive model for short-term ESLD-related mortality would appear to be generalizable to adult liver transplant candidates with ESLD residing in Atlantic Canada.

Chapter Seven: Results – Validation of the MELDNa Prognostic Model and Performance Comparison to the MELD-UNOS

A secondary objective of this study was to determine if the MELDNa prognostic model accurately estimates the risk of 90-day wait-list mortality in Atlantic Canadian adults with ESLD awaiting liver transplantation, and to compare its accuracy performance to that of the MELD-UNOS.

The serum sodium concentrations at the time of wait list registration of the 84 candidates are summarized in Figure 7.1. As can been seen, 29.8 percent of the cohort were hyponatremic (serum sodium concentration < 135 mmol/L). There was no evidence that the serum sodium concentrations of those that survived 90 days (median: 137, range: 123 – 143) differed from those that died or were delisted within 90-days (median: 138, range: 133 – 148)($p = 0.412$). Nor was there evidence that the proportion of individuals with hyponatremia differed between these two outcome groups (29.9% vs. 28.6%, $p = 1.000$). The MELDNa scores for the validation cohort of 84 candidates with non-censored 90-day wait-list outcome are displayed in Figure 7.2. Comparison of this histogram to the one for the MELD-UNOS scores (Figure 6.1) demonstrates that the inclusion of low serum sodium to the MELD predictive model has its greatest influence on the risk scores of observations with low to middle range MELD-UNOS scores (and as a consequence shifts the risk score curve to the right). This is in keeping with the findings of Heuman et al and Kim et al (see Figure 7.3).[16, 52]

Figure 7.1: Wait-listing laboratory investigation serum sodium concentration for the 84 candidates in the validation cohort (n = 84)

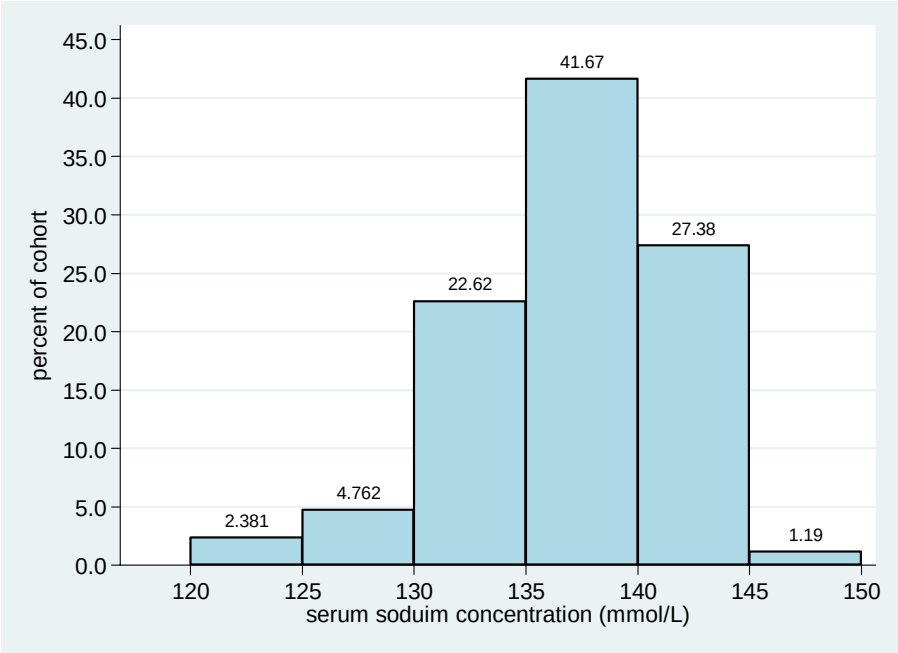


Figure 7.2: MELDNa scores from time of wait-list registration for the 84 individuals with non-censored 90-day outcomes ($n_{survivors}$ = 77, $n_{failures}$ = 7) that formed the validation cohort

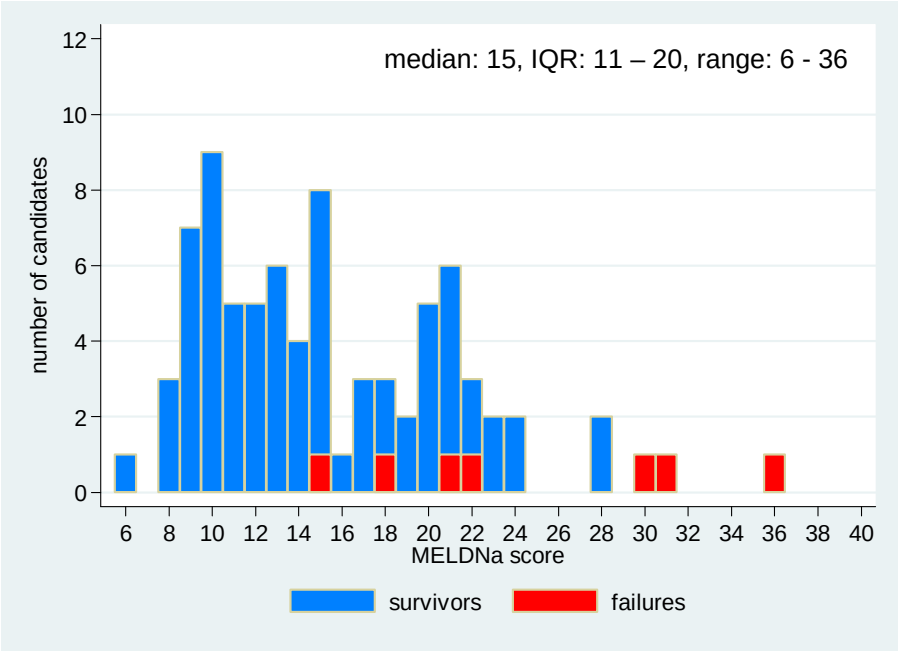
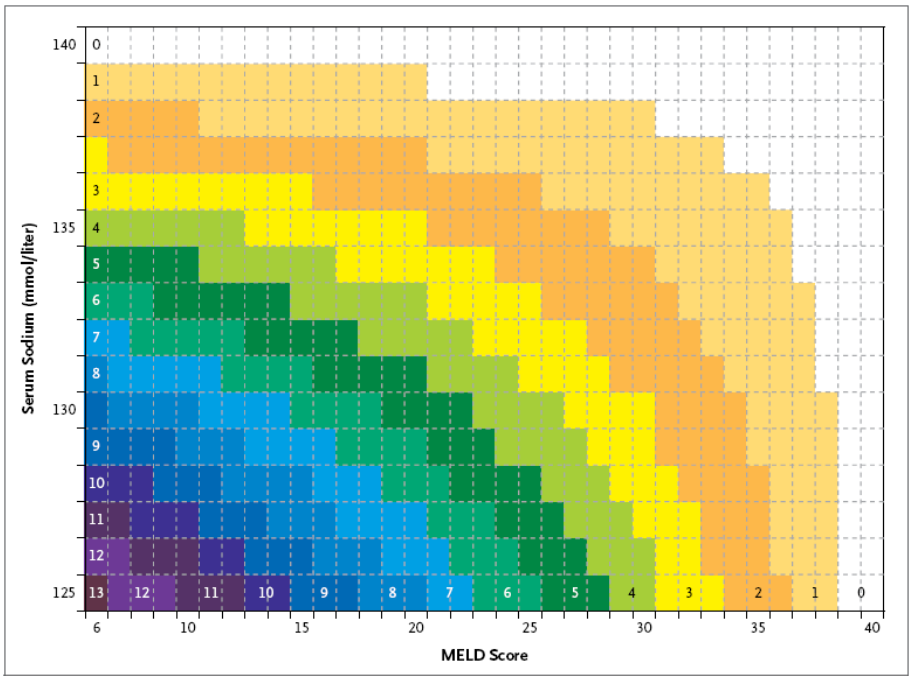


Figure 7.3: Additional risk score points granted according to the MELDNa based on interaction between serum sodium concentration and MELD-UNOS score [16] (reproduced with permission from the Massachusetts Medical Society)



7.1 Predictive Accuracy of the MELDNa

7.1.1 Calibration Accuracy of the MELDNa

The mean MELDNa generated estimate for 90-day wait-list failure for the cohort was 6.2 percent (95%CI: 3.5 – 8.8), which did not differ significantly from the observed incidence of 90-day wait-list mortality of 8.3 percent (p = 0.102). Two grouping methods were utilized for analysis of calibration goodness-of-fit and the results of these analyses are summarized in Tables 7.1 and 7.2 and Figures 7.4 and 7.5. The Hosmer-Lemeshow statistical tests for global goodness-of-fit did not produce evidence of significant lack-of-fit with either grouping strategy. As was discussed in Chapter 6 the validity of these tests can be questioned due to the low event rate. Combining the analysis of both grouping

analysis strategies suggests the MELDNa tended to underestimate mortality in the risk score range of approximately 15 to 22.

Table 7.1: MELDNa calibration analysis by risk score quartiles

MELDNa score quartile	Observed failure incidence	Estimated failure incidence	events Observed Expected	non-events Observed Expected	H-L G-o-F χ^2
6 to 10 n = 20	0.000	0.007	0 0.1	20 19.1	0.072
11 to 14 n = 20	0.000	0.014	0 0.3	20 19.7	0.326
15 to 20 n = 24	0.083	0.040	2 1.0	22 23.0	1.085
21 to 40 n = 20	0.250	0.190	5 3.8	15 16.2	0.468
total cohort n = 84	0.083	0.062	7 5.2	77 76.8	1.951*

*$\chi^2(4)$, p = 0.745

Figure 7.4: Observed versus MELDNa estimated incidence of 90-day wait-list mortality in validation cohort (n = 84) by risk score quartiles

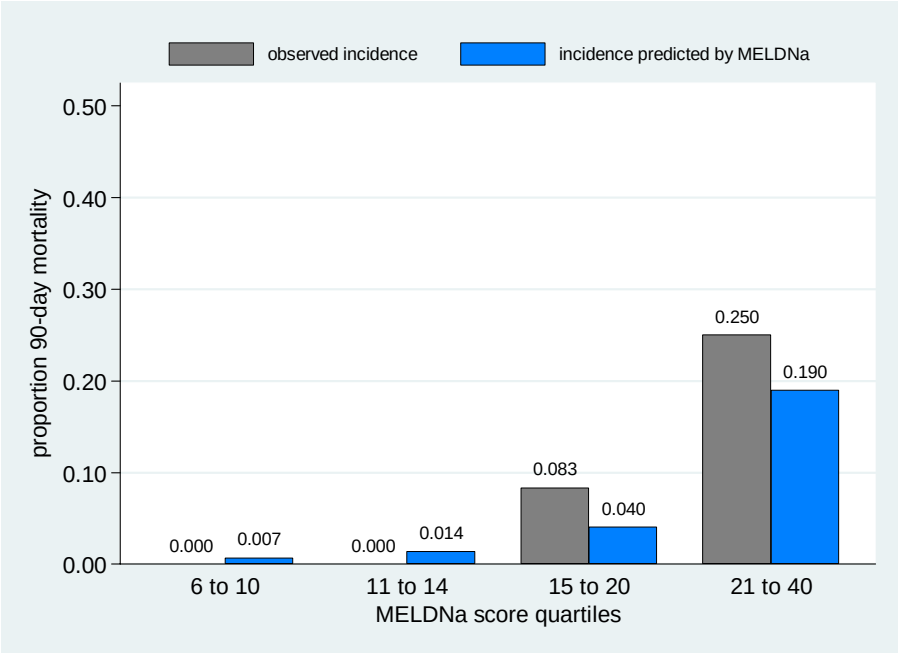


Figure 7.5: Observed versus MELDNa estimated incidence of 90-day wait-list mortality in validation cohort (n = 84) by three risk strata (low, intermediate, high)

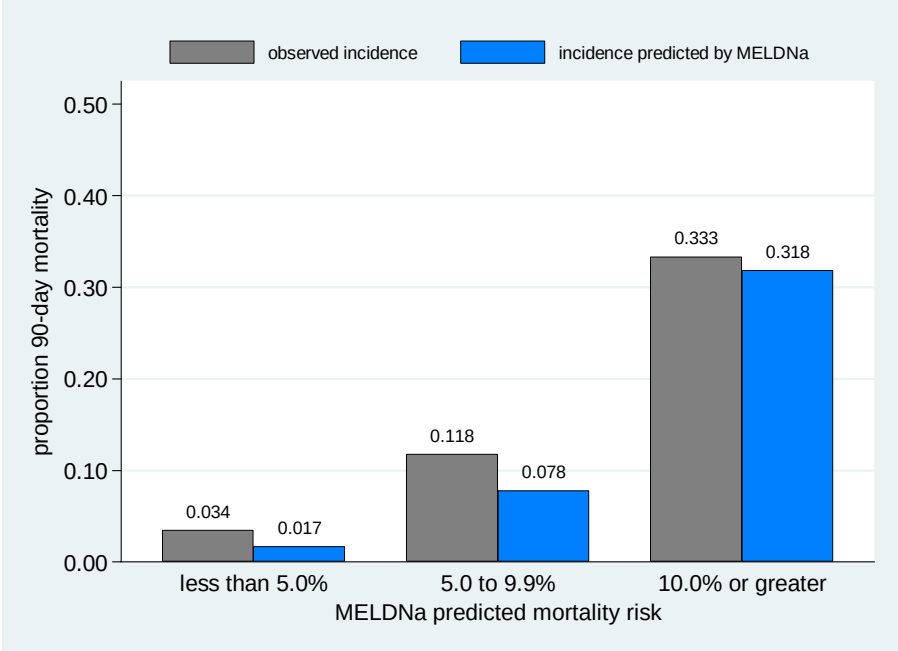


Table 7.2: MELDNa calibration analysis by three risk strata

90-day mortality risk strata	Observed failure incidence	Estimated failure incidence	events Observed Expected	non-events Observed Expected	H-L G-o-F χ^2
< 5.0% "low" n = 58	0.035	0.017	2 1.0	56 57.0	1.032
5.0 – 9.9% "intermediate" n = 17	0.118	0.078	2 1.3	15 15.7	0.401
≥ 10.0% "high" n = 9	0.333	0.318	3 2.9	6 6.1	0.005
total cohort n = 84	0.083	0.062	7 5.2	77 76.8	1.438*

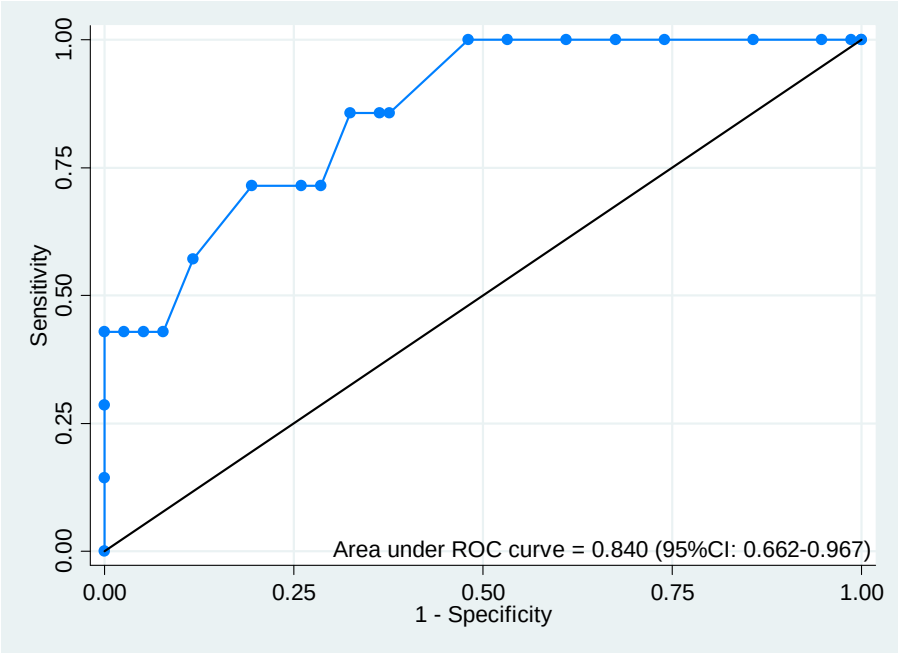
*$\chi^2(3)$, p = 0.697

7.1.2 Discrimination Accuracy of the MELDNa

Figure 7.2 demonstrates that there is reasonable separation of the MELDNa scores for observations that experienced wait-list failure versus those that did not (the lowest MELDNa score for failure observations is 15 which is greater than the median score of 14 for the individuals that survived). Displayed in Figure 7.6, the area under the non-parametric ROC curve for the MELDNa prediction of 90-day wait-list mortality, and

BcA bootstrap 95 percent confidence interval, was 0.840 (95%CI: 0.662-0.967)(the non-parametric confidence interval for this estimate, calculated by the method of DeLong, was 0.692 to 0.989). This value for the area under the ROC curve is considered to reflect excellent model discrimination.[27] It needs to be recognized however that in the "worst case" scenario, that if the true population value was close to the lower limit of the 95 percent confidence interval, model discrimination would be considered to be unacceptable.

Figure 7.6: Non-parametric receiver operating characteristic curve for the MELDNa prediction of 90-day wait-list mortality in the validation cohort (n = 84)



7.2 Comparison of Predictive Accuracy between the MELDNa and the MELD-UNOS

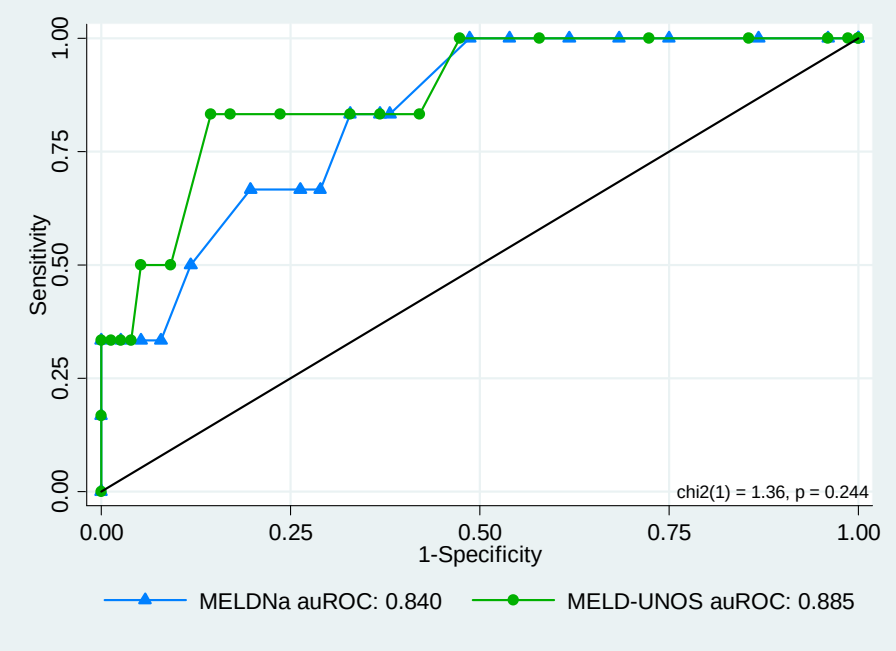
As discussed in section 2.3, the most prevalent method for comparing the performance of predictive models is via the non-parametric comparison of the respective

models' areas under the ROC curve. While this analysis permits a comparison of the discrimination between the two models, methods for comparing calibration are sparse. As the MELDNa is an augmentation of the MELD, it is considered to be a nested model and therefore comparison of global goodness-of-fit statistics can give an impression of which model is better calibrated. Additionally, risk stratification and reclassification tables can be used to characterize the relative performance of the augmented model. Reclassification tables also facilitate calculation of the Net Reclassification Improvement (NRI) and the Integrated Discrimination Improvement (IDI), which are two recently described relative measures of discrimination.[90]

As has been reported, the areas under the non-parametric ROC curve for the MELDNa and MELD-UNOS were 0.840 and 0.885, respectively (see Figure 7.7). The χ^2 statistic from the non-parametric comparison of these two areas by the method of DeLong et al was 1.36, which on a χ^2 distribution with one degree of freedom corresponds to a p-value of 0.244.[78] This result indicates there is no evidence to suggest that the MELDNa has statistically significant superior, or inferior, discriminative ability compared to the MELD-UNOS for the prediction of 90-day wait-list failure in this validation cohort.

Possibly reflecting the sensitivity of global goodness-of-fit testing to the grouping method employed that was discussed in section 2.2.1, the comparison of χ^2 statistics and related p-values from the Hosmer-Lemeshow test give contradictory results. By quartile-of-risk grouping the MELDNa would appear to be better calibrated, with a χ^2 statistic of 1.951 and p-value of 0.745 versus a χ^2 and p-value for the MELD-UNOS of 4.159 and 0.385, respectively. Yet when assessed using three clinically relevant risk strata, the

Figure 7.7: Non-parametric comparison of areas under non-parametric ROC curves for the MELDNa and MELD-UNOS predictions of the occurrence of 90-day wait-list mortality in the validation cohort (n = 84)



MELD-UNOS appears to have superior global calibration compared to the MELDNa (χ^2 = 1.216. p = 0.749 vs. χ^2 = 1.438, p = 0.697). This inconsistency may also be related to the influence of low event rate on the distributional assumptions of the Hosmer-Lemeshow goodness-of-fit test. The end result of this analysis is that no definitive conclusions can be made regarding the relative accuracy of global calibration of the MELDNa and MELD-UNOS models.

Patterns of risk reassignment as a result of the application of the MELDNa were analysed by the construction of a risk reclassification table (see Table 7.3). This method of analysis is contingent on the existence of clinically relevant risk thresholds that define the risk strata. In the case of the MELD, other than the “Share MELD 15” threshold risk

Table 7.3: Risk reclassification table MELDNa versus MELD-UNOS

	MELDNa risk groups			
MELD-UNOS risk groups	< 5.0%	5.0 – 9.9%	≥ 10%	total
< 5.0% "low"				
n total	47	3	0	50
n events	1	0	0	1
n non-events	46	3	0	49
Observed Incidence P_{event}	0.021	0.000	0.000	0.020
MELD-UNOS Expected P_{event}	0.030	0.039	0.000	0.030
MELDNa Expected P_{event}	0.013	0.079	0.000	-
5.0 – 9.9% "intermediate"				
n total	11	11	2	24
n events	1	1	0	2
n non-events	10	10	2	22
Observed Incidence P_{event}	0.091	0.091	0.000	0.083
MELD-UNOS Expected P_{event}	0.069	0.079	0.064	0.073
MELDNa Expected P_{event}	0.034	0.077	0.125	-
≥ 10% "high"				
n total	0	3	7	10
n events	0	1	3	4
n non-events	0	2	4	6
Observed Incidence P_{event}	0.000	0.333	0.429	0.400
MELD-UNOS Expected P_{event}	0.000	0.116	0.322	0.260
MELDNa Expected P_{event}	0.000	0.082	0.373	-
total				
n total	58	17	9	84
n events	2	2	3	7
n non-events	56	15	6	77
Observed Incidence P_{event}	0.035	0.118	0.333	0.083
MELD-UNOS Expected P_{event}	-	-	-	0.070
MELDNa Expected P_{event}	0.017	0.078	0.318	0.062

red indicates unfavourable reclassification, green indicates favourable reclassification

score, absolute risk thresholds are not utilized in clinical practice. Therefore, to permit the application of this method of analysis, risk thresholds of five and ten percent were selected to allow division of the cohort into "low", "intermediate", and "high" risk strata based on the observations model derived probability estimate for 90-day mortality. Analysis of the risk reclassification table reveals that none of the five individuals that

were reassigned to a higher risk stratum by the MELDNa estimate experienced a failure outcome. This is considered to be an unfavourable reassignment. Of 14 candidates that were reassigned to a lower risk stratum, 12 did not experience failure (favourable reclassifiction), and two were unfavourably reclassified as they experienced mortality. In total there were seven unfavourable reclassifications and 12 favourable reclassifications, for an absolute net favourable reassignment for five candidates out of the cohort of 84.

It is also useful to analyse reclassification from the perspective of risk score reassignment. Overall, 58 of the 84 candidates (69.0%) had their risk score increased when calculated using the MELDNa. For the seven candidates that died or were delisted, the MELDNa score was the same as the MELD-UNOS score for three (scores 18, 30 and 36), and four had their scores change by one (20 to 21), two (29 to 31), three (12 to 15) and four points (18 to 22), due to respective serum sodium concentrations of; 138, 133, 136, and 133 mmol/L. Fifteen individuals, all but one with a MELD-UNOS score of 20 or less, had larger magnitudes of score increase (range: 5 to 11) related to concentrations of sodium ranging from 123 to 134 mmol/L yet did not experience wait-list failure. To summarize this observation, in this cohort, the seven individuals that experienced 90-day mortality had at most a modest increased in their risk score, while the 15 candidates that had the greatest amount of reassignment by the MELDNa did not experience mortality. Reassignment from the score perspective therefore, did not appear to have the impact that might be expected. It must be acknowledged that when considering the data in this way that the MELD-UNOS and MELDNa have differing underlying 90-day mortality functions and therefore equivalent scores are not associated with equivalent probability estimates. For this reason, although based on somewhat arbitrary risk estimate thresholds,

the analysis based on probability estimates may be more valid. Having said this, neither analysis revealed evidence of a strong trend to favourable reassignment by the application of the MELDNa.

Data from the reclassification table was used to calculate the NRI and IDI and perform the related hypothesis tests to disclose if any change in reclassification brought about by the MELDNa, relative to the MELD, was significant (H_o: NRI or IDI = 0). The NRI was -0.195, which implies a net harmful reclassification of 19.5 percent (relative to outcome; 35.1% unfavourably reclassified vs. 15.6% favourably reclassified). There was no evidence to suggest that the MELDNa produced any significant change in net reclassification, relative to the MELD-UNOS ($z = -0.9321$, $p = 0.176$). The IDI was calculated to be 0.028, which is analogous an overall improvement in average sensitivity relative to average one minus specificity of 2.8 percent brought about by the MELDNa, relative to MELD-UNOS. This small difference was found to be insignificant on hypothesis testing ($z = 1.0628$, $p = 0.144$). Therefore, when evaluated using the NRI and IDI, two measures of change of discrimination, there was no evidence that the MELDNa offered any significant improvement in predictive performance over the MELD-UNOS.

7.3 Discussion

7.3.1 Discussion of the Predictive Accuracy of the MELDNa

With regards to the pre-specified study criteria for definition of acceptable calibration accuracy this analysis failed to produce any evidence that the MELDNa produced poorly calibrated estimates of 90-day wait-list mortality in this cohort. The MELDNa was found to accurately estimate total cohort 90-day wait-list failure. Global goodness-of-fit testing did not disclose any statistically significant lack-of-fit, although

the validity of this analysis can be challenged due to the low occurrence of wait-list failure in this cohort. Analysis of calibration within score and risk estimate defined groupings revealed that, in general, the estimates produced by the MELDNa fairly closely mirrored the observed incidence of mortality. One exception to this appeared to be for risk scores in the range of approximately 15 to 21, where the MELDNa seemed to underestimate the incidence of waiting list failure.

Considered from the perspective of the pre-specified accuracy performance criteria, the MELDNa was found to have excellent discrimination. This encouraging finding however, comes with the caveat that the 95 percent confidence interval for the estimate covers the full range of empirical qualifications of discrimination, from "unacceptable" to "excellent". As a consequence, there must be a degree of uncertainty, albeit slight, in passing down a final judgement that the discrimination accuracy of the MELDNa is completely acceptable.

Compared to the large MELDNa derivation study reported by Kim et al, the performance characterizations of the MELDNa from this study are less optimistic, and less conclusive.[16] This is no doubt at least partially related to the fact that the current study's validation cohort is considerably smaller and therefore comparisons are relatively underpowered. Kim et al found the c-statistic for the MELDNa prediction of 90-day wait-list mortality was found to be 0.883 (95%CI: not reported), which is somewhat in keeping with the point estimate generated by this study. Global goodness-of-fit testing disclosed significant lack of fit (but less than that of the MELD-UNOS); although the authors felt that due to the large cohort size even small deviations in fit would be detected.[96] Similar to their findings pertaining to the MELD-UNOS, the calibration of the MELDNa

was observed to deteriorate in the two highest risk score deciles (27 to 31 and 32 to 40). This phenomenon cannot be expanded upon by the current study as there were only five candidates with scores in this range (see Figure 7.2).

7.3.2 Discussion of the Comparison of Predictive Accuracy between the MELDNa and the MELD-UNOS

All comparisons of the predictive accuracy of the MELDNa to that of the MELD-UNOS failed to disclose any statistically significant performance difference between the two models. This lack of performance augmentation by the inclusion of the influence of low serum sodium is not surprising given that when evaluated by non-parametric univariable comparisons neither serum sodium concentration nor proportion of candidates with hyponatremia appeared to differ between the wait-list survivors and failures in this cohort. The point estimate for the area under the ROC curve for the MELDNa was slightly (0.046), but not significantly, less than that of the MELD-UNOS. These findings differ from those of Kim et al, who found that the c-statistic for the MELDNa of 0.883 was significantly greater than the c-statistic of 0.868 for the MELD-UNOS ($p < 0.001$). The clinical importance of this 0.015 absolute difference is debatable. As was discussed in section 2.3, the area under the ROC curve (and c-statistic) measure is relatively insensitive to the augmentation of a model by the introduction of a new predictor, unless the new risk factor has a large relative effect measure for the outcome (e.g. Odds Ratio ≥ 3.0).[72, 83, 89] In the case of serum sodium, Kim et al found the risk ratio for wait-list failure achieved a maximum value of approximately two when the threshold was approximately 120 mmol/L.[16]. Calculation of the NRI and IDI, both measures of change in discrimination, produced results that indicated no significant

change in discrimination associated with the application of the MELDNa, relative to the MELD-UNOS. It should be noted that the IDI is the more appropriate of these two measures for this analysis, as it does not rely on the existence of *a priori* clinically meaningful risk categories (although the measure is more difficult to interpret than the NRI). Of interest, Kim et al included calculation of the IDI in their comparative analysis of the MELDNa and MELD-UNOS and found the degree of improvement to be a marginal 1.1 percent. Although this improvement was statistically significant ($p < 0.001$), it is small and in fact less than the, non-significant, 2.8 percent performance enhancement found by this study.

Although by global goodness-of-fit testing both the MELD-UNOS and MELDNa showed significant lack-of-fit in the MELDNa derivation study, Kim et al reported that the MELDNa, by virtue of a smaller χ^2 statistic (17.8) and larger p-value (0.023) had superior calibration compared to the MELD-UNOS ($\chi^2 = 51.3$, $p < 0.001$). As was discussed, a similar method of comparison was attempted in the current study using the results from the two grouping methods for Hosmer-Lemeshow goodness-of-fit testing, which produced contradictory results. From this it is concluded that there was no consistent evidence that the global calibration accuracy of the MELDNa and MELD-UNOS differed. Subjectively, compared to the MELD-UNOS, the MELDNa appeared to better estimate the risk of wait-list failure in candidates with higher risk scores. Low event occurrence in the lower risk quartiles/strata does not permit meaningful subjective comparison of calibration over lower ranges of the risk scores.

The premise behind the MELDNa is that it more accurately estimates risk of short-term ESLD-related mortality in a subset of individuals with ESLD and

hyponatremia. The magnitude of the effect of hyponatremia on short-term ESLD-related mortality has been documented to have an inverse relationship with the MELD risk score.[16, 52] Heuman et al characterized the group at greatest risk as being individuals with serum sodium less than 135 mmol/L and MELD-UNOS score less than 21.[52] As a result of this phenomenon it would be expected that application of the MELDNa would result in reclassification of individuals with low range MELD-UNOS scores and hyponatremia to higher risk scores and risk strata. In this cohort 69.0 percent of the candidates had an increase in their risk score when it was calculated with the MELDNa, and all but three of these had MELD-UNOS scores of 20 or less. However, of the seven wait-list failures, only four had their scores increased, and this increase was, at best, modest in all of these. It is speculated that the small magnitude of the observed risk score increases brought about by application of the MELDNa for the prioritization of these four individuals would have had a minimal impact on increasing their prospect of receiving a donor liver. Also, when speculating about the potential for favourable risk score reassignment one must also consider the negative impact of the upward score reassignment in the 54 candidates that did not experience 90 day wait-list failure.

The reason for the apparent lack of influence of hyponatremia on 90-day wait-list mortality in this cohort of liver transplant candidates with ESLD is unclear. The proportion of the validation cohort with a serum sodium less than 135 mmol/L was 29.8 percent, which is very similar to the 30.9 percent reported by Kim et al from the MELDNa derivation study. It seems improbable that the clinicians caring for liver transplant candidates in Atlantic Canada have any greater success at correcting hyponatremia, and thus ameliorating this mortality risk while on the waiting list, than

clinicians elsewhere. In the absence of any more plausible clinical or physiological explanation it is possible that the low number of 90-day mortality events may be responsible for the association of hyponatremia and wait-list failure being unapparent in this study.

7.4 Conclusions

The MELDNa was found to have excellent discrimination accuracy when applied to predict the risk of 90-day mortality in this cohort of adults with ESLD awaiting liver transplantation. Analysis of calibration accuracy produced no evidence to indicate that the probability estimates generated by the MELDNa differed significantly from the observed incidence of wait-list failure. It must however be acknowledged that the analysis of model calibration was limited by low event occurrence.

Contrary to the conclusions reached by Kim et al in their first report on the MELDNa, the present study failed to demonstrate that the predictive accuracy of the MELDNa for 90-day wait-list failure was any different than that of the MELD-UNOS.[16] Overall, based on these findings, the MELDNa predictive model for short-term ESLD-related mortality would appear to be generalizable to adult liver transplant candidates with ESLD residing in Atlantic Canada; however it does not appear to offer any improvement in prognostication over the MELD-UNOS.

Chapter Eight: Results – Reintroduction of the Disease Aetiology Variable to the MELD

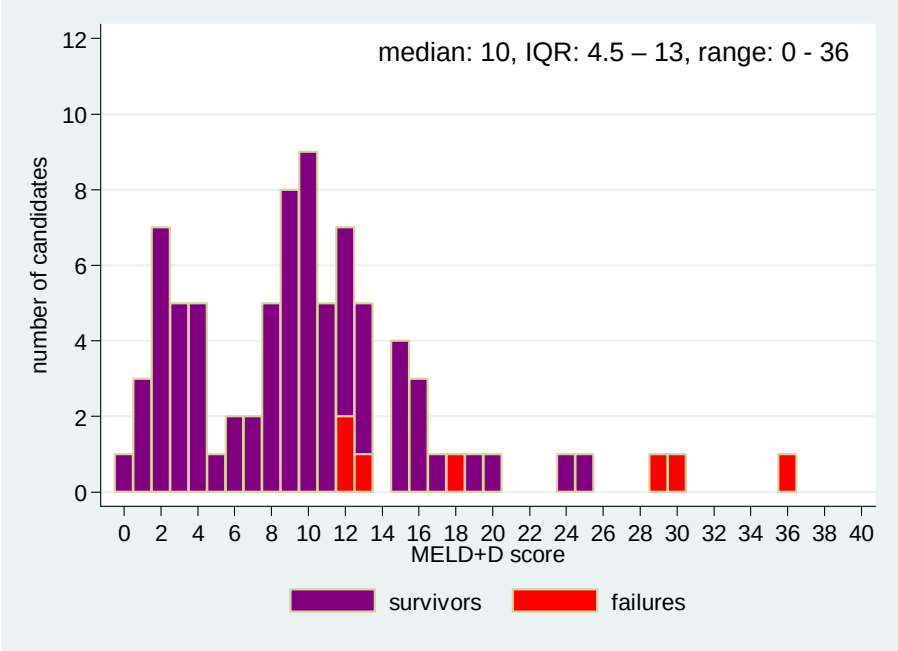
During the exploratory univariable analysis comparing characteristics between overall wait-list survivors and failures it was found that both alcohol-related ESLD and cholestatic liver disease (PBC or PSC) as primary liver diseases were relatively underrepresented in the group of individuals that experienced wait-list failure (see section 5.2.2.1 and Appendix D). This finding was congruent with the results described in the original report by Malinchoc et al on the Mayo model for ESLD, which found that cirrhotics with primary liver diseases not related to alcohol or cholestatic liver disease were at increased risk of short-term ESLD-related mortality.[26] As a consequence of this finding, the original Mayo model for ESLD contained, in addition to the three laboratory-based variables, a binary disease aetiology variable, which was coded "0" if the primary ESLD was alcohol-related or cholestatic, or "1" for all other primary liver diseases. As was discussed in section 1.5, subsequent external validation studies on the MELD carried out by the Mayo group found the influence of the disease aetiology variable on the discrimination accuracy of the model to be minimal, and therefore this variable was removed from the model.[24]

In our cohort of 116 liver transplant candidates whose primary indication for liver replacement was decompensated ESLD, the presence of non-alcohol-related, non-cholestatic liver disease as the primary liver disease diagnosis was associated with an estimated 4.7-fold increase in the incidence of overall wait-list failure (15.8% vs. 3.4%, Risk Ratio = 4.7, 95%CI: 1.05 – 20.6, p = 0.028). With an effect measure of this magnitude it was felt that the accuracy of the MELD to estimate 90-day wait-list failure

might be enhanced by reactivating the disease aetiology independent variable. This hypothesis was also motivated by the observation of substantial difference between the prevalence of cholestatic liver disease in liver transplant candidates from Atlantic Canada and those forming the cohorts of previous MELD validation studies. Individuals with cholestatic liver diseases as their primary liver disease comprised 31.0 percent of our ESLD-indication cohort (95%CI: 22.5 – 39.6), which differs markedly from the 2.9 percent and eight percent prevalences reported in the two large MELD validation studies conducted in the United States of America.[15, 16] It was speculated that this prominence of cholestatic liver disease and its apparent more favourable prognosis for wait-list survival might render the disease aetiology variable more influential in the prediction of wait-list failure in liver transplant candidates with ESLD from Atlantic Canada. A secondary objective of this study was therefore, to evaluate the predictive accuracy of a MELD which included the disease aetiology variable (MELD+D) to estimate the occurrence of 90-day mortality in Atlantic Canadians with ESLD awaiting liver transplantation, and to compare its performance to the MELD-UNOS.

Exactly one-half of the validation cohort had non-alcohol related, non-cholestatic liver disease as their primary indication for liver replacement. The wait-list MELD+D scores for the 84 candidates with non-censored 90-day wait-list outcomes are displayed in Figure 8.1. With the reactivation of the disease aetiology variable the minimum risk score becomes zero while the upper limit of its range remains 40. The median MELD+D score was 10 (range: 0 to 36). The MELD+D scores of the seven candidates that experienced 90-day wait-list failure (median: 18, range: 12 to 36) were significantly greater than the scores for individuals that survived 90 days (median: 9, range: 0 to 25)($p < 0.001$).

Figure 8.1: MELD+D scores from time of wait-list registration for the 84 individuals with non-censored 90-day outcomes ($n_{survivors}$ = 77, $n_{failures}$ = 7) that formed the validation cohort



8.1 Predictive Accuracy of the MELD+D

8.1.1 Calibration of the MELD+D

The overall MELD+D predicted 90-day wait-list mortality incidence was 5.4 percent (95%CI: 3.2 – 7.6), which was significantly less than the observed incidence of 8.3 percent (p = 0.010). Table 8.1 and Figure 8.2, and Table 8.2 and Figure 8.3, respectively provide tabular and graphical summarization of calibration goodness-of-fit analysed by risk score quartile and risk strata grouping methods. For both grouping strategies the Hosmer-Lemeshow goodness-of-fit test failed to reveal significant lack-of-fit, however the validity of this test can be questioned due to low event occurrence. Subjective evaluation of the calibration plots suggests a tendency for the MELD+D to

underestimate 90-day mortality, particularly in the middle to upper range of score/risk, a finding which concurs with the result for overall calibration.

Table 8.1: MELD+D calibration analysis by risk score quartiles

MELD+D score quartile	Observed failure incidence	Estimated failure incidence	events Observed Expected	non-events Observed Expected	H-L G-o-F χ^2
0 to 4 n = 21	0.000	0.011	0 0.2	21 20.8	0.175
5 to 9 n = 18	0.000	0.023	0 0.4	18 17.6	0.396
10 to 12 n = 23	0.087	0.035	2 0.8	21 22.2	1.854
13 to 40 n = 22	0.227	0.141	5 3.1	17 18.9	1.355
total cohort n = 84	0.083	0.054	7 4.5	77 79.5	3.780*

*$\chi^2(4)$, p = 0.437

Figure 8.2: Observed versus MELD+D predicted incidence of 90-day wait-list mortality in validation cohort (n = 84) by risk score quartiles

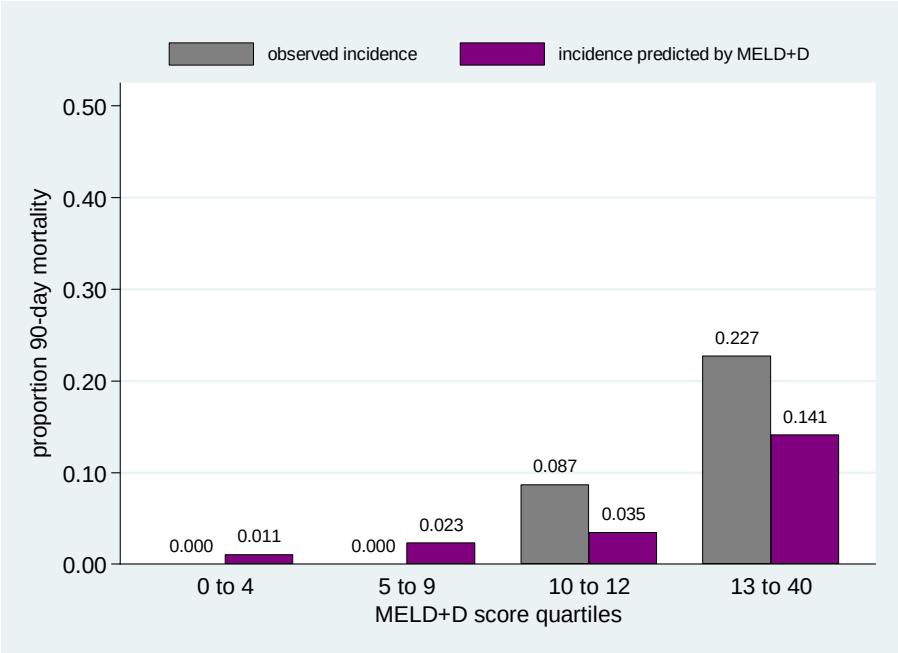


Figure 8.3: Observed versus MELD+D predicted incidence of 90-day wait-list mortality in validation cohort (n = 84) by three risk strata

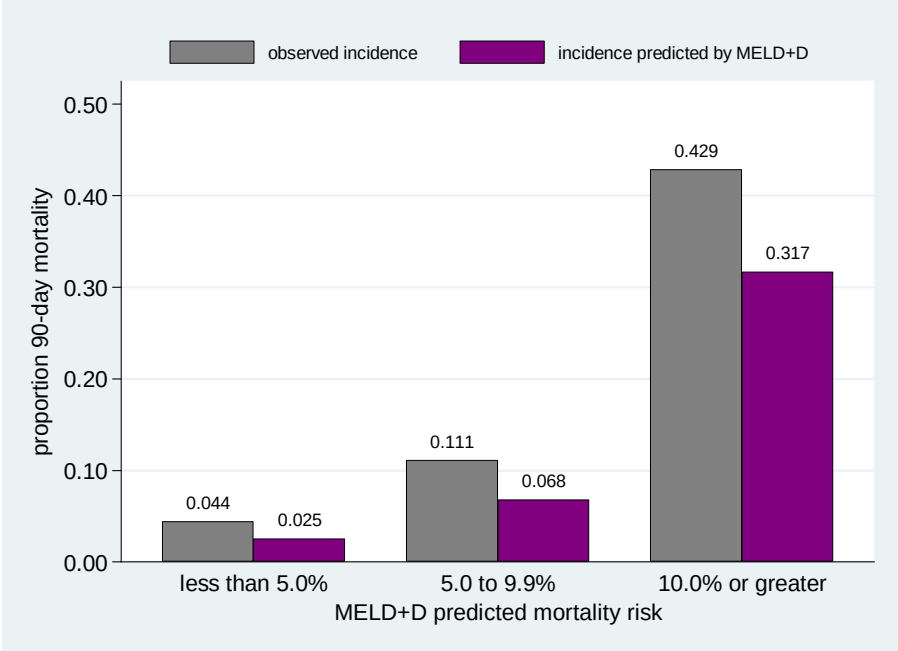


Table 8.2: MELD+D calibration analysis by three risk strata

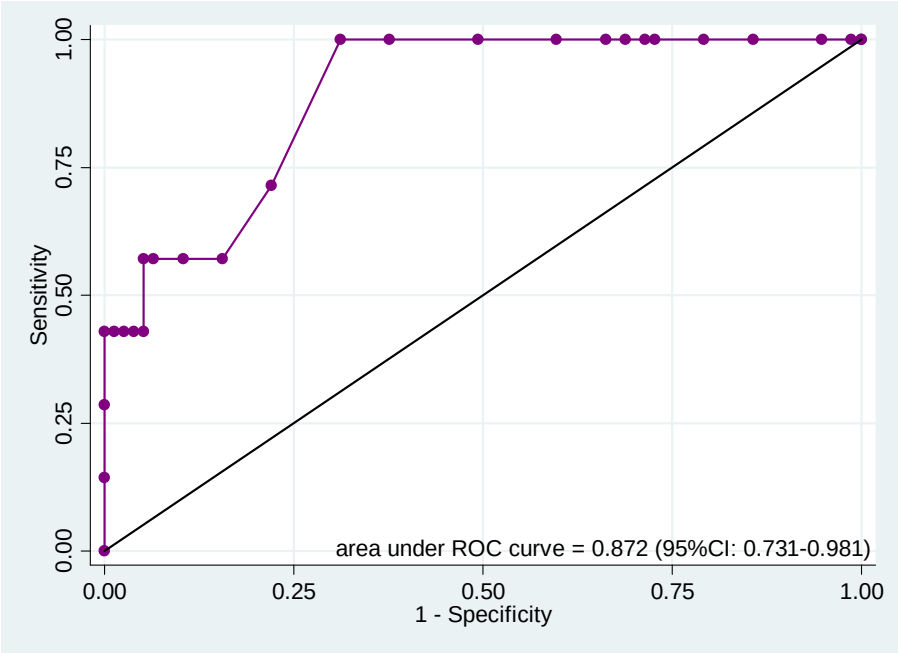
90-day mortality risk strata	Observed failure incidence	Estimated failure incidence	events Observed Expected	non-events Observed Expected	H-L G-o-F χ^2
< 5.0% "low" n = 68	0.044	0.025	3 1.7	65 66.3	1.020
5.0 – 9.9% "intermediate" n = 9	0.111	0.068	1 0.6	8 8.4	0.281
≥ 10.0% "high" n = 7	0.429	0.317	3 2.2	4 4.8	0.422
total cohort n = 84	0.083	0.054	7 4.5	77 79.5	1.723*

*$\chi^2(3)$, p = 0.632

8.1.2 Discrimination of the MELD+D

Figure 8.1 demonstrates that there is reasonable separation between the MELD+D scores of 90-day wait-list failures and survivors, with the lowest score of the failure group lying at the 75th percentile of scores for those that survived 90 days. The estimate for the area under the non-parametric ROC curve, and its BcA bootstrap confidence interval, was 0.872 (95%CI: 0.730 – 0.981)(see Figure 8.4). This value of point estimate

Figure 8.4: Non-parametric receiver operating characteristic curve for the MELD+D prediction of 90-day wait-list mortality in the validation cohort (n = 84)



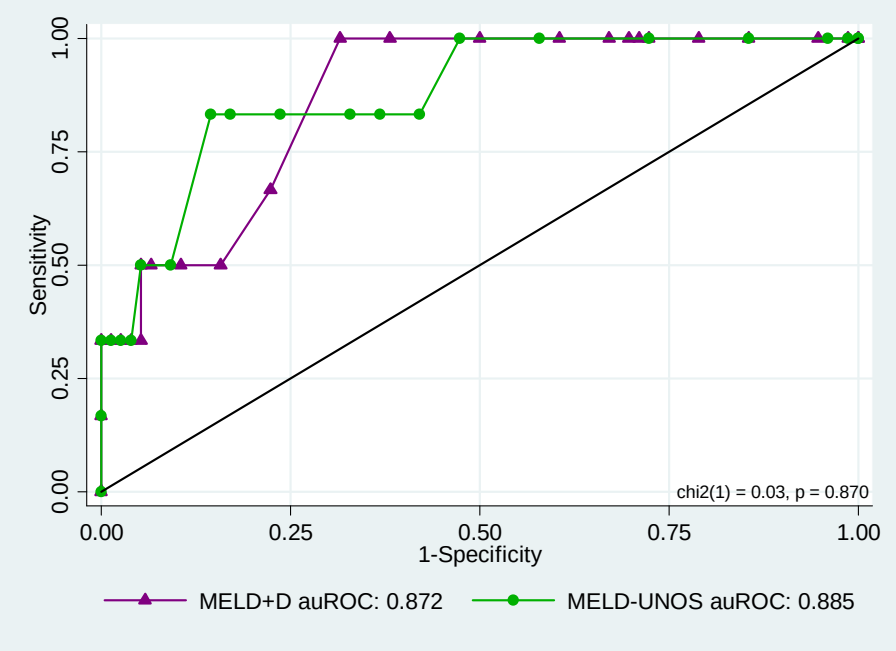
for the area under the ROC curve implies excellent discrimination accuracy for the MELD+D.[27] Even in the "worst case" scenario, where the true population value for the area under the ROC lies at the lower limit of the 95 percent confidence interval, this value is considered to reflect "acceptable" discrimination.

8.2 Comparison of Predictive Accuracy between the MELD+D and the MELD-UNOS

Displayed graphically in Figure 8.5, the comparison of the areas under the non-parametric ROC curves for the MELD+D (0.872) and MELD-UNOS (0.885) by the non-parametric method of DeLong found no statistically significant difference between the discriminative accuracy of the two models ($\chi^2(1) = 0.03$, $p = 0.870$).[78] The two measures of relative change in discrimination, the Net Reclassification Improvement

(NRI) and Integrated Discrimination Improvement (IDI), were calculated from the risk strata reclassification table (Table 8.3). The NRI was -0.065, which implies a net 6.5 percent deleterious reclassification brought about by the reincorporation of the disease aetiology variable into the MELD. This difference was however not found to be statistically significant (z = -0.311, p = 0.378). The IDI was calculated to be -0.002, which is consistent with no significant change in discriminatory performance (z = -0.1741, p = 0.431).

Figure 8.5: Non-parametric comparison of areas under the non-parametric ROC curves for the MELD+D and MELD-UNOS predictions of 90-day wait-list mortality in the validation cohort (n = 84)



On assessment of calibration the MELD+D was found to significantly underestimate the overall incidence of 90-day wait-list failure, while the MELD-UNOS estimate did not. The comparison of χ^2 statistics, and related p-values, from Hosmer-Lemeshow goodness-of-fit testing yielded contradictory results, potentially reflecting the

influence of the low event numbers on the distributional assumptions of the test. The MELD+D appeared to have slightly superior global goodness-of-fit by the quartile-of-risk score grouping strategy (χ^2 = 3780, p = 0.437 vs. χ^2 = 4.159, p = 0.385), while the MELD-UNOS appeared to be slightly superior when analysed by the three risk strata grouping method (χ^2 = 1.216, p = 0.749 vs. χ^2 = 1.723, p = 0.632). Subjectively, it appeared that the MELD+D had a general tendency to underestimate 90-day wait-list mortality, relative to the MELD-UNOS.

The risk reclassification table, Table 8.3, emphasizes that reactivation of the disease aetiology variable can only lead to downwards risk reclassification of applicable candidates. This is because the default assignment of the disease aetiology variable in the MELD-UNOS is to the exposed status (i.e. non-alcohol-related, non-cholestatic primary liver disease). Therefore the reactivation of the disease aetiology variable leads to a decrease in the risk score, and mortality risk estimate, for individuals with alcohol-related or cholestatic liver disease, and leaves all others' risk scores and estimates unchanged. From the table it is seen that two candidates that suffered wait-list failure had their risk estimates reassigned to lower risk strata, considered unfavourable reclassification, while 17 survivors were favourably reassigned to lower risk strata. The net absolute reassignment was therefore 15 candidates favourably reclassified. This apparently favourable result is tempered by the NRI result, reported above, which characterizes reclassification relative to outcome. In that analysis two of the seven failures were unfavourably reclassified (28.6%), while 17 of 77 survivors were favourably reclassified (22.1%), for a net relative incidence of unfavourable reclassification of 6.5 percent.

Table 8.3: Risk reclassification table MELD+D versus MELD-UNOS

	MELDNa risk groups			
MELD-UNOS risk groups	< 5.0%	5.0 – 9.9%	≥ 10%	total
< 5.0% "low"				
n total	50	0	0	50
n events	1	0	0	1
n non-events	49	0	0	49
Observed Incidence P_{event}	0.020	0.000	0.000	0.020
MELD-UNOS Expected P_{event}	0.030	0.000	0.000	0.030
MELD+D Expected P_{event}	0.023	0.000	0.000	-
5.0 – 9.9% "intermediate"				
n total	16	8	0	24
n events	1	1	0	2
n non-events	15	7	0	22
Observed Incidence P_{event}	0.063	0.125	0.000	0.083
MELD-UNOS Expected P_{event}	0.076	0.068	0.000	0.073
MELD+D Expected P_{event}	0.031	0.068	0.000	-
≥ 10% "high"				
n total	2	1	7	10
n events	1	0	3	4
n non-events	1	1	4	6
Observed Incidence P_{event}	0.500	0.000	0.429	0.400
MELD-UNOS Expected P_{event}	0.113	0.158	0.317	0.260
MELD+D Expected P_{event}	0.046	0.066	0.317	-
total				
n total	68	9	7	84
n events	3	1	3	7
n non-events	65	8	4	77
Observed Incidence P_{event}	0.044	0.111	0.429	0.083
MELD-UNOS Expected P_{event}	-	-	-	0.070
MELD+D Expected P_{event}	0.025	0.068	0.317	0.054

red indicates unfavourable reclassification, green indicates favourable reclassification

8.3 Discussion

Reactivation of the MELD disease aetiology variable led to predictions for 90-day wait-list mortality that were consistent with excellent accuracy of discrimination. Despite this positive finding, the discriminatory performance of the MELD+D was not found to be significantly different than that of the MELD-UNOS. This finding is consistent with

those reported by Kamath et al, who found that withdrawal of the disease aetiology variable from the MELD resulted in only a marginal decline, if any, in the c-statistic (largest change -0.02).[24] This lack of influence of the disease aetiology variable, despite an apparently large strength of association between non-alcohol-related, non-cholestatic primary liver disease and wait-list failure in this cohort, emphasizes the relative insensitivity of the area under the ROC curve measure to changes in model performance brought about by the addition of new predictors.[83, 89] It is known that if a model, by virtue of containing strong predictors, already possesses good discrimination, that it is difficult to further improve on the area under the ROC curve by adding new predictor variables, even if they are strongly associated with the outcome.[72] In this study, at 0.885, the area under the ROC curve for the MELD-UNOS may leave little room for further improvement. The result from the analysis comparing the areas under the ROC curves was supported by those of the reclassification analysis, with the NRI and IDI both failing to confirm any significant improvement in relative discrimination brought about by the reactivation of the disease aetiology variable. If anything, the NRI suggested a trend towards potentially harmful reclassification brought about by the reactivation of the disease aetiology variable.

The MELD+D produced an estimate for cohort 90-day wait-list mortality that was significantly less than the observed incidence. The Hosmer-Lemeshow goodness-of-fit test revealed no significant lack-of-fit; however the validity of that analysis can be questioned due to low event occurrence. Subjectively, the MELD+D appeared to have the tendency to underestimate the incidence of wait-list failure. By virtue of not convincingly satisfying any of the pre-specified study criteria for calibration accuracy, it must be

concluded that the calibration of the MELD+D was not acceptable. There was certainly no evidence generated that indicated that the calibration of the MELD+D was in any way superior to that of the MELD-UNOS.

8.4 Conclusions

Despite the finding of a significant association between the presence of non-alcohol-related, non-cholestatic liver disease and mortality in the cohort of Atlantic Canadian adults with ESLD awaiting liver transplantation, the reactivation of the MELD disease aetiology variable produced no significant improvement in the MELD model's ability to accurately predict 90-day wait-list mortality in the validation cohort. From the perspective of discrimination, this may be due to the finding that the discrimination of the MELD-UNOS is already very good.

Chapter Nine: Results – Comparison of Predictive Accuracy between the Child-Turcotte-Pugh Score and the MELD-UNOS

No report on the predictive accuracy of the MELD can be considered to be complete without a performance comparison to the historical gold-standard, the Child-Turcotte-Pugh (CTP) disease severity index (see Table 1.2)[22, 23]. As indicated in section 5.4, the CTP score could be calculated for 77 of the 84 individuals from the ESLD-indication cohort with a non-censored 90-day wait-list outcome (a serum albumin value was missing on seven patients). This analysis is therefore based on the comparison of the accuracy of the CTP score and the MELD-UNOS to predict 90-day mortality in this group of 77 candidates. In this group of 77 there were six mortalities (7.8%) within 90 days of wait-list assignment.

The median CTP score for this group was nine (range: 5 to 14)(see Figure 9.1). Five of the candidates (6.5%) were CTP class “A”; 37 (48.1%) were CTP “B”; and 35 (45.4%) were CTP class “C”. The CTP scores of those patients that experienced wait-list mortality within 90 days of registration (median: 12, range: 9 to 14) were significantly greater than those of individuals that survived 90 days (median: 9, range: 5 to 13)($p = 0.007$). The incidence of 90-day wait-list failure by CTP classification group was; 0.0 percent for CTP A; 2.7 percent for CTP B; and 14.3 percent for CTP C patients. The relationship between the CTP classification of the candidates and their MELD-UNOS score is shown in Figure 9.2.

As discussed in section 2.2.3, the CTP tool quantifies severity of ESLD and permits ranking of patients with regards to their ESLD severity but it does not generate a probability estimate for ESLD-related mortality. Consequently, discrimination is the only

Figure 9.1: CTP score at the time of wait-list registration for the 77 candidates from the validation cohort (n = 84) with complete data for determination of the score

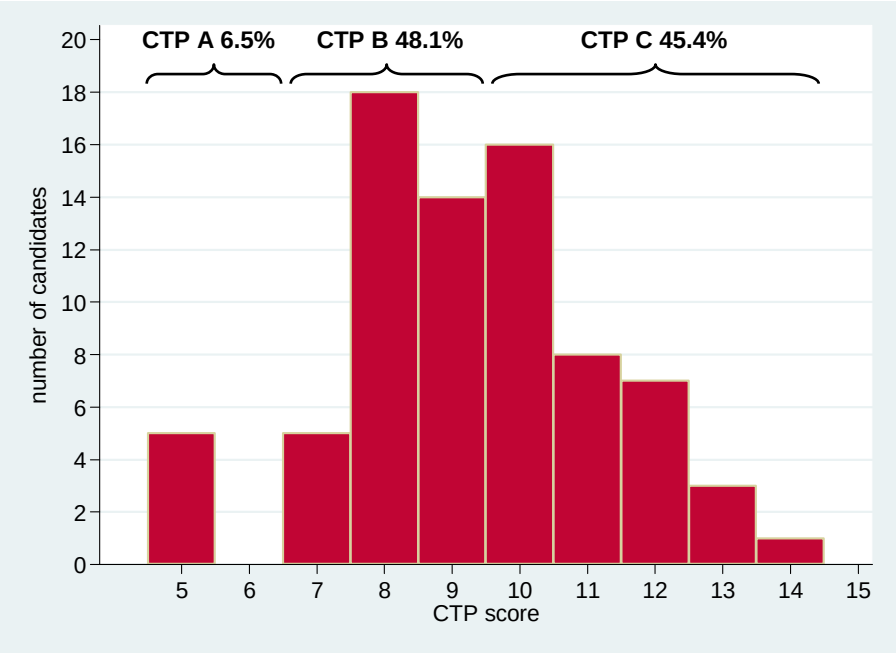
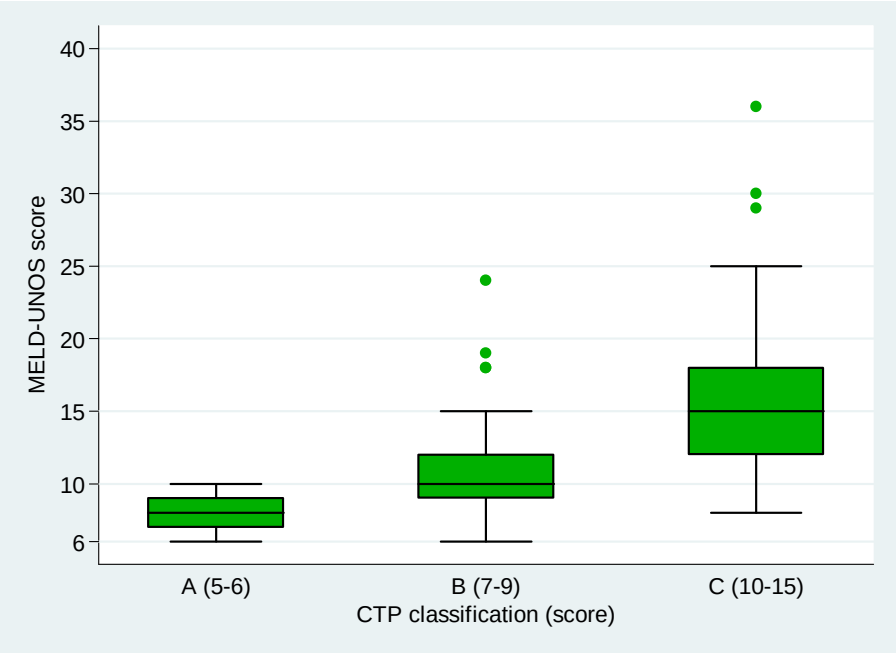
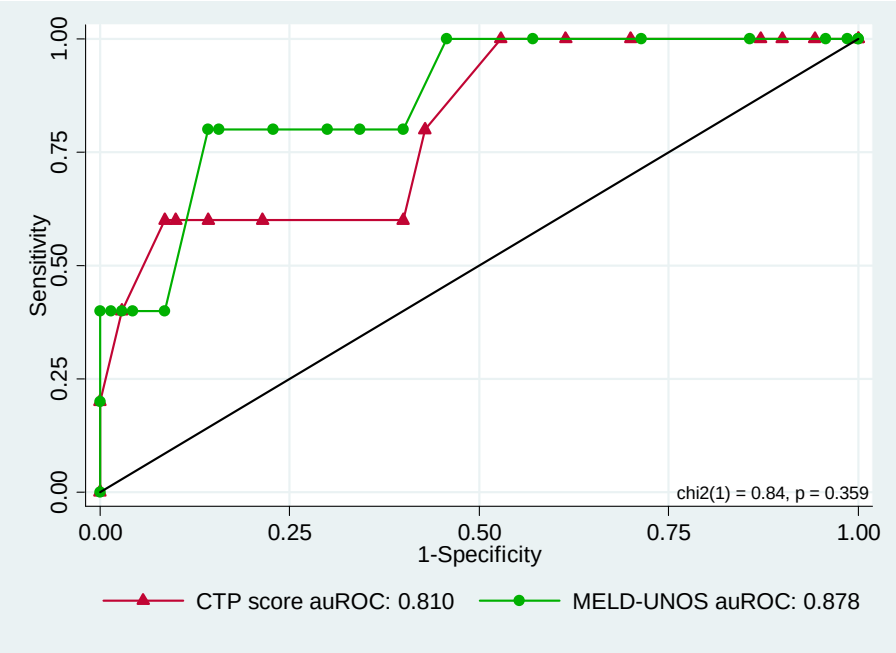


Figure 9.2: Relationship between CTP classification and MELD-UNOS score for the 77 patients in the validation cohort (n = 84) for which the CTP score could be calculated



performance characteristic that is applicable to the evaluation of the predictive accuracy of the CTP. The area under the non-parametric ROC curve for the CTP score prediction for the occurrence of 90-day mortality in the validation cohort, and its BcA bootstrap confidence interval, was 0.810 (95%CI: 0.595 – 0.967). This value would be qualified as reflecting "excellent" discrimination.[27] Although it was less than the area under the curve for the MELD-UNOS, 0.878 (95%CI: 0.700 – 1.000), non-parametric comparison of the areas revealed no statistically significant difference between the two ($\chi^2(1) = 0.84$, $p = 0.359$)(see Figure 9.3).

Figure 9.3: Comparison of areas under the non-parametric ROC curves for the CTP score and MELD-UNOS prediction of 90-day wait-list mortality in the 77 candidates with non-censored 90-day outcomes for which the CTP score could be calculated



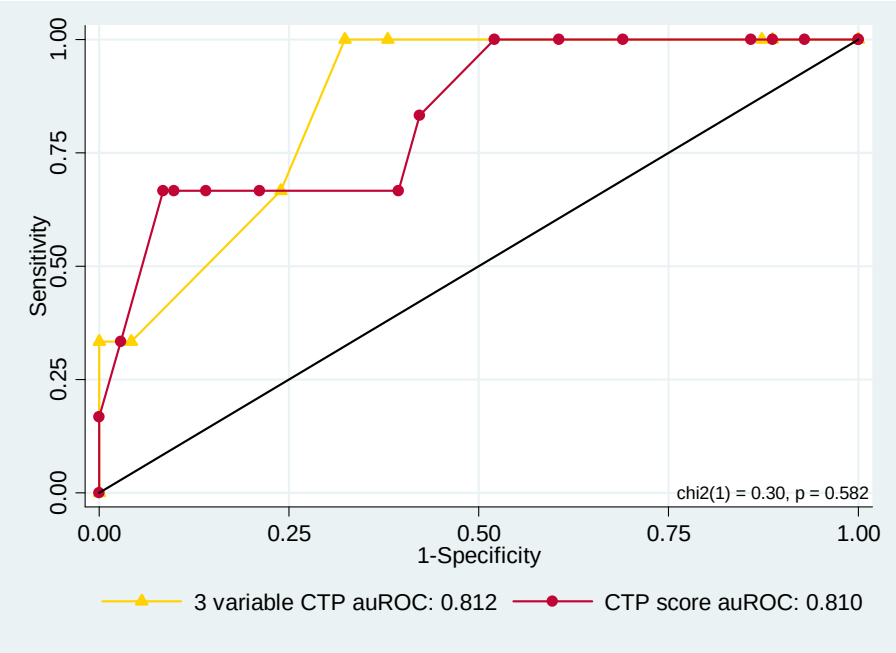
9.1 Discussion

In this cohort of patients with ESLD awaiting liver transplantation the CTP score was found to have excellent accuracy of discrimination. There was however, no indication that the CTP possessed superior discrimination to the MELD-UNOS, and if anything the absolute value of the area under the curve for the MELD-UNOS estimation of 90-day failure suggests that the MELD-UNOS may be the superior of the two. These results approximate the findings of other investigators. In their study of the MELD Kamath et al determined the c-statistic for the prediction of 3-month mortality by the MELD was 0.87 (95%CI: 0.82 – 0.92), while the c-statistic for the CTP was 0.84 (95%CI: 0.78 – 0.90).[24] The c-statistics from the two predictive instruments were not compared in that study. In the study by the UNOS Disease Severity Score Committee the area under the ROC curve for the MELD-UNOS score prediction of 3-month mortality, 0.83 (95%CI: 0.81 – 0.84), was compared to that for the CTP score, 0.76 (95%CI: 0.74 – 0.79), and was found to be significantly different ($p < 0.001$).[15] As was discussed in section 1.6, this analysis was challenged by Heuman and Mihas, who in their own analysis of UNOS data found that the c-statistics quantifying the MELD and CTP-related predictions for the occurrence of 90-day mortality, respectively 0.759 and 0.766, did not differ significantly.[39] In his study of candidates from the University of Alberta liver transplant service, Burak found the areas under the ROC curve for the CTP score and MELD score prediction of 90-day mortality to be very similar (respectively 0.78; 95%CI: 0.69 – 0.86 vs. 0.79; 95%CI: 0.70 – 0.88, $p = 0.804$).[17]

To an extent the similarity of the discriminatory performance of the MELD and CTP predictive tools is not surprising. The two instruments share two variables in

common, and the occurrence of ascites and renal dysfunction (with subsequent influence on serum creatinine) are frequently interrelated. Support for the relationship between ascites and renal dysfunction comes from the validation cohort itself, where the serum creatinine concentrations for individuals with ascites (median: 90.5, range: 16 to 680) were found to be significantly greater than those of individuals without ascites (median: 65, range: 41 to 202)(p = 0.015). Of interest, when the PSE and serum albumin concentration variables were withdrawn from the CTP index, the area under the non-parametric ROC curve for the three variable risk index's prediction of 90-day mortality was still 0.812 (95%CI: 0.644 – 0.942), which implies that these two variables contributed little to the discriminative ability of the CTP when applied to this validation cohort (see Figure 9.4).

Figure 9.4: Comparison between the areas under the non-parametric ROC curves for the prediction of the occurrence of 90-day wait-list mortality by a three variable risk index CTP score (serum total bilirubin concentration, serum INR, and ascites) and the standard CTP score



It is acknowledged that the manner in which the CTP score as calculated for this analysis may have led to some degree of constraint on its discriminative ability. Due to the retrospective nature of data collection pertaining to symptoms of decompensated ESLD, other than ascertaining whether ascites or encephalopathy was ever present, it was not possible to consistently determine the severity of ascites or PSE based on the review of the clinical documentation. Although this methodology may have somewhat limited the discrimination of the CTP, it also highlights the difficulty in assigning the status of these subjective clinically-based variables. The magnitude of this potential methodological limitation on the performance of the CTP score cannot be determined as all of the five ordinal categorical variables in the CTP index are given the same weight. This highlights another criticism of the CTP index, as it seems improbable that the survival implications of each of the five variables in the index are equal.[24] All-in-all, based on these results, and acknowledging the potential limitations of the analysis, it would appear to be safe to say that the MELD-UNOS performed, at least, as well as the CTP score in predicting the occurrence of 90-day wait-list failure in this cohort. The MELD-UNOS has the added benefit that it is not encumbered by the difficulties inherent to the subjective clinically-based variables in the CTP index.

9.2 Conclusions

This study did not produce any evidence to indicate that the CTP score predicts the occurrence of 90-day mortality in Atlantic Canadian adults with ESLD awaiting liver transplantation with any greater accuracy than the MELD-UNOS. The MELD-UNOS appears to produce survival estimates that are, at least, as accurate as the CTP score, yet it

offers the advantage that it does not rely on subjective variables to generate these estimates.

Chapter Ten: Summary of Conclusions

10.1 Primary Conclusions

This independent external validation study found that the MELD-UNOS prognostic model accurately estimates the occurrence of 90-day mortality in Atlantic Canadian adults with ESLD awaiting liver transplantation. The MELD-UNOS was found to have excellent accuracy of discrimination, which is the sole performance measure of relevance if the predictions generated by the model are used solely for ranking of liver transplant candidates. Evaluation of model calibration found no evidence that the MELD-UNOS generated inaccurate probability estimates for 90-day wait-list failure, although to an extent the analysis of calibration was hampered by low event occurrence. It appeared that the MELD-UNOS may tend to underestimate 90-day wait-list mortality in candidates with greater ESLD severity. The predictive accuracy of the MELD-UNOS was also compared to that of other prognostic instruments for ESLD; the MELDNa, the MELD with the disease aetiology variable reactivated and the CTP score. In none of these comparisons was there any evidence produced to indicate that these alternate instruments produced superior predictions than the MELD-UNOS (if anything the MELD-UNOS appeared to be superior to the alternatives). Overall, as the results of this study indicate that the estimates for 90-day wait-list survival generated by the MELD-UNOS are generalizable to adult liver transplant candidates with ESLD residing in Atlantic Canada; it would appear to be appropriate to utilize the MELD-UNOS score to prioritize patients from this population for deceased donor liver allograft allocation. Additionally, these results provide support for the adoption of the MELD-UNOS as the prioritization disease severity index in the inaugural Canadian national organ allocation system.

10.2 Secondary Conclusions

As a secondary objective, this study evaluated the predictive accuracy of the MELD model which includes serum sodium concentration as an independent variable (i.e. the MELDNa), to predict the occurrence of 90-day wait-list mortality in Atlantic Canadian adults with ESLD awaiting liver transplantation. From this analysis, the MELDNa was deemed to have excellent discrimination accuracy. Calibration analysis, which was constrained by low event occurrence, disclosed no evidence to indicate that the probability estimates generated by the MELDNa differed significantly from the observed incidence of wait-list failure. Overall, the MELDNa appeared to be generalizable to adult liver transplant candidates with ESLD residing in Atlantic Canada; however it did not appear to offer any improvement in prognostication over the MELD-UNOS.

Despite the finding of a significant association between the presence of non-alcohol-related, non-cholestatic liver disease and mortality in the cohort of individuals with ESLD awaiting liver transplantation, the reactivation of the MELD disease aetiology variable produced no significant improvement in the MELD model's ability to predict 90-day wait-list mortality in the validation cohort.

No report on a prognostic instrument for ESLD would be complete without a comparison to the Child-Turcotte-Pugh disease severity index. In a comparison of discriminatory performance, this study did not find any evidence to indicate that the CTP score predicts the occurrence of 90-day mortality in Atlantic Canadian adults with ESLD awaiting liver transplantation with any greater accuracy than the MELD-UNOS. The MELD-UNOS appears to produce survival estimates that are, at least, as accurate as the

CTP score, yet it offers the advantage that it does not rely on subjective variables to generate these estimates.

This study also provided characterization of the liver transplant referral population serviced by the Atlantic MOTP, the liver transplant assessment process, and the population of individuals assigned to the liver transplant waiting list over the first five years of operation after reactivation of the Atlantic MOTP liver transplant service on December 1st, 2004.

References

1. Starzl TE, Murase N, Marcos A, Fung JJ. History of multivisceral transplantation. In: Busattil RW, Klintmalm GK, editors. Transplantation of the Liver. 2nd ed. Philadelphia, PA: Elsevier Saunders; 2005. p. 3-22.

2. 2008 Annual Report of the U.S. Organ Procurement and Transplantation Network and the Scientific Registry of Transplant Recipients: Transplant Data 1998-2007.: Department of Health and Human Services, Health Resources and Services Administration, Healthcare Systems Bureau, Division of Transplantation, Rockville, MD; 2008. p. 3-4.

3. D'Amico G, Garcia-Tsao G, Pagliaro L. Natural history and prognostic indicators of survival in cirrhosis: a systematic review of 118 studies. J Hepatol. 2006 Jan;44(1):217-231.

4. Sears D, Davis GL. Natural history of hepatitis C. In: Busuttil RW, Klintmalm GK, editors. Transplantation of the Liver. 2nd ed. Philadelphia, PA: Elsevier Saunders; 2005. p. 132-134.

5. 2008 Annual Report of the U.S. Organ Procurement and Transplantation Network and the Scientific Registry of Transplant Recipients: Transplant Data 1998-2007.: Department of Health and Human Services, Health Resources and Services Administration, Healthcare Systems Bureau, Division of Transplantation, Rockville, MD; 2008. p. 15.

6. Evans IVR, Belle SH. U.S. trends in liver transplantation, 1988 to 2001. In: Busuttil RW, Klintmalm GK, editors. Transplantation of the Liver. 2nd ed. Philadelphia, PA: Elsevier Saunders; 2005. p. 1313-1314.

7. Canadian Institute for Health Information. Treatment of End-Stage Organ Failure in Canada, 1995 to 2004 (2006 Annual Report). Ottawa, ON: Canadian Institute for Health Information (CIHI); 2006. p. 61-73.

8. Canadian Institute for Health Information. Treatment of End-Stage Organ Failure in Canada, 1999 to 2008 - CORR Annual Report 2010. Ottawa, ON: Canadian Institute for Health Information (CIHI); 2010. p. 27-30.

9. Brown RS, Rush SH, Rosen HR, Langnas AN, Klintmalm GB, Hanto DW, Punch JD. Liver and intestine transplantation. Am J Transplant. 2004;4 Suppl 9:81-92.

10. Adam R, McMaster P, O'Grady JG, Castaing D, Klempnauer JL, Jamieson N, Neuhaus P, Lerut J, Salizzoni M, Pollard S, Muhlbacher F, Rogiers X, Garcia Valdecasas JC, Berenguer J, Jaeck D, Moreno Gonzalez E. Evolution of liver transplantation in Europe: report of the European Liver Transplant Registry. Liver Transpl. 2003 Dec;9(12):1231-1243.

11. 2005 Annual Report of the U.S. Organ Procurement and Transplantation Network and the Scientific Registry of Transplant Recipients: Transplant Data 1995-2004.: Department of Health and Human Services, Health Resources and Services Administration, Healthcare Systems Bureau, Division of Transplantation, Rockville, MD; 2005. p. 9-1, 9-16, 19-23, 19-28.

12. 2008 Annual Report of the U.S. Organ Procurement and Transplantation Network and the Scientific Registry of Transplant Recipients: Transplant Data 1998-2007.: Department of Health and Human Services, Health Resources and Services Administration, Healthcare Systems Bureau, Division of Transplantation, Rockville, MD; 2008. p. 9-1, 9-18, 19-27, 19-33.

13. Organ Procurement and Transplant Network. advanced data report (wait-list additions and removals January 1995 - February 28 2010). United Network for Organ Sharing; 2010 [updated 2010; cited 2010 May 10]; Available from: http://optn.transplant.hrsa.gov/latestData/advancedData.asp.

14. Canadian Institute for Health Information. 2007 Annual Report - Treatment of End-Stage Organ Failure in Canada, 1996 to 2005. Ottawa, ON: Canadian Institute for Health Information (CIHI); 2008. p. 55-68.

15. Wiesner R, Edwards E, Freeman R, Harper A, Kim R, Kamath P, Kremers W, Lake J, Howard T, Merion RM, Wolfe RA, Krom R. Model for end-stage liver disease (MELD) and allocation of donor livers. Gastroenterology. 2003 Jan;124(1):91-96.

16. Kim WR, Biggins SW, Kremers WK, Wiesner RH, Kamath PS, Benson JT, Edwards E, Therneau TM. Hyponatremia and mortality among patients on the liver-transplant waiting list. N Engl J Med. 2008 Sep 4;359(10):1018-1026.

17. Burak KW. Does the MELD score predict mortality before and after liver transplantation in Alberta? Calgary, AB: University of Calgary; 2005.

18. Freeman RB. Liver allocation: the U.S. model. In: Busuttil RW, Klintmalm GK, editors. Transplantation of the Liver. 2nd ed. Philadelphia, PA: Elsevier Saunders; 2005. p. 63-77.

19. Freeman RB, Jr., Edwards EB. Liver transplant waiting time does not correlate with waiting list mortality: implications for liver allocation policy. Liver Transpl. 2000 Sep;6(5):543-552.

20. Organ Procurement and Transplantation Network. Health Resources and Services Administration, HHS. Final rule. Fed Regist. 1999 Oct 20;64(202):56650-56661.

21. Freeman RB, Jr., Wiesner RH, Harper A, McDiarmid SV, Lake J, Edwards E, Merion R, Wolfe R, Turcotte J, Teperman L. The new liver allocation system: moving toward evidence-based transplantation policy. Liver Transpl. 2002 Sep;8(9):851-858.

22. Child CG, Turcotte JG. Surgery and portal hypertension. Major Probl Clin Surg. 1964;1:1-85.

23. Pugh RN, Murray-Lyon IM, Dawson JL, Pietroni MC, Williams R. Transection of the oesophagus for bleeding oesophageal varices. Br J Surg. 1973 Aug;60(8):646-649.

24. Kamath PS, Wiesner RH, Malinchoc M, Kremers W, Therneau TM, Kosberg CL, D'Amico G, Dickson ER, Kim WR. A model to predict survival in patients with end-stage liver disease. Hepatology. 2001 Feb;33(2):464-470.

25. Backman L, Forsythe JLR, Miranda B, Cuende N, Canon J, Margarit C, Gerling T, Persijn GG. Liver organ allocation: the European model. In: Busuttil RW, Klintmalm GK, editors. Transplantation of the Liver. 2nd ed. Philadelphia, PA: Elsevier Saunders; 2005. p. 79-91.

26. Malinchoc M, Kamath PS, Gordon FD, Peine CJ, Rank J, ter Borg PC. A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. Hepatology. 2000 Apr;31(4):864-871.

27. Hosmer DW, Lemeshow S. Assessing the fit of the model. Applied logistic regression. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc.; 2000. p. 143-167, 186-188.

28. Starzl TE, Marchioro TL, Vonkaulla KN, Hermann G, Brittain RS, Waddell WR. Homotransplantation of the Liver in Humans. Surg Gynecol Obstet. 1963 Dec;117:659-676.

29. United Network for Organ Sharing (UNOS). Child-Turcotte-Pugh (CTP) Score. [cited June 7, 2010]; Available from: http://www.unos.org/resources/glossary.asp#C.

30. United Network for Organ Sharing (UNOS). UNOS Policy Proposal - New UNOS/OPTN status criteria for prioritizing liver transplant candidates with pre-existing liver disease. Proposed Amendment UNOS Policy 3.6 (Allocation of Livers). March 29, 2001 [updated March 29, 2001; cited June 7, 2010]; Available from: http://www.unos.org/PublicComment/pubcommentProp_2.pdf.

31. Lucey MR, Brown KA, Everson GT, Fung JJ, Gish R, Keeffe EB, Kneteman NM, Lake JR, Martin P, McDiarmid SV, Rakela J, Shiffman ML, So SK, Wiesner RH. Minimal criteria for placement of adults on the liver transplant waiting list: a report of a national conference organized by the American Society of Transplant Physicians and the American Association for the Study of Liver Diseases. Liver Transpl Surg. 1997 Nov;3(6):628-637.

32. Trey C, Burns DG, Saunders SJ. Treatment of hepatic coma by exchange blood transfusion. N Engl J Med. 1966 Mar 3;274(9):473-481.

33. Adams RD, Foley JM. The neurological disorder associated with liver disease. Res Publ Assoc Res Nerv Ment Dis. 1953;32:198-237.

34. Fast BB, Wolfe SJ, Stormont JM, Davidson CS. Antibiotic therapy in the management of hepatic coma. AMA Arch Intern Med. 1958 Feb;101(2):467-475.

35. Conn HO, Leevy CM, Vlahcevic ZR, Rodgers JB, Maddrey WC, Seeff L, Levy LL. Comparison of lactulose and neomycin in the treatment of chronic portal-systemic encephalopathy. A double blind controlled trial. Gastroenterology. 1977 Apr;72(4 Pt 1):573-583.

36. Parsons-Smith BG, Summerskill WH, Dawson AM, Sherlock S. The electroencephalograph in liver disease. Lancet. 1957 Nov 2;273(7001):867-871.

37. Institute of Medicine. Analysis of waiting times. In Committee on Organ Transplantation. Assessing current policies and the potential impact of the DHHS final rule. Washington, DC: National Academy Press; 1999. p. 57-78.

38. Organ Procurement and Transplantation Network--HRSA. Final rule with comment period. Fed Regist. 1998 Apr 2;63(63):16296-16338.

39. Heuman DM, Mihas A. Utility of the MELD score for assessing 3-month survival in patients with liver cirrhosis: one more positive answer. Gastroenterology. 2003 Sep;125(3):992-993; author reply 994-995.

40. Kamath PS, Kim WR. The model for end-stage liver disease (MELD). Hepatology. 2007 Mar;45(3):797-805.

41. United Network for Organ Sharing (UNOS). UNOS Policy 3.6 Organ Distribution: Allocation of Livers. [November 17, 2009; cited]; Available from: http://www.unos.org/PoliciesandBylaws2/policies/pdfs/policy_8.pdf.

42. Freeman RB, Wiesner RH, Edwards E, Harper A, Merion R, Wolfe R. Results of the first year of the new liver allocation plan. Liver Transpl. 2004 Jan;10(1):7-15.

43. Olthoff KM, Brown RS, Jr., Delmonico FL, Freeman RB, McDiarmid SV, Merion RM, Millis JM, Roberts JP, Shaked A, Wiesner RH, Lucey MR. Summary report of a national conference: Evolving concepts in liver allocation in the MELD and PELD era. December 8, 2003, Washington, DC, USA. Liver Transpl. 2004 Oct;10(10 Suppl 2):A6-22.

44. Wiesner R, Lake JR, Freeman RB, Gish RG. Model for end-stage liver disease (MELD) exception guidelines. Liver Transpl. 2006 Dec;12(12 Suppl 3):S85-87.

45. Kanwal F, Dulai GS, Spiegel BM, Yee HF, Gralnek IM. A comparison of liver transplantation outcomes in the pre- vs. post-MELD eras. Aliment Pharmacol Ther. 2005 Jan 15;21(2):169-177.

46. Ioannou GN, Perkins JD, Carithers RL, Jr. Liver transplantation for hepatocellular carcinoma: impact of the MELD allocation system and predictors of survival. Gastroenterology. 2008 May;134(5):1342-1351.

47. Stone MJ, Fulmer JM, Klintmalm GK. Transplantation for primary hepatic malignancy. In: Busattil RW, Klintmalm GK, editors. Transplantation of the Liver. 2nd ed. Philadelphia, PA: Elsevier Saunders; 2005. p. 211-231.

48. Mazzaferro V, Regalia E, Doci R, Andreola S, Pulvirenti A, Bozzetti F, Montalto F, Ammatuna M, Morabito A, Gennari L. Liver transplantation for the treatment of small hepatocellular carcinomas in patients with cirrhosis. N Engl J Med. 1996 Mar 14;334(11):693-699.

49. Biggins SW, Bambha K. MELD-based liver allocation: who is underserved? Semin Liver Dis. 2006 Aug;26(3):211-220.

50. Merion RM, Schaubel DE, Dykstra DM, Freeman RB, Port FK, Wolfe RA. The survival benefit of liver transplantation. Am J Transplant. 2005 Feb;5(2):307-313.

51. Hecker R, Sherlock S. Electrolyte and circulatory changes in terminal liver failure. Lancet. 1956 Dec 1;271(6953):1121-1125.

52. Heuman DM, Abou-Assi SG, Habib A, Williams LM, Stravitz RT, Sanyal AJ, Fisher RA, Mihas AA. Persistent ascites and low serum sodium identify patients with cirrhosis and low MELD scores who are at high risk for early death. Hepatology. 2004 Oct;40(4):802-810.

53. Biggins SW, Rodriguez HJ, Bacchetti P, Bass NM, Roberts JP, Terrault NA. Serum sodium predicts mortality in patients listed for liver transplantation. Hepatology. 2005 Jan;41(1):32-39.

54. Ruf AE, Kremers WK, Chavez LL, Descalzi VI, Podesta LG, Villamil FG. Addition of serum sodium into the MELD score predicts waiting list mortality better than MELD alone. Liver Transpl. 2005 Mar;11(3):336-343.

55. Biggins SW, Kim WR, Terrault NA, Saab S, Balan V, Schiano T, Benson J, Therneau T, Kremers W, Wiesner R, Kamath P, Klintmalm G. Evidence-based incorporation of serum sodium concentration into MELD. Gastroenterology. 2006 May;130(6):1652-1660.

56. Londono MC, Cardenas A, Guevara M, Quinto L, de Las Heras D, Navasa M, Rimola A, Garcia-Valdecasas JC, Arroyo V, Gines P. MELD score and serum sodium in the prediction of survival of patients with cirrhosis awaiting liver transplantation. Gut. 2007 Sep;56(9):1283-1290.

57. Bazarah SM, Peltekian KM, McAlister VC, Bitter-Suermann H, MacDonald AS. 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 Aug;27(4):162-167.

58. O'Grady JG, Alexander GJ, Hayllar KM, Williams R. Early indicators of prognosis in fulminant hepatic failure. Gastroenterology. 1989 Aug;97(2):439-445.

59. Kondro W. Fragmented organ donation programs hinder progress. CMAJ. 2006 Oct 24;175(9):1043.

60. Kondro W, Hebert PC. They deserved better. CMAJ. 2007 May 22;176(11):1557, 1559.

61. Canadian Blood Services. Organ donation and transplantation streamlined in new national system. Edmonton, Alberta: Canadian Blood Services; August 12, 2008 [updated August 12, 2008; cited 2010 October 25]; Available from: http://www.cst-transplant.ca/cmsUploads/cst/File/News_Release_Aug_8_version_ALTA_AND_HC_APPROVED_FINAL-_ENGLISH.pdf.

62. Barshes NR, Becker NS, Washburn WK, Halff GA, Aloia TA, Goss JA. Geographic disparities in deceased donor liver transplantation within a single UNOS region. Liver Transpl. 2007 May;13(5):747-751.

63. Porta M, editor. A dictionary of epidemiology. 5th ed. New York, NY: Oxford University Press, Inc.; 2008.

64. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000 Feb 29;19(4):453-473.

65. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med. 1999 Mar 16;130(6):515-524.

66. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996 Dec;49(12):1373-1379.

67. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol. 1995 Dec;48(12):1503-1510.

68. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996 Feb 28;15(4):361-387.

69. Angermayr B, Cejna M, Karnel F, Gschwantler M, Koenig F, Pidlich J, Mendel H, Pichler L, Wichlas M, Kreil A, Schmid M, Ferlitsch A, Lipinski E, Brunner H, Lammer J, Ferenci P, Gangl A, Peck-Radosavljevic M. Child-Pugh versus MELD score in predicting survival in patients undergoing transjugular intrahepatic portosystemic shunt. Gut. 2003 Jun;52(6):879-885.

70. Schroeder RA, Marroquin CE, Bute BP, Khuri S, Henderson WG, Kuo PC. Predictive indices of morbidity and mortality after liver resection. Ann Surg. 2006 Mar;243(3):373-379.

71. MacKillop WJ, Bartfay E, Groome PA. Measuring the accuracy of prognostic judgements. In: Gospodarowicz MK, editor. Prognostic Factors in Cancer. New York: Wiley-Liss, Inc.; 2001. p. 127-145.

72. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clin Chem. 2008 Jan;54(1):17-23.

73. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Stat Med. 1997 May 15;16(9):965-980.

74. Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. Ann Intern Med. 2008 Nov 18;149(10):751-760.

75. Bartfay E, Bartfay WJ. Accuracy assessment of prediction in patient outcomes. J Eval Clin Pract. 2008 Feb;14(1):1-10.

76. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform. 2005 Oct;38(5):404-415.

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

78. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988 Sep;44(3):837-845.

79. Obuchowski NA, Lieber ML. Confidence intervals for the receiver operating characteristic area in studies with small samples. Acad Radiol. 1998 Aug;5(8):561-571.

80. Faraggi D, Reiser B. Estimation of the area under the ROC curve. Stat Med. 2002 Oct 30;21(20):3093-3106.

81. Hajian-Tilaki KO, Hanley JA. Comparison of three methods for estimating the standard error of the area under the curve in ROC analysis of quantitative data. Acad Radiol. 2002 Nov;9(11):1278-1285.

82. Diamond GA. What price perfection? Calibration and discrimination of clinical prediction models. J Clin Epidemiol. 1992 Jan;45(1):85-89.

83. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007 Feb 20;115(7):928-935.

84. Brier GW. Verification of forecasts expressed in terms of probability. Monthly Weather Review. 1950 January 1950;78(1):1-3.

85. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. Ann Intern Med. 2006 Jul 4;145(1):21-29.

86. Mackillop WJ, Quirt CF. Measuring the accuracy of prognostic judgments in oncology. J Clin Epidemiol. 1997 Jan;50(1):21-29.

87. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med. 1986 Sep-Oct;5(5):421-433.

88. Redelmeier DA, Bloch DA, Hickam DH. Assessing predictive accuracy: how to compare Brier scores. J Clin Epidemiol. 1991;44(11):1141-1146.

89. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol. 2004 May 1;159(9):882-890.

90. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008 Jan 30;27(2):157-172; discussion 207-112.

91. Bollman RD, Clemenson HA. Structure and change in Canada's rural demography: an update to 2006 with provincial detail. In: Agricultural Division, editor. Ottawa: Statistics Canada; 2008. p. 6.

92. Benson J. Re: MELD and MELDNa survival functions? Rochester, Minnesota: Statistical Programmer/Analyst, Division of Biomedical Statistics and Informatics, Mayo Clinic; April 28, 2009.

93. Pepe MS, Longton G, Janes H. Estimation and comparison of receiver operating characteristic curves (st0154). The Stata Journal. 2009;9(1):1-16.

94. Hosmer DW, Lemeshow S, May S. Descriptive methods for survival data. In: Hosmer DW, Lemeshow S, May S, editors. Applied survival analysis: regression modelling of time-to-event data. 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2008. p. 16-66.

95. Kushner RF. Chapter 75. Evaluation and Management of Obesity. In: Fauci AS, Braunwald E, Kasper DL, Hauser SL, Longo DL, Jameson JL, et al., editors. Harrison's Principles of Internal Medicine. 17th ed: McGraw-Hill Companies; 2008.

96. Therneau T. question re: N Engl J Med 2008;359:1018-26. Rochester, MN: Division of Biomedical Statistics and Informatics, Mayo Clinic; August 12, 2010.

Appendix A: Severity Indices for Portosystemic Encephalopathy

Table A.1: Stages in the onset and development of hepatic coma per Trey et al [32-34]

Stage/Grade	mental state	tremor
Stage 1 (prodrome)	• euphoria or occasionally depression • fluctuant, mild confusion • slowness of mentation and affect • slurred speech • disorder in sleep rhythm	slight tremor
Stage 2 (impending coma)	• accentuation of stage 1 • drowsiness • inappropriate behaviour • able to maintain sphincter control	present (easily elicited)
Stage 3 (stupor)	• sleeps most of the time but is rousable • speech is incoherent • confusion is marked • incontinent	usually present (if patient can co-operate
Stage 4 (deep coma)	• may or may not respond to painful stimuli	usually absent

Table A.2: West Haven criteria for grading of hepatic encephalopathy per Conn et al [35, 36]

grade	criteria
grade 0	• no abnormality detected
grade 1	• trivial lack of awareness, euphoria or anxiety • shortened attention span • impairment of addition or subtraction
grade 2	• lethargy • disorientation for time • obvious personality change • inappropriate behaviour
grade 3	• somnolence to semi-stupor • responsive to stimuli • confused • gross disorientation • bizarre behaviour
grade 4	• coma • unable to test mental state

Appendix B: King's College Criteria for Liver Transplantation in Acute Liver Failure

Table B.1: King's College criteria for liver transplantation in acute liver failure [58]

acetaminophen-induced acute liver failure:
• pH < 7.3
• **OR** all of:
  • INR > 6.5
  • serum creatinine > 300 μmol/L
  • encephalopathy ≥ grade III

non-acetaminophen-induced acute liver failure:
• INR > 6.5
• **OR** 3 or more of:
  • "unfavourable" cause (non-A non-B hepatitis, Drug-Induced Liver Injury (DILI)
  • duration of jaundice before onset of encephalopathy > 7 days
  • age < 10 years OR > 40 years
  • INR > 3.5
  • serum total bilirubin > 300 μmol/L

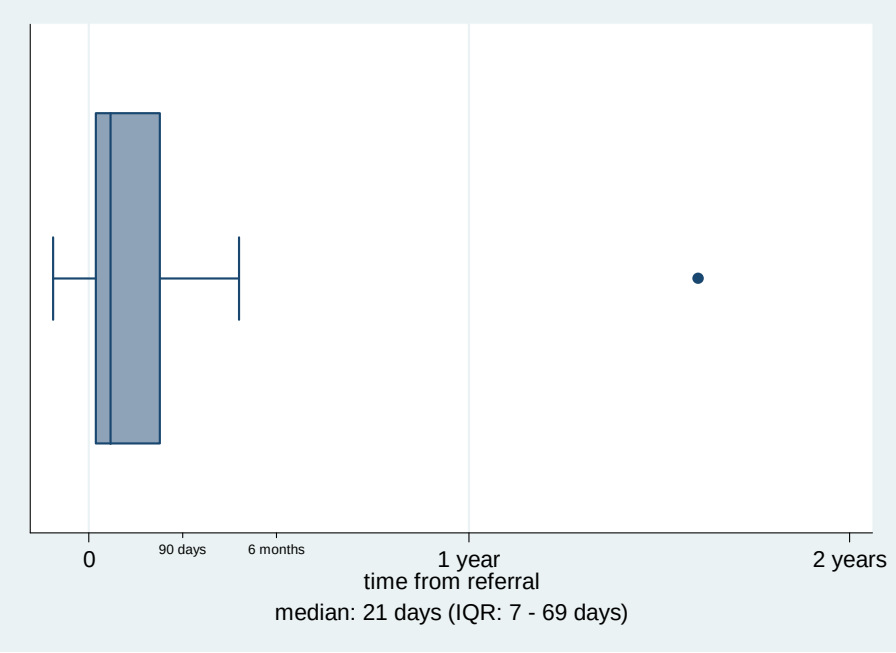
Appendix C: Analysis of Patients that Died Before or During Assessment for Liver Transplant

Of the 353 individuals that were confirmed to have been referred to the Atlantic MOTP for consideration of first liver transplant between December 1st, 2004 and November 30th, 2009, 25 (7.0%) died before (n = 16) or during (n = 6) multidisciplinary transplant assessment. It must be recognized when interpreting the data of this group that assessment of candidacy was incomplete, at best, to completely absent for most. Therefore, in addition to the implications of incomplete evaluation on the quality of the data, it also is difficult to discern if any of these deceased individuals would have actually been candidates for liver transplantation.

Compared to the 328 individuals that did not die, there were proportionately more males in the group that died, 76.0 versus 58.5 percent, although this difference was not statistically significant (p = 0.086). The median age of the group that died was 58.6 years (range: 18.3 – 66.7). Their was no evidence to suggest that the age distribution of this group was different from those that survived (p = 0.316). The pattern of province of residence did not appear to differ between those that died and those that did not (p = 0.993), with Nova Scotia being the commonest (52.0%), followed by New Brunswick (28.0%), Newfoundland and Labrador (16.0%). and Prince Edward Island (4.0%). The top three primary liver disease diagnoses recorded were: alcohol-related ESLD (28.0%), ESLD due to chronic HCV infection (24.0%) and, likely reflecting the degree of evaluation of this group, cryptogenic cirrhosis (16.0%). Only one of the 25 that died had acute liver failure. This patient had sub-acute liver failure, possibly related to autoimmune hepatitis, and died during transplant assessment from acute liver failure,

Candida septicaemia, and an acute myocardial infarction. Reflecting the illness acuity of the group as a whole, 20 of the 25 patients were referred by a gastroenterologist (n = 11) or hepatologist (n = 9).

Figure C.1: Time elapsed from referral to death before (n = 16) or during (n = 9) liver transplant candidacy assessment



A cause of death was recorded for six of the 16 patients that died before assessment by the Atlantic MOTP, and all nine of those that died during assessment. Of the 15 individuals in which a cause of death was documented, complications of decompensated ESLD were the culprit in 12, (sub) acute liver failure in one, and acute myocardial infarction was the cause of death for two patients. The median time from receipt of referral to death was 21 days (range: -34 – 585). Figure C.1 portrays the elapsed time from referral to date of death for this group. The outlying patient was proactively referred with slowly progressive ESLD due to chronic graft-versus-host

disease following bone marrow transplant and died of an acute myocardial infarction a few weeks before she was to attend her transplant assessment. Additionally, it should be noted that three patients died 34, 16 and 1 day before their referral was received by the transplant service.

For 22 of the individuals that died, laboratory investigations, obtained a median of zero days (range: -56 – 96) from the date of referral, were available for the calculation of the MELD-UNOS and MELDNa scores. The median MELD-UNOS and MELDNa scores were 19 (range: 7 – 40) and 22 (range: 8 – 40).

The high illness acuity reflected by the risk scores and the generally brief time elapsed from referral to death leads one to speculate that either these individuals suffered from rapidly progressive liver disease, or they were identified late in the course of their ESLD (only one patient had acute liver failure). In light of these findings, the implementation of methods to increase physician accessibility to, and knowledge of, the availability and indications for liver transplantation, may improve outcomes for patients with irreversible acute and chronic liver diseases residing in the Atlantic Provinces.

Appendix D: Patterns of Overall Mortality in the Total Wait-List Cohort

Table D.1 Total cohort univariable analysis of overall wait-list mortality (continuous variables expressed as median (range) categorical as proportions (frequency))

variable	overall n = 153	wait-list survivor n = 138	wait-list failure n = 15	p-value
sex – male	54.9% (84)	54.4% (75)	60.0% (9)	0.676
age at wait-listing	56.0 (21.6-69.0)	56.0 (21.6-69.0)	54.8 (26.4-68.3)	0.554
province of residence:				0.291
NS	53.6% (82)	54.4% (75)	46.7% (7)	
NB	24.8% (38)	24.6% (34)	26.7% (4)	
NL	17.7% (27)	18.1% (25)	13.3% (2)	
PE	3.9% (6)	2.9% (4)	13.3% (2)	
rural postal code	24.8% (38)	24.6% (34)	26.7% (4)	1.000
any comorbidity	53.6% (82)	52.9% (73)	60.0% (9)	0.600
cardiovasc.	28.8% (44)	28.3% (39)	33.3% (5)	0.680
pulmonary	11.8% (18)	13.0% (18)	0.0% (0)	0.219
renal	5.9% (9)	5.1% (7)	13.3% (2)	0.216
diabetes	26.8% (41)	26.8% (37)	26.7% (4)	1.000
BMI	27 (18-42)	27 (18-42)	28 (23-35)	0.652
obese (BMI≥30)	34.0% (52)	32.6% (45)	46.7% (7)	0.275
active smoking	29.4% (45)	28.3% (39)	40.0% (6)	0.343

continues on next page

variable	overall n = 153	wait-list survivor n = 138	wait-list failure n = 15	p-value
blood group:				
O	41.8% (64)	37.7% (52)	80.0% (12)	
A	41.8% (64)	44.9% (62)	13.3% (2)	**0.021**
B	11.8% (18)	12.3% (17)	6.7% (1)	
AB	4.6% (7)	5.1% (7)	0.0% (0)	
primary indication:				
ESLD	75.8% (116)	76.1% (105)	73.3% (11)	
malignancy	20.9% (32)	21.0% (29)	20.0% (3)	0.377
ALF	3.3% (5)	2.9% (4)	6.7% (1)	
primary liver disease:				
HCC	20.3% (31)	20.3% (28)	20.0% (3)	
etoh	15.0% (23)	15.9% (22)	6.7% (1)	
HCV	14.4% (22)	13.8% (19)	20.0% (3)	
PBC	12.4% (19)	13.8% (19)	0.0% (0)	
PSC	11.1% (17)	11.6% (16)	6.7% (1)	0.223
cryptogenic	6.5% (10)	5.8% (8)	13.3% (2)	
AIH	5.9% (9)	4.4% (6)	20.0% (3)	
NASH	4.6% (7)	5.1% (7)	0.0% (0)	
HBV	0.7% (1)	0.7% (1)	0.0% (0)	
other	9.2% (14)	8.7% (12)	13.3% (2)	

continues on next page

variable	overall n = 153	wait-list survivor n = 138	wait-list failure n = 15	p-value
primary liver disease non-etoh, non-cholestatic	61.4% (94)	58.7% (81)	86.7% (13)	**0.048**
prevalence of specific liver diseases:				
etoh	24.2% (37)	25.4% (35)	13.3% (2)	0.525
HCV	26.8% (41)	26.1% (36)	33.3% (5)	0.547
cholestatic	24.8% (38)	26.8% (37)	6.7% (1)	0.118
manifestations of ESLD:				
any of the below	74.5% (114)	73.9% (102)	80.0% (12)	0.761
ascites	68.0% (104)	67.4% (93)	73.3% (11)	0.776
SBP	11.8% (18)	9.4% (13)	33.3% (5)	**0.006**
PHGIH	28.1% (43)	29.0% (40)	20.0% (3)	0.559
PSE	43.1% (66)	41.3% (57)	60.0% (9)	0.165
HRS	3.3% (5)	2.9% (4)	6.7% (1)	0.407

continues on next page

variable	overall n = 153	wait-list survivor n = 138	wait-list failure n = 15	p-value
listing laboratory investigations:				
Na	137 (116-153)	137 (116-153)	138 (132-148)	0.337
hyponatremia (Na<135)	30.1% (46)	30.4% (42)	26.7% (4)	1.000
Cr	80 (16-680)	80 (36-586)	92 (16-680)	0.204
albumin	29 (13-50)	29 (13-50)	27.5 (17-41)	0.254
total bilirubin	32 (4-700)	31.5 (4-496)	49 (16-700)	**0.030**
INR	1.3 (0.9-7.2)	1.3 (0.9-7.2)	1.6 (1.1-3.8)	0.116
platelet count	104.5 (22-454)	106 (22-454)	84.5 (27-243)	0.116
platelets < 100	47.3% (70)	45.5% (61)	64.3% (9)	0.181
listing risk scores: MELD-UNOS MELDNa	 13 (6-38) 15 (6-38)	 12 (6-38) 15 (6-38)	 18 (9-36) 21 (9-36)	 **0.005** **0.025**

Appendix E: MELD Calibration Analysis by Clinically-Based "Share MELD 15" Threshold

As described in section 1.6.2, Merion et al published evidence in 2005 that suggested that the one-year survival benefit of liver transplantation was limited to individuals with MELD-UNOS scores of 15 or greater.[50] This implied that individuals whose MELD-UNOS scores were 14 or less had a greater risk of one-year mortality due to complications of liver transplantation than from ESLD-related mortality if they remained on the waiting list for a year. After consideration of this study UNOS implemented the "Share MELD 15" policy which allowed individuals with MELD scores of 15 and greater to benefit from an expanded geographic region of allograft eligibility.[41, 49] Along with the adoption of a MELD-based liver allocation system in Canada, the employment of a policy similar to Share MELD 15 has been proposed. As was discussed in section 2.2.3, when the probability estimates generated by a predictive model are utilized to make management decisions based on the relationship of the estimate to an established threshold value, calibration accuracy becomes a relevant aspect of model performance. This makes a MELD score of 15 an appropriate clinically-based threshold for the assessment of model calibration, a grouping method suggested as being more informative than the decile-based grouping strategy most commonly applied.[72, 74]

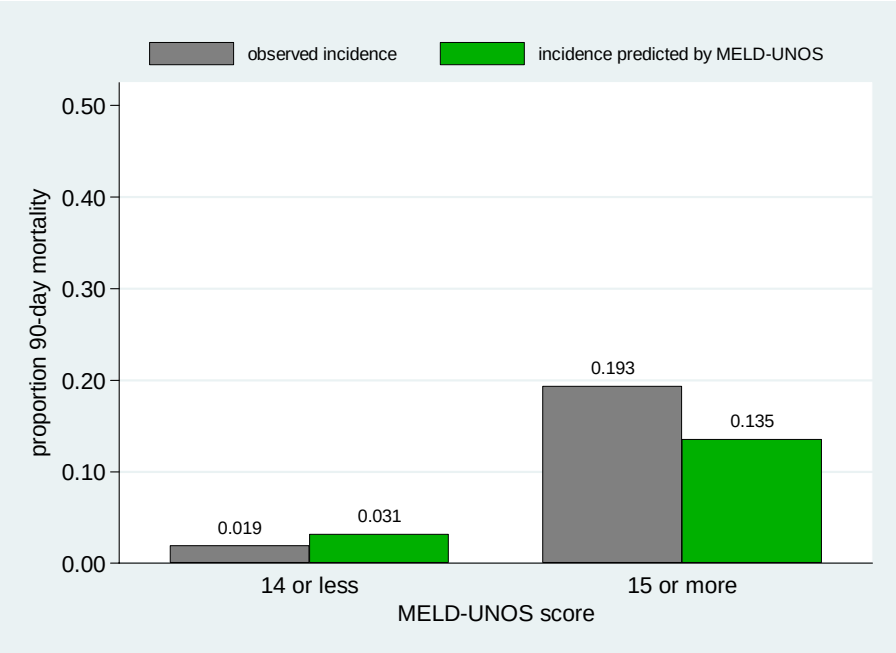
Table E.1 and Figure E.1 display the tabular and graphical results of calibration analysis based on this grouping strategy. The Hosmer-Lemeshow statistic suggests there is no evidence of significant lack-of-fit, although the validity of this analysis must be

Table E.1: MELD-UNOS calibration analysis based on "Share MELD 15" derived quantiles

MELD score category	Observed failure incidence	Estimated failure incidence	events Observed Expected	non-events Observed Expected	H-L G-o-F χ^2
6 to 14 n = 53	0.019	0.031	1 1.7	52 51.3	0.277
15 to 40 n = 31	0.194	0.135	6 4.2	25 26.8	0.904
total cohort n = 84	0.083	0.070	7 5.9	77 78.1	1.181*

*$\chi^2(2)$, p = 0.554

Figure E.1: Observed versus MELD-UNOS estimated incidence of 90-day wait-list mortality in validation cohort (n = 84) by "Share MELD 15' derived quantiles



questioned due to the low event occurrence. This grouping strategy does reveal that the probability estimates for observations with risk scores of 14 and less closely approximate the true event incidence in this group. As was seen with the quartile-of-risk and three risk strata grouping strategies, the MELD-UNOS underestimated the risk of mortality in individuals with greater ESLD severity.

The clinical implication of this finding is that, should a Share MELD 15 allocation policy be adopted in Canada, it would appear that liver transplant candidates residing in Atlantic Canada would not be disserviced by such a policy. The purpose of the Share MELD 15 policy is to increase the opportunity for individuals at higher risk of wait-list mortality to receive a deceased donor liver. This is achieved by increasing the procurement area for these individuals. Currently in Canada, an organ procured within a transplant programme's region is allocated to an individual on that programme's wait-list, and the only occasion when an organ would be transferred out of the procurement region is when there is an emergent status candidate (CanWAIT 4F, 4, or 3F) waiting in another region, or the organ is declined by the regional transplant programme. Individuals with MELD-UNOS scores of 15 and greater benefit from the Share MELD 15 policy regardless of the calibration accuracy of the MELD-UNOS; however individuals with scores of 14 or less can be potentially harmed by this policy if the model underestimates their true mortality risk. If the MELD-UNOS was found to be under calibrated for candidates with risk scores lower than the threshold or 15 these individuals could possibly suffer the double affront of having their risk for wait-list failure understated, while potentially life saving allografts were transferred out of their region. The current analysis implies that candidates with risk scores of 14 and less have their risk of 90-day

wait-list mortality well estimated by the MELD-UNOS, and therefore there should be no ill effect if this policy was adopted in conjunction with a MELD-based allocation system.