UCLA
UCLA Electronic Theses and Dissertations

Title

Integrating Machine Learning and Longitudinal Modeling to Optimize the Liver Transplant Continuum: From Waitlist Dropout to Post-Transplant Survival

Permalink

https://escholarship.org/uc/item/6px189ds

ISBN

9798247906360

Author

Melehy, Andrew

Publication Date

2026-05-19

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Integrating Machine Learning and Longitudinal Modeling to Optimize the Liver Transplant Continuum: From Waitlist Dropout to Post-Transplant Survival

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy in Medical Informatics

by

Andrew Melehy

2026

© Copyright by

Andrew Melehy

2026

ABSTRACT OF THE DISSERTATION

Integrating Machine Learning and Longitudinal Modeling to Optimize the Liver Transplant Continuum from Waitlist Dropout to Post-Transplant Survival

by

Andrew Melehy

Doctor of Philosophy in Medical Informatics

University of California, Los Angeles, 2026

Professor Alex Bui, Chair

Real-world clinical data is longitudinal in nature, owing to the multiple observations made over time when an individual interacts with the healthcare system. Despite the ubiquitous nature of longitudinal electronic heath record (EHR) data, many studies are limited to time-to-event analyses, clustering based on basic metrics, or the simple static modeling of dynamic processes. With the abundance of EHR data now available, an exploration and update of longitudinal data analysis frameworks is warranted to most accurately model patient experiences.

This dissertation focuses on the exploration of longitudinal (e.g., time series) and clustered data analysis approaches for real-world clinical data in the specific domain of liver transplantation (LT). I begin by examining traditional approaches to longitudinal analysis in the LT domain, establishing common areas for improvement using existing

methodologies, and exploring novel methods with machine and reinforcement learning (ML/RL).

This work leveraged large publicly available registry databases from the Scientific Registry of Transplant Recipients and United Network for Organ Sharing, as well as a granular multi-center dataset containing data on LT recipients. The explored traditional approaches included using linear mixed effects to model transplant center variation in waitlist mortality and logistic regression for the evaluation of novel longitudinal risk factors in transplant recipients with acute-on-chronic liver failure. The ML/RL approaches included the use of offline reinforcement learning to improve LT donor-recipient matching and the creation of a longitudinal deep survival network to better assess risk in hepatocellular carcinoma patients listed for LT.

This collection of studies demonstrates that with advanced techniques the full power of longitudinal health information can be leveraged for state-of-the-art predictive performance. Specifically, the developed waitlist dropout model for HCC patients has predictive accuracy surpassing any previous published model. This work in HCC highlights how the strong predictive performance of an ML model can be integrated into an existing clinical framework that is used and accepted by the medical community, with demonstration of the impact of implementing our model in the system using simulation. Finally, this works offers a fundamentally new framework for data-driven donor-recipient matching optimization through the use of ML/RL, providing proof-of-concept that could serve as the basis for future intelligent decision support in LT.

The dissertation of Andrew Melehy is approved.

Vatche Agopian

William Speier

Ricky Taira

Alex Bui, Committee Chair

University of California, Los Angeles

2026

DEDICATION

This thesis is dedicated to my wife Katherine, my Mom Holly, my Dad Sam, my brother, Alex, and my sister-in-law Melanie, all of whom have supported me throughout my career.

CONTENTS

ABSTRACT OF THE DISSERTATION ii

DEDICATION v

LIST OF FIGURES ix

LIST OF TABLES x

ACKNOWLEDGEMENTS xii

BIOGRAPHICAL SKETCH xiv

INTRODUCTION 1

Overview 1

Clinical Background and History 1

Clinical Problem and Motivation 2

LT Data Organization and Current use of ML 6

Core Studies 7

Center-level variations are associated with disparate waitlist mortality and probability of liver transplant in MELD 35 and greater patients (Chapter 1) 8

Liver transplantation at the extreme: analysis of outcomes and predictors of futility in acute-on-chronic liver failure grade 3 recipients (Chapter 2) 9

Liver transplant donor-recipient matching with offline reinforcement learning (Chapter 3) 10

HCC-3D: improving liver transplant allocation with the hepatocellular carcinoma dynamic-deep-drop model (Chapter 4) 11

Introduction References 12

Chapter 1 Center-level variations are associated with disparate waitlist mortality and probability of liver transplant in MELD 35 and greater patients 21
Chapter 1 Abstract 21
Chapter 1 Introduction 21
Chapter 1 Patients and Methods 23
Chapter 2 Results 27
Chapter 1 References 53
Chapter 2 Liver transplantation at the extreme: analysis of outcomes and predictors of futility in acute-on-chronic liver failure grade 3 patients on maximal life support 58
Chapter 2 Abstract 58
Chapter 2 Introduction 59
Chapter 2 Patients & Methods 61
Chapter 2 Results 65
Chapter 2 Discussion 85
Chapter 2 Limitations 89
Chapter 2 Conclusions 90
Chapter 2 References 90
Chapter 3 Liver transplant donor-recipient matching with offline reinforcement learning 96
Chapter 3 Abstract 96
Chapter 3 Introduction 96
Chapter 3 Results 99
Chapter 3 Discussion 118

Chapter 3 Methods 124
Chapter 3 Acknowledgements 139
Chapter 3 References 140
Chapter 4 HCC-3D: improving liver transplant allocation with the hepatocellular carcinoma dynamic-deep-drop model 152
Chapter 4 Abstract 152
Chapter 4 Introduction 153
Chapter 4 Patients and Methods 155
Chapter 4 Results 163
Chapter 4 Discussion 182
Chapter 4 Conclusions 186
Chapter 4 Supplementary Appendix 186
Chapter 4 Acknowledgements 193
Chapter 4 References 194
Chapter 5 Conclusions and future directions 204
Chapter 5 Summary and Discussion of Core Studies 204
Chapter 5 Key Takeaways 210

LIST OF FIGURES

Figure 1. Center and regional variation in liver transplant waitlist mortality 31
Figure 2. Cohort selection diagram. 35
Figure 3. Waitlist mortality and transplant probability across center risk groups. 39
Figure 4. Post-transplant survival across center risk groups. 43
Figure 5. Waitlist mortality and transplant probability by era. 47
Figure 6. Patient cohort selection. 65
Figure 7. Post-transplant survival across ACLF groups. 73
Figure 8. Post-transplant survival comparisons. 85
Figure 9. Overview of study and offline RL approach. 100
Figure 10. Overview of data collection and offline RL pipeline. 104
Figure 11. Feature maps. 111
Figure 12. Feature map of discarded donors. 112
Figure 13. Patient trajectories. 113
Figure 14. Post-transplant survival comparisons. 115
Figure 15. Post-transplant survival comparisons in high-risk patients 116
Figure 16. Patient cohort selection. 163
Figure 17. Cumulative viable tumor over time. 166
Figure 18. Feature importance plot. 168
Figure 19. Calibration plots. 171
Figure 20. Overview of the simulation design. 173
Figure 21. Simulation validation. 174
Figure 22. HCC-3D stratified survival by risk score group. 176
Figure 23. Simulation results. 178
Figure 24. HCC-3D differential prioritization. 180
Figure 25. HCC-3D differential prioritization in distinct subgroups. 181

LIST OF TABLES

Table 1. Univariate mixed effects Cox proportional hazards analysis 28
Table 2 Multivariate mixed effects Cox proportional hazards analysis. 29
Table 3. Baseline demographics in the total cohort as well as each risk group with comparisons among variables in the three groups. 32
Table 4. Baseline demographic comparisons in patients initially listed at MELD ≥35 among the three risk groups. 36
Table 5. Baseline demographic comparisons in patients initially listed at MELD < 35 with increase to MELD ≥35 while on the waitlist. 40
Table 6. Baseline demographic comparisons in patients with MELD ≥35 (at listing or increased on list) who underwent LT among the three risk groups. 44
Table 7. Baseline demographic comparisons between extreme and non-extreme ACLF-3 patients. 66
Table 8. Annual center transplant volume and ACLF-3 transplant volume over time with median MELD at transplant in our center and our center's DSA. 68
Table 9. Baseline demographic comparisons between ACLF Grade 3 patients with and without the futile outcome. 69
Table 10. Secondary outcomes in extreme ACLF-3 patients, comparing non-futile and futile groups and specific causes of death in the futile group. 73
Table 11. Univariate odds ratios of categorical variables. 76
Table 12. Univariable odds ratios (per SD) for continuous measures. 77
Table 13. Univariable odds ratios by quartiles of continuous measures. 78
Table 14. Results of multivariate logistic regression modeling futility. 82
Table 15. Baseline demographic and outcomes comparison in extreme ACLF-3 patients at UCLA and external MODEL Consortium cohort. 83
Table 16. Descriptive demographics of LT candidates meeting study inclusion criteria. 101
Table 17. Demographic comparisons between deceased donors used in liver transplants and discarded donor organs. 102
Table 18. Definitions of metrics. 106
Table 19. Evaluation results. 107
Table 20. Stratified evaluation results. 108
Table 21. Constrained evaluation results. 117
Table 22. Summary of reward design. 129
Table 23. Weighted importance sampling (WIS) metric of the optimal conservative Q-learning model with respect to different behavior policies. 133
Table 24. Table 23. TRIPOD-AI checklist. 137

Table 25. Proposed implementation of HCC-3D using the existing MMaT-3 framework. 161
Table 26. Baseline demographic comparisons between training and testing cohorts using features at time of listing. 164
Table 27. Model evaluation. 169
Table 28. Demographics of liver transplant patients (SRTR) used in simulation. 172
Table 29. Simulations results, with counts shown as mean (SD). 174

ACKNOWLEDGEMENTS

Chapter 1, is a reprint of the published manuscript: Andrew Melehy, Dominic Amara, Shreya Gumate, Fady M. Kaldas, Douglas G. Farmer, Alex A.T. Bui, and Vatche G. Agopian. “Center-level variations are associated with disparate waitlist mortality and probability of liver transplant in MELD 35 and greater patients.” Liver Transpl. 2025; 31(9):1100-1110. Doi: 10.1097/LVT.0000000000000616. PMID: 40172494. Reprinted with permission from Wolters Kluwer. AM conducted primary data analysis, manuscript writing and revisions, DA, SG, SE, FMK, and DGF contributed to manuscript preparation and editing, AATB advised the technical aspects of the work and contributed to critical manuscript revisions. VA conceived of the study and contributed to manuscript revisions.

Chapter 2, is the submitted version of: Andrew Melehy, Asara Thepbunchonchai, Benjamin V. Tran, Ruben Hernaez, Dominic Amara, Samer Ebaid, Daniela Markovic, Fady M. Kaldas, Igor Barjaktarevic, Victor Xia, Constantine J. Karvellas, Saro Khemichian, Lance L. Stein, Kirti Shetty, Christina C. Lindenmeyer, Justin R. Boike, Douglas A. Simonetto, Robert S. Rahimi, Prasun K. Jalal, Manhal Izzy, Michael S. Kriss, Gene Y. Im, Ming V. Lin, Janice H. Jou, Brett E. Fortune, George Cholankeril, Alexander Kuo, Vinary Sundaram, for the Multi-Organ Dysfunction and Evaluation for Liver Transplantation (MODEL) Consortium, Ronald W. Busuttil, Douglas G. Farmer, and Vatche G. Agopian. “Liver Transplantation at the Extreme: Analysis of Outcome and Predictors of Futility in Acute-on-Chronic Liver Failure Grade 3 Patients on Maximal Life Support.” JHEP Reports. In review. AM, AT, and DM performed data collection, analysis, and manuscript writing. The Consortium authors listed at the end of the

manuscript were essential collaborators in the acquisition of the multi-center MODEL Consortium data.

Chapter 3, is a reprint of the published manuscript: Andrew Melehy, Jeffrey Feng, Dominic Amara, Vatche G. Agopian, and Alex A.T. Bui. “Liver Transplant Donor-Recipient Matching with Offline Reinforcement learning.” *npj Digit. Med*. **9**. 351 (2026). PMID: 41840109. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format. http://creativecommons.org/licenses/by/4.0/. Financial support was provided by the University of California, Los Angeles, NSF Grant DEB-8227052 and a MacArthur predoctoral fellowship in International Peace and Security granted by the Social Science Research Council. AM, DA, and VA acquired the data. All authors reviewed, revised, and approved of the manuscript.

Chapter 4 is the submitted version of: Andrew Melehy, Jeffrey Feng, Dominic Amara, Carolyn Smullin, Samer Ebaid, Fady M. Kaldas, Douglas G. Farmer, Neil Mehta, Alex A.T. Bui, and Vatche G. Agopian. “HCC-3D: Improving Liver Transplant Allocation with the Hepatocellular Carcinoma Dynamic-Deep-Drop Model.” *Journal of Hepatology*. In review. AM, DA, and VA acquired the data. All authors reviewed, revised, and approved of the manuscript.

BIOGRAPHICAL SKETCH

Education:

2021, M.D. Doctor of Medicine, Vagelos College of Physicians and Surgeons, Columbia University, New York, New York

2017, B.A. Biological Sciences: Biochemistry, Dept. of Molecular Biosciences, Northwestern University, Evanston, Illinois

Work Experience:

2021-present, Resident Physician in General Surgery, University of California, Los Angeles, Department of Surgery.

2017-2021, Medical Student Researcher under the mentorship of Dr. Koji Takeda, Columbia University, Department of Surgery, Division of Cardiothoracic and Vascular Surgery.

Professional Honors:

2023, Outstanding Second Year Resident, Division of General Surgery, University of California, Los Angeles School of Medicine.

Scholarships and Sponsorships:

2023-2024, H & H Lee Surgical Research Scholars Program Award, University of California, Los Angeles, Department of Surgery, Division of Liver and Pancreas Transplantation (Vatche Agopian, PI)

2022-2024, Gerald S. Levey Surgical Research Award, University of California, Los Angeles, Department of Surgery, Division of Liver and Pancreas Transplantation (Vatche Agopian, PI)

Publications

A. Melehy, J. Feng, D. Amara, V.G. Agopian, and A.A.T. Bui (2026). “Liver Transplant Donor-Recipient Matching with Offline Reinforcement learning.” *npj Digit. Med*. **9**, 351 (2026). https://doi.org/10.1038/s41746-026-02529-1

A. Melehy, J. Feng, D. Amara, C. Smullin, S. Ebaid, F.M. Kaldas, D.G. Farmer, N. Mehta, A.A.T. Bui, and V.G. Agopian (2026). “HCC-3D: Improving Liver Transplant Allocation with the Hepatocellular Carcinoma Dynamic-Deep-Drop Model.” Gastroenterology. [In review].

D. Amara, W. Dumronggittigule, A. Melehy, D. Markovic, L. Nguyen, S. Nesbit, D.S. Lu, S. Ebaid, F.M. Kaldas, D.G. Farmer, R.W. Busuttil, and V.G. Agopian (2025). “Occult multifocal and incidental hepatocellular carcinoma: An analysis of long-term survival and risk factors at a single liver transplant center.” Liver Transpl. PMID: 40372118.

A. Melehy, D. Amara, S. Gumate, S. Ebaid, F.M. Kaldas, D.G. Farmer, A.A.T. Bui, and V.G. Agopian (2025). “Center-level variations are associated with disparate waitlist mortality and probability of liver transplant in MELD 35 and greater patients.” Liver Transpl. PMID: 40172494.

D. Amara, A. Melehy, S. Ebaid, F.M. Kaldas, D.G. Farmer, P. Stock, A.A.T. Bui, N. Mehta, and V.G. Agopian (2025). “Impact of Normothermic Machine Perfusion on Access to Liver Transplantation in Patients with Primary Hepatic Malignancies.” Clin Transplant. PMID: 40698561.

C. Zhao, Y.T. Lee, A. Melehy, M. Kim, J.Z. Yang, C. Zhang, J. Kim, R.Y. Zhang, J. Lee, H. Kim, Y. Ju, Y.J. Tsai, X.J. Zhou, S.B. Han, S. Sadeghi, R.S. Finn, S. Saab, D.S. Lu, J. Chiang, J.H. Park, T.V. Brennan, S.A. Wisel, M. Alsudaney, A. Kuo, W.S. Ayoub, H. Kim, H.D. Trivedi, Y. Wang, A. Vipani, I.K. Kim, T. Todo, J.A. Steggerda, G. Voidonikolas, K. Kosari, N.N. Nissen, R. Saouaf, A.G. Singal, M.S. Sim, D.A. Elashoff, S. You, V.G. Agopian, J.D. Yang, H.R. Tseng, and Y. Zhu (2025). “Extracellular vesicle digital scoring assay for assessment of treatment responses in hepatocellular carcinoma patients.” J Exp Clin Cancer Res. PMID: 40307890.

A. Melehy, and V.G. Agopian (2025). “Role of Liver Transplant in Primary and Secondary Liver Malignancies.” Clin Liver Dis. PMID: 40287268.

S. Athreya, A. Melehy, S.S.A. Suthahar, V. Ivezić, A. Radhachandran, V.R. Sant, C. Moleta, H. Zheng, M. Patel, R. Masamed, M. Livhits, M. Yeh, C.W. Arnold, and W. Speier (2025). “Combining Ultrasound Imaging and Molecular Testing in a Multimodal Deep Learning Model for Risk Stratification of Indeterminate Thyroid Nodules.” Thyroid. PMID: 40256961.

R. Hu, B. Tran, S. Li, M.L. Stackpole, W. Zeng, Y. Zhou, A. Melehy, S. Sadeghi, R.S. Finn, X.J. Zhou, W. Li, and V.G. Agopian (2025). “Noninvasive prognostication of hepatocellular carcinoma based on cell-free DNA methylation. PloS One. PMID: 40279344.

A. Melehy, and V.G. Agopian (2024). “Treating rare tumors with liver transplantation.” Curr Opin Organ Transplant. PMID: 37851086.

S. Mehta, J. Fried, S. Nemeth, P. Kurlansky, Y. Kaku, A. Melehy, S. Char, A. Masoumi, G. Sayer, N. Uriel, and K. Takeda (2023). “Society for Cardiovascular Angiography and Interventions Shock Classification to Stratify Outcomes of Extracorporeal Membrane Oxygenation.” ASAIO. PMID: 36730984.

S. Char, J. Fried, A. Melehy, S. Mehta, Y. Ning, P. Kurlansky, and K. Takeda (2023). “Clinical efficacy of direct or indirect left ventricular unloading during venoarterial extracorporeal membrane oxygenation for primary cardiogrenic shock. J Thorac Cardiovasc Surg. PMID: 34243933.

A. Melehy, J.E. Sanchez, S.K. Nemeth, P.A. Kurlansky, N. Uriel, G.T. Sayer, Y. Naka, and K. Takeda (2023). “National outcomes for bridge to multiorgan cardiac transplantation using mechanical circulatory support.” J Thorac Cardiovasc Surg. PMID: 33678503

INTRODUCTION

Overview

With the adoption and mandate of electronic health records (EHRs) alongside the increasing number of available medical diagnostic tests, there has been an explosion in the availability of observational clinical data. This data is longitudinal in nature with patients interacting with the healthcare system overtime, producing physiologic, laboratory, and radiologic parameters with associated time stamps. This dissertation aims to explore the analysis of longitudinal data in the liver transplant (LT) population to understand how advanced techniques using machine learning (ML) may improve predictive performance and decision support.

Clinical Background and History

The first LT was performed in the United States (US) by Thomas Starzl at the University of Colorado in 1963. This patient died, as did the subsequent four patients over the course of four months [1]. Similarly dismal results were seen at centers in Boston and Denver resulting in a moratorium on all LT-related clinical trials. It was not until 1967 that the first patient with extended survival after transplant was achieved [2]. Due to the persistence of collaborative researchers across the globe, advances in immunosuppressive and surgical techniques have led to wide adoption of LT as the gold standard treatment for end-stage liver disease with nearly 10,000 transplants performed per year in the US, representing a nearly 50% increase in the last decade [3]. While the majority of transplants were historically performed for hepatitis C cirrhosis, the introduction of direct-acting antiviral therapies for hepatitis C changed the landscape of indications for LT [4, 5]. There has been a significant increase in the incidence of

metabolic dysfunction associated steatohepatitis cirrhosis and alcoholic cirrhosis in the last decade [5]. Additionally, the accepted indications for LT have significantly expanded with an increase in transplant oncology and pathways to receive LT for hepatocellular carcinoma (HCC), cholangiocarcinoma, and others [6]. Despite the significant advances in the use of LT in the last 60 years, waitlist mortality remains as high as 25% in the US [5, 7]. The high waitlist mortality is representative of several clinical challenges including donor organ scarcity and an aging population with an increasing number of medical comorbidities. Traditionally, LTs were performed using brain-dead donors (DBD), allowing for a minimization of ischemic time and maximization of organ quality [8]. Due to the number of eligible recipients exceeding the supply of donor organs, lower quality and organs recovered after donation after circulatory death (DCD) have been increasingly utilized. The increase in the donor supply largely benefits recipients waiting for organs, though it has complicated decision making with a desire to avoid poor donor-recipient pairs that are likely to lead to early graft failure or recipient death. The persistent excess waitlist mortality and graft failures following transplant remain key challenges in LT and invite continued research to optimize outcomes for transplant waitlist registrants and recipients.

Clinical Problem and Motivation

Current liver allocation policy is based on severity of illness with the sickest patients transplanted first. Patient acuity is determined using the Model for End-Stage Liver Disease (MELD) score, which was originally calculated based on assessment of laboratory markers including international normalized ratio, total bilirubin, and creatinine [9]. The MELD score was initially proposed to predict survival in cirrhotic patients

undergoing transjugular intrahepatic portosystemic shunt procedures for refractory ascites or bleeding esophageal varices. It was later validated for the prediction of 3-month survival on the waitlist and subsequently adopted by the United Network for Organ Sharing (UNOS) for liver organ allocation priority. Since initial adoption, serum sodium was added to the MELD score in 2016 after this addition was shown to improve predictive performance [10]. Most recently, MELD 3.0 has been shown to improve waitlist prognostication with incorporation of sex and serum albumin [11]. The necessary updates to the MELD score reflect the degrading predictive ability of the model over time and despite these updates, waitlist mortality remains high. One special case of waitlist processes is in those patients who are listed for transplant due to HCC. These patients must have tumor burdens within Milan criteria (≥3 tumors each ≤3 cm, or one tumor ≤5cm) and have a mandatory waiting period of 6 months after listing prior to their being awarded a MELD score at the median MELD at transplant around the donor hospital minus three points (MMaT-3) [12-14]. These criteria have been enacted to give patients with HCC additional priority beyond liver disease measured by the MELD score, which is typically lower in these patients and doesn't account for HCC or its ability to progress and metastasize.

When a donor liver becomes available, it is offered to the transplant center with the recipient at the top of the list and an assessment of the donor-recipient pair is made by the transplant center physicians. These decisions are based on the perception of recipient and donor risk according to recipient factors such as degree of liver decompensation, need for life support, and prediction of graft survival in the recipient. Notable donor factors include type of donor (DBD, DCD, marginal, split), degree of

steatosis, donor age, and lifestyle factors (alcohol, drug, and tobacco use). The donor risk index (DRI) is a model for assessing donor risk of graft failure originally published in 2006, which includes donor age, cause of death, race, DCD status, split liver, height, organ distance from transplant center (local, regional, national), and cold ischemia time (time from donor liver perfusion with preservation solution to reperfusion in the recipient) [15]. While the DRI is commonly used as a surrogate for donor quality in research studies, it is rarely used in clinical practice with transplant providers typically making decisions based on a holistic review of compatible donor and recipient factors.

Donor liver allocation is primarily determined by the MELD score as discussed above. However, there are patients with acute fulminant hepatic failure resulting in multi-system organ failure with an extremely high risk of short-term mortality who can achieve a special status regardless of MELD score known as Status 1A. At any given time, there are approximately 50 or fewer patients listed as Status 1A in the US. Historically, transplant allocation was based on UNOS region and the organ procurement agency operating within that region, meaning that these regions operated independently, offering available donors to patients listed in that region. This, however, led to significant national disparities in waiting times on the list, median MELD score at transplant, and waitlist mortality at higher MELD ranges [16-18]. To offset these disparities, the current UNOS policy (enacted in 2020) is acuity circles, which stipulates that any available donor organ in the country is first offered to status 1A patients within a 500 nautical mile radius of the donor hospital. Following this, organs are offered to patients at the highest MELD score ranges within a 150 nautical mile radius, then 250

nautical mile radius, then 500 nautical mile radius. The organ is then offered to patients in the next MELD score range down in the same expanding radius as above.

LT 1-, 5-, and 10-year graft-failure free survival has been cited as 85-95%, 70-80%, and 60-70% in the modern era [19-21]. Early graft dysfunction (graft failure within three months to one year) has been explored extensively and is thought to be due to ischemia-reperfusion injury, donor-recipient matching, and transplant factors. Long-term graft survival (>1 year) is thought to be more related to immunosuppression, opportunistic infections, exacerbation of medical comorbidities, and social factors and thus difficult to predict from clinical factors at the time of transplant. An early effort to characterize and model early graft dysfunction was reported by Olthoff et al., in 2010, where they defined early allograft dysfunction (EAD) as a total bilirubin >10 mg/dL on the 7th day after transplant, INR >1.6 on day seven, and alanine or aspartate aminotransferase level >2,000 IU/L within the first seven days [22]. More recently, Agopian et al., proposed the Liver Graft Assessment Following Transplant risk score (L-GrAFT), which demonstrated improved accuracy in the prediction of 3-month graft failure using dynamic changes in serum levels of total bilirubin, aspartate aminotransferase, INR, and platelet count within the first 10 post-transplant days [23, 24]. This model highlighted the importance of changes in lab measures in the immediate postoperative period as well as further characterized the changes associated with allograft dysfunction due to ischemia-reperfusion injury sustained at the time of transplant. Another effort to predict early post-transplant graft survival (three month) using both recipient and donor factors at the time of transplant was undertaken with the survival outcomes following liver transplant (SOFT) score [25]. This study highlighted

the poor performance of the MELD score in predicting early post-transplant graft failure, the importance of donor factors in making this prediction, and the significant risk of early graft failure associated with recipient pre-transplant life support and history of a prior LT.

LT Data Organization and Current use of ML

LT data is by its nature longitudinal, owing to the clustering units of geographic organization and data collection during distinct phases of a patient's course (prior to listing, at time of listing, during the waitlist period, at the time of transplantation, and during the follow up period). LT centers are required to report demographic and clinical information to UNOS at the time of a patient's listing that includes relevant medical histories, MELD score components at the time of listing, liver disease etiology and indication for transplant, among many others. The centers must then continually report data to UNOS at regular intervals while on the waitlist, at time of transplant, and in the follow-up period. These requirements are for every LT patient in the US, resulting in a robust collection of longitudinal data at each of the three phases involved in an LT recipient's care (listing, transplant, post-transplant). Additionally, UNOS is broken into 11 geographic regions that share important characteristics such as median MELD at transplant, median time on the waitlist, and organ availability. Within UNOS regions, each transplant center operates with unique characteristics that stem from the experience of their hepatologists and transplant surgeons, hospital resources, critical care expertise and bed availability, and patient population.

The use of ML techniques in clinical medicine has grown, with researchers increasingly adapting such techniques to improve upon previously accepted models. LT has seen such efforts with many studies attempting to improve the prediction of waitlist

mortality and post-transplant graft failure [26-34]. Nagai et al., recently developed a neural network based approach for the prediction of 90-day LT waitlist mortality [27]. This study included over 100,000 US waitlist registrants and was able to predict 90-day waitlist mortality with an AUROC of 0.94 compared to the MELD score AUROC of 0.86, highlighting the potential for ML methods to outperform traditional linear models. Similarly, Baht et al., has recently produced the Dynamic Model of End-Stage Liver Disease for Equitable Prioritization (DynaMELD) model, which includes rates of change of time-varying laboratory biomarkers to predict 90-day waitlist mortality [26]. Using a national cohort of >50K US LT candidates, they created a neural network-based model that outperformed MELD 3.0 by a concordance of 0.5%, highlighting the advantage of leveraging dynamic changes in longitudinal measures to improve model performance. Still, while many studies have used ML models to improve the prediction of waitlist mortality and graft failure, including some that leverage longitudinal dynamics of clinical features, minimal exploration has been done in the analysis of distinct patient clusters, nor have there been published reports of ML techniques successfully performing donor-recipient matching to balance the risk of waitlist mortality with graft failure in a single model. Additionally, little work has been done in using ML techniques to leverage longitudinal data for the prediction of tumor progression and waitlist dropout in HCC patients.

Core Studies

The chapters herein detail efforts to address existing gaps in knowledge related to ML techniques as applied to longitudinal data in LT patients. The use of traditional and ML techniques for analysis of longitudinal data in LT patients is explored, leading to

the production of usable ML techniques that benefit LT patients in terms of improved waitlist candidate prioritization and donor-recipient matching.

The four main studies of this work include two traditional and two ML approaches to longitudinal data analysis. The traditional approaches included use of linear mixed models to assess transplant center variation in high-acuity LT candidates and use of regression analysis to assess dynamic risk factors for futility after LT in acute-on-chronic liver failure grade 3 (ACLF-3) patients. The ML approaches included the use of offline RL for donor-recipient matching and the use of longitudinal deep survival networks for improved waitlist prognostication in HCC patients.

Center-level variations are associated with disparate waitlist mortality and probability of liver transplant in MELD 35 and greater patients (Chapter 1)

The MELD score is currently used to determine donor liver allocation priority in the US. One limitation of the MELD score is that it does not capture the risk of waitlist mortality associated with complications from decompensated cirrhosis and portal hypertension such as bleeding esophageal varices, spontaneous bacterial peritonitis, and hepatic encephalopathy [35]. Additionally, the MELD score is a static risk assessment at a single point in time for what is truly a longitudinal patient trajectory and currently the MELD score does not take into account the rate of MELD change over time, which has itself been shown to be associated with waitlist mortality [36, 37]. The underlying premise of the MELD score is that it provides a “fair” system for organ allocation as it is entirely comprised of objective measures. An analysis of the contribution to candidate risk of mortality to transplanting center alone is needed to uncover if significant disparities in the predictive ability of the MELD score exist,

especially for patients with MELD scores in the highest range (≥35). We hypothesized that there is a higher likelihood of center variability in the highest acuity recipients because there are discrepancies in critical care capabilities and provider willingness to offer transplant.

Using mixed-effects modeling, we were able to demonstrate significant center-level variation in waitlist outcomes and probability of LT in the highest acuity candidates (MELD ≥35), especially in those listed at a lower MELD with decompensation to ≥35 after initial listing. These findings indicate that certain transplant centers may have improved ability to care for high acuity LT candidates and get them to transplant, while others fail to do so.

Liver transplantation at the extreme: analysis of outcomes and predictors of futility in acute-on-chronic liver failure grade 3 recipients (Chapter 2)

ACLF is a syndrome characterized by acute hepatic decompensation, organ failure, and high short-term mortality [38-42]. Patients with cirrhosis who develop acute decompensation are at risk for the development of ACLF with subsequent rapid organ dysfunction and an elevated mortality risk [43-45]. ACLF-3 is the most severe form of disease where patients have at least three organ failures and is associated with 28-day waitlist mortality approaching 80% [44, 46-49].

Due to the extremely high waitlist mortality of ACLF-3 patients, there is considerable interest in offering these patients early LT to afford them the greatest chance of survival [50-57]. However, this issue is balanced against the concerns about poor post-transplant outcomes and stewardship of scarce donor livers. There have been considerable efforts to predict futility following transplant in ACLF-3 recipients and to

identify risk factors for futile outcomes [44, 58-61]. One recent study was reported by Hernaez et al., where 15 US centers contributed data on 735 ACLF patients, 521 of which were considered “severe,” including ACLF-2 and ACLF-3 patients. This data was used to create the novel Sundaram ACLF-LT-Mortality (SALT-M) score, which predicts 1-year post-transplant mortality in ACLF-3 patients. There is a spectrum of disease even within ACLF-3 that has not been specifically explored. We sought to investigate a subset of ACLF-3 patients all of whom had respiratory failure requiring mechanical ventilation, renal failure requiring dialysis, and required vasopressors for maintenance of end-organ perfusion [62]. The UCLA experience in transplanting these ACLF-3 patients was examined, where novel longitudinal risk factors for futility after transplant were identified, including indication for mechanical ventilation and changes in MELD score and respiratory status from ICU admission to LT. The findings were then validated using data from the multicenter consortium used to produce the SALT-M score.

Liver transplant donor-recipient matching with offline reinforcement learning (Chapter 3)

Currently, decisions regarding suitable LT donor and recipient pairs are determined based on clinical factors and primarily assessed by the transplant team to which the organ was offered. There are existing risk models that predict early graft failure based on donor-recipient factors. These methods do not balance the risk of graft failure against the competing risk of waitlist mortality or leverage the dynamics of a patient’s trajectory on the waitlist to improve predictions. To improve upon the shortcomings of existing models we created an offline RL framework, assessing risk of waitlist dropout each time a new MELD score is recorded and allowing our system to

make a decision based on available donors at each time point to transplant one of them or allow the patient to remain on the waitlist. There are several complexities incumbent in real waitlist candidate trajectories that our model addresses: 1) rapid increase in MELD portends a high risk of waitlist mortality and should favor a decision to transplant, 2) patients with a stable lower-middle range MELD can likely safely remain on the waitlist for a period of time and a scarce donor organ is better placed in a higher acuity patient, and 3) some recipients are too critically ill to be expected to have a successful transplant regardless of the donor offered and livers should be directed away from these potential recipients.

Using our model, we were able to demonstrate that decisions regarding donor-recipient match suitability based on our framework resulted in a significantly reduced number of waitlist dropouts, while preserving the number of successful transplants with survival beyond one-year post-transplant. Additionally, use of our system may lead to a reduction in discarded donors (eligible donor organs for which no suitable recipient is found) by optimizing donor-recipient match selections.

HCC-3D: improving liver transplant allocation with the hepatocellular carcinoma dynamic-deep-drop model (Chapter 4)

Transplant-eligible HCC patients have complicated clinical trajectories during the waitlist period stemming from the typically long duration spent on the waitlist. HCC patients typically have lower MELD scores due to relatively mild liver disease. However, the MELD score is not reflective of the incumbent risk of HCC to progress and metastasize. To ensure HCC patients can receive donor Livers with adequate priority they are eligible to receive exception points according to MMaT-3. However, all HCC

patients meeting criteria (within Milan criteria, waiting period of 6 months) receive the same exception point score when some patients have a substantially increased risk of waitlist death or delisting due to decompensation of liver disease or tumor progression and should have differential priority.

Currently, risk prediction for HCC tumor progression, waitlist dropout, and post-transplant recurrence are determined primarily by static clinical and radiographic criteria. These models involve criteria relating to tumor burden (size and number of tumors), AFP cutoff values, and some basic response to treatment measures. Yet none of the existing models utilize the longitudinal measures characteristic of transplant eligible HCC patients. We demonstrate that by utilizing a longitudinal deep survival network we can leverage complex feature interactions overtime to produce more accurate estimates of HCC patient waitlist mortality risk. We then demonstrated the impact of this model on liver allocation using discrete event simulation where the number of HCC patient waitlist dropouts decreased and transplants increased without significantly affecting non-HCC waitlist dropouts or transplants.

Introduction References

[1] T. E. Starzl et al., "Evolution of liver transplantation," Hepatology, vol. 2, no. 5, pp. 614S-636S, 1982.

[2] T. E. Starzl et al., "Orthotopic homotransplantation of the human liver," (in eng), Ann Surg, vol. 168, no. 3, pp. 392-415, Sep 1968, doi: 10.1097/00000658-196809000-00009.

[3] A. J. Kwong et al., "OPTN/SRTR 2022 annual data report: liver," American Journal of Transplantation, vol. 24, no. 2, pp. S176-S265, 2024.

[4] N. J. Burstow et al., "Hepatitis C treatment: where are we now?," International journal of general medicine, pp. 39-52, 2017.

[5] T. Ivanics, P. Abreu, E. De Martin, and G. Sapisochin, "Changing trends in liver transplantation: challenges and solutions," Transplantation, vol. 105, no. 4, pp. 743-756, 2021.

[6] A. Melehy and V. Agopian, "Treating rare tumors with liver transplantation," (in eng), Curr Opin Organ Transplant, vol. 29, no. 1, pp. 30-36, Feb 1 2024, doi: 10.1097/mot.0000000000001118.

[7] A. Kwong et al., "OPTN/SRTR 2018 annual data report: liver," American Journal of Transplantation, vol. 20, pp. 193-299, 2020.

[8] M. Jiménez-Castro, J. Gracia-Sancho, and C. Peralta, "Brain death and marginal grafts in liver transplantation," Cell death & disease, vol. 6, no. 6, pp. e1777-e1777, 2015.

[9] P. S. Kamath et al., "A model to predict survival in patients with end–stage liver disease," Hepatology, vol. 33, no. 2, pp. 464-470, 2001.

[10] S. W. Biggins et al., "Evidence-based incorporation of serum sodium concentration into MELD," Gastroenterology, vol. 130, no. 6, pp. 1652-1660, 2006.

[11] W. R. Kim et al., "MELD 3.0: the model for end-stage liver disease updated for the modern era," Gastroenterology, vol. 161, no. 6, pp. 1887-1895. e4, 2021.

[12] V. Mazzaferro, S. Bhoori, C. Sposito, M. Bongini, M. Langer, R. Miceli, and L. Mariani, "Milan criteria in liver transplantation for hepatocellular carcinoma: an evidence-

based analysis of 15 years of experience," Liver transplantation, vol. 17, no. S2, pp. S44-S57, 2011.

[13] P. Pouramin, S. E. Allen, J. L. Silburt, and B. L. Gala-Lopez, "Median Meld at Transplant Minus 3 Reduces the Mortality of Non-Hepatocellular Carcinoma Patients on the Liver Transplant Waitlist," Current Oncology, vol. 31, no. 11, pp. 7051-7060, 2024.

[14] S. Bernards, R. Hirose, F. Y. Yao, C. Jin, J. L. Dodge, C. Y. Huang, and N. Mehta, "The impact of median model for end-stage liver disease at transplant minus 3 national policy on waitlist outcomes in patients with and without hepatocellular carcinoma," Liver Transplantation, vol. 28, no. 3, pp. 376-385, 2022.

[15] S. Feng et al., "Characteristics associated with liver graft failure: the concept of a donor risk index," American journal of transplantation, vol. 6, no. 4, pp. 783-790, 2006.

[16] E. Giorgakis et al., "Acuity circles allocation policy impact on waitlist mortality and donation after circulatory death liver transplantation: A nationwide retrospective analysis," Health Science Reports, vol. 6, no. 2, 2023.

[17] A. Wey et al., "The effect of acuity circles on deceased donor transplant and offer rates across model for end-stage liver disease scores and exception statuses," Liver transplantation, vol. 28, no. 3, pp. 363-375, 2022.

[18] A. M. Burton and D. S. Goldberg, "Center-level and region-level variations in liver transplantation practices following acuity circles policy change," American Journal of Transplantation, vol. 22, no. 11, pp. 2668-2674, 2022.

[19] A. K. Singal, H. Bashar, B. S. Anand, S. C. Jampana, V. Singal, and Y. F. Kuo, "Outcomes after liver transplantation for alcoholic hepatitis are similar to alcoholic

cirrhosis: exploratory analysis from the UNOS database," Hepatology, vol. 55, no. 5, pp. 1398-1405, 2012.

[20] M. Berenguer et al., "Contribution of donor age to the recent decrease in patient survival among HCV-infected liver transplant recipients," Hepatology, vol. 36, no. 1, pp. 202-210, 2002.

[21] C. Kaltenmeier et al., "The liver transplant risk score prognosticates the outcomes of liver transplant recipients at listing," HPB, vol. 23, no. 6, pp. 927-936, 2021/06/01/ 2021, doi: https://doi.org/10.1016/j.hpb.2020.10.002.

[22] K. M. Olthoff et al., "Validation of a current definition of early allograft dysfunction in liver transplant recipients and analysis of risk factors," Liver Transplantation, vol. 16, no. 8, pp. 943-949, 2010.

[23] V. G. Agopian et al., "Evaluation of early allograft function using the liver graft assessment following transplantation risk score model," JAMA surgery, vol. 153, no. 5, pp. 436-444, 2018.

[24] V. G. Agopian et al., "Multicenter validation of the liver graft assessment following transplantation (L-GrAFT) score for assessment of early allograft dysfunction," Journal of hepatology, vol. 74, no. 4, pp. 881-892, 2021.

[25] A. Rana et al., "Survival outcomes following liver transplantation (SOFT) score: a novel method to predict patient survival following liver transplantation," American Journal of Transplantation, vol. 8, no. 12, pp. 2537-2546, 2008.

[26] M. J. Cooper et al., "DynaMELD: A Dynamic Model of End-Stage Liver Disease for Equitable Prioritization," medRxiv, p. 2024.11. 19.24316852, 2024.

[27] S. Nagai et al., "Use of neural network models to predict liver transplantation waitlist mortality," Liver Transplantation, vol. 28, no. 7, pp. 1133-1143, 2022.

[28] D. Bertsimas, J. Kung, N. Trichakis, Y. Wang, R. Hirose, and P. A. Vagefi, "Development and validation of an optimized prediction of mortality for candidates awaiting liver transplantation," American Journal of Transplantation, vol. 19, no. 4, pp. 1109-1118, 2019.

[29] F. Kanwal et al., "Development, validation, and evaluation of a simple machine learning model to predict cirrhosis mortality," JAMA network open, vol. 3, no. 11, pp. e2023780-e2023780, 2020.

[30] A. Guo, N. R. Mazumder, D. P. Ladner, and R. E. Foraker, "Predicting mortality among patients with liver cirrhosis in electronic health records with machine learning," PloS one, vol. 16, no. 8, p. e0256428, 2021.

[31] A. Kwong et al., "Machine learning to predict waitlist dropout among liver transplant candidates with hepatocellular carcinoma," Cancer Medicine, vol. 11, no. 6, pp. 1535-1541, 2022, doi: https://doi.org/10.1002/cam4.4538.

[32] O. Nitski et al., "Long-term mortality risk stratification of liver transplant recipients: real-time application of deep learning algorithms on longitudinal data," The Lancet Digital Health, vol. 3, no. 5, pp. e295-e305, 2021.

[33] L. Lau, Y. Kankanige, B. Rubinstein, R. Jones, C. Christophi, V. Muralidharan, and J. Bailey, "Machine-learning algorithms predict graft failure after liver transplantation," Transplantation, vol. 101, no. 4, pp. e125-e132, 2017.

[34] Z.-L. He et al., "Application of machine learning models for predicting acute kidney injury following donation after cardiac death liver transplantation," Hepatobiliary & Pancreatic Diseases International, vol. 20, no. 3, pp. 222-231, 2021.

[35] M. Bhat, M. Rabindranath, B. S. Chara, and D. A. Simonetto, "Artificial intelligence, machine learning, and deep learning in liver transplantation," Journal of hepatology, vol. 78, no. 6, pp. 1216-1233, 2023.

[36] P. S. Kamath and R. W. Kim, "Is the change in MELD score a better indicator of mortality than baseline MELD score?," vol. 9, ed: LWW, 2003, pp. 19-21.

[37] T.-I. Huo et al., "Evaluation of the increase in model for end-stage liver disease (ΔMELD) score over time as a prognostic predictor in patients with advanced cirrhosis: risk factor analysis and comparison with initial MELD and Child–Turcotte–Pugh score," Journal of hepatology, vol. 42, no. 6, pp. 826-832, 2005.

[38] R. Moreau et al., "Acute-on-chronic liver failure is a distinct syndrome that develops in patients with acute decompensation of cirrhosis," (in eng), Gastroenterology, vol. 144, no. 7, pp. 1426-37, 1437.e1-9, Jun 2013, doi: 10.1053/j.gastro.2013.02.042.

[39] S. Piano, N. Mahmud, P. Caraceni, M. Tonon, and R. P. Mookerjee, "Mechanisms and treatment approaches for ACLF," Liver International, 2023.

[40] T. Gustot et al., "Clinical course of acute-on-chronic liver failure syndrome and effects on prognosis," Hepatology, vol. 62, no. 1, pp. 243-252, 2015.

[41] R. Moreau, B. Gao, M. Papp, R. Bañares, and P. S. Kamath, "Acute-on-chronic liver failure: a distinct clinical syndrome," Journal of hepatology, vol. 75, pp. S27-S35, 2021.

[42] V. Arroyo and R. Moreau, "Diagnosis and prognosis of acute on chronic liver failure (ACLF) in cirrhosis," Journal of Hepatology, vol. 66, no. 2, pp. 451-453, 2017.

[43] R. Jalan et al., "Acute-on chronic liver failure," Journal of hepatology, vol. 57, no. 6, pp. 1336-1348, 2012.

[44] A. Queck, N. Weiler, and J. Trebicka, "Transplantation in Acute-on-Chronic Liver Failure: Feasibility and Futility," Clinical Liver Disease, vol. 19, no. 5, pp. 191-193, 2022.

[45] J. Trebicka et al., "Rebleeding and mortality risk are increased by ACLF but reduced by pre-emptive TIPS," Journal of hepatology, vol. 73, no. 5, pp. 1082-1091, 2020.

[46] V. Arroyo, R. Moreau, R. Jalan, P. Ginès, and E.-C. C. C. Study, "Acute-on-chronic liver failure: a new syndrome that will re-classify cirrhosis," Journal of hepatology, vol. 62, no. 1, pp. S131-S143, 2015.

[47] R. Jalan et al., "Development and validation of a prognostic score to predict mortality in patients with acute-on-chronic liver failure," Journal of hepatology, vol. 61, no. 5, pp. 1038-1047, 2014.

[48] V. Sundaram et al., "Patients with acute on chronic liver failure grade 3 have greater 14-day waitlist mortality than status-1a patients," Hepatology, vol. 70, no. 1, pp. 334-345, 2019.

[49] F. Durand, O. Roux, E. Weiss, and C. Francoz, "Acute-on-chronic liver failure: Where do we stand?," Liver International, vol. 41, pp. 128-136, 2021.

[50] T. Gustot and B. Agarwal, "Selected patients with acute-on-chronic liver failure grade 3 are not too sick to be considered for liver transplantation," Journal of hepatology, vol. 67, no. 4, pp. 667-668, 2017.

[51] F. Artru et al., "Liver transplantation in the most severely ill cirrhotic patients: a multicenter study in acute-on-chronic liver failure grade 3," Journal of hepatology, vol. 67, no. 4, pp. 708-715, 2017.

[52] F. Artru, D. Goldberg, and P. S. Kamath, "Should patients with acute-on-chronic liver failure grade 3 receive higher priority for liver transplantation?," Journal of hepatology, vol. 78, no. 6, pp. 1118-1123, 2023.

[53] B. Michard et al., "Pretransplant intensive care unit management and selection of grade 3 acute-on-chronic liver failure transplant candidates," Liver Transplantation, vol. 28, no. 1, pp. 17-26, 2022.

[54] V. Sundaram et al., "Risk factors for posttransplantation mortality in recipients with grade 3 acute-on-chronic liver failure: analysis of a North American consortium," Liver Transplantation, vol. 28, no. 6, pp. 1078-1089, 2022.

[55] T. Wu and V. Sundaram, "Transplantation for acute-on-chronic liver failure," Clinical Liver Disease, vol. 14, no. 4, pp. 152-155, 2019.

[56] B. Michard et al., "Liver transplantation in critically ill patients: preoperative predictive factors of post-transplant mortality to avoid futility," Clinical Transplantation, vol. 31, no. 12, p. e13115, 2017.

[57] E. Weiss et al., "When Is a Critically Ill Cirrhotic Patient Too Sick to Transplant? Development of Consensus Criteria by a Multidisciplinary Panel of 35 International Experts," Transplantation, vol. 105, no. 3, 2021. [Online]. Available: https://journals.lww.com/transplantjournal/fulltext/2021/03000/when_is_a_critically_ill_cirrhotic_patient_too.18.aspx.

[58] C. Engelmann, K. L. Thomsen, N. Zakeri, M. Sheikh, B. Agarwal, R. Jalan, and R. P. Mookerjee, "Validation of CLIF-C ACLF score to define a threshold for futility of intensive care support for patients with acute-on-chronic liver failure," Critical Care, vol. 22, pp. 1-8, 2018.

[59] F. Figorilli et al., "Development of an organ failure score in acute liver failure for transplant selection and identification of patients at high risk of futility," PLoS One, vol. 12, no. 12, p. e0188151, 2017.

[60] T. Artzner et al., "Liver transplantation for critically ill cirrhotic patients: stratifying utility based on pretransplant factors," American Journal of Transplantation, vol. 20, no. 9, pp. 2437-2448, 2020.

[61] R. Hernaez et al., "The novel SALT-M score predicts 1-year post-transplant mortality in patients with severe acute-on-chronic liver failure," Journal of Hepatology, vol. 79, no. 3, pp. 717-727, 2023/09/01/ 2023, doi: https://doi.org/10.1016/j.jhep.2023.05.028.

[62] V. Sundaram et al., "Effect of the clinical course of acute-on-chronic liver failure prior to liver transplantation on post-transplant survival," Journal of hepatology, vol. 72, no. 3, pp. 481-488, 2020.

Chapter 1 Center-level variations are associated with disparate waitlist mortality and probability of liver transplant in MELD 35 and greater patients

Chapter 1 Abstract

Liver transplant (LT) centers may have variations in their approach to high-acuity (MELD ≥35) candidates for listing and transplantation. We investigated center-specific differences in waitlist outcomes, probability of LT, and post-LT survival in MELD ≥35 patients. Adult LT candidates (MELD ≥35) were identified from the Scientific Registry of Transplant Recipients between 01/01/10-04/01/22. Waitlist mortality was modelled with center as a random effect. Centers were grouped based on the random effect coefficient, with the highest tertile representing the highest risk group (comprised of centers with the highest risk of waitlist mortality attributable to center). Cumulative incidence of death/delisting, probability of LT, and 1-year post-LT survival were compared, with patients stratified by a listing MELD ≥35 or increase to MELD ≥35 after listing. In patients who increased to MELD ≥35, the 1-year cumulative incidence of death/delisting was highest (50%) and probability of transplant lowest (46%) in the high-risk group, compared to the low-risk group (37% and 55%, $p<0.001$ for both comparisons). For patients receiving an LT, post-transplant survival did not differ among the groups, though high-risk centers transplanted far less high-acuity patients. These variations may indicate that certain centers are associated with substantially decreased waitlist mortality for the highest acuity patients without corresponding decreases in post-transplant survival.

Chapter 1 Introduction

The Model for End-stage Liver Disease (MELD) score has been the standard for objectively determining liver transplant (LT) waitlist candidate priority for the last twenty years [1-5]. The MELD score is a validated metric to prioritize LT candidates in order of medical urgency to ensure fair distribution of scarce deceased donor livers [1,6,7]. Given changes in both the population and management of patients with end-stage liver disease over time, the MELD score has undergone several revisions to ensure that it maintains high discriminatory ability in identifying those patients at highest risk of death [8-12]. Despite these updates to the MELD score, there has been persistent interest in variations in the risk of waitlist mortality for patients with the same MELD score in different regions or centers [13,14]. Additionally, policy changes surrounding the geographic units for liver allocation have fueled conversations about fairness of MELD-based prioritization in the modern era [15]. Specifically, the Share 35 policy in 2013 and acuity circles in 2020 were implemented to increase availability of donor organs to the most critically ill patients [16-20]. However, there has been concern that the waitlist mortality associated with a given MELD score may differ among centers [21]. Some have advocated against the broad sharing of livers as doing so could shift livers out of centers serving vulnerable, disadvantaged, relatively sicker populations to large urban centers where waitlisted candidates may have a higher likelihood of survival at any MELD score [21,22].

Prior research has shown that demographic and social risk factors for waitlist mortality are asymmetrically distributed across the US and are not captured by the lab value derived MELD score [22-24]. To investigate whether these social or demographic factors contribute to center-level variation in waitlist mortality at any given physiological

MELD score was examined by Ishaque *et al* [13]. This study found that the overall variation in waitlist mortality at any given MELD score was minimal at scores from 6-40 and concluded that there are no significant differences in waitlist mortality and changes to liver allocation to capture factors not measured by MELD are not necessary. However, this group conducted subgroup analyses at different ranges of MELD scores included 31-35 and 36-40. The variation in waitlist mortality among centers in these specific subgroups was significantly greater than the average across all MELD scores. Additionally, there has been criticism regarding the prioritization of high-acuity/ high MELD patients, mainly deriving from concerns of poor post-transplant outcomes in these patients as well as potential increased wait-times for intermediate MELD patients listed at centers that have a high median MELD at transplant [25]. In this study, we analyzed center-level variation in waitlist mortality, specifically in patients with a MELD score ≥35, which has not previously been examined. We hypothesized that there is significant center-level variation in waitlist mortality in these high acuity patients that may stem from social, demographic, or transplant center behavior that is not captured by the MELD score.

Chapter 1 Patients and Methods

Data Source

This study used data from the Scientific Registry of Transplant Recipients (SRTR). The SRTR data includes data on all donor, waitlisted candidates, and transplant recipients in the US, submitted by the members of the Organ Procurement and Transplant network (OPTN). The Health Resources and Services Administration, which is part of the U.S. Department of Health and Human Services, oversees the

activities of the OPTN and SRTR contractors. The data reported here have been supplied by the Hennepin Healthcare Research Institute (HHRI) as the contractor for the SRTR. The interpretation and reporting of these data are the responsibility of the authors and in no way should be seen as an official policy of or interpretation by the SRTR or the U.S. Government. This study involved secondary analysis of de-identified registry data and was granted exemption from IRB approval at our institution.

Study Population

We identified adult (age > 18), first-time, liver-only waitlist candidates from 112 LT centers from January 1, 2010 to April 1, 2022. We excluded 40 LT centers that performed fewer than an average of five LTs per year during the study period or were missing either a center code or center ID. To assess the medical urgency of LT, waitlist candidates were assigned a physiologic MELD score from 6 to 40 using serum creatinine, bilirubin, and INR (prior to 2016), and serum creatinine, bilirubin, INR, and sodium (after 2016). We excluded patients who received any MELD exception points and those listed as status 1A at any point while waitlisted as these factors affect organ allocation and disease severity independent of MELD score.

Primary and Secondary Outcomes

The primary outcome of interest among patients who received a physiologic MELD ≥35 on the waitlist was one-year waitlist mortality or delisting due to candidate condition deterioration or being “too sick for transplant.” Additionally, if a patient was removed from the waitlist within one year for reason “other” and they subsequently had a documented date of death, then they were considered to have experienced the outcome of interest. Secondary outcomes included cumulative incidence of candidate

death or delisting due to medical deterioration, cumulative incidence of LT, and post-transplant one-year mortality or graft failure in those patients who received LT.

Statistical Analysis

All analysis were conducted using R version 4.1.0 (R Foundation for Statistical Computing, Vienna, Austria). Any P values less than 0.05 were considered significant.

To stratify centers based on the risk of waitlist mortality or delisting due to disease progression among patients with a MELD score ≥35, we created a mixed effects Cox proportional hazards model using one-year waitlist death or delisting as the outcome and LT center as a random effect. The mixed effects Cox proportional hazards model is known as a "frailty" model and allows center-level variation in the risk of the outcome while controlling for other covariates.26,27 We conducted univariate analyses of all candidate variables with clinical relevance and with no significant missing data (< 5%). Variables that were significantly associated with the outcome in univariate analysis and were deemed to be clinically relevant were included in a multivariate model. We included an "era" variable that split patients into groups based on time of listing (1/1/2010-6/1/2013, 6/1/2013-2/1/2020, 2/1/2020-4/1/2022) to capture any effects of allocation policy changes that have affected availability of donor organs for patients with MELD ≥35. To produce the coefficients for the center-level random effect variable, the baseline hazard of all patients without clustering by center was set to zero with each center given a random effect coefficient, forming a normal distribution with mean zero. The random effect coefficients for each center were plotted as a normal distribution and groups were formed by taking tertiles. The low-risk group had the lowest values of the center-level random effect coefficient, followed by the middle-risk group, and high-risk

group with the highest values. The hazard ratios associated with each center-level random effect coefficient were calculated using the exponential function. The hazard ratios with confidence intervals were displayed in a caterpillar plot. All variables included in modeling had fewer than 5% missing data and missing data was handled via list-wise deletion. A map of the US showing each center's geographic location and which group it belongs to was created. Differences in UNOS region representation among the three groups was tested using the Chi-squared test.

Baseline demographic comparisons were made among the three groups. Categorical variables are displayed as number (%) and compared using Chi-squared. Continuous variables were displayed as median (interquartile range) and compared using the Kruskal-Wallis test.

The cumulative incidence function was used to create cumulative incidence curves with one-year waitlist death or delisting as the primary outcome and transplant as a competing risk. Separate cumulative incidence curves of waitlist death or delisting and transplant were created for patients who were initially listed at a MELD ≥35 and those who decompensated to a MELD ≥35 after listing. Significance between differences in cumulative incidence were assessed via Fine and Gray competing risk regression.

For the post-transplant analysis, we included patients receiving LT who either had 1) initial listing MELD > 35 or 2) had listing MELD < 35 who subsequently decompensated to a MELD ≥35. Patients were censored at time of last follow-up and one-year post-transplant graft-failure free survival was compared among the three groups using Kaplan-Meier curves. Survival differences were tested for significance

using the log-rank test. Demographic and pre-LT acuity measures were also compared among the three risk groups. Hypothesis testing of demographic comparisons was performed as described above.

We performed a sensitivity analysis to further examine any effects of “era” on the identified risk groups by re-creating the three risk-groups using the methodology described above for each era separately. The centers changing group affiliation over time were noted and waitlist death or delisting and probability of transplant were plotted in cumulative incidence curves for each time period.

Additionally, we examined drivers of risk-group differences using several approaches. Differences in transplant volume, median MELD at transplant, and UNOS regional representation in the three groups were assessed. We then created multivariate Cox proportional hazards models (no random effects) with one-year waitlist death or delisting as the outcome and the same covariates included in the mixed-effects modeling used to create the groups. This effort was focused on identifying differences in significant features associated with the outcome between the three groups. To further

identify possible drivers of group outcome differences, we compared causes of death in patients who experienced waitlist mortality between the three groups.

Chapter 2 Results

Univariate factors associated with an increased risk of waitlist mortality/delisting in MELD ≥35 patients are shown in Table 1. On multivariate analysis (Table 2), final MELD prior to event (HR 1.08, 95% CI [1.07-1.09]), male sex (HR 0.89, 95% CI [0.85-0.94]), primary diagnosis, ascites, encephalopathy, age (HR 1.01, 95% CI [1.01-1.02]), mechanical ventilation (HR 5.23, 95% CI [4.81-5.69]), hemodialysis (HR 1.07, 95% CI

[1.01-1.13]), and era were identified as significant independent covariates associated with one year waitlist death or delisting. The center-specific adjusted hazard ratios based on the mixed effects Cox proportional hazards model are shown in a caterpillar plot, with the risk groups stratified by tertile of adjusted center-specific risk of one-year waitlist death or delisting. Risk groups are shown in green for low-risk, blue for middle-risk, and red for high-risk groups (Figure 1A). To assess any geographic associations with membership in the risk groups, we created a map of the US with every center plotted in its geographic location and colored by group affiliation (Figure 1B). There were no apparent geographic associations with membership in any one group and no difference in UNOS region representation among the groups (p = 0.17).

Table 1. Univariate mixed effects Cox proportional hazards analysis

Variable	HR	LL	UL	P
Age, y	1.01	1.00	1.01	<0.001
Sex, M	0.89	0.85	0.94	0.005
Primary Diagnosis				
AIH/PSC/PBC	1.00	Ref	Ref	Ref
Alcoholic Cirrhosis	1.45	1.32	1.59	<0.001
Alcoholic Hepatitis	1.64	1.31	2.05	<0.001
Hepatic Malignancy/ Benign Tumor	1.75	1.42	2.15	<0.001
Hepatitis B	2.01	1.66	2.42	<0.001
Hepatitis C	1.20	1.08	1.33	<0.001
Metabolic Disease	1.04	0.83	1.29	0.76
NASH	1.21	1.09	1.34	<0.001
Other	1.81	1.32	1.59	<0.001
BMI	1.00	0.99	1.00	0.91

MELD at Listing	1.11	1.10	1.11	<0.001
MELD Last	1.09	1.08	1.10	<0.001
Life Support	0.94	0.13	6.78	0.95
Mechanical Ventilation	5.47	5.06	5.91	<0.001
Bacterial Peritonitis	1.35	1.25	1.46	<0.001
Portal Vein Thrombosis	0.99	0.86	1.06	0.41
TIPSS	0.81	0.73	0.89	<0.001
Previous Abdominal Surgery	0.95	0.90	1.00	0.070
Last Serum Albumin	0.97	0.94	1.00	0.10
Ascites				
Absent	1.00	Ref	Ref	Ref
Slight	0.84	0.77	0.93	<0.001
Moderate	1.12	1.02	1.22	0.019
Encephalopathy				
None	1.00	Ref	Ref	Ref
Grade 1-2	1.26	1.15	1.39	<0.001
Grade 3-4	2.56	2.34	2.80	<0.001
Dialysis Prior Week	1.36	1.29	1.44	<0.001
Era				
2013-2020	1.00	Ref	Ref	Ref
2010-2013	0.96	0.90	1.01	0.14
2020-2022	1.24	1.15	1.33	<0.001

HR: hazard ratio; LL: lower bound of 95% confidence interval; UL: upper bound of 95% confidence interval; MELD: model for end-stage liver disease; AIH: autoimmune hepatitis; PSC: primary sclerosing cholangitis; PBC: primary biliary cholangitis; NASH: nonalcoholic steatohepatitis; TIPSS: Transjugular intrahepatic portosystemic shunt; BMI: body mass index

Table 2 Multivariate mixed effects Cox proportional hazards analysis.

Variable	HR	LL	UL	P
MELD Last	1.08	1.07	1.09	<0.001

Age, y	1.01	1.01	1.02	<0.001
Gender = M	0.89	0.85	0.94	<0.001
Primary Diagnosis				
AIH/PSC/PBC	1.00	Ref	Ref	Ref
Alcoholic Cirrhosis	1.43	1.29	1.57	<0.001
Alcoholic Hepatitis	1.85	1.47	2.32	<0.001
Hepatic Malignancy/ Benign Tumor	1.67	1.36	2.06	<0.001
Hepatitis B	1.98	1.63	2.39	<0.001
Hepatitis C	1.15	1.03	1.28	0.025
Metabolic Disease	1.14	0.91	1.41	0.11
NASH	1.08	0.98	1.20	0.15
Other	1.63	1.47	1.81	<0.001
Ascites				
Absent	1.00	Ref	Ref	Ref
Slight	0.77	0.70	0.85	<0.001
Moderate	0.82	0.74	0.91	0.019
Encephalopathy				
None	1.00	Ref	Ref	Ref
Grade 1-2	1.28	1.16	1.41	<0.001
Grade 3-4	2.18	1.98	2.41	<0.001
Age, y	1.01	1.01	1.02	<0.001
Mechanical Ventilator	5.23	4.81	5.69	<0.001
Dialysis Prior Week = Y	1.07	1.01	1.13	0.027
Era (ref 2013-2020)				
2013-2020	1.00	Ref	Ref	Ref
2010-2013	0.92	0.87	0.98	0.008
2020-2022	1.22	1.13	1.32	<0.001

HR: hazard ratio; LL: lower bound of 95% confidence interval; UL: upper bound of 95% confidence interval; MELD: Model for end-stage liver disease; AIH: autoimmune hepatitis; PSC: primary sclerosing cholangitis; PBC: primary biliary cholangitis; NASH: nonalcoholic steatohepatitis

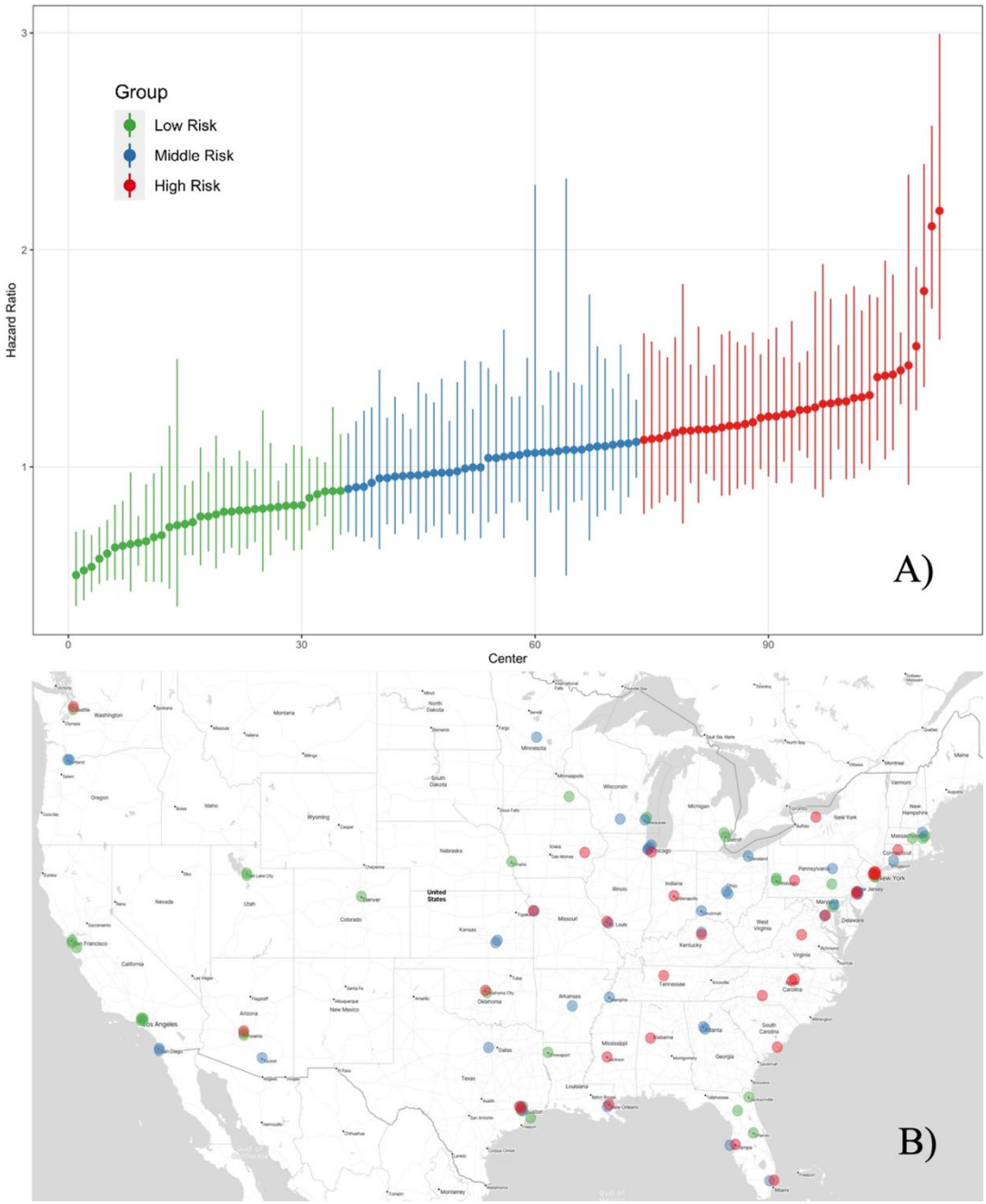


Figure 1. Center and regional variation in liver transplant waitlist mortality

A) Caterpillar plot showing the adjusted center-specific hazard ratios for waitlist death or delisting. B) Map of the United States showing the regional distribution of centers in

each risk group. This map was created using the “ggmap” R package utilizing a template provided by Stadia Maps (stadiamaps.com).

While risk-score groups were constructed based on waitlist outcomes in only the MELD ≥35 patients, outcomes on all waitlisted and transplanted patients by risk-score group were first evaluated. There were 95,121 total patients listed for LT who met the study inclusion criteria. In the total cohort. There were 35, 38, and 39 centers with 34,720 (37%), 29,799 (31%), and 30,602 (32%) patients listed for LT in the low-, middle-, and high-risk groups. Center-specific hazard ratios for waitlist mortality/delisting did not appear to be associated with overall center volume, as there was a median of 37.30 (IQR: [21.70-53.50]), 36.00 (IQR: [14.20-54.80]), and 29.80 (IQR: [19.00-51.10]) LTs performed per year in the low-, middle-, and high-risk groups, respectively (p=0.68). The median MELD at transplant for the overall population was 28.00 (IQR: [24.00-31.00]), 26.00 (IQR: [24.00-29.00]), and 24.00 (IQR: [22.00-31.00]) in the low-, middle-, and high-risk groups, respectively (p=0.010). Complete patient-level demographic information for the entire cohort is displayed in Table 3 and a flow-chart diagraming inclusion and exclusion criteria resulting in our patient cohorts for analysis is shown in Figure 2.

Table 3. Baseline demographics in the total cohort as well as each risk group with comparisons among variables in the three groups.

Variable	Total (n=95,121)	Low Risk (n=34,720)	Middle Risk (n=29,799)	High Risk (n=30,602)	P
Age, y (median [IQR])	56.00 [48.00, 62.00]	56.00 [49.00, 63.00]	56.00 [48.00, 62.00]	56.00 [48.00, 62.00]	0.004
Gender, M (%)	57088 (60)	20700 (60)	17896 (60)	18492 (60)	0.10

Primary Diagnosis (%)					<0.001
AIH/PSC/PBC	10691 (11)	3837 (11)	3137 (11)	3717 (12)	
Alcoholic cirrhosis	32733 (34	12426 (36)	10706 (36)	9601 (31)	
Alcoholic hepatitis	1148 (1)	403 (1)	426 (1)	319 (1)	
Hepatic malignancy/ benign tumor	2392 (3)	866 (3)	790 (3)	736 (2)	
Hepatitis B	1446 (2)	560 (2)	427 (1)	459 (2)	
Hepatitis C	13776 (15)	5023 (15)	4356 (15)	4397 (14)	
Metabolic disease	1750 (2)	607 (2)	544 (2)	599 (2)	
NASH	17707 (19)	6058 (18)	5414 (18)	6235 (20)	
other	13450 (14)	4928 (14)	3992 (13)	4530 (15)	
BMI (median [IQR])	28.00 [25.00, 33.00]	28.00 [24.00, 33.00]	28.00 [25.00, 33.00]	28.00 [25.00, 33.00]	<0.001
MELD at Listing (median [IQR])	19.00 [14.00, 27.00]	18.00 [13.00, 27.00]	19.00 [14.00, 27.00]	19.00 [14.00, 26.00]	<0.001
MELD Last (median [IQR])	23.00 [15.00, 32.00]	23.00 [15.00, 34.00]	23.00 [16.00, 32.00]	22.00 [16.00, 31.00]	<0.001
MELD ≥35 at Listing (%)	11755 (12)	4637 (13)	3725 (13)	3393 (11)	<0.001
MELD Increased to ≥35(%)	10178 (11)	4398 (13)	2945 (10)	2835 (9)	<0.001
Life Support (%)	4632 (5)	2038 (6)	1588 (5)	1006 (3)	<0.001
Mechanical Ventilator (%)	3249 (3)	1572 (5)	946 (3)	731 (2)	<0.001
Bacterial Peritonitis (%)	9480 (10)	3494 (10)	2965 (10)	3021 (10)	<0.001
Portal Vein Thrombosis (%)	6683 (7)	2536 (7)	1853 (6)	2294 (8)	<0.001

Previous Abdominal Surgery (%)	38859 (41)	13659 (39)	12649 (42)	12551 (41)	<0.001
Ascites (%)					<0.001
absent	18137 (19)	6259 (18)	5608 (19)	6270 (21)	
slight	43060 (45)	15572 (45)	13656 (46)	13832 (45)	
moderate	33908 (36)	12882 (37)	10531 (35)	10495 (34)	
Hepatic Encephalopathy (%)					<0.001
None	28006 (29)	9363 (27)	8732 (29)	9911 (32)	
Grade 1-2	51181 (54)	18335 (53)	16370 (55)	16476 (54)	
Grade 3-4	15918 (17)	7015 (20)	4693 (16)	4210 (14)	
Dialysis Prior Week (%)	17959 (19)	8099 (23)	5540 (19)	4320 (14)	<0.001
Liver Transplant (%)	54002 (57)	18112 (52)	17844 (60)	18046 (59)	<0.001
Death or Delisting (%)	20670 (22)	8152 (24)	5965 (20)	6553 (21)	<0.001
One-year Death or Delisting (%)	13700 (14)	4967 (14)	4152 (14)	4581 (15)	<0.001
Time to Event (median [IQR])	101.00 [14.00, 458.00]	150.00 [18.00, 615.00]	79.00 [12.00, 381.00]	85.00 [14.00, 371.00]	<0.001
Era by Listing (%)					0.72
2010-2013	23901 (25)	8755 (25)	7406 (25)	7740 (25)	
2013-2020	50793 (53)	18529 (53)	15940 (54)	16324 (53)	
2020-2022	20427 (22)	7436 (21)	6453 (22)	6538 (21)	

AIH: autoimmune hepatitis; PSC: primary sclerosing cholangitis; PBC: primary biliary cholangitis; NASH: non-alcoholic steatohepatitis; BMI: body mass index; MELD: model for end-stage liver disease; IQR: interquartile range; SD: standard deviation

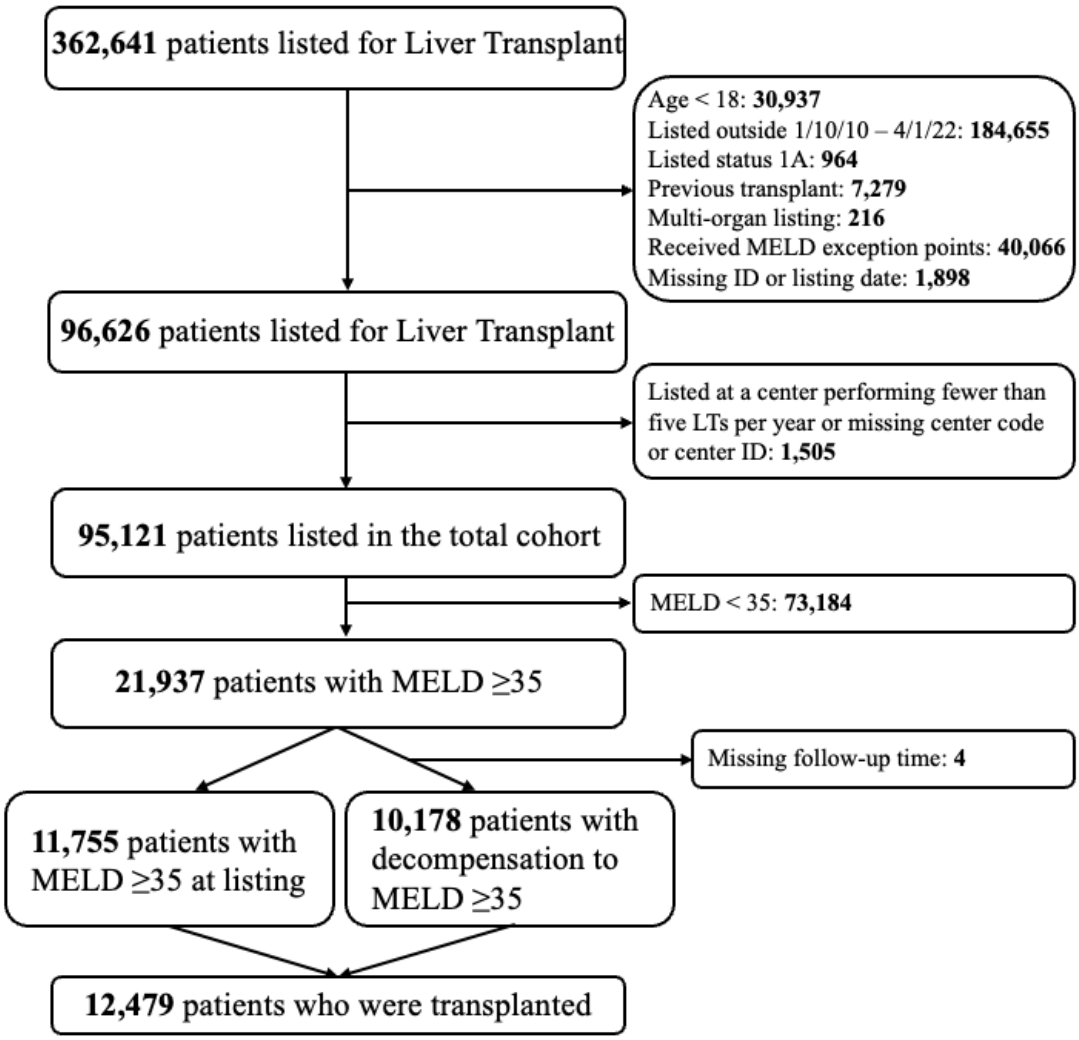


Figure 2. Cohort selection diagram.

Flow diagram showing inclusion and exclusion criteria for this study including the number of patients in each cohort for analysis (MELD ≥35 at listing, decompensation to MELD ≥35 after listing, MELD ≥35 at any point who received a transplant. MELD, Model for End-stage Liver Disease.

MELD ≥35 at time of Listing.

Among patients listed with an initial MELD ≥35, the low-risk group had more LT candidates on life support (27% low-risk, 25% middle-risk, 17% high-risk; p<0.001), more candidates on mechanical ventilation (22% low-risk, 16% middle-risk, 13% high-risk; p<0.001), more candidates with grade 3-4 encephalopathy (41% low-risk, 36% middle-risk, 33% high-risk; p<0.001), and more candidates with moderate ascites (57% low-risk, 51% middle-risk, 47% high-risk; p<0.001). Complete demographic comparisons of MELD ≥35 patients at listing among the three risk groups are shown in Table 4. The cumulative incidence of death or delisting among patients initially listed at a MELD ≥35 was significantly lower in the low-risk group, compared to the high-risk group (21% vs 24%, respectively, p<0.001; Figure 3A). The cumulative incidence of LT in patients with an initial MELD ≥35 was not significantly different among the three risk groups (78% low-risk, 78% middle-risk, 75% high-risk, p=0.65; Figure 3B).

Table 4. Baseline demographic comparisons in patients initially listed at MELD ≥35 among the three risk groups.

Variable	Low Risk (n=4637)	Middle Risk (n=3725)	High Risk (n=3393)	P
Age, y (median [IQR])	53.00 [43.00, 60.00]	52.00 [42.00, 60.00]	52.00 [40.00, 59.00]	<0.001
Sex = M (%)	2697 (58.2)	2199 (59.0)	1976 (58.2)	0.69
Primary Diagnosis (%)				<0.001
AIH/PSC/PBC	274 (5.9)	237 (6.4)	260 (7.7)	

Alcoholic Cirrhosis	1813 (39.1)	1486 (39.9)	1172 (34.6)	
Alcoholic Hepatitis	236 (5.1)	254 (6.8)	186 (5.5)	
Hepatic Malignancy/ Benign Tumor	62 (1.3)	54 (1.5)	43 (1.3)	
Hepatitis B	121 (2.6)	100 (2.7)	97 (2.9)	
Hepatitis C	461 (10.0)	374 (10.0)	307 (9.1)	
Metabolic Disease	85 (1.8)	74 (2.0)	64 (1.9)	
NASH	556 (12.0)	400 (10.7)	327 (9.6)	
Other	1025 (22.1)	745 (20.0)	935 (27.6)	
BMI (median [IQR])	29.00 [25.00, 34.00]	29.00 [25.00, 34.00]	30.00 [26.00, 35.00]	0.017
MELD at Listing (median [IQR])	40.00 [37.00, 40.00]	40.00 [37.00, 40.00]	39.00 [37.00, 40.00]	0.058
MELD Last (median [IQR])	40.00 [36.00, 40.00]	40.00 [36.00, 40.00]	40.00 [36.00, 40.00]	0.29
Life Support (%)	1269 (27.4)	941 (25.3)	592 (17.4)	<0.001
Mechanical Ventilator (%)	999 (21.5)	593 (15.9)	430 (12.7)	<0.001
Bacterial Peritonitis (%)	677 (14.6)	568 (15.2)	449 (13.2)	0.001
Portal Vein Thrombosis (%)	256 (5.5)	180 (4.8)	167 (4.9)	0.34
Previous Abdominal Surgery (%)	1453 (31.3)	1250 (33.6)	1045 (30.8)	0.005
Ascites (%)				<0.001
absent	512 (11.0)	458 (12.3)	601 (17.7)	
slight	1482 (32.0)	1351 (36.3)	1208 (35.6)	

moderate	2643 (57.0)	1915 (51.4)	1584 (46.7)	
Hepatic Encephalopathy (%)				<0.001
None	589 (12.7)	594 (16.0)	637 (18.8)	
Grade 1-2	2127 (45.9)	1797 (48.3)	1631 (48.1)	
Grade 3-4	1921 (41.4)	1333 (35.8)	1125 (33.2)	
Dialysis Prior Week (%)	3002 (64.7)	2029 (54.5)	1474 (43.4)	<0.001
Liver Transplant = 1 (%)	3557 (76.7)	2820 (75.7)	2451 (72.2)	<0.001
Death or Delisting = 1 (%)	916 (19.8)	765 (20.5)	763 (22.5)	0.011
One-year Death or Delisting = 1 (%)	911 (19.6)	763 (20.5)	761 (22.4)	0.009
Time to Event (median [IQR])	5.00 [3.00, 10.00]	4.00 [2.00, 9.00]	4.00 [2.00, 8.00]	<0.001
Era by Listing (%)				0.002
2010-2013	935 (20.2)	841 (22.6)	719 (21.2)	
2013-2020	2543 (54.8)	2010 (54.0)	1765 (52.0)	
2020-2022	1159 (25.0)	874 (23.5)	909 (26.8)	

AIH: autoimmune hepatitis; PSC: primary sclerosing cholangitis; PBC: primary biliary cholangitis; NASH: non-alcoholic steatohepatitis; BMI: body mass index; MELD: model for end-stage liver disease; IQR: interquartile range; SD: standard deviation

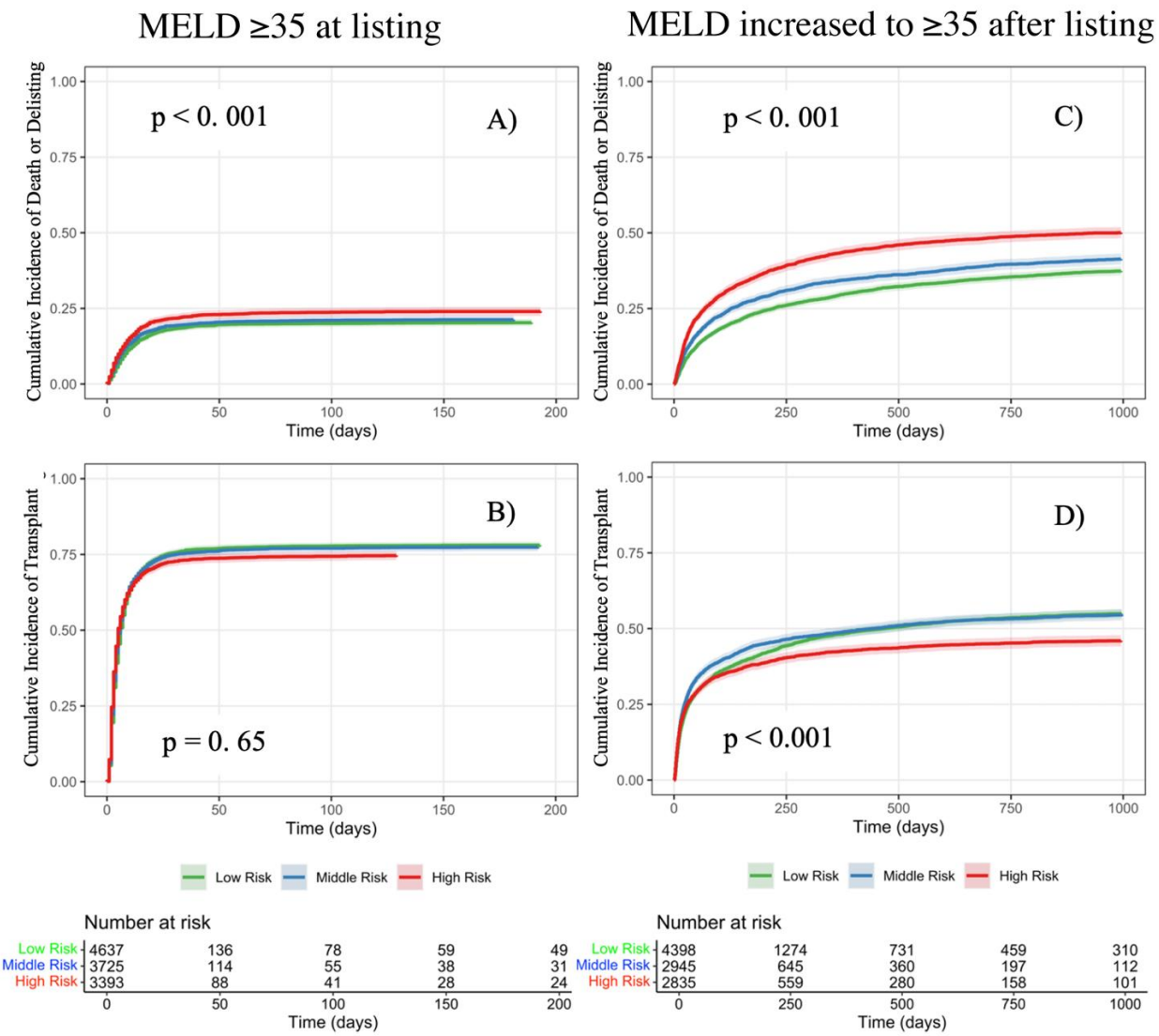


Figure 3. Waitlist mortality and transplant probability across center risk groups.

Cumulative incidence curves comparing the three risk groups in patients initially listed at a MELD ≥35 in terms of waitlist death or delisting (A) and transplant (B) as well as patients initially at a MELD <35 with increase to ≥35 on the waitlist in terms of death or delisting with (C) and transplant (D). MELD: model for end stage liver disease.

MELD < 35 at listing with subsequent decompensation to MELD ≥35 on the waitlist.

Among patients with a listing MELD < 35 with decompensation to MELD ≥35, there were 4398, 2945, and 2835 patients in the low-, middle-, and high-risk groups, respectively. The low-risk group had more LT candidates on life support (6.3% low-risk, 5.8% middle-risk, and 4.3% high-risk, p=0.003), more candidates on mechanical ventilation (4.9% low-risk, 3.7% middle-risk, and 3.6% high-risk, p=0.008), more candidates with grade 3-4 encephalopathy (41% low-risk, 35% middle-risk, and 31% high-risk, p<0.001), and more candidates with moderate ascites (59% low-risk, 55% middle-risk, and 53% high-risk, p<0.001). Comparisons of baseline demographics among the groups restricted to patients initially listed at a MELD < 35 with decompensation to MELD ≥35 are shown in Table 5. The cumulative incidence of death or delisting among patients who decompensated to a MELD ≥35 on the list was significantly lower in the low-risk group compared to the middle and high-risk groups (37%, 41%. 50% in the low-, middle-, and high-risk groups, respectively, p<0.001; Figure 3C). Patients with a listing MELD < 35 who decompensated to MELD ≥35 were significantly more likely to receive LT in the low (55%) and middle (55%) risk groups compared to the high-risk group (46%, p<0.001; Figure 3D).

Table 5. Baseline demographic comparisons in patients initially listed at MELD < 35 with increase to MELD ≥35 while on the waitlist.

Variable	Low Risk	Middle Risk	High Risk	P
Age, y (median [IQR])	57.00 [50.00,	57.00 [49.00,	57.00 [49.00,	0.051
Gender = M (%)	2425 (55.1)	1672 (56.8)	1590 (56.1)	0.371

Primary Diagnosis (%)				<0.001
AIH/PSC/PBC	274 (5.9)	237 (6.4)	260 (7.7)	
Alcoholic Cirrhosis	1813 (39.1)	1486 (39.9)	1172 (34.6)	
Alcoholic Hepatitis	236 (5.1)	254 (6.8)	186 (5.5)	
Hepatic Malignancy/ Benign	62 (1.3)	54 (1.5)	43 (1.3)	
Hepatitis B	121 (2.6)	100 (2.7)	97 (2.9)	
Hepatitis C	461 (10.0)	374 (10.0)	307 (9.1)	
Metabolic Disease	85 (1.8)	74 (2.0)	64 (1.9)	
NASH	556 (12.0)	400 (10.7)	327 (9.6)	
Other	1025 (22.1)	745 (20.0)	935 (27.6)	
BMI (median [IQR])	29.00 [25.00,	29.00 [25.00,	29.00 [25.00,	0.061
MELD at Listing (median [IQR])	22.00 [16.00,	24.00 [17.00,	24.00 [18.00,	<0.001
MELD Last (median [IQR])	40.00 [37.00,	39.00 [36.00,	39.00 [37.00,	<0.001
Life Support (%)	279 (6.3)	172 (5.8)	123 (4.3)	0.003
Mechanical Ventilator = 1 (%)	214 (4.9)	109 (3.7)	101 (3.6)	0.008
Bacterial Peritonitis (%)	499 (11.3)	362 (12.3)	310 (10.9)	0.008
Portal Vein Thrombosis (%)	331 (7.5)	189 (6.4)	221 (7.8)	0.036
Previous Abdominal Surgery (%)	1686 (38.3)	1190 (40.4)	1146 (40.4)	<0.001
Ascites (%)				<0.001
absent	349 (7.9)	278 (9.4)	335 (11.8)	
slight	1465 (33.3)	1055 (35.8)	1003 (35.4)	
moderate	2583 (58.7)	1611 (54.7)	1497 (52.8)	
Encephalopathy (%)				<0.001
None	532 (12.1)	457 (15.5)	610 (21.5)	
Grade 1-2	2083 (47.4)	1447 (49.2)	1341 (47.3)	
Grade 3-4	1782 (40.5)	1040 (35.3)	884 (31.2)	
Dialysis Prior Week (%)	2678 (60.9)	1429 (48.5)	1109 (39.1)	<0.001
Liver Transplant = 1 (%)	2495 (56.7)	1634 (55.5)	1309 (46.2)	<0.001
Death or Delisting = 1 (%)	1814 (41.2)	1253 (42.5)	1457 (51.4)	<0.001

One-year Death or Delisting = 1	1278 (29.1)	992 (33.7)	1203 (42.4)	<0.001
Time to Event (median [IQR])	81.00 [19.00,	49.00 [14.00,	45.00 [13.00,	<0.001
Time from Initial MELD to Last	72.00 [15.00,	44.00 [12.00,	40.00 [11.00,	<0.001
Era by Listing (%)				0.004
2010-2013	1245 (28.3)	920 (31.2)	881 (31.1)	
2013-2020	2490 (56.6)	1646 (55.9)	1534 (54.1)	
2020-2022	663 (15.1)	379 (12.9)	420 (14.8)	

AIH: autoimmune hepatitis; PSC: primary sclerosing cholangitis; PBC: primary biliary cholangitis; NASH: non-alcoholic steatohepatitis; BMI: body mass index; MELD: model for end-stage liver disease; IQR: interquartile range; SD: standard deviation.

MELD 35 patients (MELD ≥35 at listing and decompensated to MELD ≥35) who received LT.

Of patients with MELD ≥35 at any point in the waitlist, 5,250, 3,877, and 3,352 patients underwent LT in the low-, middle-, and high-risk groups, respectively. The median MELD at transplant did not significantly differ among the groups (39.00, IQR: [36.00-40.00], 39.00, IQR: [36.00-40.00], 39.00, IQR [36.00-40.00] in the low, middle, and high-risk groups, respectively, p=0.40). The one-year post-transplant graft-failure free survival was not significantly different among the three groups in patients initially listed at MELD ≥35 or in those who had an increase to MELD ≥35 on the waitlist (p=0.87 and p=0.26, respectively). Kaplan-Meier curves comparing graft-failure free survival of patients initially listed at MELD ≥35 and those who increased to MELD ≥35 among the three groups are displayed in Figure 4A & B, respectively. While survival outcomes and MELD scores in the transplanted patients did not differ among the low, middle, and high-risk groups, there were significant differences in the pretransplant characteristics, with the low-risk group transplanting much higher-acuity patients.

Compared to patients transplanted in the middle- and high-risk group centers, LT recipients in the low-risk group centers were significantly more likely to be mechanically ventilated at time of LT (22% vs 14% and 13%, respectively; p<0.001), be hospitalized in the ICU at transplant (62% vs 45% and 43%, respectively, p<0.001), and have received dialysis in the week prior to transplant (59% vs 45% and 35%, respectively; p<0.001). Complete demographic comparisons in MELD ≥35 patients at transplant among the three risk groups is shown in Table 6.

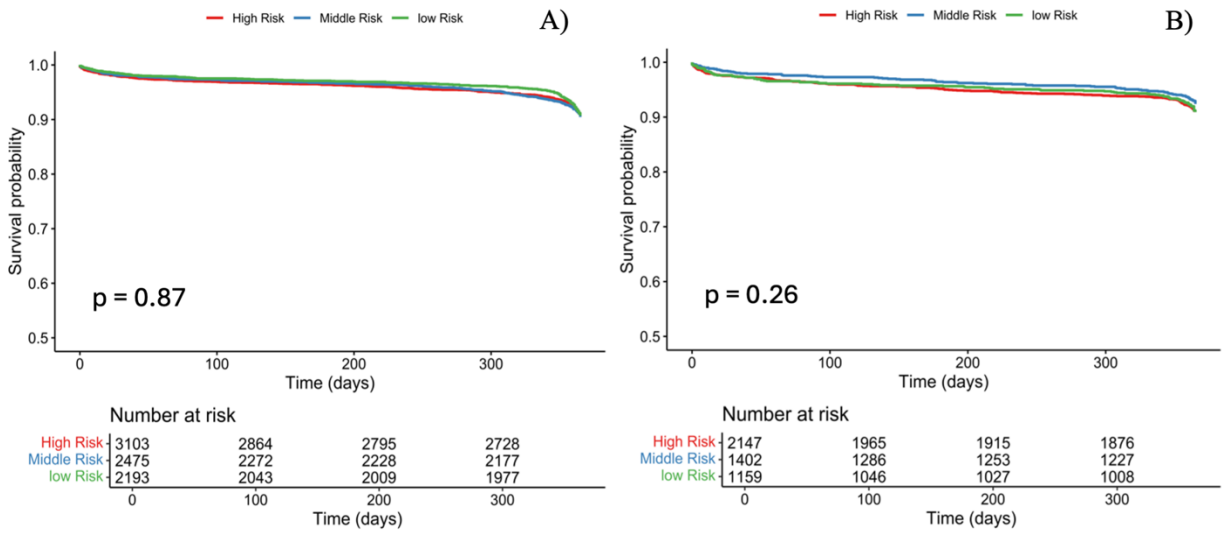


Figure 4. Post-transplant survival across center risk groups.

Kaplan-Meier curves comparing one-year death or graft failure following liver transplant in the three risk groups in patients initially listed at a MELD >=35 (A) and in patients listed at a MELD < 35 with subsequent increase to MELD >=35 (B). Displayed p-values correspond to the log-rank test for significance in the difference between mortality in the three groups. MELD: model for end-stage liver disease.

Table 6. Baseline demographic comparisons in patients with MELD ≥35 (at listing or increased on list) who underwent LT among the three risk groups.

Variable	Low Risk (n=5250)	Middle Risk (n=3877)	High Risk (n=3352)	P
Age at Transplant, y (median [IQR])	54.00 [44.83, 60.73]	52.83 [43.17, 59.92]	51.91 [41.67, 59.33]	<0.001
Gender = M (%)	3017 (57.5)	2328 (60.0)	2014 (60.1)	0.014
Primary Diagnosis (%)				<0.001
AIH/PSC/PBC	478 (9.1)	384 (9.9)	349 (10.4)	
alcoholic cirrhosis	1907 (36.4)	1444 (37.2)	1123 (33.5)	
alcoholic hepatitis	181 (3.5)	220 (5.7)	151 (4.5)	
hepatic malignancy/ benign tumor	148 (2.8)	87 (2.2)	52 (1.6)	
hepB	128 (2.4)	99 (2.6)	93 (2.8)	
hepC	629 (12.0)	432 (11.1)	388 (11.6)	
metabolic disease	125 (2.4)	94 (2.4)	87 (2.6)	
NASH	703 (13.4)	460 (11.9)	366 (10.9)	
other	946 (18.0)	657 (16.9)	742 (22.1)	
BMI (median [IQR])	28.45 [24.37, 33.30]	28.74 [24.81, 33.62]	29.27 [25.31, 34.33]	<0.001
MELD at Transplant (median [IQR])	39.00 [36.00, 40.00]	39.00 [36.00, 40.00]	39.00 [36.00, 40.00]	0.400
Life Support (%)	2197 (41.8)	1072 (27.7)	747 (22.3)	<0.001

Mechanical Ventilation (%)	1153 (22.0)	536 (13.8)	450 (13.4)	<0.001
Recipient Medical Condition (%)				<0.001
ICU	3235 (61.7)	1738 (44.9)	1443 (43.0)	
Hospitalized (Ward)	1718 (32.7)	1807 (46.6)	1576 (47.0)	
Not Hospitalization	293 (5.6)	330 (8.5)	333 (9.9)	
Portal Vein Thrombosis (%)	576 (11.0)	345 (8.9)	283 (8.4)	<0.001
Previous Abdominal Surgery (%)	1828 (34.8)	1527 (39.4)	1247 (37.2)	<0.001
Ascites (%)				<0.001
Absent	509 (9.7)	445 (11.5)	463 (13.8)	
Slight	1822 (34.7)	1468 (37.9)	1252 (37.4)	
Moderate	2919 (55.6)	1963 (50.6)	1637 (48.8)	
Encephalopathy (%)				<0.001
None	718 (13.7)	708 (18.3)	652 (19.5)	
Grade 1-2	2730 (52.0)	2088 (53.9)	1767 (52.7)	
Grade 3-4	1802 (34.3)	1080 (27.9)	933 (27.8)	
Dialysis Prior Week = Y (%)	3106 (59.2)	1740 (44.9)	1188 (35.4)	<0.001
Donor Risk Index (median [IQR])	1.24 [1.07, 1.49]	1.21 [1.05, 1.46]	1.20 [1.05, 1.46]	<0.001

One-year Post-Transplant Death or Graft failure (%)	450 (8.6)	321 (8.3)	292 (8.7)	0.79
Follow-up Time (median [IQR])	1450.50 [673.25, 2557.00]	1444.00 [676.00, 2555.00]	1416.00 [645.75, 2562.00]	0.54
Candidate Wait time (median [IQR])	7.00 [3.00, 27.00]	6.00 [3.00, 17.00]	5.00 [2.00, 14.00]	<0.001
AIH: autoimmune hepatitis; PSC: primary sclerosing cholangitis; PBC: primary biliary cholangitis; NASH: non-alcoholic steatohepatitis; BMI: body mass index; MELD: model for end-stage liver disease; IQR: interquartile range; SD: standard deviation; ICU: intensive care unit				

Sensitivity analysis examining the effect of era.

The low-risk group was made up of 30 centers across the total timeframe, 8 (27%) of which remained in the low-risk group in each era. There were 16 centers (53%) who changed from the low to middle-risk group in at least one of the three eras and 6 (20%) who changed from low to high during the three eras. In the middle-risk group, there were 49 total centers, of which 17 (37%) remained in this group in each era. There were 19 (39%) who changed from middle to high-risk in at least one era and 13 (27%) who changed from middle to the low-risk group in at least one era. The high-risk group consisted of 29 centers, of which 6 (21%) stayed in the high group during all three eras, 9 (31%) changed from high to low-risk in at least one era, and 14 (48%) changed from high to middle-risk. Despite the fluctuations in group membership during the different eras, the cumulative incidence of waitlist death or delisting in patients initially listed at a MELD ≤35 with decompensation to ≥35 remained highest in the high-risk group, followed by the middle-risk, and lowest in the low-risk group in all eras Figure 5.

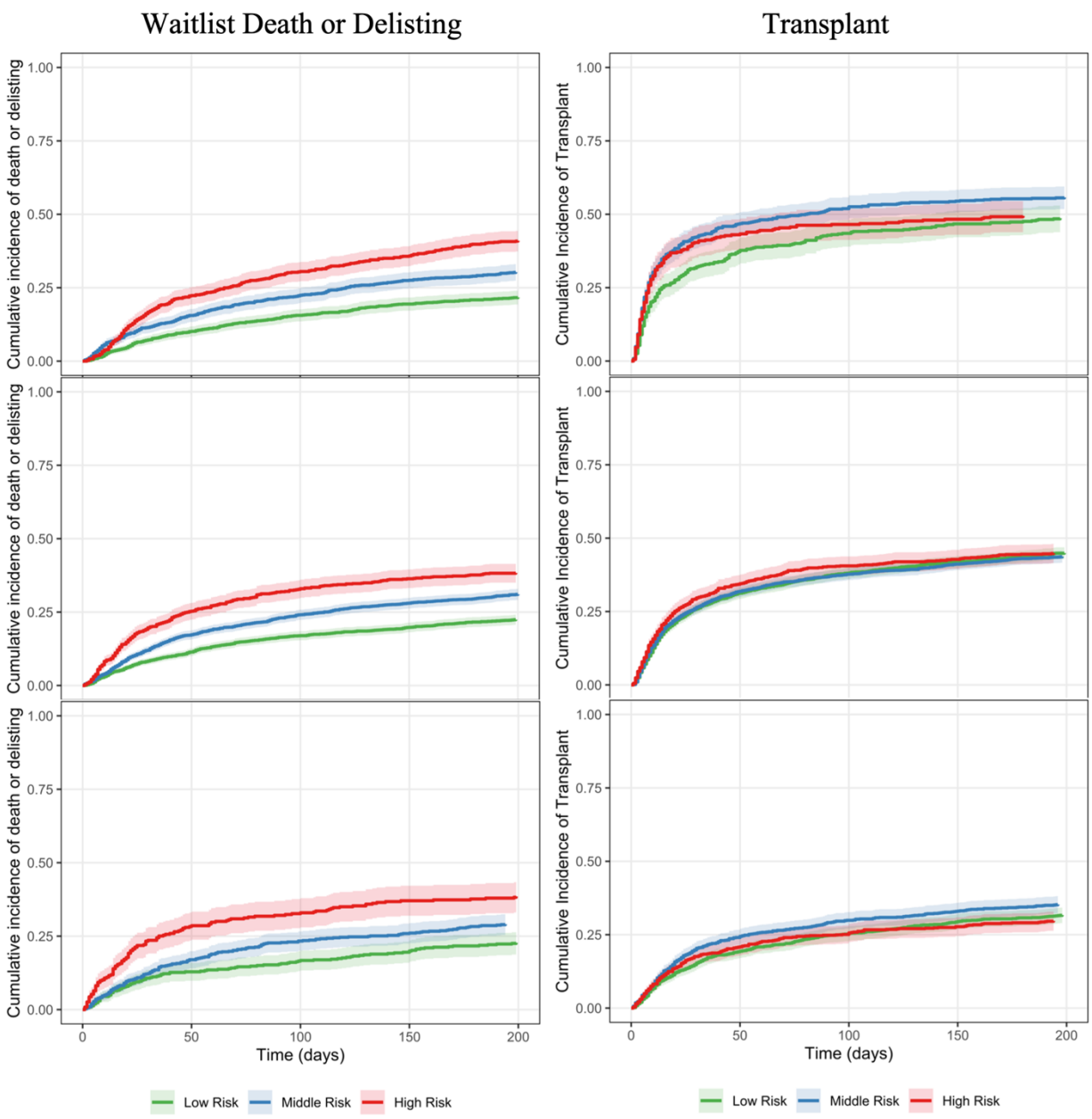


Figure 5. Waitlist mortality and transplant probability by era.

Cumulative incidence and probability of transplant in patients who decompensated to a MELD ≥35 on the waitlist in the low, middle, and high-risk groups in the three different eras, A) 2010-2013, B) 2013-2020, C) 2020-2022.

Analysis of drivers of waitlist mortality differences between center risk-groups.

There were no significant differences in median annual transplant volume, median MELD at transplant, or UNOS regional representation among the three groups. On multivariate Cox proportional hazards modeling of patients who decompensated to a MELD ≥35 within each group, some differences in significant features among the three groups were revealed. Sex was not significantly associated waitlist mortality in the low or high-risk groups, but male sex was protective in the middle-risk group (HR 0.85, $p<0.001$). Additionally, hepatitis C as an etiology of cirrhosis was significantly associated with mortality exclusively in the middle risk group (HR 1.15, $p=0.025$). Non-alcoholic steatohepatitis (NASH) as an etiology of cirrhosis was significantly associated with mortality exclusively in the high-risk group (HR 1.33, $p=0.005$) and receipt of hemodialysis was significantly associated mortality exclusively in the low-risk group (HR 1.07, $p=0.027$). Notably, the vast majority of waitlist mortalities in all three groups were attributed to multi-system organ failure (50%, 39%, and 39%) in the low, middle, and high-risk groups, respectively.

Chapter 1 Discussion

In this multi-center retrospective cohort study, we analyzed the center-level variation in waitlist mortality and probability of receiving LT in patients with a MELD score ≥35. We found center-specific differences after accounting for variables known to affect waitlist mortality (MELD, gender, ventilation status, encephalopathy, need for dialysis, etc.). Compared to low-risk centers, patients listed at high-risk centers showed higher waitlist mortality in MELD ≥35 patients and those listed initially at MELD <35 with subsequent decompensation to MELD ≥35. The likelihood of LT was significantly lower

in the high-risk group compared to the low-risk group, particularly among patients initially listed with a MELD < 35 who later progressed to MELD ≥35. The outcome differences between groups could not be explained by differences in LT volume or UNOS region among the groups. Specifically, the map of centers shows that the risk groups were not associated with traditionally high or low MELD regions with multiple risk groups represented in individual UNOS regions and even individual cities. The disparities in death or delisting among the three risk groups were most pronounced in patients who were initially listed at a MELD <35 with subsequent increase to 35 on the waitlist, with the highest risk group having a 50% waitlist mortality compared to 37% in the low-risk group. Analogously, the likelihood of receiving LT in these patients was much lower in the high-risk group (46%) compared to the low-risk group (55%).

These results suggest that centers in the high-risk group are more selective in listing patients when they are already at MELD ≥35 at listing. This selectiveness is strongly supported by differences in patient acuity characteristics among the three groups. In patients initially listed at a MELD ≥35, candidates at centers in the high-risk group had a significantly lower likelihood of requiring mechanical ventilation, having significant ascites, hepatic encephalopathy, or the need for hemodialysis, which suggests that centers in the high-risk group may be carefully selecting the MELD ≥35 patients they choose to list that have other favorable characteristics compared to high MELD patients listed at low-risk centers. Similarly, while the overall one-year graft-failure free survival after LT in the three groups was not significantly different, there were notable differences in acuity of the LT recipients not captured by MELD. Patients who underwent LT at centers in the low-risk group (compared to the middle and high-

risk groups) tended to be older, have a significantly higher likelihood of being on mechanical ventilation at time of LT, a higher likelihood of having portal vein thrombosis, more likely to be in the ICU at time of transplant, more likely to have significant ascites, significant encephalopathy, and require hemodialysis. The lack of difference in one-year post-transplant survival is thus significant as patients transplanted at centers in the low-risk group were significantly more ill despite all patients having a MELD ≥35. The lack of difference in one-year survival post-transplant among the three groups suggests that centers in the middle and high-risk groups may be overly selective about which MELD ≥35 patients they are willing to transplant, which likely explains the significantly increased waitlist mortality in these patients compared to the low-risk group. Furthermore, despite higher acuity in LT candidates at low-risk centers, the preservation of excellent post-transplant outcomes strongly argues that stringent selectivity displayed by high-risk centers in listing and transplanting high-MELD patients is not necessary and is associated with a significantly greater waitlist mortality for MELD ≥35 patients.

To our knowledge, this is the first comprehensive analysis of center-level variation in the waitlist and post-transplant outcomes for MELD ≥35 patients. One recent study examining center-level variations in the predicted waitlist mortality associated with the MELD score was reported by Ishaque *et al* [13]. This study determined a “MELD correction factor” for each center which represents the predicted waitlist mortality risk associated with a MELD score at a specific center compared to the national average (i.e. correction factor +1 means a MELD score of 18 at that center would be associated with the likelihood of waitlist mortality predicted by a MELD score of 19). These authors found that for the majority of MELD scores, the MELD correction factor was within 1-2

points at every center with the exception of MELD scores from 31-40 where the variance ranged from 2-4. The increased variation in the predicted mortality of the MELD at certain centers at the higher range of MELD scores was not further explored and is analyzed in depth in our analysis, which shows that when grouped according to adjusted risk of waitlist death or delisting, low-risk centers have significantly improved waitlist outcomes compared to the other groups. This analysis may suggest that the prioritization of high-MELD/ high-acuity patients by some centers is beneficial to the LT community in general as some centers are capable of treating these high-acuity patients with improved outcomes.

The factors that lead to a center's categorization into the low, middle, or high-risk group warrants some discussion. The groupings are not associated with significant differences in median annual transplant volume, median MELD at listing, or median MELD at transplant, suggesting that the adjusted risk of waitlist mortality among MELD ≥35 patients is independent of annual center transplant volume and UNOS region/geography. The analysis of factors most predictive of one-year waitlist mortality within each group revealed a few features with outsized importance in one group versus the others (low-risk: dialysis; middle-risk: Sex and Hepatitis C; high-risk: NASH. However, these differences do not paint a clear picture of the true center-characteristics that are driving the group outcome differences. The examination of causes of death in waitlisted candidates in each group revealed that multi-system organ failure accounted for the majority of mortalities in all groups. The overwhelming presence of multi-system organ failure as a cause of death in this population may suggest that these mortality differences are in part reflective of critical illness management expertise in the different

centers, with low-risk centers likely being those with robust critical care systems and possibly dedicated liver ICUs. Additionally, organ availability at each center could affect the outcomes of MELD ≥35 patients as high-risk donor organs are typically not chosen for use in high-MELD/ high acuity transplant recipients, thus centers with fewer high quality donor organ offers could see worse waitlist outcomes for MELD ≥35 patients. The allocation policy changes such as Share 35 and acuity circles should improve organ availability for MELD ≥35 patients, though these policies were enacted at different times during the timeframe examined in this study. To ensure the controlling for any effect of these allocation changes on waitlist mortality, we included the “era” variable which groups patients into three groups based on the times these allocation changes were made. The era variable was included in the final multivariate model to control for these effects and thus our final separation of centers into risk groups should be unbiased by these allocation changes. Additionally, we performed a sensitivity analysis, showing that while some centers did change risk-groups during the different eras, the waitlist mortality differences across the three groups remained constant.

Another potential driver of MELD ≥35 waitlist outcomes is LT center behavior where certain centers prioritize the care for high MELD/ high acuity patients on the waitlist and are comfortable treating them with LT, while other centers may have concerns about utilizing scarce donor organs to transplant high MELD/ high acuity patients that they perceive to have a low likelihood of success post-transplant. Because this study is limited by data availability in the SRTR, we cannot definitively determine steps that high-risk centers could take to address the outcome disparities in MELD ≥35 patients. However, each center has its own procedures for determining if a patient

should be precluded from transplant for being “too sick” and the data presented could prompt centers to re-examine these thresholds given that certain centers may be transplanting patients with similar acuity with appropriate post-transplant outcomes.

This study has several limitations including the retrospective design, which limits our ability to account for all possible confounding variables, preventing determination of any causal relationships. As a registry-based study, we are limited by the information that is collected by the organization responsible for database maintenance.

In conclusion, we demonstrate center-level variation in waitlist outcomes and probability of LT in the highest-acuity candidates (MELD ≥35), particularly for patients with initial listing MELD < 35 who decompensate to higher MELD. These variations may indicate that certain centers are associated with substantially decreased waitlist mortality for the highest acuity patients without corresponding decreases in post-transplant survival.

Chapter 1 References

[1] Freeman Jr RB, Wiesner RH, Harper A, et al. The new liver allocation system: moving toward evidence-based transplantation policy. Liver Transplantation. 2002;8(9):851-858.

[2] Kamath PS, Kim WR. The model for end-stage liver disease (MELD). Hepatology. Mar 2007;45(3):797-805. doi:10.1002/hep.21563

[3] Lee E, Johnston CJ, Oniscu GC. The trials and tribulations of liver allocation. Transplant International. 2020;33(11):1343-1352.

[4] Pillai A, Couri T, Charlton M. Liver allocation policies in the USA: past, present, and the future. Digestive diseases and sciences. 2019;64:985-992.

[5] Jung G, Encke J, Schmidt J, Rahmel A. Model for end-stage liver disease: new basis of allocation for liver transplantations. Der Chirurg. 2008;79:157-163.

[6] Institute of Medicine Committee on Organ P, Transplantation P. Organ Procurement and Transplantation: Assessing Current Policies and the Potential Impact of the DHHS Final Rule. National Academies Press (US)

Copyright 1999 by the National Academy of Sciences. All rights reserved.; 1999.

[7] Freeman RB, Jr., Edwards EB. Liver transplant waiting time does not correlate with waiting list mortality: implications for liver allocation policy. Liver Transpl. Sep 2000;6(5):543-52. doi:10.1053/jlts.2000.9744

[8] 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. Apr 2000;31(4):864-71. doi:10.1053/he.2000.5852

[9] Biggins SW, Kim WR, Terrault NA, et al. Evidence-based incorporation of serum sodium concentration into MELD. Gastroenterology. May 2006;130(6):1652-60. doi:10.1053/j.gastro.2006.02.010

[10] Kim WR, Mannalithara A, Heimbach JK, et al. MELD 3.0: The Model for End-Stage Liver Disease Updated for the Modern Era. Gastroenterology. Dec 2021;161(6):1887-1895.e4. doi:10.1053/j.gastro.2021.08.050

[11] Locke JE, Shelton BA, Olthoff KM, et al. Quantifying sex-based disparities in liver allocation. JAMA surgery. 2020;155(7):e201129-e201129.

[12] Singh SP, Maiwall R. MELD 3.0: A better score for liver allocation? Liver Transplantation. 2023;29(10):1017-1018.

[13] Ishaque T, Kernodle AB, Motter JD, et al. MELD is MELD is MELD? Transplant center–level variation in waitlist mortality for candidates with the same biological MELD. American Journal of Transplantation. 2021;21(10):3305-3311.

[14] Hirose R. State-Based Liver Allocation: Is It a Viable Alternative? livEr transplantation. 2019;25(4):535-537.

[15] Procurement O. Organ Procurement and Transplantation Network (OPTN) policies. 2021. Online: https://optn transplant hrsa gov/media/1200/optn_policies pdf, Stand. 2021;9

[16] Giorgakis E, Ivanics T, Wallace D, et al. Acuity circles allocation policy impact on waitlist mortality and donation after circulatory death liver transplantation: A nationwide retrospective analysis. Health Sci Rep. Feb 2023;6(2):e1066. doi:10.1002/hsr2.1066

[17] Goldberg DS, Levine M, Karp S, Gilroy R, Abt PL. Share 35 changes in center-level liver acceptance practices. Liver Transpl. May 2017;23(5):604-613. doi:10.1002/lt.24749

[18] Nicolas CT, Nyberg SL, Heimbach JK, et al. Liver transplantation after share 35: Impact on pretransplant and posttransplant costs and mortality. Liver Transpl. Jan 2017;23(1):11-18. doi:10.1002/lt.24641

[19] Wey A, Noreen S, Gentry S, et al. The effect of acuity circles on deceased donor transplant and offer rates across model for end-stage liver disease scores and exception statuses. Liver Transplantation. 2022;28(3):363-375.

[20] Burton AM, Goldberg DS. Center-level and region-level variations in liver transplantation practices following acuity circles policy change. American journal of transplantation. 2022;22(11):2668-2674.

[21] Baliga P, Sade RM. New Liver Allocation Policy: Flawed Moral and Empirical Foundations. J Law Med Ethics. Jun 2019;47(2):320-322. doi:10.1177/1073110519857288

[22] Lynch RJ, Ye F, Sheng Q, Zhao Z, Karp SJ. State-Based Liver Distribution: Broad Sharing With Less Harm to Vulnerable and Underserved Communities Compared With Concentric Circles. Liver Transpl. Apr 2019;25(4):588-597. doi:10.1002/lt.25425

[23] Ross K, Patzer RE, Goldberg DS, Lynch RJ. Sociodemographic Determinants of Waitlist and Posttransplant Survival Among End-Stage Liver Disease Patients. Am J Transplant. Nov 2017;17(11):2879-2889. doi:10.1111/ajt.14421

[24] Lynch RJ, Magliocca JF, Hundley JC, Karp SJ. Moving past "think local, act global": A perspective on geographic disparity. Am J Transplant. Jul 2019;19(7):1907-1911. doi:10.1111/ajt.15079

[25] Petrowsky H, Rana A, Kaldas FM, et al. Liver transplantation in highest acuity recipients: identifying factors to avoid futility. Ann Surg. Jun 2014;259(6):1186-94. doi:10.1097/sla.0000000000000265

[26] Pankratz VS, De Andrade M, Therneau TM. Random-effects Cox proportional hazards model: General variance components methods for time-to-event data. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society. 2005;28(2):97-109.

[27] Vaida F, Xu R. Proportional hazards model with random effects. Statistics in medicine. 2000;19(24):3309-3324.

Chapter 2 Liver transplantation at the extreme: analysis of outcomes and predictors of futility in acute-on-chronic liver failure grade 3 patients on maximal life support

Chapter 2 Abstract

Background & Aims: While patients with grade 3 acute-on-chronic liver failure (ACLF-3), have high mortality, there is a subset requiring maximal life support that have not been examined. We report post liver transplant (LT) outcomes and identify predictors of futility in extreme ACLF-3 (eACLF-3) recipients, all requiring intubation, vasopressors, and dialysis at LT.

Methods: We identified adult eACLF-3 LT recipients between January 2010 and December 2021. Pre-LT characteristics were compared among eACLF-3 recipients with and without the futile outcome (one-year or in-hospital mortality), and predictors of futility were identified. Findings were validated using the Multi-Organ Dysfunction and Evaluation for Liver Transplantation (MODEL) Consortium dataset.

Results: Of 1,608 adult LT recipients at our center, 177 (11%) were eACLF-3. Thirty-five (20%) eACLF-3 recipients experienced futility. Predictors of futility included previous upper abdominal surgery (OR=4.78, P=0.003), sepsis within 14 days of LT (OR= 6.68, P= 0.015), intubation due to pneumonia or acute respiratory distress syndrome (OR= 3.843, P= 0.008), increasing model for end-stage liver disease (MELD-Na) score (OR= 3.07, P=0.029), worsening PaO2:FiO2 ratio (OR= 4.26, P=0.016), pre-LT base deficit (OR=1.33, P=0.003), and donor risk index (OR=2.542, P=0.008). When compared to the MODEL Consortium eACLF-3 patients, our cohort had improved three-, six-, and twelve-month survivals of 89%, 85%, and 80%, compared to 83%, 79%, and 70% (p=0.048); however overall 5-year survival was similar (63% vs 59%, p=0.14).

Conclusions: We report the largest single-center experience of LT for eACLF-3 patients, demonstrating acceptable short and long-term survival despite high acuity, validated in the external MODEL Consortium dataset. We propose novel risk factors for futility which may aid in appropriate selection of those ACLF-3 patients requiring maximal life support prior to LT.

Chapter 2 Introduction

Acute on chronic liver failure (ACLF) is a syndrome characterized by acute hepatic decompensation, organ failure, and high short-term mortality [1-5]. Patients with cirrhosis who develop an acute decompensation are at risk for the development of ACLF with subsequent rapid organ dysfunction and an elevated mortality risk [6-8]. The European CANONIC trial defined ACLF in three grades according to the number of associated organ failures [9, 10]. ACLF can be stratified using the CLIF-Organ failure score, which reflects the number and degree of organ failures and the CLIF-C ACLF-score which predicts outcomes in patients with ACLF and may better assess risk of waitlist mortality in ACLF patients than the model for end-stage liver disease (MELD) score [11]. ACLF grade 3 (ACLF-3) is the most severe form of disease where patients have at least three organ failures and is associated with 28-day waitlist mortality approaching 80% [7, 11-14]. In an analysis of the United Network for Organ Sharing (UNOS) database, ACLF-3 recipients had worse waitlist and post-transplant survival compared to status 1A recipients, traditionally considered the be the highest acuity liver transplant (LT) population [13].

Due to the extremely high waitlist mortality of ACLF-3 patients, there is considerable interest in offering these patients early LT to afford them the greatest chance of survival

[15-22]. However, this is balanced against the concerns about poor post-transplant outcomes and stewardship of scarce donor livers. There have thus been considerable efforts to predict futility following transplant in ACLF-3 recipients and to identify risk factors for futile outcomes [7, 23-26]. In one large multicenter ACLF-3 study, five European centers contributed data on 152 ACLF-3 LT recipients and created the "transplantation for ACLF-3 model" (TAM-score) to predict one year post-transplant mortality [25]. This study identified risk factors for one-year post transplant mortality including advanced age, lactate, respiratory failure, and increasing WBC count and were able to predict one year mortality with an area under the receiver operating characteristic curve (AUROC) of 0.88. Another recent study was reported by Hernaez et al where 15 US centers contributed data on 735 ACLF patients, 521 of which were considered "severe," including ACLF 2 and 3 patients [26]. This data was used to create the novel Sundaram ACLF-LT-Mortality (SALT-M) score, which predicts 1-year post-transplant mortality in this group of severe ACLF patients with a reported AUROC of 0.80.

While the TAM and SALT-M scores represent significant contributions to the ability to discriminate patients with severe ACLF who may be at risk for one-year post-transplant mortality, there is a spectrum of disease even within ACLF-3 that has not been specifically explored. The European definition of ACLF, based on the CLIF-C organ failure scoring system, assesses function of 6 organ systems (liver, kidney, brain, coagulation, circulation, and respiration) [11]. Based on this system, patients can be considered to have ACLF-3 with evidence of failure in any three of these organ systems. It is likely that a patient with acutely decompensated liver function, disordered

coagulation, and hepatic encephalopathy would have a significantly different risk profile compared to a patient with respiratory failure requiring mechanical ventilation, circulatory failure requiring vasopressors, and renal failure requiring dialysis [27]. In the original TAM score study, 73% of the cohort was mechanically ventilated, 70% on pressors, and 65% receiving dialysis at time of transplant. In the cohort used to create the SALT-M score, 48% were receiving mechanical ventilation, 52% were on vasopressors, and 84% were receiving dialysis at time of LT. We hypothesized that ACLF-3 patients requiring concomitant mechanical ventilation, vasopressors, and dialysis – all forms of critical life support – represent a hyper-acute subgroup of extreme ACLF-3 (eACLF-3) transplant recipients who have not specifically been analyzed in prior studies. We sought to characterize our single-center experience in transplanting these patients, with a specific goal of identifying risk factors for futility after transplant. Additionally, we sought to externally validate our findings using data from the Multi-Organ Dysfunction and Evaluation for Liver Transplantation (MODEL) Consortium, which was used to develop the SALT-M score [26].

Chapter 2 Patients & Methods

Patient Cohorts

We identified adult (age ≥ 18), first-time, non-status 1A, liver-only or liver-kidney transplant recipients at UCLA medical center from January 1, 2010 to December 31, 2021 through retrospective chart review. After exclusion, we identified patients with decompensation with one or more additional organ failures and stratified them based on ACLF score (1, 2, or 3). Within the ACLF-3 patients, we specifically identified recipients with acute decompensation of cirrhosis with multiple organ failures resulting in the need

for mechanical ventilation, vasopressors, and renal replacement therapy (eACLF-3) at the time of transplantation. This study was approved by the UCLA institutional review board (IRB #11-001932).

Data collection for the MODEL Consortium has been previously described [26], approval from contributing authors was obtained and data had been collected under IRB approvals at each participating institution. From this cohort, we identified eACLF-3 LT recipients requiring vasopressors, on dialysis, and mechanically ventilated at time of LT.

Data Collection and Definition of eACLF-3

Patients meeting the criteria for ACLF (grades 1-3) were identified and categorized at the time of LT based on the European Association for the Study of the Liver-Chronic Liver Failure (EASL-CLIF) criteria. The severe ACLF-3 patients were then further stratified into eACLF-3, which included only those who were mechanically ventilated, receiving vasopressors, and receiving renal replacement therapy at the time of LT. If patients were extubated, weaned off pressors, or weaned off dialysis prior to LT, they were not included as part of this eACLF-3 cohort. For the eACLF-3 cohort, pre-transplant clinical characteristics were collected at time of ICU admission and at time of transplant. Several variables were calculated using the change in value from ICU admission to transplant, including change in MELD score and change in the ratio of the partial pressure of oxygen in arterial blood to the fraction of inspiratory oxygen concentration (PaO2:FiO2 ratio). The indication for mechanical ventilation was recorded as either strictly for airway protection or due to pulmonary pathology including pneumonia and acute respiratory distress syndrome (ARDS). The indication for ventilation was determined through manual chart review with physician review of chest

X-rays, X-ray radiology reports, and any relevant microbiological data including sputum and blood cultures. Sepsis was considered if patients had clinical suspicion for sepsis with an identified (culture proven) infectious source (pneumonia, spontaneous bacterial peritonitis, urinary tract infection, or other source). We included a sepsis duration variable that signified if a patient was diagnosed with sepsis greater than two weeks prior to transplant, within two weeks, or not diagnosed with sepsis.

Outcomes

The primary outcome of interest was futility following transplant, defined as one-year post-transplant all-cause mortality or in-hospital mortality. We included in-hospital mortality in the definition to capture patients who receive prolonged intensive care and life support following transplant, ultimately experiencing mortality in the hospital, which should be considered a futile outcome even if it did not occur within one-year. Extensive secondary outcomes were evaluated and reported, including five-year overall survival, post-LT ICU and total length of hospitalization, total time of intubation, number of reoperations, planned versus unplanned reoperations, time to reoperation, postoperative respiratory failure (requiring re-intubation), neurological failure (ischemic stroke or cerebral hemorrhage), cardiac failure (heart failure leading to low output syndrome), sepsis, portal vein thrombosis, hepatic artery thrombosis, primary graft non-function, discharge to skilled nursing facility, discharge to long-term acute care hospital,

readmission within one month after discharge, dialysis dependence following transplant, tracheostomy-dependence, graft failure requiring re-transplant, and cause of death.

Statistical Analysis

Analyses were conducted using SAS and R version 4.1.0 (R Foundation for Statistical Computing, Vienna, Austria). Any p values less than 0.05 were considered significant.

Baseline characteristics at time of ICU admission and at LT were compared between the futile and non-futile groups. Characteristics were also compared between eACLF-3 and non-eACLF-3 patients at our center as well as eACLF-3 patients at our center and in the MODEL Consortium. Group medians were displayed with interquartile ranges and compared using the Mann-Whitney U test. Categorical variables were displayed as number (%) and compared using the chi-squared test. We compared survivals between the futile and non-futile eACLF-3 patients as well as the eACLF-3 patients at our center and in the MODEL Consortium using the Kaplan-Meier method. While our focus was on eACLF-3 patients, we also compared overall survival between ACLF 1, 2, and 3 patients using the Kaplan-Meier method. Survival differences were assessed for statistical significance using the log-rank test in all comparisons.

Univariate logistic regression analysis with futility modeled as the outcome was conducted for all collected variables. For features that were calculated based on change in values from ICU admission to LT, we rigorously tested different binarized threshold value cut-offs and determined quantities of change most predictive of futility empirically. Features for the final multivariate logistic regression model were selected using a combination of forward stepwise selection based on the Akaike information criterion (AIC) and clinical relevance. Missing values for variables in the final multivariate model

were imputed using multiple imputation by chained equations using the “MICE” R package [28].

Chapter 2 Results

Of 2,024 total LTs performed at our center, 1,608 met study inclusion criteria, of which 182 (11%) were ACLF-1, 250 (16%) were ACLF-2, and 581 (36%) were for ACLF-3 (Figure 6).

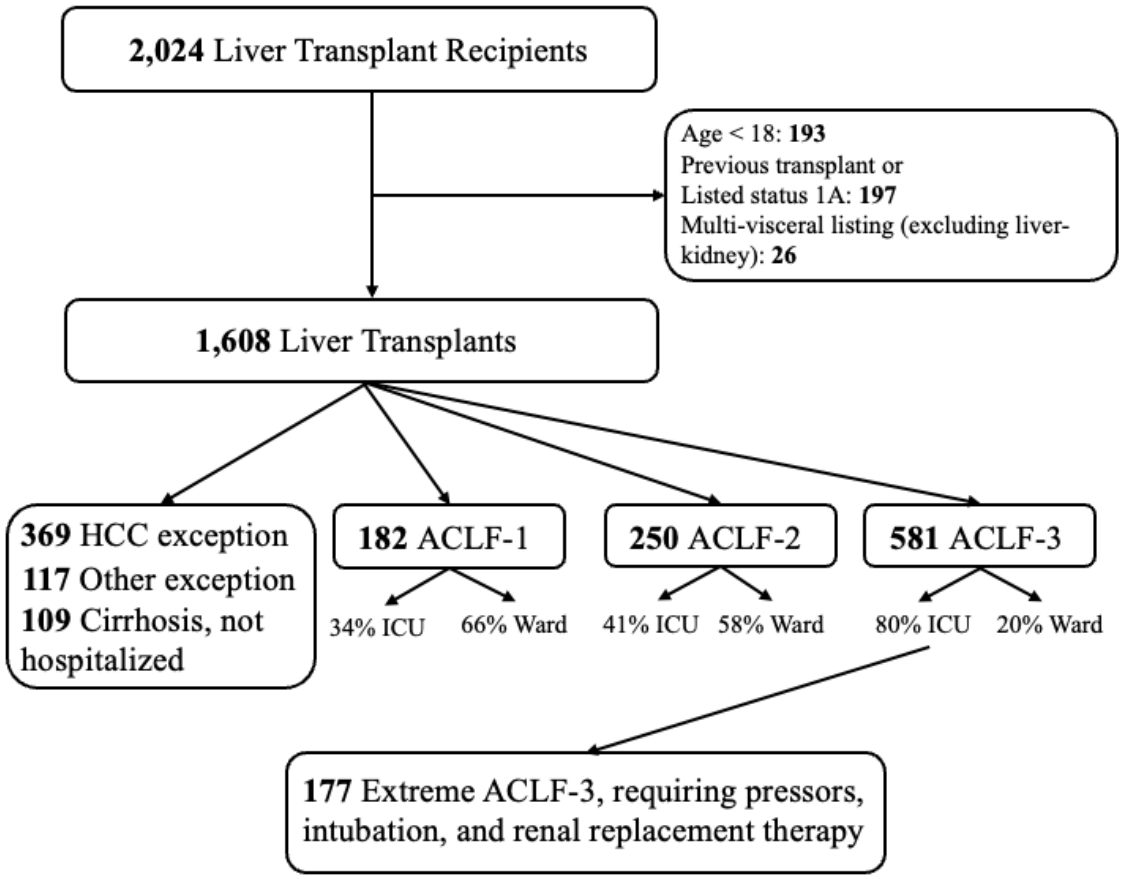


Figure 6. Patient cohort selection.

The remaining 595 (37%) patients were those with hepatocellular carcinoma (369, 23%), non-HCC related MELD exception points (117, 7%), or decompensated cirrhosis not requiring hospitalization prior to transplant (109, 7%). Of the 581 ACLF-3

patients, 177 (30%) represented eACLF-3 patients all requiring mechanical ventilation, vasopressors, and receiving renal replacement therapy at time of LT. Baseline demographic comparisons between non-extreme ACLF-3 and eACLF-3 patients are shown in Table 6**.**

Table 7. Baseline demographic comparisons between extreme and non-extreme ACLF-3 patients.

Variable	Total (n=581)	Non-extreme (n=404)	Extreme (n=177)	P
Demographic				
Age (y), median (IQR)	57.0 (50.0-63.0)	57.0 (50.0-62.3)	58.0 (49.6-63.9)	0.86
Gender = M (%)	325 (56)	236 (58)	89 (50)	0.06
BMI (kg/m^2), median (IQR)	27.4 (23.7-31.8)	27.7 (23.8-31.9)	26.6 (23.6-31.7)	0.19
Primary Diagnosis (%)				0.24
Alcohol	220 (38)	155 (38)	65 (37)	
MAFLD	115 (20)	70 (17)	45 (25)	
Hepatitis C	89 (15)	55 (14)	34 (19)	
Autoimmune/ cholestasis	54 (9)	33 (8)	21 (12)	
Other	103 (18)	91 (23)	12 (7)	
Previous abdominal surgery	200 (34)	144 (36)	56 (32)	0.49
Diabetes mellitus	163 (28)	109 (27)	54 (31)	0.16
Pre-LT Acuity				
MELD-Na at LT, median (IQR)	40.0 (37.0-40.0)	40.0 (38.0-40.0)	40.0 (37.0-40.0)	0.35
Mechanical Ventilation at LT, n (%)	313 (54)	136 (38)	177 (100)	<0.001
Dialysis at LT, n (%)	462 (80)	285 (71)	177 (100)	<0.001

Pre-LT Hospitalization Characteristics				
Listing date to LT (d), median (IQR)	17.0 (6.0-127.0)	16.0 (6.0-124.8)	18.0 (6.0-157.0)	0.70
Duration of preoperative admission (d), median (IQR)	17.0 (10.0-28.0)	16.0 (9.0-25.0)	21.0 (14.0-31.0)	<0.001
ICU admission at LT, n (%)	462 (80)	285 (71)	177 (100)	<0.001
Donor and Operative Characteristics				
Donor Risk Index, median (IQR)	1.3 (1.1-1.6)	1.3 (1.1-1.6)	1.3 (1.1-1.6)	0.26
DCD donor	19 (3)	15 (4)	4 (2)	0.77
Cold ischemic time (h), median (IQR)	7.0 (6.0-8.3)	7.2 (6.0-8.4)	7.0 (6.0-8.0)	0.57

ACLF, acute-on-chronic liver failure; IQR, interquartile range; BMI, body mass index; MAFLD, metabolic-associated fatty liver disease; LT, liver transplant; MELD, Model for End-stage Liver Disease; ICU, intensive care unit; DCD, donation after circulatory death.

Annual transplant center volume, number (%) of transplants for ACLF-3, and transplant center and donor service area median MELD at transplant during successive years of the study period are shown in Table 7**.** There were 35 eACLF-3 patients who experienced the futile outcome (20%), with baseline characteristics shown in Table 8. Compared to non-futile eACLF-3 patients, futile eACLF-3 were significantly more likely to have hepatitis C (43% vs 21%, p = 0.012), previous upper abdominal surgery (13% vs 11%, p=0.042), had a greater base deficit at time of transplant (median, IQR: 0.0, -2.0-1.0 vs 1.0, -1.0-2.0, p =0.004), were more likely to be intubated at LT for pneumonia or ARDS (77% vs 54%, p=0.011), and were more likely to have sepsis within 14 days of transplant (77% vs 54%, p=0.040). The two groups did not significantly differ in MELD or CLIF-C ACLF scores at ICU admission or at transplant, and had similar waitlist times, duration of pre-transplant total and ICU admission, and similar proportion of outside

hospital admission and duration of admission prior to transferring care to our center. Regarding donor and operative characteristics, futile eACLF-3 patients received liver allografts with greater median donor risk index (median, IQR: 1.4, 1.2-2.3 vs 1.3, 1.1-1.6, p = 0.011), received more packed red blood cell transfusions during transplant (median, IQR: 31.0 (18.0-50.0) vs 23.0 (15.0-36.0), p = 0.049), and were more likely to have required venovenous bypass during transplant (86% vs 65%, p=0.017) compared to the non-futile eACLF-3 cohort.

Table 8. Annual center transplant volume and ACLF-3 transplant volume over time with median MELD at transplant in our center and our center's DSA.

Year	Number of liver transplants	Number (%) of ACLF-3 transplants	Number (%) of ACLF-3 transplants for hepatitis C	Median MELD at transplant in southern California donor service area	Median MELD at transplant at our center
2010	164	57 (35)	10 (18)	28	31
2011	180	66 (37)	19 (29)	28	33
2012	144	40 (28)	10 (25)	27	27
2013	159	51 (32)	10 (20)	30	30
2014	155	55 (35)	9 (16)	30	33
2015	139	56 (40)	8 (14)	31	34
2016	151	62 (41)	4 (6)	34	37
2017	135	40 (30)	4 (10)	30	33
2018	153	49 (32)	0 (0)	30	35
2019	157	42 (27)	0 (0)	23	33
2020	158	37 (23)	4 (11)	28	32
2021	136	26 (19)	1 (4)	29	33
ACLF, acute-on-chronic liver failure; MELD, model for end-stage liver disease; DSA, donor service area.					

Table 9. Baseline demographic comparisons between ACLF Grade 3 patients with and without the futile outcome.

Variable	Total (n=177)	Non-futile (n=142)	Futile (n=35)	P
Demographic				
Age (y), median (IQR)	58.0 (49.6-64.0)	57.6 (51.9-63.3)	58.3 (49.1-64.1)	0.70
Gender = M (%)	89 (50)	72 (51)	17 (49)	0.82
BMI (kg/m^2), median (IQR)	26.6 (23.6-31.7)	26.5 (23.4-31.8)	27.8 (24.1-31.4)	0.79
Primary Diagnosis (%)				0.012
Alcohol	65 (37)	54 (38)	11 (31)	
MAFLD	45 (25)	38 (27)	7 (20)	
Hepatitis C	34 (19)	20 (14)	14 (40)	
Autoimmune/ cholestasis	21 (12)	19 (14)	2 (6)	
Other	12 (7)	10 (7)	2 (6)	
Prior abdominal surgery	70 (40)	50 (35)	20 (57)	0.018
Type of abdominal surgery				0.042
Previous upper abdominal surgery	23 (13)	15 (11)	8 (23)	
Previous lower abdominal surgery	33 (19)	26 (19)	7 (20)	
Previous upper and lower abdominal surgery	13 (7)	8 (6)	5 (14)	

Diabetes mellitus	54 (31)	46 (32)	8 (23)	0.27
Hypertension	54 (31)	46 (32)	8 (23)	0.27
Hyperlipidemia	32 (18)	29 (20)	3 (9)	0.10
Metabolic syndrome	39 (22)	36 (25)	3 (9)	0.032
Coronary artery disease	26 (15)	20 (14)	6 (17)	0.65
Cardiac arrythmia	38 (21)	27 (19)	11 (31)	0.11
Pre-LT Acuity				
MELD-Na on ICU admission, median (IQR)	38.4 (35.4-41.5)	38.9 (35.9-41.5)	37.8 (33.6-41.9)	0.26
CLIF-C ACLF on ICU admission, median (IQR)	58.5 (54.0-66.8)	58.4 (53.6-64.1)	58.2 (52.4-65.9)	0.79
MELD-Na at LT, median (IQR)	41.0 (38.5-42.8)	41.0 (38.3-43.0)	40.7 (39.1-42.8)	0.90
CLIF-C ACLF at LT, median (IQR)	68.0 (62.8-73.0)	67.8 (62.8-71.9)	68.7 (65.7-72.5)	0.34
Base excess at LT, median (IQR)	0.0 (-1.0-2.0)	1.0 (-1.0-2.0)	0.0 (-2.0-1.0)	0.004
Reason for Mechanical Ventilation, n (%)				0.011
Lung pathology (Pneumonia/ARDS)	103 (58)	76 (54)	27 (77)	
Hepatic Encephalopathy/ airway protection	74 (42)	66 (47)	8 (23)	
Pre-LT Sepsis Characteristics				
Sepsis within one month before LT, n (%)	105 (59)	84 (60)	21 (60)	0.96

Interval from sepsis to LT, n (%)				0.040
No infection	39 (22)	35 (25)	4 (11)	
≤ 14 days	103 (58)	76 (54)	27 (77)	
> 14 days	35 (20)	31 (22)	4 (11)	
Infection with MRDO, n (%)	17 (10)	12 (9)	5 (14)	0.18
Pre-LT Hospitalization Characteristics				
Listing date to LT (d), median (IQR)	18.0 (6.0-157.0)	18.0 (6.0-146)	31.0 (7.0-228)	0.54
Duration of preoperative admission (d), median (IQR)	21.0 (14.0-31.0)	21.0 (15.0-32.0)	17.0 (11.0-29.0)	0.21
Duration of preoperative ICU admission (d), median (IQR)	17.0 (9.0-23.0)	17.0 (9.0-23.0)	13.0 (9.0-24.0)	0.74
ICU admission prior to listing, n (%)	91 (51)	74 (52)	17 (49)	0.85
OSH admission, n (%)	131 (74)	108 (76)	23 (66)	0.21
Duration of OSH admission (d), median (IQR)	4.0 (0.0-11.0)	5.0 (0.0-12.0)	0.0 (0.0-5.0)	0.07
Donor and Operative Characteristics				
Donor risk index, median (IQR)	1.3 (1.1-1.6)	1.3 (1.1-1.6)	1.4 (1.2-2.3)	0.011
DCD donor	4 (2)	3 (2)	1 (9)	1.00
Cold ischemic time (h), median (IQR)	7.0 (6.0-8.0)	7.2 (6.2-8.3)	7.2 (6.0-8.2)	0.68
PRBC 71ransfusión (u), median (IQR)	23.0 (16.0-36.0)	23.0 (15.0-36.0)	31.0 (18.0-50.0)	0.049

Venovenous bypass, n (%)	122 (69)	92 (65)	30 (86)	0.017
Worst base excess, median (IQR)	-8.7 (-11.9- -6.5)	-8.3 (-11.0- -6.0)	-9.4 (-13.1- -7.0)	0.033
Base excess at end of LT, median (IQR)	-2.0 (-4.0-1.1)	-1.3 (-3.6-1.5)	-3.6 (-5.4-0.0)	0.013

ACLF, acute-on-chronic liver failure; IQR, interquartile range; BMI, body mass index; MAFLD, metabolic-associated fatty liver disease; LT, liver transplant; MELD, Model for End-stage Liver Disease; CLIF-C, Chronic Liver Failure Consortium; ICU, intensive care unit; ARDS, acute respiratory distress syndrome; MRDO, multi-drug resistant organism; OSH, outside hospital; DCD, donation after circulatory death; PRBC, packed red blood cells.

Categorical variables are displayed as number (%) and differences were assessed using the Chi-squared test (significant when $p<0.05$). Continuous variables are displayed as median (IQR) and differences were assessed using the Mann-Whitney U test (significant when $p<0.05$).

Overall survival was decreased in ACLF-3 patients compared to ACLF-1 and ACLF-2 (though not statistically significantly), with survivals of 87%, 90%, and 83% at one year, 74%, 81%, and 76% at three years, and 70%, 74%, and 69% at five years for ACLF-1, ACLF-2, and ACLF-3 patients, respectively (Figure 7A**,** overall p=0.19). In a sub-analysis of ACLF-3 patients, eACFL-3 patients had lower one-, three, and five-year survival compared to non-eACLF-3 (81%, 73%, and 63% vs 84%, 77%, and 71%, respectively; Figure 7B, p=0.013).

Details of the causes of death and comparisons of secondary outcomes between the eACLF-3 non-futile and futile groups are shown in Table 9. Patients in the futile group had a longer median time of post-LT intubation (45.0 days, IQR: [7.0-94.0] vs 8.0 days, IQR: [3.0-27.5], p=0.003), and a larger proportion of patients with two or greater unplanned reoperations (20% vs 11%, p=0.025). Additionally, patients in the futile group were more likely to have had postoperative respiratory, neurological, and cardiac failure (97%, 60%, 69% vs 59%, 9%, 10%, p<0.001 for all comparisons). In the futile group, the

most common cause of death within one year was sepsis with multisystem organ failure (20%), followed by hepatic artery thrombosis (11%).

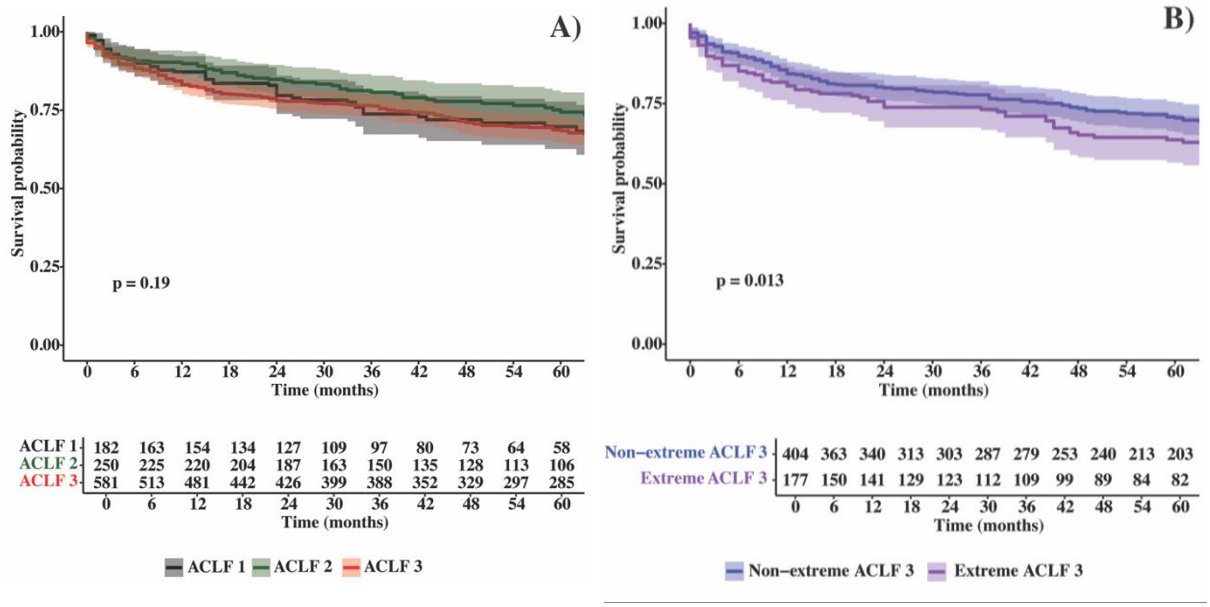


Figure 7. Post-transplant survival across ACLF groups.

A) Survival comparisons between ACLF 1, 2, and 3 patients. B) Survival comparison between non-extreme ACLF-3 and extreme ACLF-3 patients.

Table 10. Secondary outcomes in extreme ACLF-3 patients, comparing non-futile and futile groups and specific causes of death in the futile group.

Variable	Total (n=177)	Non-futile (n=142)	Futile (n=35)	P
Post-LT ICU length of stay (d), median (IQR)	29.0 (14.0-54.0)	26.0 (15.0-49.0)	44.0 (10.0-73.0)	0.41
Post-LT length of stay (d), median (IQR)	53.0 (32.0-78.5)	53.0 (33.0-80.0)	56.0 (14.0-80.0)	0.65

Time of intubation post-LT	9.0 (3.0-44.0)	8.0 (3.0-27.5)	45.0 (7.0-94.0)	0.003
Number of reoperations				0.071
0	76 (43)	64 (45)	12 (34)	
1	61 (35)	51 (36)	10 (29)	
≥2	40 (23)	27 (19)	13 (37)	
Unplanned reoperations				0.025
0	120 (68)	103 (73)	17 (49)	
1	34 (19)	23 (16)	11 (31.)	
≥2	23 (13)	16 (11)	7 (20)	
Time to first reoperation		8.0 (4.0-16.0)	5.0 (2.0-12.0)	0.25
Postoperative respiratory failure	118 (67)	84 (59)	34 (97)	<0.001
Postoperative neurological failure	33 (19)	12 (9)	21 (60)	<0.001
Postoperative cardiac failure	38 (22)	14 (10)	24 (69)	<0.001
Postoperative sepsis	95 (54)	76 (54)	19 (54)	0.013
Portal vein thrombosis	2 (1)	1 (1)	1 (3)	0.28
Hepatic artery thrombosis	9 (5)	5 (4)	4 (11)	0.057
Primary graft non-function	6 (3)	3 (2)	3 (9)	0.171

Discharge to SNF, n (%)	10 (6)	8 (6)	2 (6)	<0.001
Discharge to LTACH	11 (6)	11 (8)	0 (0)	<0.001
Post-LT readmission within one month, n (%)	43 (24)	39 (28)	4 (11)	0.046
Post-LT dialysis dependence, n (%)	22 (12)	16 (11)	6 (17)	0.35
Post-LT tracheostomy, n (%)	47 (27)	34 (24)	13 (37)	0.11
Re-LT, n (%)	10 (6)	8 (6)	2 (6)	0.57
Discharged following index LT	-	-	9 (26)	-
Cause of death				
Cerebral Hemorrhage	-	-	3 (9)	-
HAT	-	-	4 (11)	-
MSOF – no sepsis	-	-	2 (6)	-
PE	-	-	1 (3)	-
Pneumonia with sepsis, MSOF	-	-	3 (9)	-
PNF	-	-	3 (9)	-
Delayed Graft Function with MSOF			3 (9)	
Sepsis with MSOF	-	-	7 (20)	-
Other			6 (17)	

Unknown	-	-	3 (9)	-

ACLF, acute-on-chronic liver failure; IQR, interquartile range; LT, liver transplant; ICU, intensive care unit; SNF, skilled nursing facility; LTACH, long-term acute care hospital; HAT, hepatic artery thrombosis; MSOF, mutli-system organ failure; PE, pulmonary embolism; PNF, primary non-function.

Univariate modeling results evaluating all candidate predictors of the futile outcome are detailed in Table 10 (categorical variables) and Table 11 (continuous variables). Tested thresholds for continuous variables converted to categorical predictors are shown in Table 12. The results of our final multivariate logistic regression model are shown in Table 13. Independent predictors of post-LT futility included previous upper abdominal surgery (OR=4.78; 95% CI, 1.71-13.30, P=0.003), sepsis within 14 days before LT (OR= 6.68; 95% CI, 1.45-30.79, P= 0.015), requiring ventilatory support due to pneumonia or ARDS (OR= 3.843; 95% CI, 1.43-10.35, P= 0.008), increasing MELD-Na score $\geq$ 6% per week (OR= 3.07; 95% CI, 1.12-8.40, P=0.029), worsening PaO2:FiO2 $\geq$ 12% per week (OR= 4.26; 95% CI, 13.81-1.31, P=0.016), worst pre-LT base deficit (OR=1.33 per unit decrease; 95% CI, 1.10-1.61; P=0.003), and higher donor risk index (OR=2.542 per unit increase; 95% CI, 1.28-5.04, P=0.008). When applied to the training data of 177 extreme ACLF-3 patients, the final model resulted in an AUROC of 0.83, sensitivity of 0.63, and specificity of 0.90.

Table 11. Univariate odds ratios of categorical variables.

Variable	OR	Lower 95% CI	Upper 95% CI	P
	.	.	.	.
HCV	6.649	1.330	33.236	0.021
NASH/Cryptogenic/HBV	1.750	0.331	9.250	0.51
EtOH	1.935	0.393	9.534	0.42
Cholestatic/AIH	1.000	.	.	.
Female	1.089	0.520	2.283	0.82

Male	1.000	.	.	.
Previous Lower Abdominal Surgery	1.633	0.602	4.429	0.34
Previous Upper Abdominal Surgery	3.429	1.433	8.203	0.006
No Prior Abdominal Surgery	1.000	.	.	.
Diabetes	0.618	0.261	1.467	0.28
Hyperlipidemia	0.365	0.104	1.277	0.12
Metabolic Syndrome	0.276	0.080	0.956	0.042
Coronary Artery Syndrome	1.262	0.465	3.424	0.65
Sepsis within 1 month	1.018	0.478	2.166	0.96
2+ Sites of infection	2.386	0.708	8.047	0.16
1 Site of infection	2.639	0.832	8.373	0.10
No sights of infection	1.000	.	.	.
Multi-drug resistant infection	1.806	0.591	5.513	0.30
3 or more pressors pre-LT	0.755	0.086	6.638	0.80
2 pressors pre-LT	1.177	0.536	2.585	0.69
No pressors pre-LT	1.000	.	.	.
Intubated for pulmonary pathology	2.931	1.246	6.892	0.014
Intubated for airway protection	1.000	.	.	.
Infection duration > 14 days	1.193	0.276	5.169	0.81
Infection duration 7-14 days	3.303	0.938	11.636	0.06
Infection duration < 7 days	3.418	1.059	11.036	0.04
No infection	1.000	.	.	.

Table 12. Univariable odds ratios (per SD) for continuous measures.

Variable	OR	Lower CL	Upper CL	P
Time on waitlist, d	1.257	0.890	1.776	0.19
Age, y	1.023	0.704	1.486	0.90
BMI	0.977	0.673	1.420	0.90
Days admitted prior to LT, d	0.892	0.603	1.320	0.57
Days admitted in ICU prior to LT, d	1.067	0.746	1.527	0.72
PaO_2:FiO_2 ratio at ICU admission	0.921	0.631	1.345	0.67
PEEP at ICU admission	0.596	0.278	1.276	0.18
Log WBC at ICU admission	0.935	0.645	1.354	0.72
Log Plts at ICU admission	1.020	0.704	1.478	0.92
MELD at ICU admission	0.861	0.593	1.250	0.43
MELD Na at ICU admission	0.827	0.567	1.207	0.33
CLIF-C-ACLF at ICU admission	1.094	0.753	1.589	0.64

PaO_2:FiO_2 ratio pre-LT	0.830	0.571	1.205	0.33
PEEP pre-LT	0.989	0.663	1.473	0.96
Base excess pre-LT	0.601	0.414	0.872	0.007
WBC pre-LT	0.984	0.679	1.426	0.93
Plt pre-LT	0.988	0.683	1.432	0.95
MELD pre-LT	1.064	0.731	1.549	0.75
MELD Na pre-LT	1.062	0.729	1.546	0.75
CLIF-C-ACLF pre-LT	1.191	0.813	1.745	0.37
DRI, per unit	1.867	1.069	3.261	0.028
Cold ischemia time	0.920	0.626	1.351	0.67
Change in MELD Na*	1.115	0.753	1.653	0.59
Change in MELD*	1.083	0.744	1.576	0.68
Change in CLIF-C-ACLF*	0.868	0.572	1.317	0.51
Change in WBC*	0.829	0.578	1.188	0.31
Change in Na*	0.806	0.437	1.487	0.49
Change in Plt*	0.848	0.594	1.212	0.37
Change in P:F ratio*	1.071	0.718	1.597	0.74
Percent change in MELD Na	1.147	0.794	1.656	0.47
Increase in MELD Na by at least 6%	2.124	0.944	4.781	0.07
Percent change in MELD	1.099	0.769	1.570	0.61
Percent change in CLIF-C-ACLF	0.858	0.562	1.311	0.48
Percent change WBC	0.937	0.635	1.383	0.74
Percent change Na	0.802	0.432	1.487	0.48
Percent change Plt	0.967	0.646	1.451	0.87
Percent change PaO_2:FiO_2 ratio	1.096	0.769	1.561	0.61
Decrease in PaO_2:FiO_2 ratio by at least 12%	0.464	0.188	1.142	0.10
*Variables denoted by “change” were calculated per SD per day from ICU admission to LT.				

Table 13. Univariable odds ratios by quartiles of continuous measures.

Variable	OR	LL	UL	P
Time on waitlist, q4	1.5641	0.50221	4.8713	0.44
Time on waitlist, q3	1.5641	0.50221	4.8713	0.44
Time on waitlist, q2	1.15384	0.35457	3.7549	0.81
Time on waitlist, q1	1	.	.	.
Age, q4	0.84416	0.26583	2.6806	0.77

Age, q3	1.47887	0.51488	4.2477	0.47
Age, q2	1.92309	0.68558	5.3944	0.21
Age, q1	1	.	.	.
BMI, q4	1	0.34384	2.9083	1.00
BMI, q3	1.78373	0.6623	4.804	0.25
BMI, q2	0.72499	0.23425	2.2438	0.58
BMI, q1	1	.	.	.
Days admitted prior to LT, q4	0.41316	0.15457	1.1044	0.08
Days admitted prior to LT, q3	0.44946	0.17272	1.1696	0.10
Days admitted prior to LT, q2	0.14607	0.04105	0.5198	0.003
Days admitted prior to LT, q1	1	.	.	.
Days admitted in ICU prior to LT, q4	1.52922	0.51263	4.5618	0.45
Days admitted in ICU prior to LT, q3	0.60673	0.16558	2.2233	0.45
Days admitted in ICU prior to LT, q2	2.77552	0.97317	7.9159	0.06
Days admitted in ICU prior to LT, q1	1	.	.	.
PaO_2:FiO_2 ratio at ICU admission, q4	0.69422	0.20778	2.3195	0.55
PaO_2:FiO_2 ratio at ICU admission, q3	1.67489	0.58737	4.7759	0.33
PaO_2:FiO_2 ratio at ICU admission, q2	1.48359	0.5132	4.2889	0.47
PaO_2:FiO_2 ratio at ICU admission, q1	1	.	.	.
PEEP at ICU admission, q4	0.3012	0.05174	1.7536	0.18
PEEP at ICU admission, q1	1	.	.	.
WBC at ICU admission, q4	0.65714	0.22842	1.8906	0.44
WBC at ICU, admission, q3	1.12794	0.42625	2.9848	0.81
WBC at ICU, admission, q2	0.65714	0.22842	1.8906	0.44
WBC at ICU, admission, q1	1	.	.	.
Platelets at ICU admission, q4	0.87927	0.32149	2.4048	0.80
Platelets at ICU admission, q3	0.85518	0.31319	2.3351	0.76
Platelets at ICU admission, q2	0.65714	0.22842	1.8906	0.44
Platelets at ICU admission, q1	1	.	.	.
CLIF-C-ACLF at LT, q4	1.04638	0.39686	2.7589	0.93
CLIF-C-ACLF at LT, q3	0.62239	0.20389	1.8999	0.41
CLIF-C-ACLF at LT, q2	1.92762	0.71794	5.1755	0.19
CLIF-C-ACLF at LT, q1	1	.	.	.
MELD at ICU admission, q4	0.62441	0.23681	1.6464	0.34
MELD at ICU admission, q3	0.31686	0.10468	0.9591	0.042
MELD at ICU admission, q2	0.54337	0.20134	1.4664	0.23
MELD at ICU admission, q1	1	.	.	.
CLIF-C at ICU admission, q4	1.12794	0.42625	2.9848	0.81

CLIF-C at ICU admission, q3	0.55473	0.18596	1.6548	0.29
CLIF-C at ICU admission, q2	0.76516	0.27356	2.1402	0.61
CLIF-C at ICU admission, q1	1	.	.	.
PaO_2:FiO_2 ratio at LT, q4	0.72632	0.24781	2.1288	0.56
PaO_2:FiO_2 ratio at LT, q3	0.62948	0.20697	1.9145	0.41
PaO_2:FiO_2 ratio at LT, q2	1.50852	0.57229	3.9764	0.41
PaO_2:FiO_2 ratio at LT, q1	1	.	.	.
PEEP at LT, q4	0.88093	0.2014	3.8532	0.87
PEEP at LT, q1	1	.	.	.
Base excess at LT, q4	0.18744	0.06102	0.5758	0.004
Base excess at LT, q3	0.23636	0.07602	0.7349	0.012
Base excess at LT, q2	0.38314	0.14961	0.9812	0.046
Base excess at LT, q1	1	.	.	.
WBC at LT, q4	1.30691	0.46725	3.6555	0.61
WBC at LT, q3	1.43139	0.52003	3.9399	0.49
WBC at LT, q2	0.72499	0.23425	2.2438	0.58
WBC at LT, q1	1	.	.	.
Platelets at LT, q4	1.14996	0.40271	3.2837	0.79
Platelets at LT, q3	1.00447	0.35441	2.8469	0.99
Platelets at LT, q2	1.14996	0.40271	3.2837	0.79
Platelets at LT, q1	1	.	.	.
MELD at LT, q4	1	0.34384	2.9083	1.00
MELD at LT, q3	0.97333	0.33516	2.8266	0.96
MELD at LT, q2	1.47412	0.5346	4.0647	0.45
MELD at LT, q1	1	.	.	.
CLIF-C-ACLF, q4	1.80267	0.6043	5.3775	0.29
CLIF-C-ACLF, q3	1.54161	0.50912	4.668	0.44
CLIF-C-ACLF, q2	1.80267	0.6043	5.3775	0.29
CLIF-C-ACLF, q1	1	.	.	.
Cold ischemia time, q4	0.84433	0.29517	2.4152	0.75
Cold ischemia time, q3	1.13928	0.41353	3.1387	0.80
Cold ischemia time, q2	0.62798	0.20583	1.9159	0.41
Cold ischemia time, q1	1	.	.	.
MELD change ICU to LT, q4	1.92309	0.68558	5.3944	0.21
MELD change ICU to LT, q3	1.13333	0.37978	3.382	0.82
MELD change ICU to LT, q2	1.16438	0.38963	3.4797	0.79
MELD change ICU to LT, q1	1	.	.	.
CLIF-C ACLF change ICU to LT, q4	1.28281	0.47593	3.4577	0.62

CLIF-C ACLF change ICU to LT, q3	0.5203	0.16364	1.6544	0.27
CLIF-C ACLF change ICU to LT, q2	1.13731	0.41583	3.1105	0.80
CLIF-C ACLF change ICU to LT, q1	1	.	.	.
WBC change ICU to LT, q4	1.33802	0.45725	3.9154	0.60
WBC change ICU to LT, q3	1.92309	0.68558	5.3943	0.21
WBC change ICU to LT, q2	1	0.32587	3.0687	1.00
WBC change ICU to LT, q1	1	.	.	.
Na change ICU to LT, q4	0.65714	0.22842	1.8906	0.44
Na change ICU to LT, q3	0.52381	0.16614	1.6515	0.27
Na change ICU to LT, q2	1.18286	0.46225	3.0268	0.73
Na change ICU to LT, q1	1	.	.	.
Platelets change ICU to LT, q4	1.26375	0.48346	3.3034	0.63
Platelets change ICU to LT, q3	0.85518	0.31319	2.3351	0.76
Platelets change ICU to LT, q2	0.36502	0.10911	1.2211	0.10
Platelets change ICU to LT, q1	1	.	.	.
PaO_2:FiO_2 ratio change ICU to LT, q4	0.76259	0.27075	2.1479	0.61
PaO_2:FiO_2 ratio change ICU to LT, q3	0.76259	0.27075	2.1479	0.61
PaO_2:FiO_2 ratio change ICU to LT, q2	0.65392	0.22579	1.8939	0.43
PaO_2:FiO_2 ratio change ICU to LT, q1	1	.	.	.
Percent change MELD ICU to LT, q4	1.92309	0.68558	5.3944	0.21
Percent change MELD ICU to LT, q3	0.97403	0.31783	2.985	0.96
Percent change MELD ICU to LT, q2	1.33802	0.45725	3.9154	0.60
Percent change MELD ICU to LT, q1	1	.	.	.
Percent change CLIF-C-ACLF ICU to LT, q4	1.28281	0.47593	3.4577	0.62
Percent change CLIF-C-ACLF ICU to LT, q3	0.6309	0.20805	1.9132	0.42
Percent change CLIF-C-ACLF ICU to LT, q2	1	0.35917	2.7842	1.00
Percent change CLIF-C-ACLF ICU to LT, q1	1	.	.	.
Percent change WBC ICU to LT, q4	1.52174	0.52894	4.378	0.44
Percent change WBC ICU to LT, q3	1.92309	0.68558	5.3943	0.21
Percent change WBC ICU to LT, q2	0.84416	0.26583	2.6806	0.77
Percent change WBC ICU to LT, q1	1	.	.	.
Percent change Na, ICU to LT, q4	0.65714	0.22842	1.8906	0.44
Percent change Na, ICU to LT, q3	0.52381	0.16614	1.6515	0.27
Percent change Na, ICU to LT, q2	1.18286	0.46225	3.0268	0.73
Percent change Na, ICU to LT, q1	1	.	.	.
Percent change platelets, ICU to LT, q4	0.87927	0.32149	2.4048	0.80
Percent change platelets, ICU to LT, q3	0.85518	0.31319	2.3351	0.76
Percent change platelets, ICU to LT, q2	0.65714	0.22842	1.8906	0.44

Percent change platelets, ICU to LT, q1	1	.	.	.
Percent change PaO_2:FiO_2 ratio ICU to LT, q4	0.77676	0.286	2.1096	0.62
Percent change PaO_2:FiO_2 ratio ICU to LT, q3	0.48779	0.16464	1.4452	0.20
Percent change PaO_2:FiO_2 ratio ICU to LT, q2	0.57868	0.20255	1.6533	0.31
Percent change PaO_2:FiO_2 ratio ICU to LT, q1	1	.	.	.

Table 14. Results of multivariate logistic regression modeling futility.

Variable	OR	95% CI	P
Previous upper abdominal surgery	4.8	1.7-13.3	0.003
Sepsis Duration (ICU to LT)			
None	1.0	-	-
≤ 14 days	6.7	1.5-10.4	0.015
> 14 days	2.8	0.4-19.0	0.29
Intubation Indication			
Airway Protection	1.0	-	-
Pneumonia/ ARDS	3.8	1.4-10.4	0.008
Increase in MELD ≥ 6% per week from ICU to LT	3.1	1.1-8.4	0.029
Decrease in PaO2:FiO2 ≥ 12% per week from ICU to LT	4.3	1.3-13.8	0.016
Base excess at LT	0.8	0.6-0.9	0.003
Donor Risk Index	2.5	1.3-5.0	0.008

OR, odds ratio; CI, confidence interval; ICU, intensive care unit; LT, liver transplant; ARDS, acute respiratory distress syndrome; MELD, model for end-stage liver disease

Displayed p-values are determined using the Wald test, $p<0.05$ was considered significant.

Comparison to external MODEL Consortium cohort

There were 169 patients meeting criteria for eACLF-3 in the MODEL Consortium dataset. Complete comparisons of demographics and outcomes is shown in Table 14. Compared to the UCLA cohort eACLF-3 patients, the MODEL eACLF-3 cohort had lower MELD scores at ICU admission and at LT (36 and 38 versus 38 and 41, p <

0.001), fewer patients intubated for pulmonary pathology (43% versus 58%, p = 0.009), lower median PaO_2:FiO_2 ratios at LT (259.0 versus 303.3, p = 0.003), shorter median listing date to LTs (7 days versus 18 days, p<0.001), and shorter median pre-LT ICU admission lengths (9 days versus 17 days, p < 0.001).

Table 15. Baseline demographic and outcomes comparison in extreme ACLF-3 patients at UCLA and external MODEL Consortium cohort.

Variable	UCLA (n=177)	MODEL (n=169)	P
Demographic			
Age (y), median (IQR)	58.00 [49.6, 63.9]	55.00 [46.0, 61.3]	0.016
Gender = M (%)	89 (50)	78 (46.2)	0.51
BMI (kg/m^2), median (IQR)	26.6 [23.6, 31.7]	29.9 [24.9, 36.0]	0.001
Primary Diagnosis (%)			0.39
Alcohol	65 (37)	65 (39)	
MAFLD	45 (25)	27 (16)	
Hepatitis C	34 (19)	29 (17)	
autoimmune/cholestasis	21 (12)	25 (15)	
other	12 (7)	23 (14)	
Prior abdominal surgery (%)	70 (40)	70 (42)	0.77
Diabetes mellitus (%)	54 (31)	40 (24)	0.20
Pre-LT Acuity/ Hospitalization Characteristics			
MELD-Na on ICU admission, median (IQR)	38.4 [35.4, 41.5]	36.0 [32.0, 40.0]	<0.001
MELD-Na at LT, median (IQR)	41.0 [38.5, 42.8]	37.7 [34.8, 41.3]	<0.001
Reason for Mechanical Ventilation, n (%)			
Lung pathology (Pneumonia/ARDS)	103 (58)	71 (43)	0.009
Hepatic Encephalopathy/ airway protection	74 (42)	98 (58)	
PaO_2:FiO_2 ratio at ICU admission, median (IQR)	247.1 [165.0, 334.2]	255.0 [161.6, 349.3]	0.92
PaO_2:FiO_2 ratio at LT, median (IQR)	303.3 [230.0, 387.5]	259.0 [211.4, 330.6]	0.003
Preoperative sepsis, n (%)	105 (60)	96 (57)	0.66
Listing date to LT (d), median (IQR)	18.0 [6.0, 157.0]	7.0 [3.0, 43.5]	<0.001
Duration of preoperative ICU admission (d), median (IQR)	17.0 [9.0, 23.0]	9.0 [5.0, 16.0]	<0.001
Donor and Operative Characteristics			
Donor age (y), median (IQR)	34.0 [22.0, 49.0]	35.5 [26.0, 49.8]	0.14
Cold ischemic time (h), median (IQR)	7.0 [6.0, 8.0]	5.9 [4.4, 7.9]	0.001
DCD (%)	4 (2)	18 (11)	0.003
Outcomes			

Post-LT ICU length of stay (d), median (IQR)	29.0 [14.0, 54.0]	9.0 [4.0, 18.0]	<0.001
Post-LT length of stay (d), median (IQR)	53.0 [32.0, 78.5]	22.0 [14.0, 36.0]	<0.001
Post-LT duration of intubation (median [IQR])	9.0 [3.0, 44.0]	2.0 [1.0, 5.3]	<0.001
Extubated after LT, n (%)	20 (11)	7 (4)	0.25
PNF, n (%)	6 (3)	5 (3)	0.81
HAT, n (%)	9 (5)	2 (1)	0.08
Discharge location, n (%)			0.007
Home	76 (53)	92 (63)	
LTAC	11 (6)	20 (14)	
Rehab or SNF	57 (40)	35 (24)	
Post-LT readmission within one month, n (%)	43 (24)	60 (36)	0.030
Biliary complications, n (%)	29 (17)	30 (18)	0.84
Re-LT, n (%)	10 (6)	1 (1)	0.017
Futility, n (%)	35 (20)	50 (30)	0.046

ACLF, acute-on-chronic liver failure; MODEL, Multi-Organ Dysfunction and Evaluation for Liver Transplantation; IQR, interquartile range; BMI, body mass index; MAFLD, metabolic-associated fatty liver disease; LT, liver transplant; MELD, Model for End-stage Liver Disease; CLIF-C, Chronic Liver Failure Consortium; ICU, intensive care unit; ARDS, acute respiratory distress syndrome; OSH, outside hospital; DCD, donation after circulatory death

Categorical variables are displayed as number (%) and differences were assessed using the Chi-squared test (significant when $p<0.05$). Continuous variables are displayed as median (IQR) and differences were assessed using the Mann-Whitney U test (significant when $p<0.05$).

In terms of outcomes, MODEL patients had shorter median post-LT ICU lengths of stay (9 days versus 29 days, $p < 0.001$), shorter median total lengths of stay (22 days versus 53 days, $p < 0.001$), shorter median duration of post-LT intubation (2 days versus 9 days, $p < 0.001$), fewer re-transplants (1% versus 6%, $p = 0.017$) compared to the UCLA cohort. However, MODEL eACFL-3 patients experienced more readmissions one month after LT (36% versus 24%, $p=0.030$), with a larger proportion of patients experiencing the futile outcome (30% versus 20%, $p = 0.046$).

MODEL cohort patients had worse three-, six-, and twelve-month survivals of 83%, 79%, and 70%, compared to the UCLA cohort with 89%, 85%, and 80% (Figure 8A, $p = 0.048$). However, overall survival was equivalent, with the MODEL cohort

patients having one-year, three-year, and 5-year overall survivals of 70%, 65%, and 59%, compared to 80%, 73%, and 63% in the UCLA cohort (Figure 8B, p = 0.14).

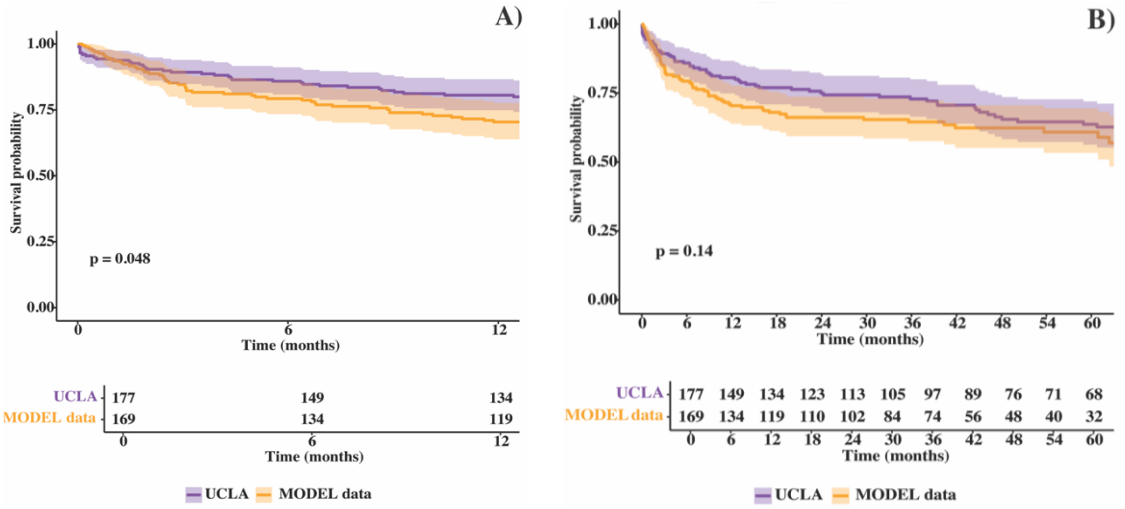


Figure 8. Post-transplant survival comparisons.

One-year (A) and overall (B) survival comparisons between extreme ACLF-3 patients at UCLA (purple) and in the MODEL Consortium (orange).

Chapter 2 Discussion

In this study, we report the largest single-center experience of ACLF-3 (n=581) and the only study to define a unique subset of extreme ACLF-3 patients receiving maximal life support (mechanical ventilation, vasopressors, and renal replacement therapy) at time of LT. Twenty percent of our eACLF-3 cohort experienced a futile outcome, making the overall 5-year survival 63%, which may be considered an unacceptably high short-term mortality to justify the broad use of scarce donor livers. However, considering the high-acuity of these eACLF-3 patients, it was remarkable to see an overall 5-year survival of 79%, conditioned on lack of futility, suggesting that ACLF-3 patients on maximal life support can achieve acceptable long-term survival with

appropriate selection (29). The 177 eACLF-3 patients examined at our center had overall consistent patient characteristics and outcomes when compared to the 169 patients in the MODEL Consortium external validation cohort.

As it pertains to this unique population of eACLF-3 patients, significant risk factors for futility in this study included a decrease in PaO2:FiO2 ratio and increase in MELD-Na during their ICU stay prior to LT, sepsis diagnosed within 14 days of LT, and the indication for mechanical ventilation. In a 2021 expert consensus panel on futility in LT for ACLF-3 patients, Weiss et al reported their prescribed criteria for futility in ACLF-3 including persistent fever or < 72 hours of antimicrobial therapy in the case of ongoing sepsis, PaO2:FiO2 ratio < 150 at LT, levophed dose > 1 μg/kg per minute, serum lactate > 9 mmol/ L. Based on our data, the receipt of antimicrobial therapy for 72 hours prior to LT in the case of sepsis may be inadequate for eACLF-3 patients on maximal life support, with a longer observation period between sepsis onset and LT allowing for optimal treatment of sepsis and affording the greatest reduction in risk of futility. The decision to proceed with LT in a septic patient must balance the concern about delaying transplant too long, resulting in further decompensation and an inability to perform surgery safely with the risk of infection recrudescence following LT with the addition of immunosuppression. Additionally, the incorporation of acuity measure changes from ICU admission to LT may be better suited to this population with a decrease in PaO2:FiO2 ratio and increase in MELD-Na score from ICU to LT being more predictive of futility than a specific value cutoff at LT. Given that ACLF is associated with a rapid decompensation leading to organ failure, the rate of change of decompensation may

better inform prognosis and the likelihood of futility following LT than acuity measures at LT alone (30, 31).

To our knowledge, this is the first study that specifically analyzed the indication for mechanical ventilation in ACLF-3 patients as a predictor of post-LT outcomes. Determining indications for intubation can be a difficult task, requiring manual physician chart review with analysis of chest X-rays and culture data. Neither the TAM score nor the SALT-M score included indication for ventilation as a risk factor in the final model and this data is not available in national transplant registries, highlighting the significance of the elucidation of this specific risk factor in this study. ACLF-3 patients can require mechanical ventilation for a variety of reasons, and it is not surprising that those intubated for airway protection due to severe hepatic encephalopathy or esophageal variceal bleeding with no underlying pulmonary pathology have a lower risk of futility than those intubated for respiratory failure due to pneumonia or ARDS. In fact, among the 35 patients who experienced futility, 29 (83%) of them were intubated for pulmonary pathology preoperatively, with 21 (60%) patients never extubated postoperatively and expiring while intubated. In those futile patients who were extubated postoperatively, 8 (23%) were reintubated prior to discharge or mortality.

In addition to indication for mechanical ventilation, worsening base deficit at LT was identified as a significant risk factor for futility, which is consistent with prior reports of increasing lactate level (another marker of tissue ischemia and systemic inflammation) being associated with futility in ACLF-3 patients (22, 25). Given that nearly 35% of the futile patients had multi-system organ failure as a cause of death, base deficit may be an important indicator of unrecoverable global tissue ischemia that

portends futile outcomes even with LT. Additionally, within the 35% who died as a result of multi-system organ failure, 29% had associated sepsis again signifying the prognostic importance of sepsis in this population. Shorter duration sepsis with increased base deficit may suggest overwhelming sepsis leading to multi-system organ failure that cannot be salvaged even with LT.

An increasing donor risk index was independently associated with futility, validating that high-risk donor livers are ill suited to such critically ill ACLF-3 patients on maximal life support (32, 33). In fact, in our cohort only four eACLF-3 patients received donation after circulatory death (DCD) livers, showing that our institutional practice has been to prioritize higher quality organs for the most significantly ill recipients. However, in the current era of machine perfusion, the need to avoid organs traditionally considered marginal for the highest acuity recipients may not be as relevant due to the afforded ability to mitigate ischemia-reperfusion injury and assess organ function prior to transplantation with normothermic machine perfusion. Going forward, whether it is suitable to increasingly utilize DCD livers for these high acuity recipients in the era of widespread machine perfusion will certainly be an area of active investigation (34).

Our multivariate logistic regression model predicted futility in eACLF-3 patients with an AUROC of 0.83, incorporating dynamic changes in MELD score and oxygenation status along with the novel risk factor of pulmonary pathology as an indication for intubation. Our final model had high specificity (90%) with lower sensitivity (63%), which would afford a minimization of false positive futility predictions and a tolerance of relatively more false negative results. In the context of selecting eACLF-3 patients for LT, a relatively lower number of incorrect futility predictions would be

desired to not exclude consideration of a patient for receiving a potentially life-saving LT. Given the relatively low number of events in our cohort, our current model may be overfit to the training data and external validation in an external dataset is required prior to any clinical adoption to predict futility in this population. While we were able to obtain the MODEL Consortium external validation data, the granularity of data necessary to validate our model was not present. However, key takeaways from this data are that eACLF-3 patients exist in the LT population with broadly similar characteristics to our cohort, though are rare (n=169 in 15 centers). Additionally, a similar survival trend was seen where the vast majority of mortality is seen in the short term, with futile outcomes in eACLF-3 observed in 20% of the UCLA cohort and 30% of the MODEL cohort. However, eACLF-3 patients surviving that first year have acceptable long-term survival, emphasizing that with appropriate patient selection these patients can be transplanted successfully.

Chapter 2 Limitations

Due to the retrospective design of this study, we cannot account for all potential confounding variables, preventing determination of any causal relationships. We focused on eACLF-3 patients who were transplanted and analyzed risk factors, though we must acknowledge that by design this presents a possible selection bias, whereby the specific subset of listed eACLF-3 patients surviving to and following transplant differed from eACFL-3 patients experiencing pre-LT mortality. Additionally, our findings regarding post-transplant survival were broadly corroborated in an external validation cohort. While the model itself cannot be externally validated due to missing data in the MODEL Consortium data, it should be noted that the main contribution of our

manuscript is a description of outcomes in this unique subset of extreme ACLF-3 recipients on maximal pre-LT life support. Our model certainly adds insight into some of the factors that may portend a poor outcome in this subset of patients but will require additional prospective data for validation and testing moving forward. While our center experience with transplantation of many ACLF-3 and eACLF-3 patients may not be directly generalizable to other centers with lower acuity patient populations, the similar outcomes observed in the eACFL-3 MODEL population support that acceptable outcomes are achievable with careful patient selection. Futility in LT is not clearly defined and can have differing meanings depending on context (35). In this work, we defined futility as one-year post-transplant death or in-hospital mortality, which is appropriate when considering transplantation for a hyperacute population.

Chapter 2 Conclusions

We report the largest single-center experience of LT for ACLF-3 patients, and define a unique hyperacute cohort of extreme ACLF-3 patients all of whom required mechanical ventilation, vasopressors, and dialysis prior to LT. In this subset of eACLF-3 patients, we demonstrate excellent short and long-term survival considering the high-acuity of these patients, which is validated in an external cohort of patients from the MODEL Consortium. We propose novel risk factors for futility in this population which may aid in appropriate selection of eACLF-3 patients requiring maximal life support for lifesaving LT.

Chapter 2 References

[1] Moreau R, Jalan R, Gines P, et al. Acute-on-chronic liver failure is a distinct syndrome that develops in patients with acute decompensation of cirrhosis. Gastroenterology. 2013;144(7):1426-37, 37.e1-9.

[2] Piano S, Mahmud N, Caraceni P, et al. Mechanisms and treatment approaches for ACLF. Liver International. 2023.

[3] Gustot T, Fernandez J, Garcia E, et al. Clinical course of acute-on-chronic liver failure syndrome and effects on prognosis. Hepatology. 2015;62(1):243-52.

[4] Moreau R, Gao B, Papp M, et al. Acute-on-chronic liver failure: a distinct clinical syndrome. Journal of hepatology. 2021;75:S27-S35.

[5] Arroyo V, Moreau R. Diagnosis and prognosis of acute on chronic liver failure (ACLF) in cirrhosis. Journal of Hepatology. 2017;66(2):451-3.

[6] Jalan R, Gines P, Olson JC, et al. Acute-on chronic liver failure. Journal of hepatology. 2012;57(6):1336-48.

[7] Queck A, Weiler N, Trebicka J. Transplantation in Acute-on-Chronic Liver Failure: Feasibility and Futility. Clinical Liver Disease. 2022;19(5):191-3.

[8] Trebicka J, Gu W, Ibáñez-Samaniego L, et al. Rebleeding and mortality risk are increased by ACLF but reduced by pre-emptive TIPS. Journal of hepatology. 2020;73(5):1082-91.

[9] Moreau R, Jalan R, Gines P, et al. Acute-on-chronic liver failure is a distinct syndrome that develops in patients with acute decompensation of cirrhosis. Gastroenterology. 2013;144(7):1426-37. E9.

[10] Trebicka J, Fernandez J, Papp M, et al. The PREDICT study uncovers three clinical courses of acutely decompensated cirrhosis that have distinct pathophysiology. Journal of hepatology. 2020;73(4):842-54.

[11] Jalan R, Saliba F, Pavesi M, et al. Development and validation of a prognostic score to predict mortality in patients with acute-on-chronic liver failure. Journal of hepatology. 2014;61(5):1038-47.

[12] Arroyo V, Moreau R, Jalan R, et al. Study E-CCC. Acute-on-chronic liver failure: a new syndrome that will re-classify cirrhosis. Journal of hepatology. 2015;62(1):S131-S43.

[13] Sundaram V, Shah P, Wong RJ, et al. Patients with acute on chronic liver failure grade 3 have greater 14-day waitlist mortality than status-1a patients. Hepatology. 2019;70(1):334-45.

[14] Durand F, Roux O, Weiss E, et al. Acute-on-chronic liver failure: Where do we stand? Liver International. 2021;41:128-36.

[15] Gustot T, Agarwal B. Selected patients with acute-on-chronic liver failure grade 3 are not too sick to be considered for liver transplantation. Journal of hepatology. 2017;67(4):667-8.

[16] Artru F, Louvet A, Ruiz I, et al. Liver transplantation in the most severely ill cirrhotic patients: a multicenter study in acute-on-chronic liver failure grade 3. Journal of hepatology. 2017;67(4):708-15.

[17] Artru F, Goldberg D, Kamath PS. Should patients with acute-on-chronic liver failure grade 3 receive higher priority for liver transplantation? Journal of hepatology. 2023;78(6):1118-23.

[18] Michard B, Artzner T, Deridder M, et al. Pretransplant intensive care unit management and selection of grade 3 acute-on-chronic liver failure transplant candidates. Liver Transplantation. 2022;28(1):17-26.

[19] Sundaram V, Patel S, Shetty K, et al. Risk factors for posttransplantation mortality in recipients with grade 3 acute-on-chronic liver failure: analysis of a North American consortium. Liver Transplantation. 2022;28(6):1078-89.

[20] Wu T, Sundaram V. Transplantation for acute-on-chronic liver failure. Clinical Liver Disease. 2019;14(4):152-5.

[21] Michard B, Artzner T, Lebas B, et al. Liver transplantation in critically ill patients: preoperative predictive factors of post-transplant mortality to avoid futility. Clinical Transplantation. 2017;31(12):e13115.

[22] Weiss E, Saner F, Asrani SK, et al. When Is a Critically Ill Cirrhotic Patient Too Sick to Transplant? Development of Consensus Criteria by a Multidisciplinary Panel of 35 International Experts. Transplantation. 2021;105(3).

[23] Engelmann C, Thomsen KL, Zakeri N, et al. Validation of CLIF-C ACLF score to define a threshold for futility of intensive care support for patients with acute-on-chronic liver failure. Critical Care. 2018;22:1-8.

[24] Figorilli F, Putignano A, Roux O, et al. Development of an organ failure score in acute liver failure for transplant selection and identification of patients at high risk of futility. PloS One. 2017;12(12):e0188151.

[25] Artzner T, Michard B, Weiss E, et al. Liver transplantation for critically ill cirrhotic patients: stratifying utility based on pretransplant factors. American Journal of Transplantation. 2020;20(9):2437-48.

[26] Hernaez R, Karvellas CJ, Liu Y, et al. The novel SALT-M score predicts 1-year post-transplant mortality in patients with severe acute-on-chronic liver failure. Journal of Hepatology. 2023;79(3):717-27.

[27] Sundaram V, Kogachi S, Wong RJ, et al. Effect of the clinical course of acute-on-chronic liver failure prior to liver transplantation on post-transplant survival. Journal of hepatology. 2020;72(3):481-8.

[28] Azur MJ, Stuart EA, Frangakis C, et al. Multiple imputation by chained equations: what is it and how does it work? International journal of methods in psychiatric research. 2011;20(1):40-9.

[29] Kwong AJ, Ebel NH, Kim WR, et al. OPTN/SRTR 2021 annual data report: liver. American Journal of Transplantation. 2023;23(2):S178-S263.

[30] Györi GP, Silberhumer GR, Rahmel A, et al. Impact of dynamic changes in MELD score on survival after liver transplantation–a Eurotransplant registry analysis. Liver International. 2016;36(7):1011-7.

[31] Cholankeril G, Li AA, Dennis BB, et al. Pre-operative delta-MELD is an independent predictor of higher mortality following liver transplantation. Scientific Reports. 2019;9(1):8312.

[32] Nemes B, Gelley F, Zádori G, et al., editors. Outcome of liver transplantation based on donor graft quality and recipient status. Transplantation proceedings; 2010: Elsevier.

[33] Winter A, Féray C, Antoine C, et al. Matching Graft Quality to Recipient's Disease Severity Based on the Survival Benefit in Liver Transplantation. Sci Rep. 2020;10(1):4111.

[34] Parente A, Wehrle CJ, Schold JD, et al. Frequent Use of Machine Perfusion Enables Utilization of Higher-Risk DCD Livers in the United States: Are We There Yet? Journal of Hepatology. 2024.

[35] Rolak S, Elhawary A, Diwan T, et al. Futility and poor outcomes are not the same thing: A clinical perspective of refined outcomes definitions in liver transplantation. Liver Transplantation. 2024;30(4)

Chapter 3 Liver transplant donor-recipient matching with offline reinforcement learning

Chapter 3 Abstract

Despite advances in liver transplantation (LT), deciding when to transplant a patient within the context of high waitlist mortality, organ scarcity, and risk of graft failure, remains an ongoing challenge. Existing approaches focus on the static prediction of successful LT donor-recipient pairs without weighing the competing interests such as the risk of graft failure against the risk of waitlist mortality and how these risks change over time. Instead, we used an offline reinforcement learning (RL) approach to represent the problem as the optimization of the series of decisions to wait, delist, or transplant a candidate at different timepoints. Using waitlist trajectories for LT candidates from the national Scientific Registry of Transplant Recipients (SRTR) database, we trained a model resulting in the avoidance of 73% of donor-recipient pairs that led to graft failure or death, preservation of 93% of successful transplants, and potentially suitable donors were found for 47% of those patients that died on the waitlist. Notably, the analysis of decisions and post-transplant survival revealed that our model learned features suggestive of successful donor-recipient pairs. Overall, we demonstrate how RL-based approaches better portray real-world LT donor-recipient matching decisions, illustrating their potential as useful clinical tools.

Chapter 3 Introduction

Liver transplantation (LT) is the gold standard treatment for end-stage liver disease [1]. Despite significant advances in the use of LT, waitlist mortality remains as high as 25% [2, 3]. This high mortality is representative of several challenges including donor scarcity and an aging recipient population with additional medical comorbidities

that increase waitlist mortality risk. The Model for End-stage Liver Disease (MELD) score, calculated from a candidate's serum sodium, bilirubin, INR, and creatinine, is the standard for determining LT priority for patients with decompensated end-stage liver disease in the current United Network for Organ Sharing allocation policy [4-9]. Allocation policies (Share 35 in 2013, acuity circles in 2020) determine the offering order and are designed to increase availability of organs to the most critically ill patients [10-14].

When a donor organ is offered to a candidate, the patient's transplant providers holistically review recipient and donor factors to determine the match suitability and mitigate the risk of graft failure. Despite an overall shortage of donor livers, many available donor livers are discarded or rejected by transplant teams due to a variety of reasons including poor quality of the donor organ, concerns about donor-recipient size mismatch, or other logistical concerns [15]. Graft failure can be considered to include malfunction of the graft necessitating a new liver or death of the recipient. In either case, graft failure is a significant negative clinical event in LT due to the compromised health of that patient and the depletion of that organ from the donor pool. Graft failures occurring within the first year after transplant are thought to be due to donor-recipient match and operative characteristics as many early graft failures are due to vascular compromise of the graft resulting from ischemia reperfusion injury, poor donor organ quality, recipient critical illness, among others [16, 17]. Graft failures occurring later are more often due to organ rejection, complications resulting from immunosuppression, or recurrence of liver disease. Existing models can predict 3-month graft failure like the survival outcomes following liver transplant (SOFT) and the balance of risk (BAR) score

(area under the receiver operating curve; AUROC=0.89) [18-20]. Likewise, several machine learning (ML) techniques predict early post-transplant graft failure with AUROCs from 0.90-0.96 [21-26].

However, existing predictive models only use data from a single time-point – often from donor and recipient features at time of transplant (e.g., SOFT and BAR) – to predict post-transplant mortality. These outcome-oriented models fail to evaluate dynamic changes (e.g., in MELD score and donor-recipient match suitability) over the course of the waitlist period and only consider transplant recipients – negating those patients who had waitlist mortality, an important aspect of real donor-recipient matching decisions.

Instead of conventional pattern recognition to predict outcomes from labeled data, offline reinforcement learning (RL) uses retrospective data to optimize the decisions throughout the LT donor-recipient matching process given short- and long-term rewards [27-30]. Another advantage of offline RL is the ability to learn improved decisions when given retrospective examples of suboptimal trajectories, which is common in real-world data and in other domains (e.g., robotic control, self-driving, recommendation systems) [29, 31-34]. Crucially, this approach better resembles the real decision-making of transplant providers compared to standard outcomes prediction; and considers the competing risks and benefits of multiple decisions at any given time for a donor-recipient pair such as transplant resulting in early post-LT graft failure or successful transplant vs. remaining on the waitlist resulting in recovery or mortality. While RL has been explored in many domains [35-48], offline RL for medicine is an emerging field that has never been applied to optimize LT donor-recipient matching [49-

59]. This application of offline RL has the potential to transcend pattern recognition towards a more accurate portrayal of real-world decision-making through the optimization of donor-recipient matching when accounting for multiple competing potential outcomes.

Chapter 3 Results

This study used the Scientific Registry of Transplant Recipients (SRTR) database, with a full description of definitions, processing, and inclusion criteria in Methods “Overview of the Scientific Registry of Transplant Recipients”. Of the 45,635 LT candidates who met the study inclusion criteria (Figure 9a), 7,616 (17%) died on the waitlist; 28,587 (63%) were transplanted, of whom 2,119 (7%) had graft failure or died within 1-year post-transplant. Candidates’ demographic information satisfying inclusion criteria are shown in Table 15. After excluding cases for missingness (<5%), there were 43,595 candidates, 26,790 used donor organs, and 16,272 discarded organs available for analysis, resulting in 214,865 states after processing the episodes. Demographic comparisons between used and discarded donor livers are shown in Table 16.

We processed each candidate’s MELD scores over time to generate episodes of an LT Markov Decision Process (MDP) (Figure 9b), with complete details in Methods “Episode processing details”. A MELD score indicated a step in the MDP, with the candidate and potential donor’s features constituting the continuous state observations at that step, and the potential actions of waiting, delisting, or transplanting at each step (Figure 9c). Rewards were associated with the state-action pairs and the outcomes of terminal transitions (details in Methods “Reward function design”). We used Conservative Q-learning (CQL), a popular standard in offline RL (Figure 10) [60, 61],

with full details of the methods and experimental approach outlined in Methods “Conservative Q-learning” and Methods “Experimental pipeline”.

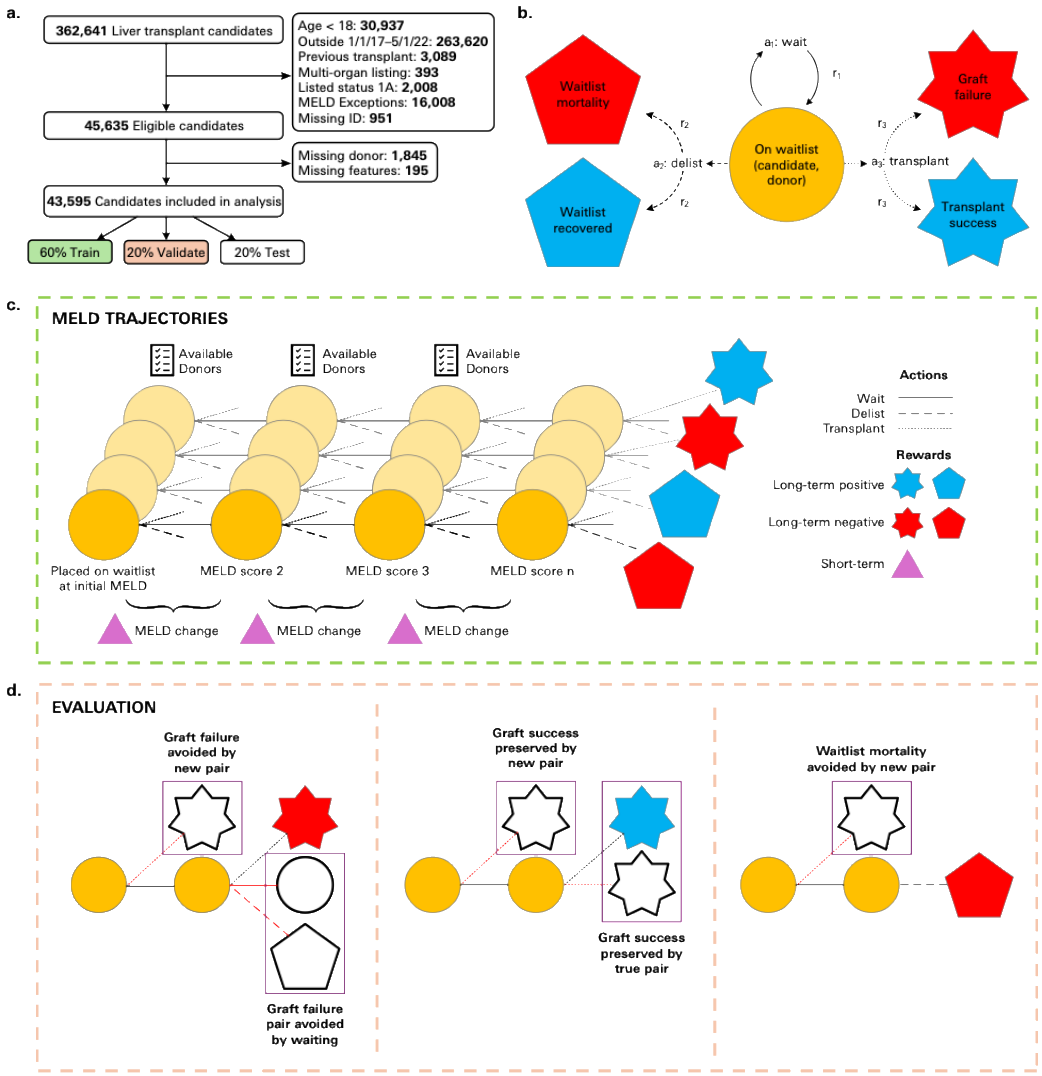


Figure 9. Overview of study and offline RL approach.

A) Inclusion and exclusion criteria for arrival at the final cohort of patients used for analysis. B) Markov decision process showing states derived from retrospective data (yellow: candidate and donor features with possible actions (transplant, delist, wait) that lead to discrete outcomes (blue: removal from the waitlist due to improved status

resulting in a transplant no longer being required and transplant success; red: waitlist mortality and graft failure). C) Training process diagram including visualization of MELD trajectories in sequence with possible donors at each point and potential actions available, which lead to one of four potential outcomes. The reward function is based on short-term (MELD change) and long-term (outcomes) rewards that are weighed together to optimize the decision-making process at each point in the trajectory. d) Evaluation process diagram displaying the avoidance of graft failure by either transplanting that candidate to a different donor or point in the trajectory or by taking a new action (left panel), graft success preservation achieved by transplanting to true successful pair or the same candidate with a new pair (middle panel) and correct intervention where a waitlist mortality is avoided by selecting a potential donor for transplant (right panel).

Table 16. Descriptive demographics of LT candidates meeting study inclusion criteria.

Variable	Total cohort (n=45,635)
Age, y (median [IQR])	56.00 [48.00, 63.00]
Sex = M (%)	27,494 (60.2)
Height, cm (median [IQR])	171.50 [162.60, 179.07]
Weight, kg (median [IQR])	83.70 [70.60, 98.88]
BMI (median [IQR])	28.00 [25.00, 33.00]
Liver Disease Etiology (%)	
Alcoholic Cirrhosis	19,044 (41.7)
Autoimmune	1,508 (3.3)
Biliary Disease	3,792 (8.3)
Cirrhosis – Other	527 (1.2)
Hepatitis B	597 (1.3)
Hepatitis C	2,978 (6.5)
Malignancy	1,144 (2.5)
Metabolic Disease	758 (1.7)
MASLD	10,708 (23.5)
Other	3,599 (7.9)

Diabetes (%)	12645 (27.7)
Blood Type, ABO (%)	
A	17,339 (38.0)
AB	1,813 (4.0)
B	5,402 (11.8)
O	21,081 (46.2)
Mechanical Ventilation (%)	808 (1.8)
Bacterial Peritonitis (%)	5,350 (11.7)
Portal Vein Thrombosis (%)	3,813 (8.4)
TIPSS (%)	4,040 (8.9)
Ascites (%)	
Absent	7,124 (15.6)
Moderate	15,804 (34.7)
Slight	22,676 (49.7)
Encephalopathy (%)	
Grade 1-2	27,127 (59.5)
Grade 3-4	4,658 (10.2)
None	13,820 (30.3)
Dialysis (%)	5,403 (11.8)
MELD listing (median [IQR])	19.00 [14.00, 27.00]
MELD last (median [IQR])	22.00 [15.00, 31.00]
INR listing (median [IQR])	1.57 [1.30, 2.08]
Sodium listing (median [IQR])	136.00 [132.00, 139.00]
Creatinine listing (median [IQR])	1.08 [0.80, 1.67]
Bilirubin listing (median [IQR])	3.50 [1.78, 8.60]
Albumin listing (median [IQR])	3.10 [2.70, 3.60]
Outcomes	
Transplanted (%)	28,587 (62.6)
Death or Delisting (%)	7,616 (16.7)
Delisting, recovered (%)	2,678 (6.2)

Table 17. Demographic comparisons between deceased donors used in liver transplants and discarded donor organs.

Variable	Used (n=26,790)	Discard (n=16,272)	P-value
Age, y (median [IQR])	40.00 [29.00, 53.00]	49.00 [37.00, 57.00]	<0.001
Blood Type, ABO (%)			<0.001
A	10186 (38.0)	6100 (37.5)	
AB	764 (2.9)	728 (4.5)	

B	3018 (11.3)	1889 (11.6)	
O	12822 (47.9)	7555 (46.4)	
Height, cm (median [IQR])	172.72 [165.00, 180.00]	172.00 [165.00, 178.00]	<0.001
Weight, kg (median [IQR])	80.90 [68.80, 95.10]	86.70 [72.60, 105.00]	<0.001
Creatinine (median [IQR])	1.10 [0.77, 1.84]	1.00 [0.70, 1.77]	<0.001
Blood urea nitrogen (median [IQR])	22.00 [15.00, 36.00]	22.00 [15.00, 35.00]	0.025
Total bilirubin (median [IQR])	0.60 [0.40, 1.00]	0.70 [0.40, 1.30]	<0.001
SGOT/AST (median [IQR])	41.00 [24.00, 82.00]	59.00 [33.00, 124.00]	<0.001
SGPT/ALT (median [IQR])	38.00 [21.00, 78.00]	46.00 [25.00, 101.00]	<0.001
Sodium (median [IQR])	148.00 [143.00, 154.00]	146.00 [141.00, 152.00]	<0.001
INR (median [IQR])	1.22 [1.10, 1.40]	1.20 [1.10, 1.40]	0.846
PH (median [IQR])	7.41 [7.37, 7.45]	7.40 [7.35, 7.45]	<0.001
Hematocrit (median [IQR])	28.10 [25.00, 32.00]	29.70 [25.60, 34.60]	<0.001
Cause of death (%)			<0.001
Anoxia	11920 (44.5)	7669 (47.1)	
Stroke	6939 (25.9)	4430 (27.2)	
Head Trauma	7289 (27.2)	3321 (20.4)	
CNS Tumor	86 (0.3)	45 (0.3)	
Other	556 (2.1)	807 (5.0)	
Anit-HIV I/II (%)	38 (0.1)	130 (0.8)	<0.001
Anti-HTLV I/II (%)	14 (0.1)	6 (0.0)	<0.001
RPR-VDRL (%)	324 (1.2)	194 (1.2)	<0.001
Anti-CMV (%)	16968 (63.3)	10070 (61.9)	0.001
HbsAg (%)	42 (0.2)	63 (0.4)	<0.001
Anti-HBC (%)	1352 (5.0)	1166 (7.2)	<0.001
Anti-HCV (%)	2591 (9.7)	2059 (12.7)	<0.001
HbsAb (%)	1204 (4.5)	686 (4.2)	<0.001
EBV IgG (%)	24103 (90.0)	14903 (91.6)	<0.001
EBV IgM (%)	364 (1.4)	276 (1.7)	<0.001
Clinical infection (%)	20646 (77.1)	11456 (70.4)	<0.001
Inotropic support (%)	10881 (40.6)	4481 (27.5)	<0.001
Infection source: blood (%)	3351 (12.5)	2459 (15.1)	<0.001
Cocaine use (%)	6383 (23.8)	3908 (24.0)	<0.001
% Micro vesicular fat (median [IQR])	5.00 [0.00, 10.00]	10.00 [0.00, 21.00]	<0.001
% Macro vesicular fat (median [IQR])	5.00 [0.00, 10.00]	20.00 [5.00, 45.00]	<0.001

Final MAP, mmHg (median [IQR])	88.00 [79.00, 98.00]	87.00 [78.00, 99.00]	0.530
Initial MAP, mmHg (median [IQR])	84.00 [73.00, 97.00]	81.50 [73.00, 92.00]	0.177
Maximum creatinine (median [IQR])	1.10 [0.78, 1.85]	1.01 [0.70, 1.77]	<0.001
High risk by CDC guideline (%)	7116 (26.6)	3805 (23.4)	<0.001
Donation after circulatory death (%)	2586 (9.7)	8713 (53.5)	<0.001

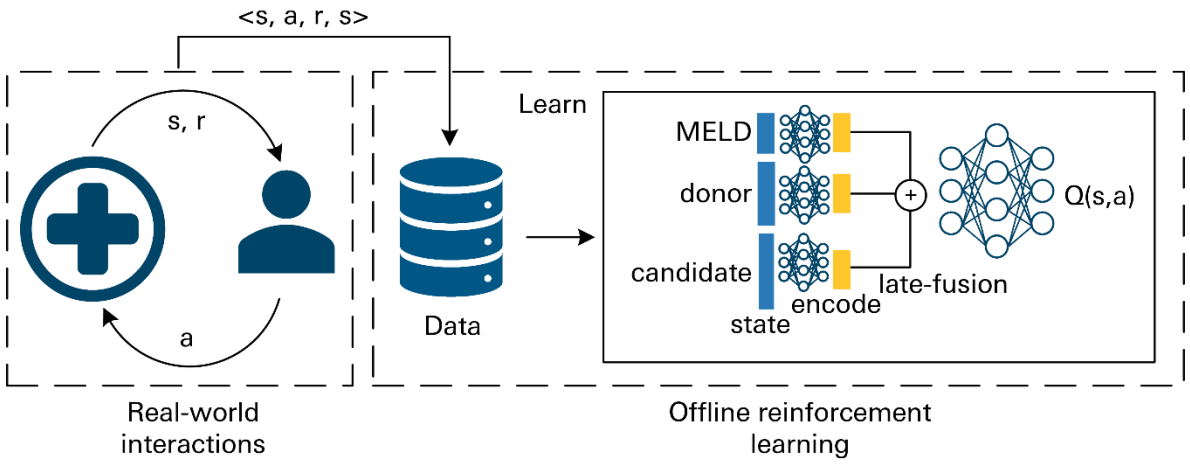


Figure 10. Overview of data collection and offline RL pipeline.

In offline RL, data is collected from sampling real-world interactions between an agent and an environment. Trajectories are formed as the agents perform actions (a) in the environment, resulting in changes of state (s) and a reward ®. Transitions (s, a, r, s) are extracted from these trajectories to form the final dataset for offline RL. Here, we illustrate the general schematic of CQL – one approach to offline RL – that is based on deep Q-learning. An action-value function Q is parameterized by a deep neural network that learns the value given states and actions to ultimately choose the action that maximizes the expected return. The inputs to the Q function are the embedded observations pertaining to the potential donor-recipient pairs.

Evaluation of reinforcement learning model performance

We aimed to understand how each model aligned with pragmatic clinical objectives of simultaneously preventing graft failures while maintaining transplants that led to (or could lead to) successful transplants and delisting candidates that recovered. Accordingly, we developed metrics contingent on observable actions and outcomes (Figure 9d, Table 17), with full definitions in Methods “Evaluation metrics”. To assess the ability to circumvent poor transplant outcomes, we measured “graft failure prevention rate” (GFPR) as the percent of real-world graft failure pairs avoided by the decision to not transplant and “graft failure intervention rate” (GFIR) as the percent avoided by transplanting to a different pair. The retention of positive transplant outcomes was measured using “graft success preservation rate” (GSPR) as the percent of successful transplant candidates who were transplanted with their real-world or an earlier donor. The percent of recovered candidates who were chosen to be delisted was measured as the “correct removal rate” (CRR) and the percent of candidates who experienced waitlist mortality but were recommended earlier transplants was measured as the “correct intervention rate” (CIR). For each metric, we compared our learned model against baseline methods (Methods “Baseline comparisons to a MELD threshold” and “Baseline comparisons to static predictive models”) ranging from existing prioritization standards (i.e., MELD threshold model such as “MELD 15” that transplants any time a candidate’s MELD score exceeds 15) to static predictive models.

Table 18. Definitions of metrics.

Name	Abbreviation	Description
Graft failure prevention rate	GFPR	Percent of candidates that experienced graft failure that the model decided not to transplant given their true pair
Graft failure intervention rate	GFIR	Percent of candidates that experienced graft failure the model decided to transplant to an earlier pair
Graft success preservation rate	GSPR	Percent of candidates that experienced successful transplant that the model decided to transplant given their true pair or an earlier pair
Correct removal rate	CRR	Percent of candidates that recovered on the waitlist that the model decided to delist at any time
Correct intervention rate	CIR	Percent of candidates that experienced waitlist mortality that the model decided to transplant before death
Action match rate	--	Percent of actions decided on by the model that matched the real-life actions recorded in the data

We selected a final offline RL model that was more conservative, indicating an ability to choose more optimal donor-recipient pairs, as opposed to more aggressive models attempting to maximize rewards by over-transplanting and ultimately selecting suboptimal pairs. In the holdout set, there were 5,334 patients who received transplants in reality. The CQL model performed a mean total of 6,927 (SD=671.57) transplants in the holdout set compared to the indiscriminate MELD 15 model with 33,247 (SD=0.00) transplants. Of the graft failure patients, the CQL model avoided the true donor-recipient pair in 73% (GFPR; SD=2%) and selected a new donor-recipient pair for transplant in 55% (GFIR; SD=3%). The model preserved 93% (GSPR; SD=1%) of successful transplants by transplanting the true successful pair or by pairing the candidate to a new donor. The model correctly delisted 38% (CRR; SD=3%) of patients who were delisted due to no longer requiring a transplant and transplanted 47% (CIR; SD=3%) of patients who experienced waitlist dropout. Full metrics are reported in Table 18, with further

subgroup analysis across demographic factors (candidate sex, location, and race) in Table 19.

Table 19. Evaluation results.

Columns represent evaluation metrics: action match rate is the percentage of actions taken by the model that match what occurred in reality; action: wait is the number of times a decision was made to wait; action: delist is the number of times a decision was made to delist; action: transplant is the number of times a decision was made to transplant; GFPR is graft failure prevention rate or of the graft failure patents in what percentage of them did the model select a different action; GFIR is graft failure intervention rate or of the graft failure patients in what percentage of them did the model choose to transplant to a new donor; GSPR is the graft success preservation rate or in the successful transplant patients what percentage of them were transplanted by the model; CRR is the correct removal rate of the patients delisted due to no longer needing a transplant what percentage of them did the model choose to delist; CIR is correct intervention rate or of the waitlist mortality patients what percentage of them did the model choose to transplant. Rows represent model tested: CQL, a final conservative Q-learning based offline reinforcement learning model; PM, static predictive models trained on actions, outcomes, or transplant outcomes using logistic regression (LGR) or eXtreme gradient boosting (XGB); MELD #, meld threshold models using thresholds 15, 20, 24, and 30.

Model	Action match rate, % (SD)	Action: Wait, n (SD)	Action: Delist, n (SD)	Action: Transplant, n (SD)	GFPR, % (SD)	GFIR, % (SD)	GSPR, % (SD)	CRR, % (SD)	CIR, % (SD)
CQL	73.05 (0.97)	34,847.40 (733.34)	1,040.20 (116.31)	6,927.40 (671.57)	73.18 (1.82)	54.74 (3.02)	93.09 (1.17)	37.55 (3.38)	46.83 (3.01)

Action PM: LGR	82.24 (0.00)	42,220.20 (0.84)	103.80 (0.45)	491.00 (0.71)	90.76 (0.00)	9.24 (0.00)	10.74 (0.02)	4.05 (0.00)	2.78 (0.00)
Outcome PM: LGR	34.49 (0.04)	10,582.20 (10.33)	3,323.40 (8.93)	28,909.40 (15.04)	14.69 (0.00)	97.63 (0.00)	99.87 (0.01)	73.88 (0.32)	91.62 (0.04)
Graft failure PM: LGR	51.61 (6.62)	24,359.20 (4121.84)	0.00 (0.00)	18,455.80 (4121.84)	45.02 (10.54)	76.35 (8.08)	87.81 (4.73)	0.00 (0.00)	78.94 (7.86)
Action PM: XGB	83.97 (0.11)	39,142.40 (36.20)	867.80 (34.84)	2,804.80 (29.41)	64.74 (0.81)	40.24 (0.78)	79.68 (1.44)	36.20 (1.37)	10.81 (0.46)
Outcome PM: XGB	46.96 (0.30)	17,011.00 (170.59)	3,224.20 (19.51)	22,579.80 (187.39)	17.49 (1.36)	95.73 (0.87)	99.63 (0.06)	78.61 (0.38)	78.46 (0.84)
Graft failure PM: XGB	65.62 (8.51)	33,898.60 (4623.46)	0.00 (0.00)	8,916.40 (4623.46)	72.18 (13.87)	55.02 (17.55)	78.35 (11.73)	0.00 (0.00)	58.50 (16.43)
MELD 15	38.86 (0.00)	9,568.00 (0.00)	0.00 (0.00)	33,247.00 (0.00)	7.11 (0.00)	95.73 (0.00)	94.69 (0.00)	0.00 (0.00)	90.87 (0.00)
MELD 20	58.72 (0.00)	18,545.00 (0.00)	0.00 (0.00)	24,270.00 (0.00)	23.22 (0.00)	84.36 (0.00)	84.22 (0.00)	0.00 (0.00)	76.11 (0.00)
MELD 24	69.23 (0.00)	25,860.00 (0.00)	0.00 (0.00)	16,955.00 (0.00)	39.57 (0.00)	70.14 (0.00)	71.56 (0.00)	0.00 (0.00)	62.54 (0.00)
MELD 30	76.80 (0.00)	34,162.00 (0.00)	0.00 (0.00)	8,653.00 (0.00)	61.61 (0.00)	46.68 (0.00)	48.90 (0.00)	0.00 (0.00)	41.03 (0.00)

Table 20. Stratified evaluation results.

Stratified evaluation results for the final conservative Q-learning based offline reinforcement model stratified by demographic factors including: candidate sex, the location of the candidate defined by the 11 United Network for Organ Sharing (UNOS) geographic regions, and candidate race. Metrics are consistent with those reported in Table 18.

Variable	n (%)	Action match rate, % (SD)	Action: Wait, n (SD)	Action: Delist, n (SD)	Action: Transplant, n (SD)	GFPR, % (SD)	GFIR, % (SD)	GSPR, % (SD)	CRR, % (SD)	CIR, % (SD)
Sex										
Female	3376 (38.72)	73.86 (1.31)	14,186.80 (347.03)	423.20 (43.78)	2,681.00 (320.58)	73.46 (4.25)	55.47 (6.33)	92.84 (1.34)	38.48 (2.75)	47.98 (2.90)
Male	5343 (61.28)	72.51 (0.80)	20,660.60 (395.96)	617.00 (75.06)	4,246.40 (356.37)	73.00 (0.85)	54.30 (3.34)	93.24 (1.10)	36.92 (3.89)	45.93 (3.41)
Region										
1	450 (5.16)	74.87 (1.21)	2,201.80 (53.68)	107.80 (17.80)	393.40 (45.24)	74.44 (9.30)	55.56 (10.39)	95.05 (1.17)	45.84 (5.29)	52.04 (3.67)
2	1030 (11.81)	72.40 (0.80)	4,018.60 (92.86)	167.80 (16.33)	800.60 (78.46)	71.74 (5.95)	58.70 (5.10)	93.16 (1.66)	44.20 (3.59)	45.99 (2.24)
3	1291 (14.81)	71.91 (1.03)	4,610.40 (94.43)	106.60 (12.97)	941.00 (90.09)	74.32 (2.70)	53.51 (4.23)	92.89 (1.06)	33.39 (4.62)	45.71 (4.93)

4	1091 (12.51)	72.77 (0.90)	4,173.40 (70.82)	128.60 (10.81)	844.00 (63.02)	72.31 (3.34)	54.36 (9.66)	94.12 (1.68)	37.76 (3.55)	43.27 (4.07)
5	1368 (15.69)	74.01 (1.01)	6,136.40 (143.84)	180.60 (26.61)	1,214.00 (143.71)	77.31 (3.86)	51.94 (5.91)	93.64 (1.04)	38.64 (5.89)	53.36 (2.26)
6	216 (2.48)	73.67 (1.53)	904.80 (20.07)	19.00 (3.39)	207.20 (18.20)	77.78 (17.57)	51.11 (12.67)	97.07 (1.36)	37.50 (5.10)	51.00 (8.02)
7	686 (7.87)	73.57 (1.09)	2,817.00 (70.10)	62.60 (18.37)	596.40 (60.13)	62.93 (8.52)	58.05 (8.86)	93.15 (1.15)	31.63 (6.80)	54.46 (5.64)
8	471 (5.40)	72.71 (0.55)	1,642.80 (38.11)	35.60 (3.91)	318.60 (37.59)	76.36 (7.47)	38.18 (7.61)	93.75 (2.32)	27.76 (6.56)	39.66 (7.00)
9	473 (5.42)	73.07 (1.41)	2,086.80 (59.18)	70.60 (5.77)	450.60 (58.35)	75.45 (9.96)	59.09 (4.55)	93.20 (0.72)	45.23 (2.79)	55.11 (5.29)
10	811 (9.30)	71.59 (1.33)	2,872.80 (55.54)	94.80 (11.12)	491.40 (50.16)	74.36 (4.05)	44.62 (4.66)	88.72 (2.94)	38.96 (4.31)	34.43 (5.09)
11	832 (9.54)	73.60 (1.40)	3,382.60 (69.43)	66.20 (9.52)	670.20 (66.32)	71.56 (5.31)	60.00 (8.89)	93.73 (1.16)	28.94 (3.64)	42.03 (3.28)
Race										
Asian	261 (2.99)	71.17 (1.92)	996.40 (18.02)	40.00 (5.52)	218.60 (22.43)	80.00 (7.82)	48.57 (5.98)	89.35 (2.09)	39.13 (4.07)	47.76 (5.88)
Black	527 (6.04)	71.81 (1.47)	1,953.60 (53.39)	72.40 (14.81)	417.00 (51.05)	67.86 (6.19)	55.00 (5.42)	92.33 (1.46)	43.81 (5.45)	48.45 (6.73)
Other	149 (1.71)	74.51 (1.96)	620.60 (16.46)	19.40 (5.37)	121.00 (13.66)	64.00 (18.17)	56.00 (18.17)	96.05 (2.03)	40.00 (6.12)	46.96 (8.36)
White	7782 (89.25)	73.16 (0.91)	31,276.80 (648.00)	908.40 (100.83)	6,170.80 (586.96)	73.57 (2.37)	54.38 (3.11)	93.20 (1.18)	37.00 (3.51)	46.71 (2.80)

Features indicative of successful donor-recipient pairs

We compared the feature distributions of the counterfactual donor-recipient matches with respect to successful donor-recipient pairs to interpret their clinical feasibility. [62] These comparisons were conducted in three subgroups of candidates who (in reality): 1) experienced graft failure; 2) experienced waitlist mortality; or 3) were transplanted successfully, but the model selected a new donor-recipient pair. We compared the feature distributions of discarded organs that the model chose to transplant vs. not transplant to observe if our model was tending towards certain donor dispositions in counterfactual transplants, with statistical tests described in Methods “Feature comparisons”.

The counterfactual pairs selected by the CQL model did not significantly differ from true successful donor-recipient pairs with respect to number of donations after circulatory death (DCD) donors, donor pH, sodium, and transaminase levels, candidate

encephalopathy, need for dialysis, and need for mechanical ventilation (Figure 11). Pairs selected by the model tended to have lower candidate MELD scores. In contrast, the MELD threshold models chose pairs with higher MELD scores, more candidates requiring mechanical ventilation and dialysis, candidates more likely to have portal vein thrombosis and bacterial peritonitis, and poor-quality donors in general (more DCD, higher transaminase, lower sodium, lower pH, higher bilirubin). Remarkably, the final CQL model utilized 461 (23%) discarded organs that shared characteristics of relatively higher donor quality compared to other discards (fewer DCD, less likely to have an infection, lower transaminase levels, higher sodium, and pH; Figure 12). We also illustrate examples of simulated case-level trajectories that indicate transplant decisions being made when there is low candidate MELD, higher donor PH, lower donor creatinine, lower donor age, and a closer match between donor and candidate weight and height (Figure 13).



Figure 11. Feature maps.

Comparing variables of new donor-recipient pairs selected by various models (columns) in: true graft failure candidates to true successful donor-recipient pairs (left); true waitlist mortality candidates to true successful donor-recipient pairs (center); and true graft success candidates (new selected donor/ point in MELD trajectory) to true successful donor-recipient pairs (right). If a box is white there was no significant difference in that feature (reference: true successful donor-recipient pairs). Red indicates a positive feature difference, blue a negative, and gray a non-directional effect (from comparing multi-category variables). Circle size indicates the magnitude of the effect size.

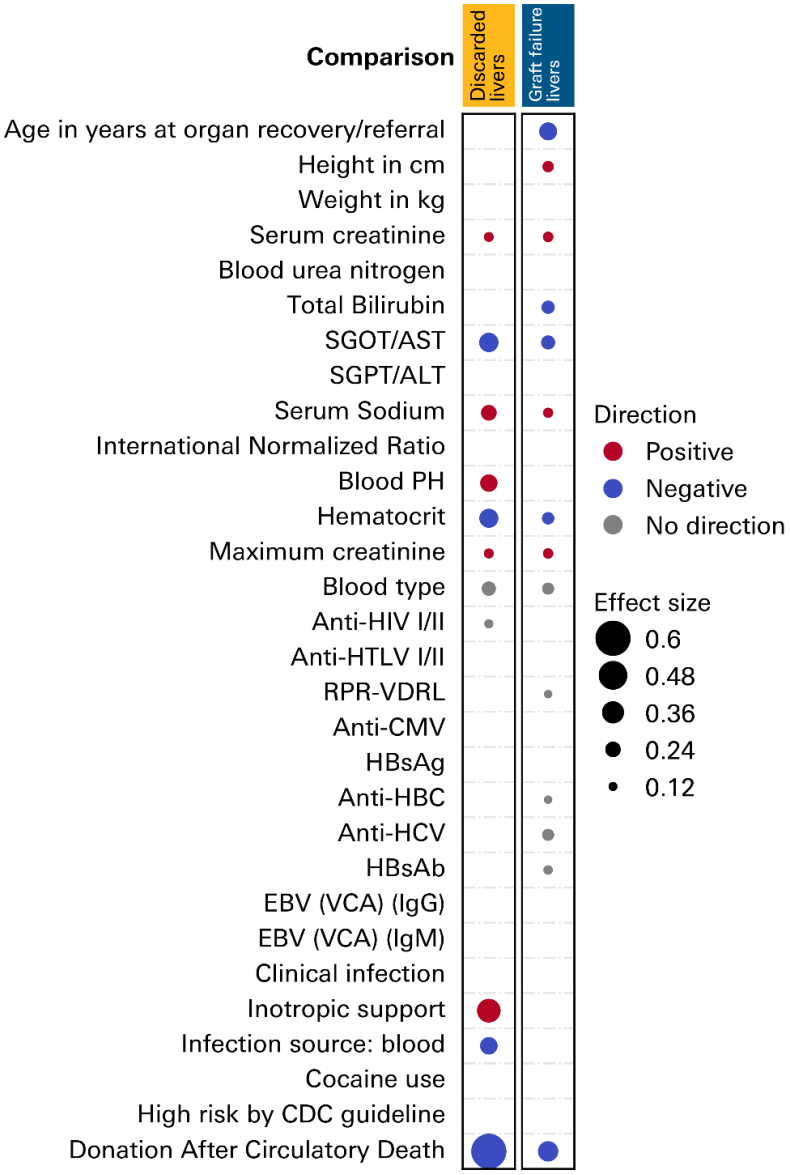


Figure 12. Feature map of discarded donors.

Feature map comparing: discarded donor livers that were selected for transplant by the final CQL model to discarded livers that were not selected (left); and donor livers that led to true graft failures that were selected for transplant by the final CQL model to donor livers that led to true graft failures that were not selected (right).

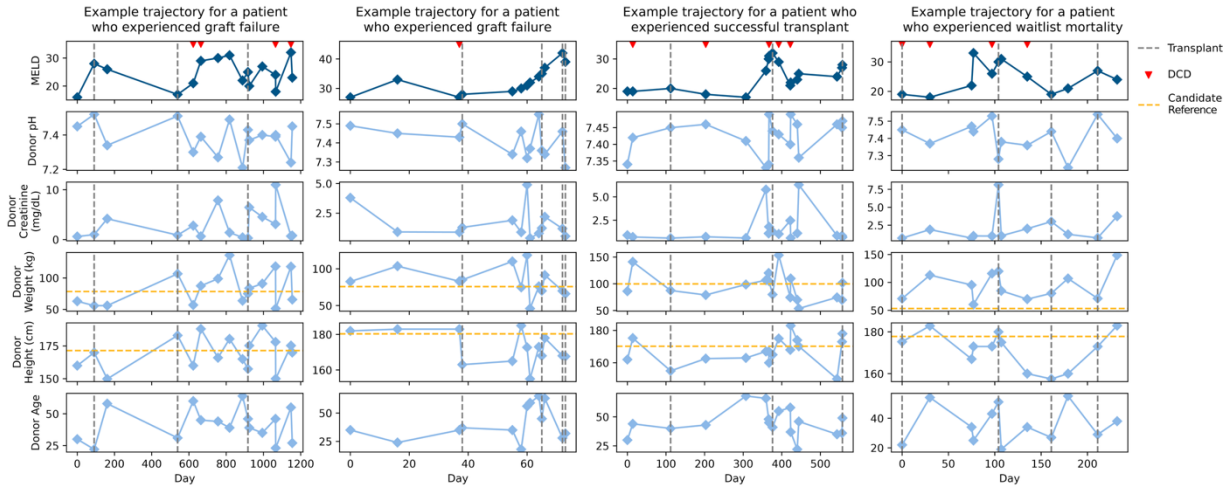


Figure 13. Patient trajectories.

Examples of patient trajectories throughout the donor-recipient matching process, and transplant decisions by the optimal CQL model. Transplant suggestions by the CQL model are indicated by the gray vertical lines. Note that these trajectories illustrate a simulated scenario where a candidate could be transplanted at multiple points in time, as a way to demonstrate the full waitlist period for a given patient. Realistically, they would be transplanted at their first indicated match. The two left columns illustrate patients who experienced graft failure. The second column to the right illustrates a patient who received a successful transplant. The rightmost column illustrates a patient who experienced waitlist mortality. Donations after circulatory death (DCD) donors are indicated by the red triangles. Select features are shown, including: candidate MELD, donor PH, donor creatinine, donor weight, donor height, and donor age over time.

Differentiation of survival in real-world donor-recipient pairs

We compared how our model discriminated post-transplant survival for those who received a transplant in real-life vs. SOFT (high-moderate, high, and futile vs. low

and low-moderate) and BAR (BAR score ≥18 vs. <18) scores. For patients indexed within the high-risk category of the SOFT/BAR scores, we determined if the patients that our model chose to transplant vs. not transplant further differed in survival. In real-world donor-recipient pairs, there was a significant post-transplant survival advantage in patients that the model chose to transplant vs. those it chose not to ($p<0.001$) (Figure 14a). In comparison, high-risk by SOFT was associated with decreased post-transplant survival in our cohort ($p=0.01$), while the BAR score was not ($p=0.11$) (Figure 14b-c). Even within high-risk patients by these scores, our model transplanted pairs with improved post-transplant survival (SOFT, $p=0.004$; BAR, $p=0.04$) (Figure 15a-b). Notably, the median SOFT and BAR scores between patients transplanted vs. not transplanted by the model in the high-risk subset of each score did not significantly differ (SOFT $p=0.20$; BAR $p=0.80$). Note that this analysis was limited to the true donor-recipient pairs that occurred; in a full simulation, the model may transplant these candidates earlier in their trajectories and/or to different donors rather than not transplant them.

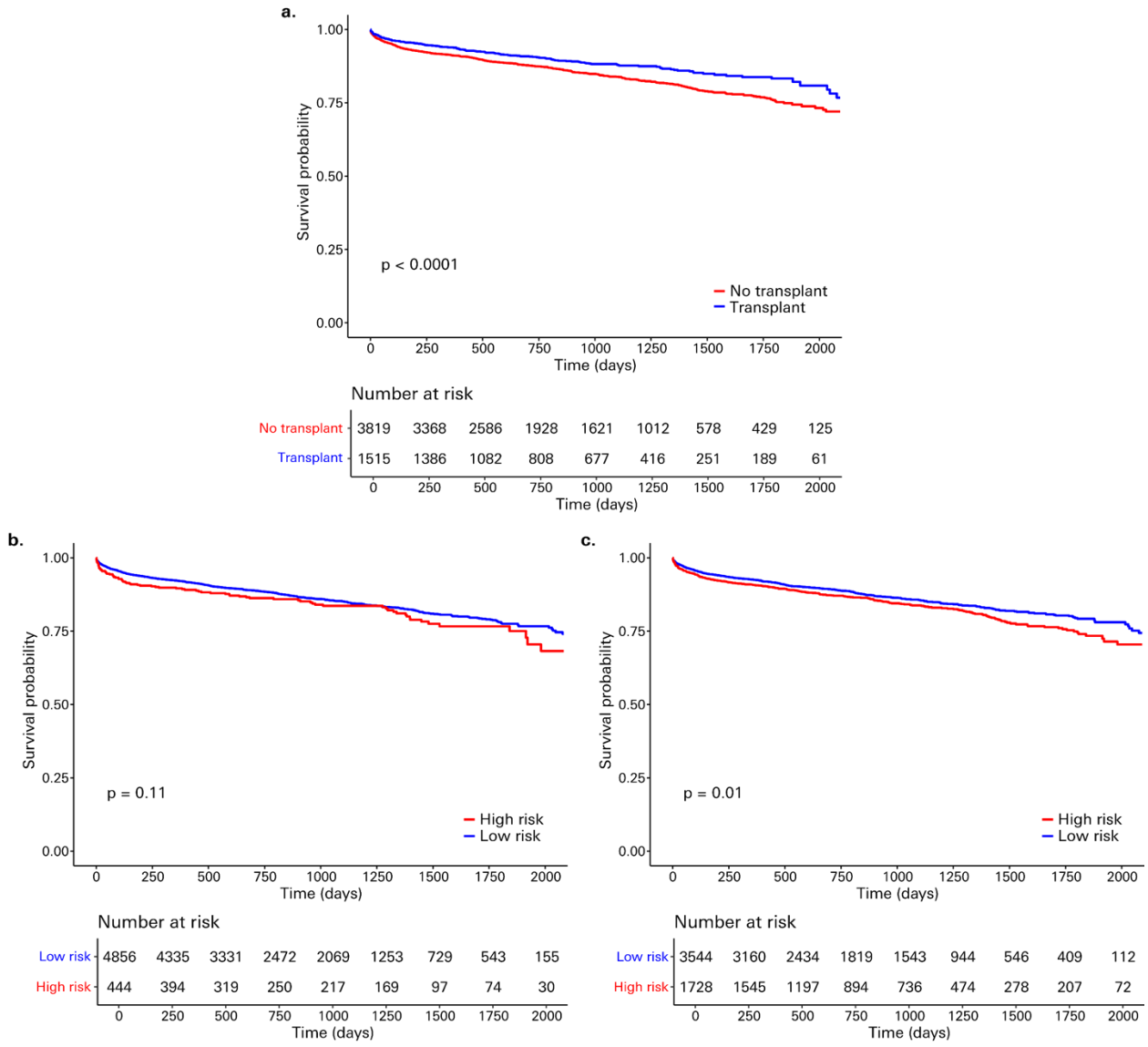


Figure 14. Post-transplant survival comparisons.

Kaplan-Meier curves comparing post-transplant survival in: a) true donor-recipient pairs selected by the CQL model for transplant to those not selected; b) low and high-risk patients based on a BAR score threshold of 18; and c) low and high-risk patients based on SOFT score designations of high-moderate, high, and futile.

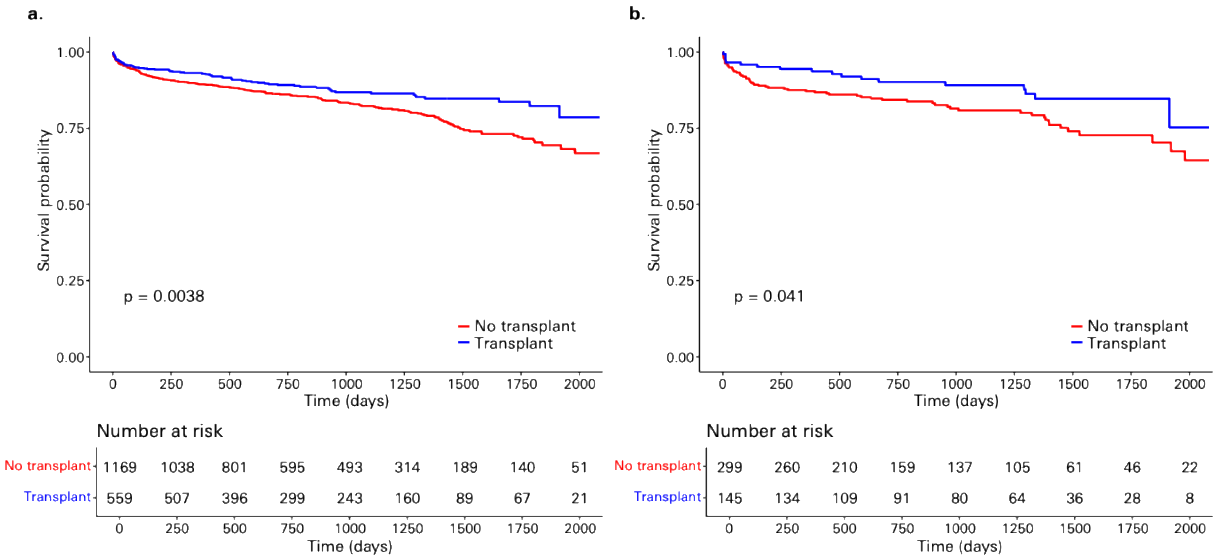


Figure 15. Post-transplant survival comparisons in high-risk patients

Kaplan-Meier curves comparing post-transplant survival in: a) true donor-recipient pairs within those designed high-moderate, high, and futile by the SOFT score; and b) true donor-recipient pairs within those designed high-risk by the BAR score (BAR ≥18).

Application in a constrained setting

We also simulated the models in a setting with more realistic resource constraints (Methods “Evaluation in a constrained setting”). In this constrained setting where each candidate and donor would be removed from analysis following transplant, death, or removal, the model transplanted 86% of graft failure patients to a different donor. Furthermore, 85% of patients who faced waitlist mortality were transplanted to possible donors, and 90% of candidates who had successful donor-recipient pairs were transplanted (Table 20). The MELD 15 threshold model transplanted 95% of graft failure patients to new donors, 91% of waitlist mortality patients, and 95% of successful transplant patients. But it achieved these results by naively transplanting nearly every

patient (i.e., when MELD ≥ 15) and chose suboptimal donor-recipient combinations (Figure 11).

Table 21. Constrained evaluation results.

Columns represent evaluation metrics with total number of actions (standard deviation) taken by each model (left panel), specific actions in graft failure patients including the percentage chosen for transplant (Tx) and delisting. Potential intervention refers to the percentage of graft failure patients transplanted to a new donor and incorrect intervention the percentage of patients transplanted to the same donor that led to the graft failure (middle panel). The percentage of patients transplanted and delisted by each model in successfully transplanted patients, those who were delisted and recovered and those who experienced waitlist mortality is shown in the three right panels. Rows represent model tested: CQL, a final conservative Q-learning based offline reinforcement learning model; PM, static predictive models trained on actions, outcomes, or transplant outcomes using logistic regression (LR) or eXtreme gradient boosting (XGB); MELD #, meld threshold models using thresholds 15, 20, 24, and 30.

	Action			Graft failures				Successful transplant		Waitlist recovered		Waitlist mortality	
Model	Wait, n (SD)	Delist, n (SD)	Tx, n (SD)	Tx, % (SD)	Delist, % (SD)	Potential intervention, % (SD)	Incorrect intervention , % (SD)	Tx, % (SD)	Delist, % (SD)	Tx , % (SD)	Delist, % (SD)	Tx, % (SD)	Delist, % (SD)
CQL	784.20 (116.89)	553.00 (26.55)	7,381.80 (122.99)	89.67 (1.36)	2.27 (0.70)	86.45 (1.70)	3.22 (0.74)	89.82 (1.10)	2.22 (0.17)	67.12 (2.84)	22.98 (1.41)	84.59 (1.90)	5.92 (0.61)
Action PM: LGR	8,032.20 (0.84)	91.00 (0.00)	595.80 (0.84)	10.43 (0.00)	0.00 (0.00)	3.32 (0.00)	7.11 (0.00)	9.23 (0.02)	0.14 (0.00)	0.47 (0.00)	6.31 (0.00)	5.59 (0.04)	0.46 (0.00)
Outcome PM: LGR	48.00 (0.00)	244.80 (4.76)	8,426.20 (4.76)	98.91 (0.13)	0.85 (0.13)	98.44 (0.13)	0.47 (0.00)	99.12 (0.03)	0.72 (0.03)	88.96 (0.15)	10.56 (0.15)	95.63 (0.07)	2.85 (0.07)
Graft failure PM: LGR	1,455.80 (630.84)	0.00 (0.00)	7,263.20 (630.84)	88.91 (4.64)	0.00 (0.00)	86.97 (4.94)	1.94 (0.31)	82.00 (7.93)	0.00 (0.00)	80.32 (9.07)	0.00 (0.00)	91.33 (4.12)	0.00 (0.00)

Action PM: XGB	2,435.60 (154.56)	668.60 (24.28)	5,614.80 (177.72)	75.26 (2.41)	2.84 (0.75)	65.26 (2.26)	10.00 (0.81)	75.12 (1.42)	1.60 (0.15)	31.26 (3.64)	35.54 (1.21)	60.97 (2.35)	8.33 (0.28)
Outcome PM: XGB	70.00 (2.45)	466.80 (3.35)	8,182.20 (3.03)	97.91 (0.26)	1.99 (0.13)	97.35 (0.39)	0.57 (0.13)	98.05 (0.05)	1.71 (0.02)	79.08 (0.38)	19.57 (0.36)	93.95 (0.33)	4.67 (0.23)
Graft failure PM: XGB	2,074.80 (1062.03)	0.00 (0.00)	6,644.20 (1062.03)	81.66 (9.17)	0.00 (0.00)	80.24 (9.26)	1.42 (0.29)	73.80 (12.46)	0.00 (0.00)	75.54 (15.05)	0.00 (0.00)	84.65 (9.17)	0.00 (0.00)
MELD 15	1,048.00 (0.00)	0.00 (0.00)	7,671.00 (0.00)	95.97 (0.00)	0.00 (0.00)	95.26 (0.00)	0.71 (0.00)	94.61 (0.00)	0.00 (0.00)	61.39 (0.00)	0.00 (0.00)	90.73 (0.00)	0.00 (0.00)
MELD 20	2,334.00 (0.00)	0.00 (0.00)	6,385.00 (0.00)	84.60 (0.00)	0.00 (0.00)	83.41 (0.00)	1.18 (0.00)	84.18 (0.00)	0.00 (0.00)	32.02 (0.00)	0.00 (0.00)	76.04 (0.00)	0.00 (0.00)
MELD 24	3,594.00 (0.00)	0.00 (0.00)	5,125.00 (0.00)	70.38 (0.00)	0.00 (0.00)	68.96 (0.00)	1.42 (0.00)	71.50 (0.00)	0.00 (0.00)	13.28 (0.00)	0.00 (0.00)	62.41 (0.00)	0.00 (0.00)
MELD 30	5,394.00 (0.00)	0.00 (0.00)	3,325.00 (0.00)	46.68 (0.00)	0.00 (0.00)	44.55 (0.00)	2.13 (0.00)	48.88 (0.00)	0.00 (0.00)	2.64 (0.00)	0.00 (0.00)	40.97 (0.00)	0.00 (0.00)

Chapter 3 Discussion

In this work, we approached the optimization of decisions in LT donor-recipient matching using offline RL. Our approach utilized the interplay of the many factors that affect the risk of graft failure including donor-recipient characteristics, candidate trajectories, and transplant timing. In contrast, the static predictive models could prevent poor outcomes (graft failure in the action prediction model; waitlist mortality in the outcome prediction model), but only at the expense of another outcome (increased waitlist mortality in the action prediction model; graft failure in the outcome prediction model). The graft failure prediction model balanced the identification of successful transplants and avoidance of graft failures, but performance was still worse than the offline RL model. Overall, static predictive models are limited to the correlation of features with an outcome, as opposed to accounting for multiple competing considerations involved in longitudinal processes of real-world decision-making. Similarly, in the MELD threshold models, we observed that with increasing threshold, graft failure was reduced by transplanting less and with a decreasing threshold, waitlist

mortality was reduced by transplanting more; the outcomes cannot be optimized simultaneously only based on candidate acuity.

As graft failure could be entirely avoided by choosing to never transplant, the prevention of graft failure must be contextualized alongside the preservation of successfully transplanted candidates. Thus, our model's ability to preserve 93% of successfully transplanted candidates while avoiding graft failure in 73% of patients is significant. Additionally, waitlist dropout challenges LT, with many candidates becoming ineligible due to severe clinical decompensation, and others being severely ill at time of listing such that transplant would not result in an acceptable outcome regardless of timing or donor. Of patients who experienced waitlist death or delisting, our model selected 47% to receive a transplant, suggesting that it identified donors and times in candidate trajectories that would maximize the likelihood of successful transplant. Markedly, existing models to predict graft failure (SOFT and BAR) are limited in their examination of only true donor-recipient pairs, precluding their application to real-world LT and considerations of waitlist mortality. While we cannot directly compare to the SOFT and BAR scores, our model differentiates survival in recipients within the high-risk classification of SOFT and BAR. Remarkably, these survival differences were not associated with differences in the values of either score, indicating that our model did not simply identify patients with lower SOFT and BAR scores for transplant, but recognized complexities indicative of improved survival that the SOFT and BAR scores were unable to discern.

A goal of offline RL is to learn counterfactual decisions (e.g., donor-recipient pairs that did not occur). The MELD threshold models serve as an important control in

this light given that they know nothing about donor-recipient pair quality and consistently transplant when the MELD score of a candidate exceeds a threshold. This analysis shows that the pairs selected by the model were of better disposition compared to the pairs selected by the MELD threshold models, with fewer feature differences with successful donor-recipient pairs. Our model also identified discarded donors that were more likely to be suitable for transplant. Notably, the reward function design assumed that these organs were non-randomly discarded and were on average worse than those used in transplants. The model thus determined that the likelihood of a successful transplant for a given candidate outweighed the risk of using a discarded organ, demonstrating a learned complexity regarding the optimization of several considerations when selecting a donor-recipient pair. Moreover, this finding highlights the potential for such a model to theoretically reduce the number of discarded livers by assisting in the selection of likely successful pairs.

In our primary analysis (Table 18), there was potential for the same donor to be offered to multiple candidates. The constrained analysis removes donors when they are transplanted and candidates after they are delisted or transplanted. Although this analysis lacks true real-world constraints, it offers important insight by more closely approximating reality. The constrained model chose to transplant 90% of those experiencing graft failure in reality, 90% of those who had true successful transplants, and 85% of those who experienced waitlist mortality. These results suggest that the model continues to optimize donor-recipient matching to reduce both graft failure and waitlist mortality and that, with an ample supply of donors, successfully matching nearly

every patient is theoretically possible. Still, an important caveat is that even in this "constrained" analysis, the donor pool was far less restricted than in reality.

The results are promising in terms of reducing waitlist mortality and early post-transplant mortality and are robust with minimal variability across cross-validation folds. Additionally, the correlation with real post-transplant survival differences between patients it selects for transplant having better post-transplant survival than those it does not improves our confidence in this approach. Our work demonstrates promising preliminary insight in the potential for an offline RL approach to address the LT donor-recipient matching problem, but it is not intended for operational use in its current form. There are several limitations that we summarize into four main themes for future work that are important requisites to improving real-world utility:

1. Incorporating additional real-world constraints such as those related to SRTR regions. To maximize the learning of meaningful donor-recipient matches, we did not incorporate the geographical restrictions considered in the true allocation policy. [10-14] However, differences in donor availability, demographics, and transplant practices across SRTR regions suggest a potential trade-off between optimizing a global model vs. tailoring to specific locations. Additionally, techniques like normothermic machine perfusion will continue to change the landscape of organ availability and allocation policy, particularly as it pertains to DCD donor allografts. [3] Due to the study period ending in May 2022, machine perfusion did not have a significant impact on the present study (1.5% of transplanted donors and <1% of discarded donors), but could be explored in future work once the data reflects wide adoption of this practice. There is also opportunity to explore inter-site variability,

optimize allocation in the context of multiple health systems in a unified network, and validate the offline RL approach in countries with different allocation policies. While we demonstrated that results remained largely consistent with slight variation across regions and demographics, future work will incorporate these factors and constraints, especially regional constraints to reflect real-world allocation and geographical limitations, as we move towards creating a comprehensive and pragmatic clinical tool.

2. Collecting high-quality prospective data that better reflects real-world decisions over time. Given the limitations of the SRTR dataset and assumptions required in this study, it will be necessary to collaborate with clinicians and collect prospective data that ascertains the real-life patterns of sequential decision making in LT donor-recipient matching. A core assumption in this work was the existence of unobserved decisions – that offers were made but declined throughout the duration of a candidate's time on the waitlist. While reasonable to some degree, the extent to which this assumption holds is impossible to verify in the SRTR dataset. The added perspective of how clinicians balance short- and long-term risks when making decisions will allow us to refine training and improve evaluation to ensure efficacy and alignment with clinical goals. Furthermore, granular documentation of reasons for outcomes such as waitlist death would also provide better real-world context for clinicians' decisions.
3. Exploring alternative algorithms and their interpretability. To illustrate the value of the RL approach, we maintained simplicity in elements such as the Q-function network architecture, reward function, and MDP. There remains a vast optimization space to

be fully explored and the potential for additional hyperparameter tuning in future work. There is also strong clinical motivation for accompanying explanations to assess the trustworthiness of model-based decisions. Along with these modeling advancements, future work will explore the ability to improve case-level interpretability through approaches such as feature-based explanation methods and Shapley values for RL [63].

4. Aligning with real-world settings to improve decision-making. The model is currently framed as a decision support tool where when a specific candidate is offered a potential donor liver, a recommendation to transplant, wait, or delist is given. This assessment is based on the candidate trajectory and candidate and donor features, from which the model has learned to minimize waitlist mortality and early post-transplant mortality simultaneously. However, we assumed that decisions were made independently between single candidate-donor pairs. Modeling how one organ can potentially be offered to multiple candidates (e.g., in a multi-agent system) will better represent the donor scarcity problem. Such an approach can also incorporate inter-candidate dependency, as clinicians do not reach conclusions in isolation, and decisions made for one candidate inevitably impact the potential outcomes for other patients. In reality, donor scarcity and disparities in access also exacerbate concerns related to fairness and equity of allocation. In future work, we will extend our stratified analysis to explore potential disparities and solutions through calibration and fair RL methods [64].

Overall, we applied offline RL to represent the clinical decision-making process involved in LT donor-recipient matching, improving upon existing methods for organ

allocation by reducing the rate of graft failure while preserving the rate of successful transplants. While this work provides preliminary insight, re-examining how data-driven methods are applied to donor-recipient matching through the lens of RL shows potential in advancing the design and implementation of clinical decision-making tools. Given the promising directions of current findings, we believe this motivates the exploration of next steps that bring the developed models closer to real-world use.

Chapter 3 Methods

Overview of the Scientific Registry of Transplant Recipients

This study used the Scientific Registry of Transplant Recipients (SRTR) database, which includes data on all donor, waitlisted candidates, and transplant recipients in the United States (US) (full description in Data Availability). We identified adult (age >18), first-time, non-status 1A, liver-only waitlist candidates from 1/1/2017-5/1/2022 (Figure 9a). A subset of candidates who received MELD exception points to increase their prioritization (most commonly for hepatocellular carcinoma) were excluded from analysis. The defined outcomes were waitlist death or delisting due to medical deterioration or being “too sick for transplant,” delisting due to medical condition improvement with no further need for LT, successful transplant, or graft failure (need for re-transplant) or mortality within one year of LT. Graft failure or all-cause mortality within one year is a well-established post-LT endpoint [20, 21, 65-67]. Patients who remained on the waitlist during the study period were included in analysis with no defined outcome. Clinical, laboratory, and demographic features were captured at time of listing and all reported MELD scores were collected from listing to waitlist outcome. Additionally, discarded donors with discard date within the timeframe above were

included in this study. Candidates were excluded if there was a documented date of transplant with no corresponding donor information or they had missing features. This study involved secondary analysis of de-identified registry data and was granted exemption from Institutional Review Board (IRB) approval at the University of California, Los Angeles. All research was done in accordance with the Declaration of Helsinki.

Episode processing details

In this study, each of the 43,595 candidates exhibited a unique trajectory of MELD scores over time, resulting in the generation of 43,595 processed episodes with a mean of 4.93 (SD=3.36) MELD scores per episode. Thus, the total number of unique candidate-MELD instances was 214,865, with each MELD score indicating a step in a LT MDP (Figure 9b). The possible actions at each step were to: 1) wait (i.e., the candidate remaining on the waitlist); 2) delist; or 3) transplant. Delist/transplant were terminal actions that marked the end of an episode. In the SRTR dataset, only terminal actions were observed at the final MELD scores for each candidate, and donor-recipient matches were indicated for transplants that occurred. Therefore, we assumed that obtaining an earlier MELD score was associated with the decision to keep a candidate on the waitlist and we inferred state transitions based on potential available donors at that time. For every non-transplant step in the MDP, we identified the set of actual donors and discarded organs with a recovery date closest to the time of each MELD score; if there were multiple potential donors, one was randomly sampled to represent the state antecedent to the action at that step. The set of potential donors was restricted to those that matched the blood type of the candidate. Multiple patients with the same blood type and MELDs recorded on the same day could be offered the same donor.

Altogether, the observations at each state throughout a candidate's episode were represented by their static listing features, changing MELD over time, and a new potential donor organ at each step.

One potential donor randomly sampled at each time of MELD for each candidate resulted in 214,865 as the exact number of states and transitions to be used in the offline RL analysis. However, this subset was only one instantiation of the total number of possibilities. At each time of MELD, there was a mean of 8.09 (SD=4.60) potential donors including discarded organs. Therefore, the total number of possible states including all combinations of candidates, times of MELD, and potential donors was on the order of 1.74M. When considering all possible states, the number of transitions was on the order of 13.3M. There were limitations in how our sampling applied a simplifying assumption on the dataset that deviated from the real-life decision-making process (i.e., we sampled one possible decision with one potential match at a time) and pose future work in the direction of representations that better align with reality.

Reward function design

While the design of reward functions is understood to often require a trial-and-error approach, [68-70] we followed best practices and exemplars from existing RL examples in healthcare [37, 40, 43, 48, 51]. The rewards were designed to reflect a clinician's perspective of a patient's condition while on the waitlist, as well as their final outcome. As illustrated in Figure 1b, there were three components to the reward function, corresponding to the possible actions at each step in an episode. The reward for deciding to wait and remain on the waitlist is described in Equation 1, where ®

indicates the step in the episode. A decrease/increase in MELD associated with improving/worsening patient condition was used as a positive/negative reward. The magnitude of MELD change was scaled by a factor of how close the MELD was to 24 (the median MELD score in the study population). The 1_A (x) denotes the indicator function that provided additional reward of β_1 and β_2 respectively when waiting on potential donors that led to eventual graft failures or were eventually discarded. These rewards employed the assumption that on average, such organs can be expected to be lower quality than those that were used, thus guiding the model to learn more optimal donor-recipient matches. As there are many reasons for discarding organs, we asserted that β_2<β_1 to reflect that they should be generally avoided in transplants, but less so than organs that definitively resulted in graft failure. In general, graft failure is the worst outcome due to the effective loss of at least two lives, as the decision to transplant is nearly never independent of the decision to not transplant another patient.

$$r_{1,i} = (\text{MELD}_i - \text{MELD}_{i+1}) \times \frac{min(\text{MELD}_{i+1}, 24)}{24} + \beta_1 \mathbf{1}_{\text{GF donor}}(\textbf{donor}) + \beta_2 \mathbf{1}_{\text{Discarded donor}}(\textbf{donor}) \quad (1)$$

The reward for delisting a patient from the waitlist was dependent on the outcome for the patient (Equation 2). If patients were delisted due to medical deterioration or being too sick to transplant, the action was penalized with β_3 to steer the model to alternative outcomes. Delisting for medical condition improvement was rewarded with β_4 as a positive outcome.

$$r_2 = \begin{cases} \beta_3 \text{ if died} \\ \beta_4 \text{ if recovered} \end{cases} \quad (2)$$

The reward for transplant was based on the outcome of the operation (Equation 3). Transplants that resulted in a 1-year graft failure were penalized with β_5 and transplants that led to successful transplants were rewarded with β_6.

$$r_3 = \begin{cases} \beta_5 \text{ if graft failure} \\ \beta_6 \text{ if successful transplant} \end{cases} \quad (3)$$

The final reward function was the sum of the constituents in Equations 1-3. Overall, positive rewards from terminal actions were associated with successful transplants and delisting due to candidate recovery; negative rewards were attributed to graft failures and death on the waitlist. Remaining on the waitlist resulted in a reward dependent on the subsequent change in MELD, as well as a general notion of donor quality; remaining on the waitlist when potentially matched with a discarded organ was associated with a positive reward; and a greater reward when with a donor that resulted in graft failure. We examined the sensitivity of reward function parametrizations by performing a grid search over: β_1={1.0,0.5}, β_2={0,0.1,0.25,0.5}, β_3={-1.0}, β_4={1.0}, β_5={-2.0,-4.0,-8.0}, β_6={1.0,2.0}. Our main interests (and hence focus on β_2 and β_5) were exploring the extent of contribution of including the discarded organs in the analysis (which has not been explored in this domain), and the level of penalization to apply to graft failures. The optimal parameterization was: β_1=1.0, β_2=0.25, β_3=-1.0, β_4=1.0, β_5=-4.0, β_6=1.0. Details of reward design and components are summarized in Table 21.

Table 22. Summary of reward design.

Component	Final Coefficient	Description
$(MELD_i - MELD_{i+1}) \times \frac{min(MELD_{i+1}, 24)}{24}$	--	Decrease in MELD is a positive reward. Increase in MELD is a negative reward. The magnitude of MELD change was scaled by a factor of how close the MELD was to the median MELD score in the study population, which was 24.
$\beta_1 \mathbf{1}_{\text{GF donor}}(\boldsymbol{donor})$	1.0	Positive reward for waiting on potential donors that led to eventual graft failure.
$\beta_2 \mathbf{1}_{\text{Discarded donor}}(\boldsymbol{donor})$	0.25	Positive reward for waiting on potential donors that were eventually discarded.
β_3	-1.0	Negative reward for delisting due to medical deterioration or being too sick to transplant.
β_4	1.0	Positive reward for delisting for medical condition improvement.
β_5	-4.0	Negative reward for transplants resulting in a 1-year graft failure.
β_6	1.0	Positive reward for transplants resulting in a successful transplant.

Conservative Q-learning

Offline RL methods share the problem of extrapolation error due to distribution shift between the learned model and the underlying behavior policy that collected the training data [28-30, 71-73]. Despite new approaches in both of the two key branches of techniques for offline RL (model-free vs. model-based), CQL remains a popular standard due to its strong performance on conventional benchmarks and practical implementation. The CQL algorithm combines the conventional Q-learning objective with an additional term for regularization controlled by a hyperparameter, α (Equation 4). Intuitively, the regularization term minimizes large Q-values and thereby avoids the over-estimation of out-of-distribution actions. CQL extends conventional RL with deep Q-networks (DQN), which uses neural networks to model the action-value function, allowing for continuous states. Not only do we utilize double deep Q-networks (DDQN) to mitigate overestimation errors, but we also use a modified encoder of the state space

that applies late fusion to candidate and donor embeddings generated by upstream neural networks. We used a simple architecture consisting of two hidden layers with 256 units and rectified linear unit (ReLU) activation functions. To consolidate the different candidate and donor features, we incorporated separate upstream encoders that embed the different feature vectors into a denser representation and applied late fusion to form the final embedding for Q-learning (Figure 10).

$$L(\theta) = \alpha E_{s\sim\mathcal{D}}\left[\log\sum_{a}\exp\big(Q(s,a)\big) - E_{a\sim\mathcal{D}}[Q(s,a)]\right] + L_{\text{Double Deep Q-learning}}(\theta) \quad (4)$$

We trained all models with 1.25M iterations at a batch size of 256. Given that α represents how conservative the learned model should be, we tuned this parameter across α={0,0.001,0.01,0.1,0.5,1.0}, resulting in an optimal value of α=0.1.

Experimental pipeline

To derive the final model, we applied a systematic approach of learning and validation. First, we split the dataset into 80/20 training and holdout test cohorts. On the training cohort, we performed 5-fold cross validation to identify the optimal hyperparameters as well as assess sensitivity across different randomizations of the training data. The models trained on each fold were evaluated on the holdout set to obtain standard deviations of the performance metrics. For the downstream interpretation of the models, we retrained a single model with the optimal configuration identified during cross-validation. This final model was used to perform any analysis on the feature distributions and mortality within the holdout set. Finally, we used the optimal configuration to assess sensitivity to reward function parameterizations.

Evaluation metrics

A challenge of offline RL is evaluating the learned models, as a given result can be counterfactual and non-existent in the retrospectively collected data. To conduct offline RL evaluation, we relied on metrics that offered clinical interpretations of the models. These domain-centric metrics better exhibit the value of the models with respect to clinical objectives and expectations.

As the only observable negative outcome of a transplant was graft failures, we explicitly evaluated whether the model decided against transplanting the true pairs to provide insight into the learning of “bad” matches. Specifically, given the true donor-recipient pairs from the candidates of the holdout set that experienced graft failure, we defined “graft failure prevention rate” (GFPR) as the percent that the model decided not to transplant. Beyond explicitly evaluating the true pair, we also assessed the ability to “avoid” a graft failure by transplanting a new pair resulting in potential improvements to clinical outcomes. We defined “graft failure intervention rate” (GFIR) as those who were matched with earlier donors, summarized as the percent of candidates who experienced graft failure that were matched with an alternative donor. Conversely, observable positive outcomes were successful transplants, and we evaluated if candidates from these pairs were transplanted with their true donor or an earlier donor, assuming that they were amenable to other potential donors. Given real-life or earlier pairs from the candidates of the holdout set that experienced a successful transplant, we defined “graft success preservation rate” (GSPR) as the percent that the model decided to transplant. Both GFPR and GSPR are forms of sensitivity.

For candidates who experienced waitlist mortality, the percent that were recommended alternative interventions as earlier transplants was defined as the

“correct intervention rate” (CIR). The percent of recovered delisted patients that the model also decided to delist was summarized as the “correct removal rate” (CRR). For candidates who recovered, we assessed whether they were still delisted, potentially at an earlier timepoint. Given such metrics reflect the identification of counterfactual actions, their absolute values were less important than the insight they provided in exposing deviation from real-life decisions.

In the decision of selecting an “optimal” model, we used the metrics that best embodied the goals of clinicians. The main objectives were to simultaneously maintain a low number of transplants, avoid the transplant decisions that led to graft failures, and maintain the transplants that led to (or could lead to) successful transplants. Concretely, we selected an optimal model that maintained a high GFPR, a high GSPR, and a low (but non-zero) number of transplant actions.

In addition to the clinically-focused metrics, we also utilized more conventional measures in RL to assess overfit during training. We observed the temporal-difference loss, conservative loss, and average value estimation over the duration of the training process. [74, 75] Indicators of overfit included converging loss with diverging value estimations, suggesting that the model learned to overestimate out-of-distributions action-values. Ideally, the loss should decrease with a slow increase in value estimation. We observed this trend with $\alpha=0.1$, while also observing signs of overestimation with lower α and underfitting with higher α – all indicative of the utility of CQL and the existence of an optimal α that balances unseen value estimation and conservation.

While other popular metrics exist for off-policy evaluation, we elected for more clinically-driven metrics to improve interpretability and correspondence to clinical goals, which is desired in the field of RL for healthcare. [62, 76] For example, weighted importance sampling (WIS) is used to provide an estimation of the value of a target with respect to the true behavior policy. However, in practice, the behavior policy is often unknown and dependent on additional estimations through methods such as regression or behavior cloning to learn a distribution of actions on the states. [62, 76, 77] For example, we demonstrate superiority of the CQL model with respect to different behavior policies when measuring performance with WIS (Table 22). But even with an ideal behavior policy, interpretation is challenging when trajectories deviate significantly, and thus evaluation is difficult in complex problems such as donor-recipient matching. [62, 76, 78] Instead, inspired by WIS, we quantified the extent to which decisions matched the actions recorded in the data, but agnostic to the outcomes and value estimation of the counterfactual decisions; we defined the “action match rate” as the percent of total actions from the model that matched the decisions made in real-life. We also summarized the counts of actions across the holdout dataset to provide insight into the overall tendencies of each model.

Table 23. Weighted importance sampling (WIS) metric of the optimal conservative Q-learning model with respect to different behavior policies.

Behavior Policy	WIS (SD)
Logistic regression	3.06 (0.41)
XGBoost	2.12 (0.69)
Behavior Cloning	1.27 (0.31)

Baseline comparisons to a MELD threshold

We compared against naïve models denoted by “MELD k” where k represents a predefined MELD score threshold. These models simulated the scenario of an organ being transplanted to a candidate any time their MELD score exceeded the threshold. Given that there is currently no true reference for donor-recipient matching over time, we used the MELD score as it is the existing standard for candidate prioritization. While deviating from reality, the MELD threshold models provided an important reference to a decision-making process that only used candidate disposition and predictions of waitlist mortality without considering the quality of the donor-recipient pair or the longitudinal trajectory of the candidate. Multiple thresholds of k ={15,20,24,30} were used to encapsulate a spectrum of conservative to non-conservative models based solely on MELD score. These thresholds were chosen primarily from a clinical expertise, as a MELD of 15 is the minimum score typically needed for placement on the LT waitlist and above 30 is considered high-risk for 90-day waitlist mortality. A MELD of 24 was the median score in the study population, while a score of 20 was chosen arbitrarily as an additional point of granularity to observe model behavior when forced to transplant between the high and low thresholds.

Baseline comparisons to static predictive models

We also contrasted the RL approach against more traditional static predictive models that focused on the supervised learning of output variables. These models employed pattern recognition to map input features to output values at single timepoints, without considering the short- and long-term risks that are characteristic of the RL approach. We used multivariate logistic regression (LGR) as a simpler statistical

method and eXtreme gradient boosting (XGB) as a more complex nonlinear model. [79] For each architecture, we trained three separate models with different objectives. First, we used actions as the dependent variable to learn “action prediction models” that predict actions given states. Next, we trained on terminal outcomes (i.e., successful transplant, delist recovery, graft failure/mortality), representing an “outcome prediction model” that predicts outcomes supervised using only the true, observable data. Lastly, we limited the dataset to the donor-recipient matches that occurred in real life along with their corresponding outcomes, to create “graft failure prediction models” that illustrate the limitations of approaches in the literature that only focus on predicting graft failure. All predictive models used the same dataset splits and hyperparameter tuning pipeline as the RL approach.

Feature comparisons

To statistically compare features, we used the two-sided Student’s t-test/Mann-Whitney U test for normally/non-normally distributed continuous variables (using the Shapiro-Wilk test to test for normality), the Fischer exact test for discrete binary variables, and the Chi-squared test for discrete multi-category variables. Effect sizes were respectively calculated using Hedge’s g, Cohen’s h, and Cramer’s V.

Evaluation in a constrained setting

In the primary analysis, it was possible for the same donor to be offered to multiple candidates even if a decision was made to transplant it. Decisions made on a single candidate in their trajectory would also be independent of all other decisions, meaning that a candidate could be chosen to be delisted or transplanted at one point and subsequently could be transplanted or delisted again. This paradigm was

necessary during training to preserve independence assumptions, and during evaluation to force the model to make a decision on the terminal state that would include the true pair if the patient was transplanted. The analysis of the true pair was essential to assess the model's decision making on the real-life pairs that had associated outcomes.

As this approach deviates from reality, we designed a subsequent evaluation in a constrained setting that better reflects the decision-making process that occurs in real-life. We imposed a limitation that a donor could only be transplanted once and was removed as soon as they were transplanted. If a donor was matched with multiple candidates, then it was offered to the closest candidate in time. We also enforced that candidates were removed from the waitlist if they were transplanted or delisted, and not eligible for any future decisions. The collected metrics included transplant and removal rates for the subgroups of candidates with different outcomes who were transplanted or delisted.

Data availability

The data that support the findings of this study are available from the SRTR, which leverages an Internet-based system called Unet (SM) to for data collection. Restrictions apply to the availability of these data, which were used under licence for the current study, and so are not publicly available. Data are however available upon reasonable request to the SRTR and compliance with their data use agreements.

Code availability

The corresponding code used in this work and examples are publicly available at https://github.com/uclamii/offline_liver_transplant. To the best of our ability, we adhered to the TRIPOD-AI guidelines for the development of prediction models that use ML methods (Table 24).

Table 24. Table 23. TRIPOD-AI checklist.

Section/Topic	Item	Checklist Item	Reported
Title	1	Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted	1
Abstract	2	See TRIPOD+AI for Abstracts checklist	2
Background	3a	Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model, including references to existing models	3
	3b	Describe the target population and the intended purpose of the prediction model in the context of the care pathway, including its intended users (e.g., healthcare professionals, patients, public)	3
	3c	Describe any known health inequalities between sociodemographic groups	--
Objectives	4	Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both)	5
Data	5a	Describe the sources of data separately for the development and evaluation datasets (e.g., randomized trial, cohort, routine care or registry data), the rationale for using these data, and representativeness of the data	5
	5b	Specify the dates of the collected participant data, including start and end of participant accrual; and, if applicable, end of follow-up	5
Participants	6a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including the number and location of centers	5
	6b	Describe the eligibility criteria for study participants	5
	6c	Give details of any treatments received, and how they were handled during model development or evaluation, if relevant	5
Data preparation	7	Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups	6
Outcome	8a	Clearly define the outcome that is being predicted and the time horizon, including how and when assessed, the rationale for choosing this outcome, and whether the method of outcome assessment is consistent across sociodemographic groups	6
	8b	If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors	6
	8c	Report any actions to blind assessment of the outcome to be predicted	6
Predictors	9a	Describe the choice of initial predictors (e.g., literature, previous models, all available predictors) and any pre-selection of predictors before model building	6
	9b	Clearly define all predictors, including how and when they were	6

		measured (and any actions to blind assessment of predictors for the outcome and other predictors)	
	9c	If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors	6
Sample size	10	Explain how the study size was arrived at (separately for development and evaluation), and justify that the study size was sufficient to answer the research question. Include details of any sample size calculation	5
Missing data	11	Describe how missing data were handled. Provide reasons for omitting any data	5
Analytical methods	12a	Describe how the data were used (e.g., for development and evaluation of model performance) in the analysis, including whether the data were partitioned, considering any sample size requirements	6,20
	12b	Depending on the type of model, describe how predictors were handled in the analyses (functional form, rescaling, transformation, or any standardization)	6
	12c	Specify the type of model, rationale, all model-building steps, including any hyperparameter tuning, and method for internal validation	14-20
	12d	Describe if and how any heterogeneity in estimates of model parameter values and model performance was handled and quantified across clusters (e.g., hospitals, countries)	12
	12e	Specify all measures and plots used (and their rationale) to evaluate model performance (e.g., discrimination, calibration, clinical utility) and, if relevant, to compare multiple models	18
	12f	Describe any model updating (e.g., recalibration) arising from the model evaluation, either overall or for particular sociodemographic groups or settings	--
	12g	For model evaluation, describe how the model predictions were calculated (e.g., formula, code, object, application programming interface)	23
Class imbalance	13	If class imbalance methods were used, state why and how this was done, and any subsequent methods to recalibrate the model or the model predictions	--
Fairness	14	Describe any approaches that were used to address model fairness and their rationale	--
Model output	15	Specify the output of the prediction model (e.g., probabilities, classification). Provide details and rationale for any classification and how the thresholds were identified	6
Training vs evaluation	16	Identify any differences between the development and evaluation data in healthcare setting, eligibility criteria, outcome, and predictors	12
Ethical approval	17	Name the institutional research board or ethics committee that approved the study and describe the participant-informed consent or the ethics committee waiver of informed consent	5
Funding	18a	Give the source of funding and the role of the funders for the present study	30
Conflicts of interest	18b	Declare any conflicts of interest and financial disclosures for all authors	30
Protocol	18c	Indicate where the study protocol can be accessed or state that a protocol was not prepared	--
Registration	18d	Provide registration information for the study, including register name and registration number, or state that the study was not registered	--
Data sharing	18e	Provide details of the availability of the study data	24
Code sharing	18f	Provide details of the availability of the analytical code	24
Patient & Public	19	Provide details of any patient and public involvement during the design,	30

Involvement		conduct, reporting, interpretation, or dissemination of the study or state no involvement	
Participants	20a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful	--
	20b	Report the characteristics overall and, where applicable, for each data source or setting, including the key dates, key predictors (including demographics), treatments received, sample size, number of outcome events, follow-up time, and amount of missing data. Report any differences across key demographic groups	--
	20c	For model evaluation, show a comparison with the development data of the distribution of important predictors (demographics, predictors, and outcome)	--
Model development	21	Specify the number of participants and outcome events in each analysis (e.g., for model development, hyperparameter tuning, model evaluation)	5
Model specification	22	Provide details of the full prediction model (e.g., formula, code, object, API) to allow predictions in new individuals and to enable third-party evaluation and implementation	24
Model performance	23a	Report model performance estimates with confidence intervals, including for any key subgroups (e.g., sociodemographic). Consider plots to aid presentation	8
	23b	If examined, report results of any heterogeneity in model performance across clusters	--
Model updating	24	Report the results from any model updating, including the updated model and subsequent performance	--
Interpretation	25	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies	10
Limitations	26	Discuss any limitations of the study (such as a non-representative sample, sample size, overfitting, missing data) and their effects on any biases, statistical uncertainty, and generalizability	12-13
Usability of the model	27a	Describe how poor quality or unavailable input data should be assessed and handled when implementing the prediction model	13
	27b	Specify whether users will be required to interact in the handling of the input data or use of the model, and what level of expertise is required of users	13
	27c	Discuss any next steps for future research, with a specific view to applicability and generalizability of the model	13

Chapter 3 Acknowledgements

AATB was supported by NIH R01 EB027650 and NIH R01 CA226079; JF was supported by NIH T32 EB016640. This study used data from the SRTR. The SRTR data system includes data on all donor, wait-listed candidates, and transplant recipients in the US, submitted by the members of the Organ Procurement and Transplantation network (OPTN). The Health Resources and Services Administration (HRSA), U.S.

Department of Health and Human Services provides oversight to the activities of the OPTN and SRTR contractors. The data reported here have been supplied by the Hennepin Healthcare Research Institute (HHRI) as the contractor for the SRTR. The interpretation and reporting of these data are the responsibility of the authors and in no way should be seen as an official policy of or interpretation by the SRTR or the U.S. Government.

Chapter 3 References

[1] A. J. Kwong et al., “OPTN/SRTR 2022 annual data report: liver,” American Journal of Transplantation, vol. 24, no. 2, pp. S176-S265, 2024, doi: 10.1016/j.ajt.2024.01.014.

[2] A. Kwong et al., “OPTN/SRTR 2018 annual data report: liver,” American Journal of Transplantation, vol. 20, pp. 193-299, 2020, doi: 10.1111/ajt.15674.

[3] T. Ivanics, P. Abreu, E. De Martin, and G. Sapisochin, “Changing trends in liver transplantation: challenges and solutions,” Transplantation, vol. 105, no. 4, pp. 743-756, 2021, doi: 10.1097/TP.0000000000003454.

[4] R. B. Freeman Jr et al., “The new liver allocation system: moving toward evidence-based transplantation policy,” Liver Transplantation, vol. 8, no. 9, pp. 851-858, 2002, doi: 10.1053/jlts.2002.35927.

[5] P. S. Kamath and W. R. Kim, “The model for end-stage liver disease (MELD),” Hepatology, vol. 45, no. 3, pp. 797-805, Mar 2007, doi: 10.1002/hep.21563.

[6] E. Lee, C. J. Johnston, and G. C. Oniscu, “The trials and tribulations of liver allocation,” Transplant International, vol. 33, no. 11, pp. 1343-1352, 2020, doi: 10.1111/tri.13710.

[7] A. Pillai, T. Couri, and M. Charlton, “Liver allocation policies in the USA: past, present, and the future,” Digestive diseases and sciences, vol. 64, pp. 985-992, 2019, doi: 10.1007/s10620-019-05549-y.

[8] G. Jung, J. Encke, J. Schmidt, and A. Rahmel, “Model for end-stage liver disease: new basis of allocation for liver transplantations,” Der Chirurg, vol. 79, pp. 157-163, 2008, doi: 10.1007/s00104-008-1463-4.

[9] S. W. Biggins et al., “Evidence-based incorporation of serum sodium concentration into MELD,” Gastroenterology, vol. 130, no. 6, pp. 1652-60, May 2006, doi: 10.1053/j.gastro.2006.02.010.

[10] E. Giorgakis et al., “Acuity circles allocation policy impact on waitlist mortality and donation after circulatory death liver transplantation: A nationwide retrospective analysis,” Health Sci Rep, vol. 6, no. 2, p. e1066, Feb 2023, doi: 10.1002/hsr2.1066.

[11] D. S. Goldberg, M. Levine, S. Karp, R. Gilroy, and P. L. Abt, “Share 35 changes in center-level liver acceptance practices,” Liver Transpl, vol. 23, no. 5, pp. 604-613, May 2017, doi: 10.1002/lt.24749.

[12] C. T. Nicolas et al., “Liver transplantation after share 35: Impact on pretransplant and posttransplant costs and mortality,” Liver Transpl, vol. 23, no. 1, pp. 11-18, Jan 2017, doi: 10.1002/lt.24641.

[13] A. Wey et al., “The effect of acuity circles on deceased donor transplant and offer rates across model for end-stage liver disease scores and exception statuses,” Liver Transplantation, vol. 28, no. 3, pp. 363-375, 2022, doi: 10.1002/lt.26286.

[14] A. M. Burton and D. S. Goldberg, “Center-level and region-level variations in liver transplantation practices following acuity circles policy change,” American journal of transplantation, vol. 22, no. 11, pp. 2668-2674, 2022, doi: 10.1111/ajt.17131.

[15] J. Torabi et al., “A decade of liver transplantation in the United States: Drivers of discard and underutilization,” Transplantation direct, vol. 10, no. 6, p. e1605, 2024, doi: 10.1097/TXD.0000000000001605.

[16] T. Berger et al., “Post-liver transplantation anemia and its correlation with mortality and graft failure,” Digestive Diseases and Sciences, vol. 65, no. 10, pp. 3040-3051, 2020, doi: 10.1007/s10620-019-06021-7.

[17] L. Haddad et al., “Factors associated with mortality and graft failure in liver transplants: a hierarchical approach,” PloS one, vol. 10, no. 8, p. e0134874, 2015, doi: 10.1371/journal.pone.0134874.

[18] P. Dutkowski et al., “Are there better guidelines for allocation in liver transplantation?: A novel score targeting justice and utility in the model for end-stage liver disease era,” Annals of surgery, vol. 254, no. 5, pp. 745-754, 2011, doi: 10.1097/SLA.0b013e3182365081.

[19] A. Rana et al., “Survival Outcomes Following Liver Transplantation (SOFT) Score: A Novel Method to Predict Patient Survival Following Liver Transplantation,”

American Journal of Transplantation, vol. 8, no. 12, pp. 2537-2546, 2008/12/01/ 2008, doi: 10.1111/j.1600-6143.2008.02400.x.

[20] S. Feng et al., “Characteristics associated with liver graft failure: the concept of a donor risk index,” American journal of transplantation, vol. 6, no. 4, pp. 783-790, 2006, doi: 10.1111/j.1600-6143.2006.01242.x.

[21] L. R. Wingfield, C. Ceresa, S. Thorogood, J. Fleuriot, and S. Knight, “Using artificial intelligence for predicting survival of individual grafts in liver transplantation: a systematic review,” Liver transplantation, vol. 26, no. 7, pp. 922-934, 2020, doi: 10.1002/lt.25772.

[22] V. F. Hughes, D. G. Melvin, M. Niranjan, G. A. Alexander, and A. K. Trull, “Clinical validation of an artificial neural network trained to identify acute allograft rejection in liver transplant recipients,” Liver Transplantation, vol. 7, no. 6, pp. 496-503, 2001, doi: 10.1053/jlts.2001.24642.

[23] J. Rajanayagam, E. Frank, R. Shepherd, and P. Lewindon, “Artificial neural network is highly predictive of outcome in paediatric acute liver failure,” Pediatric transplantation, vol. 17, no. 6, pp. 535-542, 2013, doi: 10.1111/petr.12100.

[24] A. Zare et al., “A neural network approach to predict acute allograft rejection in liver transplant recipients using routine laboratory data,” Hepatitis Monthly, vol. 17, no. 12, 2017, doi: 10.5812/hepatmon.55092.

[25] M. Dorado-Moreno, M. Pérez-Ortiz, P. A. Gutiérrez, R. Ciria, J. Briceño, and C. Hervás-Martínez, “Dynamically weighted evolutionary ordinal neural network for solving

an imbalanced liver transplantation problem," Artificial Intelligence in Medicine, vol. 77, pp. 1-11, 2017, doi: 10.1016/j.artmed.2017.02.004.

[26] M. D. Ayllón et al., "Validation of artificial neural networks as a methodology for donor-recipient matching for liver transplantation," Liver Transplantation, vol. 24, no. 2, pp. 192-203, 2018, doi: 10.1002/lt.24870.

[27] M. A. Wiering and M. Van Otterlo, "Reinforcement learning," Adaptation, learning, and optimization, vol. 12, no. 3, p. 729, 2012, doi: 10.1007/978-3-642-27645-3.

[28] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," arXiv, 2020, doi: 10.48550/arXiv.2005.01643.

[29] J. Fu, A. Kumar, O. Nachum, G. Tucker, and S. Levine, "D4RL: Datasets for Deep Data-Driven Reinforcement Learning," arXiv, 2021/02/06/ 2021, doi: 10.48550/arXiv.2004.07219.

[30] R. F. Prudencio, M. R. O. A. Maximo, and E. L. Colombini, "A Survey on Offline Reinforcement Learning: Taxonomy, Review, and Open Problems," Ieee T Neur Net Lear, vol. 35, no. 8, pp. 10237-10257, Aug 2024, doi: 10.1109/Tnnls.2023.3250269.

[31] S. Fujimoto, D. Meger, and D. Precup, "Off-policy deep reinforcement learning without exploration," arXiv, pp. 2052-2062, 2019, doi: 10.48550/arXiv.1812.02900.

[32] A. Kumar, J. Fu, M. Soh, G. Tucker, and S. Levine, "Stabilizing off-policy q-learning via bootstrapping error reduction," arXiv, 2019, doi: 10.48550/arXiv.1906.00949.

[33] A. Kumar, J. Hong, A. Singh, and S. Levine, “When should we prefer offline reinforcement learning over behavioral cloning?,” arXiv, 2022, doi: 10.48550/arXiv.2204.05618.

[34] Y. Wu, G. Tucker, and O. Nachum, “Behavior regularized offline reinforcement learning,” arXiv, 2019, doi: 10.48550/arXiv.1911.11361.

[35] C. Barata et al., “A reinforcement learning model for AI-based decision support in skin cancer,” (in en), Nat Med, pp. 1-6, 2023/07/27/ 2023, doi: 10.1038/s41591-023-02475-5.

[36] X. Feng et al., “Reinforcement Learning Model for Managing Noninvasive Ventilation Switching Policy,” IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 8, pp. 4120-4130, 2023/08// 2023, doi: 10.1109/JBHI.2023.3274568.

[37] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, “The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care,” (in en), Nat Med, vol. 24, no. 11, pp. 1716-1720, 2018/11// 2018, doi: 10.1038/s41591-018-0213-5.

[38] Z. Lu, J. Liu, R. Luo, and C. Li, “Reinforcement Learning with Balanced Clinical Reward for Sepsis Treatment,” (in en), Artificial Intelligence in Medicine: 22nd International Conference, pp. 161-171, 2024 2024, doi: 10.1007/978-3-031-66538-7_17.

[39] S. Nemati, M. M. Ghassemi, and G. D. Clifford, “Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach,” 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2978-2981, 2016/08// 2016, doi: 10.1109/EMBC.2016.7591355.

[40] A. Peine et al., “Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care,” (in en), npj Digit. Med., vol. 4, no. 1, pp. 1-12, 2021/02/19/ 2021, doi: 10.1038/s41746-021-00388-6.

[41] A. Raghu, M. Komorowski, I. Ahmed, L. Celi, P. Szolovits, and M. Ghassemi, “Deep Reinforcement Learning for Sepsis Treatment,” arXiv, 2017/11/27/ 2017, doi: 10.48550/arXiv.1711.09602.

[42] X. Sun et al., “Effective Treatment Recommendations for Type 2 Diabetes Management Using Reinforcement Learning: Treatment Recommendation Model Development and Validation,” J Med Internet Res, vol. 23, no. 7, p. e27858, 2021/07/22/ 2021, doi: 10.2196/27858.

[43] G. Wang et al., “Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial,” (in en), Nat Med, pp. 1-10, 2023/09/14/ 2023, doi: 10.1038/s41591-023-02552-9.

[44] X. Wu, R. Li, Z. He, T. Yu, and C. Cheng, “A value-based deep reinforcement learning model with human expertise in optimal treatment of sepsis,” (in en), npj Digit. Med., vol. 6, no. 1, pp. 1-12, 2023/02/02/ 2023, doi: 10.1038/s41746-023-00755-5.

[45] A. Yala et al., “Optimizing risk-based breast cancer screening policies with reinforcement learning,” (in en), Nat Med, vol. 28, no. 1, pp. 136-143, 2022/01// 2022, doi: 10.1038/s41591-021-01599-w.

[46] C. Yu, J. Liu, S. Nemati, and G. Yin, “Reinforcement Learning in Healthcare: A Survey,” ACM Comput. Surv., vol. 55, no. 1, pp. 5:1-5:36, 2021/11/23/ 2021, doi: 10.1145/3477600.

[47] L. Yucheng et al., “Development and validation of a deep reinforcement learning algorithm for auto-delineation of organs at risk in cervical cancer radiotherapy,” (in en), Sci Rep, vol. 15, no. 1, p. 6800, 2025/02/25/ 2025, doi: 10.1038/s41598-025-91362-9.

[48] K. Zhang et al., “An interpretable RL framework for pre-deployment modeling in ICU hypotension management,” (in en), npj Digit. Med., vol. 5, no. 1, pp. 1-10, 2022/11/18/ 2022, doi: 10.1038/s41746-022-00708-4.

[49] P. Ghasemi, M. Greenberg, D. A. Southern, B. Li, J. A. White, and J. Lee, “Personalized decision making for coronary artery disease treatment using offline reinforcement learning,” npj Digit. Med., vol. 8, no. 1, p. 99, 2025, doi: 10.1038/s41746-025-01498-1.

[50] T. Zhu, K. Li, and P. Georgiou, “Offline deep reinforcement learning and off-policy evaluation for personalized basal insulin control in type 1 diabetes,” IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 10, pp. 5087-5098, 2023, doi: 10.1109/JBHI.2023.3303367.

[51] M. Nambiar, S. Ghosh, P. Ong, Y. E. Chan, Y. M. Bee, and P. Krishnaswamy, “Deep Offline Reinforcement Learning for Real-world Treatment Optimization Applications,” Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 4673-4684, 2023, doi: 10.1145/3580305.3599800.

[52] P. Jayaraman, J. Desman, M. Sabounchi, G. N. Nadkarni, and A. Sakhuja, “A Primer on Reinforcement Learning in Medicine for Clinicians,” npj Digit. Med., vol. 7, no. 1, Nov 26 2024, doi: 10.1038/s41746-024-01316-0.

[53] Z. Wang, H. Zhao, P. Ren, Y. Zhou, and M. Sheng, “Learning Optimal Treatment Strategies for Sepsis Using Offline Reinforcement Learning in Continuous Space,” arXiv, 2022/07/14/ 2022, doi: 10.48550/arXiv.2206.11190.

[54] T. Beolet, A. Adenis, E. Huneker, and M. Louis, “End-to-end Offline Reinforcement Learning for Glycemia Control,” arXiv, 2023/10/16/ 2023, doi: 10.48550/arXiv.2310.10312.

[55] F. Kondrup et al., “Towards Safe Mechanical Ventilation Treatment Using Deep Offline Reinforcement Learning,” arXiv, 2022/10/05/ 2022, doi: 10.48550/arXiv.2210.02552.

[56] H. Emerson, M. Guy, and R. McConville, “Offline Reinforcement Learning for Safer Blood Glucose Control in People with Type 1 Diabetes,” arXiv, 2023/05/05/ 2023, doi: 10.48550/arXiv.2204.03376.

[57] S. Joshi, S. Parbhoo, and F. Doshi-Velez, “Learning-to-defer for sequential medical decision-making under uncertainty,” arXiv, 2022/12/05/ 2022, doi: 10.48550/arXiv.2109.06312.

[58] R. Liu, J. L. Greenstein, J. C. Fackler, J. Bergmann, M. M. Bembea, and R. L. Winslow, “Offline reinforcement learning with uncertainty for treatment strategies in sepsis,” arXiv, 2021/09/22/ 2021, doi: 10.48550/arXiv.2107.04491.

[59] D. H. Choi et al., “Individualized decision making in on-scene resuscitation time for out-of-hospital cardiac arrest using reinforcement learning,” (in en), npj Digit. Med., vol. 7, no. 1, pp. 1-14, 2024/10/09/ 2024, doi: 10.1038/s41746-024-01278-3.

[60] A. Kumar, A. Zhou, G. Tucker, and S. Levine, “Conservative Q-Learning for Offline Reinforcement Learning,” arXiv, 2020/08/19/ 2020, doi: 10.48550/arXiv.2006.04779.

[61] T. Yu, A. Kumar, R. Rafailov, A. Rajeswaran, S. Levine, and C. Finn, “COMBO: Conservative Offline Model-Based Policy Optimization,” arXiv, 2022/01/27/ 2022, doi: 10.48550/arXiv.2102.08363.

[62] O. Gottesman et al., “Evaluating Reinforcement Learning Algorithms in Observational Health Settings,” arXiv, 2018/05/31/ 2018, doi: 10.48550/arXiv.1805.12298.

[63] D. Beechey, T. M. Smith, and Ö. Şimşek, “Explaining reinforcement learning with shapley values,” arXiv, pp. 2003-2014, 2023, doi: 10.48550/arXiv.2306.05810.

[64] P. Gajane, A. Saxena, M. Tavakol, G. Fletcher, and M. Pechenizkiy, “Survey on fair reinforcement learning: Theory and practice,” arXiv, 2022, doi: 10.48550/arXiv.2205.10032.

[65] M. Zhang et al., “Mortality risk after liver transplantation in hepatocellular carcinoma recipients: a nonlinear predictive model,” Surgery, vol. 151, no. 6, pp. 889-97, Jun 2012, doi: 10.1016/j.surg.2011.12.034.

[66] K. Shimata et al., “A novel scoring system to predict short-term mortality after living donor liver transplantation for acute liver failure,” Am J Transplant, vol. 24, no. 10, pp. 1857-1867, Oct 2024, doi: 10.1016/j.ajt.2024.04.016.

[67] L. Shaffer, S. Abu-Gazala, D. E. Schaubel, P. Abt, and N. Mahmud, “Performance of risk prediction models for post-liver transplant patient and graft survival

over time," Liver Transpl, vol. 30, no. 7, pp. 689-698, Jul 1 2024, doi: 10.1097/LVT.0000000000000326.

[68] A. Abouelazm, J. Michel, and J. M. Zoellner, "A Review of Reward Functions for Reinforcement Learning in the context of Autonomous Driving," arXiv, 2024/04/12/ 2024, doi: 10.48550/arXiv.2405.01440.

[69] S. Ibrahim, M. Mostafa, A. Jnadi, and P. Osinenko, "Comprehensive Overview of Reward Engineering and Shaping in Advancing Reinforcement Learning Applications," arXiv, 2024/07/22/ 2024, doi: 10.48550/arXiv.2408.10215.

[70] H. Sowerby, Z. Zhou, and M. L. Littman, "Designing Rewards for Fast Learning," arXiv, 2022/05/30/ 2022, doi: 10.48550/arXiv.2205.15400.

[71] S. Fujimoto and S. S. Gu, "A Minimalist Approach to Offline Reinforcement Learning," (in en), arXiv, 2021, doi: 10.48550/arXiv.2106.06860.

[72] B. Kang, X. Ma, Y. Wang, Y. Yue, and S. Yan, "Improving and Benchmarking Offline Reinforcement Learning Algorithms," arXiv, 2023/06/01/ 2023, doi: 10.48550/arXiv.2306.00972.

[73] Y. Fu, D. Wu, and B. Boulet, "A Closer Look at Offline RL Agents," (in en), Proceedings of the 36th International Conference on Neural Information Processing Systems, vol. 625, pp. 8591 – 8604, 2022, doi: 10.5555/3600270.3600895.

[74] A. Kumar, A. Singh, S. Tian, C. Finn, and S. Levine, "A Workflow for Offline Model-Free Robotic Reinforcement Learning," arXiv, 2021/09/23/ 2021, doi: 10.48550/arXiv.2109.10813.

[75] T. L. Paine et al., “Hyperparameter Selection for Offline Reinforcement Learning,” arXiv, 2020/07/17/ 2020, doi: 10.48550/arXiv.2007.09055.

[76] S. Tang and J. Wiens, “Model Selection for Offline Reinforcement Learning: Practical Considerations for Healthcare Settings,” arXiv, 2021/07/23/ 2021, doi: 10.48550/arXiv.2107.11003.

[77] S. Doroudi, P. S. Thomas, and E. Brunskill, “Importance Sampling for Fair Policy Selection,” Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp. 5239 – 5243, 2017, doi: 10.5555/3304652.3304742.

[78] M. Uehara, C. Shi, and N. Kallus, “A Review of Off-Policy Evaluation in Reinforcement Learning,” arXiv, 2022/12/13/ 2022, doi: 10.48550/arXiv.2212.06355.

[79] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” J Mach Learn Res, vol. 12, pp. 2825-2830, Oct 2011, doi: 10.5555/1953048.2078195.

Chapter 4 HCC-3D: improving liver transplant allocation with the hepatocellular carcinoma dynamic-deep-drop model.

Chapter 4 Abstract

Background: The allocation of deceased donor livers to hepatocellular carcinoma (HCC) patients relies on the granting of Model-for-End-Stage Liver Disease (MELD) exception points according to the median MELD at transplant in the area around the donor hospital minus three points (MMaT-3). The current allocation system prioritizes all patients equally and fails to differentiate HCC patients on their variable risks of waitlist dropout due to tumor progression or liver decompensation. We propose the novel HCC Dynamic-Deep-Drop (HCC-3D) model and evaluate its application in a simulated liver allocation model.

Methods: Adult liver transplant candidates with HCC and at least one submitted MELD exception form from 10/08/2015 to 12/31/2023 were evaluated from the United Network for Organ Sharing database. HCC-3D was trained and evaluated with Harrell's C-index using 5-fold cross-validation and a holdout test set. The effect of implementing HCC-3D on the rates of waitlist dropout and LT in all LT candidates was assessed using discrete event simulation.

Results: There were 20,240 HCC patients available for analysis, representing 64,610 MELD exception forms. HCC-3D predicted waitlist dropout at three and six months in the holdout test set with c-indices of 0.80, 95% CI: [0.77-0.82] and 0.78, 95% CI: [0.76-0.80]. After post hoc calibration, HCC-3D had agreement between expected and observed predicted event probabilities with a mean absolute error of 0.03. In

simulation, use of HCC-3D resulted in an average of 27 fewer waitlist dropouts for HCC patients ($p<0.001$) with no significant difference in non-HCC waitlist dropouts ($p=0.78$).

Conclusions: We demonstrate state-of-the-art discriminative performance in the prediction of waitlist dropout in transplant-eligible HCC patients using a longitudinal deep survival model (HCC-3D). In simulation, HCC-3D led to fewer HCC patient waitlist dropouts and more transplants with no significant change to the dropout or transplant rate of non-HCC candidates. Implementation of HCC-3D could improve allocation of donor livers to HCC patients.

Chapter 4 Introduction

Hepatocellular carcinoma (HCC) is the most common primary liver malignancy and accounts for the third leading cause of cancer-related mortality worldwide [1-3]. Liver transplantation (LT) is the primary curative treatment modality for select patients with unresectable disease [4-8]. The Milan Criteria (MC) have been adopted for HCC patient LT eligibility and are associated with 75-80% five-year post-transplant survival [9-16]. Patients with liver-confined tumor burdens beyond MC are also eligible for Model for End-Stage Liver Disease (MELD) exception points, provided the viable tumor burden can be downstaged with the use of locoregional (LRT) or systemic therapy. Under the current paradigm, patients must have tumor burdens within MC (≥3 tumors each ≤3 cm, or one tumor ≤5cm) and undergo a mandatory six-month waiting period after listing before receiving MELD exception points at the median MELD at transplant in the area of the donor hospital minus three points (MMaT-3) [17-19]. Exception points based on MMaT-3 have helped equalize transplant rates for HCC and non-HCC patients [20].

The MMaT-3 policy has reduced waitlist dropout for HCC patients overall, but it fails to differentiate candidates with the same MMaT-3 score despite variability in tumor aggressiveness and underlying liver dysfunction that impact rates of waitlist dropout due to tumor progression or liver decompensation, respectively [21, 22]. Prior studies have consistently identified alpha-fetoprotein (AFP), tumor burden, and liver dysfunction as key predictors of waitlist death or delisting in HCC patients [23-29]. Recently, the Model of Urgency for Liver Transplantation in HCC (Multi-HCC) score was proposed to better distinguish 90-day waitlist death or delisting risk [28]. Although Multi-HCC demonstrated promising discrimination (c-index of 0.71), like many previous approaches, it is limited to static risk prediction based on initial listing tumor burden and laboratory biomarkers. The prognostic importance of evolving tumor characteristics and treatment response is well established, yet most models remain limited to static, single time point prediction [30-32]. These approaches fail to utilize the longitudinal changes HCC patients experience on the waitlist, require measurements at prespecified times (i.e., six-month landmark), and depend on a priori, engineered features such as AFP response to treatment.

Machine learning approaches have been used to predict LT waitlist death or delisting among HCC LT candidates, including a random forest model with a c-index 0.74 [33]. More advanced approaches including the use of deep learning (DL) or neural networks with multiple layers have been applied to LT waitlist prognostication [34]. Despite the use of increasingly complex models, the reliance on static risk assessment using a single time-point persists. Recurrent neural networks (RNNs) are a class of deep learning models that can model sequences over time [35, 36]. Additionally, DL has been applied to survival analysis and competing risks analyses, leveraging the

complexity of deep neural network architectures to model time-to-event outcomes commonly studied in clinical medicine [37-40]. Recently, Dynamic-DeepHit was proposed for cystic fibrosis, combining an RNN to capture temporal dependencies with a deep competing-risks survival network and achieving state-of-the-art c-indices [38].

The development of models aimed at improving LT allocation is substantially strengthened by evaluating their performance in a simulated allocation framework, such as the Liver-Simulated Allocation Model (LSAM). In this work, we propose HCC Dynamic-Deep-Drop (HCC-3D), an adaptation of the Dynamic-DeepHit architecture for predicting waitlist dropout in HCC patients with the aim of providing improved differential assessments of risk compared to current policy where exception points are assigned as the MMAT-3. Additionally, we aimed to evaluate the potential impact of HCC-3D on allocation of donor livers to all waitlist patients including its projected impact on waitlist dropout and LT using a discrete event simulation framework based on the Liver-Simulated Allocation Model (LSAM).

Chapter 4 Patients and Methods

Model training and evaluation

Adult (age > 18) patients with HCC listed for LT (excluding resection) with a MELD exception form filed from 10/08/2015-12/31/2023 in the United Network for Organ Sharing (UNOS) database were retrospectively reviewed. Exception forms were excluded if tumor size information was missing. Patients were excluded from analysis if the original presenting tumor burden information was missing and the first exception form indicated a cumulative tumor diameter of zero, indicating treatment received that was not captured in the data. In patients with no original presenting tumor burden, but

with viable tumor reported on the index form, this was considered the original presenting tumor burden. Additional exclusions were made for duplicate registrations, duplicate exception forms, previous transplant, and living donor transplant. This study involved secondary analysis of de-identified registry data and was considered IRB exempt at our institution.

Static features used in the model included sex, original presenting tumor number, original presenting cumulative tumor diameter, and original presenting tumor burden (within MC, within UNOS, all-comers). Temporal features collected at each exception form were cumulative viable tumor diameter, tumor number, AFP, LRT received, creatinine, international normalized ratio (INR), albumin, total bilirubin, and sodium. Temporal changes in cumulative viable tumor and AFP would be akin to dynamic response to LRT and AFP response as reported in previous studies using dynamic feature changes in LT [41, 42]. The outcome of interest in this study was waitlist dropout, defined as waitlist mortality or delisting due to medical decompensation (tumor progression beyond MC or liver decompensation). The UNOS database does not definitively report tumor progression as a reason for waitlist dropout so reasons of “medically unsuitable”, “died”, and “candidate condition deteriorated” were used [27]. Time to event was defined as the number of days from the exception form to transplant, dropout, or last follow-up.

Data processing was performed using R version 4.4.2. UNOS requires that MELD exception forms are submitted every three months. A longitudinal sequence was created from the exception forms for each patient up to a configurable maximum of two exception forms for model training and evaluation. If a patient only had one form

available, a second form was added where all variables had a value of zero (zero-padded). A separate variable was created to inform the model that some observations contain zeros for every feature and should be ignored. Continuous features were normalized using Z-score normalization (training set parameters used to normalize all data). Missing continuous features within an observed form (meaning the form was present but a specific data point was not recorded) were imputed with a non-informative constant and a separate variable was created to inform the model that this datapoint should be ignored. Categorical features were one-hot encoded with a separate missing classification. Follow-up time was discretized into one-month bins and the time to outcome was measured as the time from the last exception form in the sequence. If the sequence only included the first exception form, the outcome time would be the time from listing to outcome. If the sequence had two exception forms, the follow-up time would be the time from the second exception form to the outcome or last follow-up.

Modeling was performed using PyTorch (version 1.12.0). The architecture was adapted from the published Dynamic-DeepHit model and detailed in the Supplementary Appendix: Model Architecture and Loss Function Design [38]. Patients were split into 80% training and 20% test, stratified by outcome (waitlist dropout and transplant). Baseline demographics were compared between training and test sets; medians were compared using the Mann-Whitney U test and categorical features with the Chi-squared test. Models were trained using 5-fold cross-validation stratified by outcome using the 80% training set. Hyperparameters were tuned using results from cross-validation runs. The Adam optimizer was used with a learning rate of 0.002 and a batch size of 64. Models were run for 100 epochs with early stopping based on validation c-index

increase with a patience of 20. Learning rate scheduling with reduce on plateau was used with a factor of 0.7 and patience of 5.

Feature importance was assessed using Shapley additive explanations (SHAP). For all performance estimates, the risk score was the predicted CIF (cumulative probability of event) up to the time of interest (3-, 6-, and 12-months). Performance metrics included Harrell's C-index and the integrated Brier score. The best model in each fold was selected based on the highest 6-month c-index and metrics from that fold were averaged and reported with standard deviations. HCC-3D was compared to several baselines including a model based on DeepHit (the static version of Dynamic-DeepHit with no longitudinal encoding), a random survival forest model using the same features as the deep learning models, the MELD 3.0 individual components in a Cox proportional hazards model, the calculated MELD 3.0 score in a Cox model, and the Multi-HCC model. We also used published risk score formulas for the MELD-Matched Multi-HCC risk score [28], Mehta et al [27], HALT-HCC [25], deMELD [29], HCC-MELD [26], and MELD-Eq [43] and evaluated these scores without re-training. We compared our model to baselines using only the first observation of data (time of listing, negating the longitudinal aspect) as well as using sequences of length 2 (HCC-3D, DeepHit, and random survival forest models used two observations of data; Cox models used the second observation when available). As a sensitivity analysis, we also evaluated a version of HCC-3D that allowed up to three exception forms instead of two as in the primary model. All models were tested using 5-fold cross-validation with the same splits used for the assessment of each model.

The final HCC-3D model was selected from the cross-validation fold that produced the highest 6-month c-index and evaluated on the hold-out test set using the same set of performance metrics as above. Confidence intervals for test set performance were produced using 100 bootstrapped samples with replacement. To evaluate HCC-3D in a simulated allocation model, the calibration accuracy of predicted three-month waitlist dropout CIFs was assessed for all patients in the test set. The predicted CIFs were binned into quartiles. The observed waitlist dropout CIF was determined using the Aalen-Johansen estimator [44]. Calibration plots were produced comparing predicted and observed 3-month CIFs with average differences assessed using mean absolute error. A post hoc cubic smoothing spline calibrator was fit using the predicted and observed waitlist mortality CIFs in the training set. This post hoc calibrator was then applied to the test set predictions to improve agreement between predicted and observed 3-month waitlist dropout CIFs.

Allocation simulation

The Liver Simulated Allocation Model (LSAM) is a discrete event simulation software developed and maintained by the Scientific Registry of Transplant Recipients (SRTR), serving as the standard tool for assessing the impact of proposed changes to liver allocation [45-49]. However, the LSAM source code is not publicly available, making it difficult to test new concepts that require modifications beyond LSAM's capabilities. An open-source version of LSAM, called LivSim, was developed in 2018 and is publicly available [49]. However, it is not updated with MMaT-3 nor Acuity Circles [50] and is based on the donor service area (DSA) system, which is no longer relevant in the modern era. To fully assess the impact of our model on organ allocation, we

constructed a discrete event simulation model based on LSAM and LivSim. Both LSAM and LivSim use the SRTR database to reflect real-world impact, so we similarly utilized the SRTR standard analysis files for this simulation, focusing on adult liver-only candidates listed on or after 02/04/2020 to ensure all patients were subject to Acuity Circles and MMAT-3 allocation policies.

To implement our model in the simulation, sequences were extracted from HCC patient exception form data to ensure the collection of all necessary features needed as model input. There was no truncation of sequences, each patient's full set of exception form data was preserved. Patients with at least one submitted exception form received predicted 3-month waitlist mortality CIFs from the model in a sliding window fashion with a new score calculated every three months. The model used the same static features for each prediction and always the two most recent exception forms starting from the first form submitted to the last. These model outputs were used to derive MELD equivalent (M_{eq}) scores. This derivation depends on the underlying survival function inherent to MELD 3.0, with a full explanation in Supplementary Appendix: Derivation of the MELD Equivalent. Simulation with M_{eq} scores reflects the implementation of our model's continuous score and its impact when integrated with current allocation policies. As described in the Multi-HCC paper, we also wanted to provide a method for using HCC-3D to modify the existing MMaT framework. To do this, we assigned M_{eq} scores to patients in the holdout test set (UNOS) and separated them into tertiles. Recurrence-free-survival and cumulative incidence of waitlist dropout were compared among the three risk groups using the Kaplan-meier method and cumulative incidence functions and differences were assessed using the log-rank test and Fine and Gray's competing

risks regression, respectively. The M_{eq} score ranges corresponding to each risk group were used to assign exception points according to Table 24.

Table 25. Proposed implementation of HCC-3D using the existing MMaT-3 framework.

Risk score groups		Median MELD at Transplant points awarded			
		0 months	3 months	6 months	9 months
HCC-3D	Tertile 1 HCC-3D < 27	Lab MELD 3.0	Lab MELD 3.0	MMaT - 5	MMaT - 3
	Tertile 2 HCC-3D 27-29		Lab MELD 3.0	MMaT - 3	MMaT - 3
	Tertile 3 HCC-3D ≥ 30		MMAT - 3*	MMaT - 1	MMaT - 1

HCC, hepatocellular carcinoma; MELD, Model for End-Stage Liver Disease; MMaT, median MELD at transplant

* Two separate simulations were run: one allowing exception points only after 6 months and the second allowing exception points at three months for patients in the highest tertile.

The simulation was run for five years from the starting time using existing allocation policies (Acuity Circles and MMaT-3) where waitlist dropouts and transplants were compared to the actual event rates in the patients included in the simulation to ensure simulated outcomes align with actual outcomes at baseline. The simulation was then run using our derived continuous M_{eq} score where patients received their M_{eq} if they met clinical criteria (within MC or downstaged to within MC) after the 6-month waiting period. Furthermore, we simulated the scenario where patients received

exception points based on stratified M_{eq} score ranges that would modify MMaT-based point allocation (Table 24), as a potential allocation scheme that would incorporate the risk stratification by M_{eq} while preserving the equity in access to transplants provided by the current system for HCC patients. During this simulation patients could move up in risk tertile, but would never move down, preventing any loss of exception points. Given that our model only requires 3 months on the waitlist to achieve optimal discriminative performance, we simulated the use of stratified M_{eq} where modified MMaT-based exception points could only be received after 6 months and, separately, where the highest risk patients could receive points after 3 months. The simulation using MMaT-3 was compared to the actual outcomes in the simulated patients. Use of M_{eq} was compared to MMaT-3 by running 10 simulations for each policy and averaging the results. Statistical differences in simulation results were determined using Student's t-test between the simulation results for each policy. Implementation details related to the allocation simulation framework are in the Supplementary Appendix: Liver Simulated Allocation Model Analysis and Design.

To visualize how the continuous M_{eq} score would compare to the real-life allocation MELD in select groups of patients, cross tabulations were performed. These analyses examined patients in the simulation who had HCC and received exception points at the MMaT-3. The three analyses were restricted to those who had waitlist dropout, restricted to those with low-risk HCC (MELD <15, single lesion <3 cm, AFP ≤ 20 ng/mL) [51], and those patients not considered low-risk. For all analyses, allocation MELD and M_{eq} scores were binned into ranges (< 20, 20-24, 25-29, 30-34, and ≥ 35) and plots were created of allocation MELD ranges versus MELD M_{eq} score ranges.

Chapter 4 Results

Model training and evaluation

After exclusion, there were 20,240 HCC patients for analysis, representing 64,610 exception forms (Figure 16). The training/validation and test sets were well-balanced in terms of demographic information, tumor characteristics, and outcomes (Table 25). There was a median of three exception forms per patient (IQR: 2.0-4.0). The number of patients contributing data from each exception form and mean cumulative viable tumor diameter is shown in Figure 17. All patients had a non-zero original presenting tumor burden, and all patients had at least one form of data. The overall incidence of LT was 62% and waitlist dropout 20%.

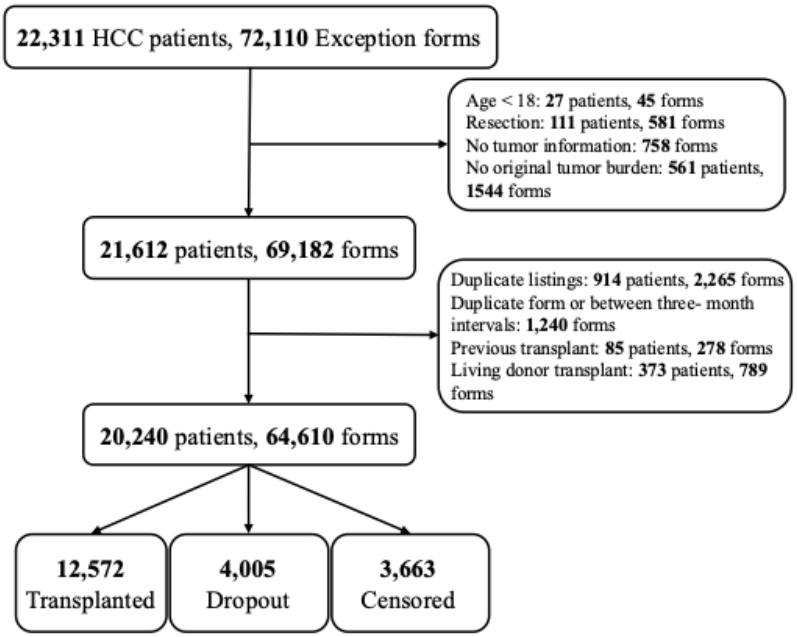


Figure 16. Patient cohort selection.

Flow diagram showing how the final cohort was selected for analysis in the United Network for Organ Sharing database including all inclusion and exclusion criteria.

Table 26. Baseline demographic comparisons between training and testing cohorts using features at time of listing.

Variable	Total (n=20,240)	Train (n=16,192)	Test (4,048)	P
Candidate demographic information				
Age, y (median [IQR])	63.00 [58.00, 66.00]	63.00 [58.00, 66.00]	63.00 [58.00, 66.00]	0.23
Sex = M (%)	15416 (76)	12340 (76)	3076 (76)	0.78
BMI, kg/m^2 (median [IQR])	28.95 [25.66, 32.87]	28.93 [25.62, 32.84]	29.03 [25.82, 33.02]	0.18
Liver disease etiology (%)				0.57
AIH/PSC/PBC	544 (3)	435 (3)	109 (3)	
Alcoholic Cirrhosis	3,778 (21)	3,060 (21)	718 (20)	
Hepatitis B	869 (5)	684 (5)	185 (5)	
Hepatitis C	6,950 (39)	5,588 (39)	1,362 (38)	
Metabolic Disease	190 (1)	153 (1)	37 (1)	
MAFLD	3,863 (21)	3,068 (21)	795 (22)	
Other	1,849 (10.2)	1,457 (10)	392 (10.9)	
MELD Score (median [IQR])	10.00 [8.00, 14.00]	10.00 [8.00, 14.00]	10.00 [8.00, 14.00]	0.91
Child-Pugh Score (median [IQR])	7.00 [5.00, 8.00]	7.00 [5.00, 8.00]	6.00 [5.00, 8.00]	0.57
Albumin, g/dL (median [IQR])	3.50 [3.00, 3.90]	3.50 [3.00, 3.90]	3.50 [3.00, 3.90]	0.69
Total Bilirubin, mg/dL (median [IQR])	1.20 [0.73, 2.00]	1.20 [0.80, 2.00]	1.20 [0.70, 2.00]	0.46
INR (median [IQR])	1.20 [1.10, 1.32]	1.20 [1.10, 1.32]	1.20 [1.10, 1.31]	0.29
Creatinine, mg/dL (median [IQR])	0.86 [0.72, 1.05]	0.86 [0.72, 1.05]	0.87 [0.71, 1.06]	0.69
Sodium, mEq/L (median [IQR])	138.00 [136.00, 140.00]	138.00 [136.00, 140.00]	138.00 [136.00, 140.00]	0.28
Ascites (%)				0.83
Absent	10,717 (53)	8,564 (53)	2,153 (53)	
Slight	7,136 (35)	5,702 (35)	1,434 (35)	
Moderate	1,426 (7)	1,148 (7)	278 (7)	
N/A	961 (5)	778 (5)	183 (5)	
Encephalopathy (%)				0.14
None	12388 (61)	9876 (61)	2512 (62)	
Grade 1-2	6518 (32)	5252 (32)	1266 (31)	

Grade 3-4	370 (2)	283 (2)	87 (2)	
N/A	964 (5)	781 (5)	183 (5)	
Dialysis (%)	267 (1)	214 (1)	53 (1)	0.88
HCC characteristics				
Exception Approved	18772 (93)	15021 (93)	3751 (93)	0.25
Original Viable Cumulative Tumor Diameter, cm (median [IQR])	2.90 [2.20, 4.10]	2.90 [2.20, 4.10]	2.80 [2.20, 4.10]	0.34
Original Number of Tumors, n (median [IQR])	1.00 [0.00, 1.00]	1.00 [0.00, 1.00]	1.00 [0.00, 1.00]	0.98
Original Tumor Burden (%)				0.51
Within Milan	18,062 (89)	14,456 (89)	3,606 (89)	
Within UNOS downstage	1,674 (8)	1343 (8)	331 (8)	
All-comers	504 (3)	393 (2)	111 (3)	
Tumor Burden (T2 criteria) (%)				0.85
T2 <3	10,026 (64)	8,506 (64)	1,520 (63)	
T2 3-5	2,221 (14)	1,872 (14)	349 (14)	
T2 multi	1,918 (12)	1,620 (12)	298 (12)	
Downstaged	1,582 (10)	1,332 (10)	250 (10)	
AFP, ng/mL (median [IQR])	6.00 [4.00, 16.00]	6.00 [4.00, 16.00]	6.00 [4.00, 17.00]	0.79
Locoregional Treatment (%)				0.85
Chemical Ablation	272 (1)	220 (1)	52 (1)	
External Beam Radiation	377 (2)	302 (2)	75 (2)	
Y90	2887 (14)	2298 (14)	589 (15)	
TACE	8022 (40)	6427 (40)	1595 (39)	
Thermal Ablation	2544 (13)	2014 (12)	530 (13)	
None	6138 (30)	4931 (31)	1207 (30)	
Outcomes				
Transplant	12,572 (62)	10,058 (62)	2,514 (62)	-
Waitlist Dropout	4,005 (20)	3,204 (20)	801 (20)	-
Censored	3,663 (18)	2,930 (18)	733 (18)	-
Waitlist Follow-up Time, d (median [IQR])	237.00 [159.00, 370.00]	237.00 [161.00, 371.00]	235.00 [156.00, 364.00]	0.242
Post-LT HCC Recurrence	542 (3)	426 (3)	116 (3)	-

Post-LT mortality	1,977 (10)	1,597 (10)	380 (9)	-
Post-LT censored	17,721 (88)	14,169 (88)	3,552 (88)	-
Post-LT follow-up Time, d (median [IQR])	1,106.00 [382.00, 1880.25]	1,110.00 [385.00, 1875.00]	1,095.00 [369.25, 1902.75]	0.39

IQR, interquartile range; AIH, autoimmune hepatitis; PSC, primary sclerosing cholangitis; PBC, primary biliary cirrhosis; MAFLD, metabolic dysfunction-associated fatty liver disease; MELD, Model for End Stage Liver Disease; INR, international normalized ratio; UNOS, United Network for Organ Sharing; TACE, transarterial chemoembolization; LT, liver transplant; HCC, hepatocellular carcinoma.

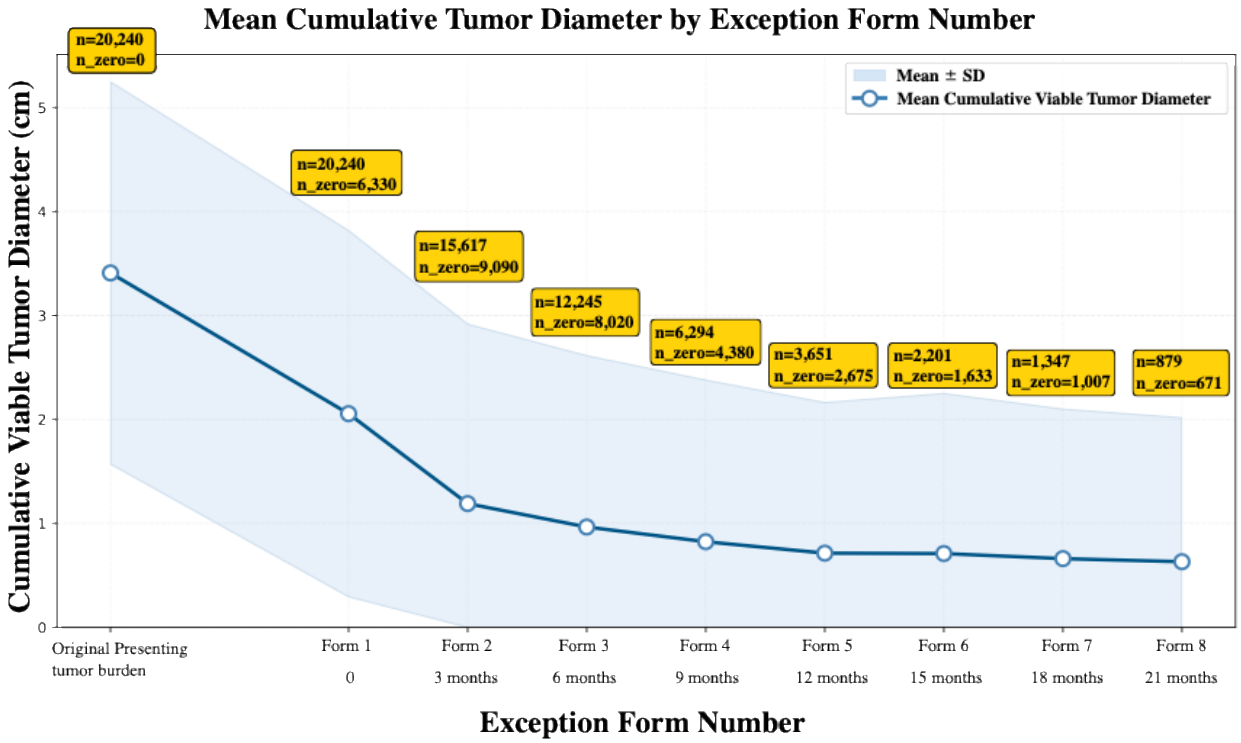


Figure 17. Cumulative viable tumor over time.

Mean cumulative viable tumor diameter (+/- 1 SD) at each observation or submitted exception form. Observation -1 denotes the original presenting cumulative tumor diameter. The yellow boxes denote the number of patients with available data at each time point and the number of patients with a reported cumulative tumor diameter of zero at each time point.

Model feature importance is shown in the SHAP plot in Figure 18, which highlights both static features (shown in bold) and dynamic features over time (shown in italics). The HCC-3D and DeepHit models performed similarly in cross validation when input was restricted to the first observation (mean 6-month c-index ± SD: 0.73 ± 0.01 versus 0.74 ± 0.01, respectively). These two models performed significantly better than all others tested in terms of 3- and 6-month c-indices when the first observation of data was used. HCC-3D and DeepHit performed similarly when there is only one observation of data to draw from. This result is comparable to Multi-HCC, where only one timepoint of data is used, and even when restricted to data from one time point, our models show improved performance. The main benefit of HCC-3D is that the vast majority of patients have more than one time point of data and when longitudinal data becomes available, the HCC-3D model is far superior to all models tested (HCC-3D mean c-index at 3-, 6-, and 12-months ± SD: 0.82 ± 0.02, 0.80 ± 0.01, 0.78 ± 0.01). Excellent performance was maintained in the hold-out test set (c-index at 3-, 6-, and 12-months [95% CI]: 0.80 [0.77-0.82], 0.78 [0.76-0.80], 0.77 [0.75-0.79]). This performance is compared to the Multi-HCC model with inferior results in cross-validation at three months (mean c-index ± SD: 0.71 ± 0.02) and six months (mean c-index ± SD: 0.70 ± 0.01). Complete model evaluation results are shown in Table 26. Calibration plots of observed versus predicted 3-month dropout CIFs demonstrated a systematic under-predicting of 3-month dropout CIF by HCC-3D with a mean absolute error of 0.20. After calibration, there was substantial improvement with a mean absolute error of 0.03 (Figure 19).

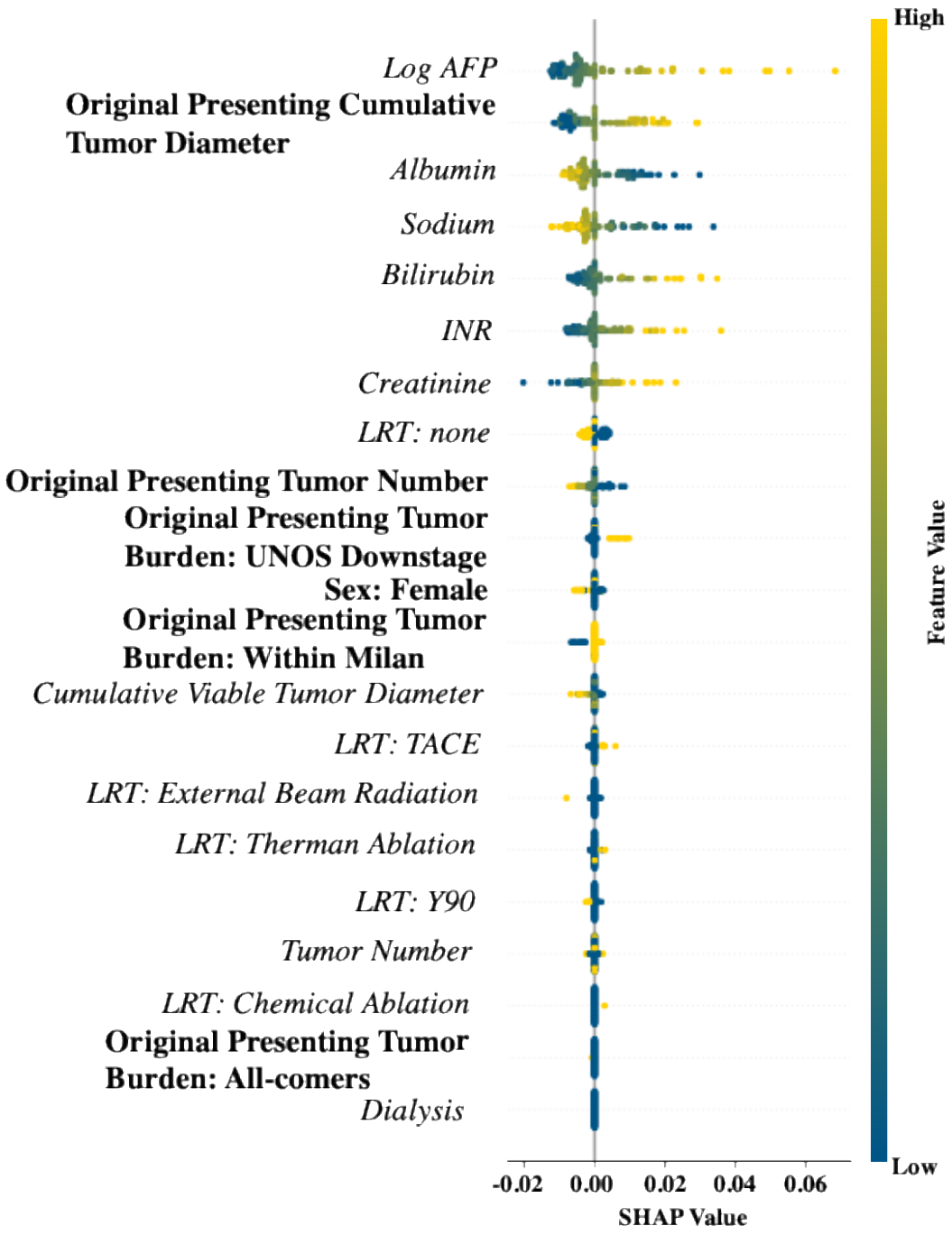


Figure 18. Feature importance plot.

Shapley additive explanations plot showing feature importance in **static** and *longitudinal* features.

Table 27. Model evaluation.

Model	Harrell's C-index (mean ± SD)			Integrated Brier Score (mean ± SD)		
	3-month	6-month	12-month	3-month	6-month	12-month
First observation only						
HCC-3D	0.74 ± 0.02	0.73 ± 0.01	0.72 ± 0.01	0.01 ± 0.00	0.03 ± 0.00	0.07 ± 0.00
Deep Hit	0.74 ± 0.02	0.74 ± 0.01	0.72 ± 0.01	0.01 ± 0.00	0.03 ± 0.00	0.06 ± 0.00
RSF	0.70 ± 0.01	0.71 ± 0.01	0.72 ± 0.01	0.03 ± 0.00	0.07 ± 0.00	0.12 ± 0.00
Cox MELD 3.0 Components	0.68 ± 0.05	0.65 ± 0.03	0.63 ± 0.01	0.01 ± 0.00	0.03 ± 0.00	0.08 ± 0.00
Cox MELD 3.0	0.65 ± 0.05	0.61 ± 0.03	0.58 ± 0.02	0.01 ± 0.00	0.03 ± 0.00	0.08 ± 0.00
Multi-HCC [28]	0.73 ± 0.03	0.69 ± 0.02	0.66 ± 0.01	0.01 ± 0.00	0.03 ± 0.00	0.09 ± 0.01
Multi-HCC Risk Score[†]	0.72 ± 0.02	0.68 ± 0.02	0.65 ± 0.01			
Mehta et al [27]	0.70 ± 0.02	0.66 ± 0.01	0.64 ± 0.01			
HALT-HCC [25, 53, 54]	0.68 ± 0.02	0.66 ± 0.01	0.64 ± 0.01			
deMELD [29]	0.66 ± 0.04	0.62 ± 0.02	0.60 ± 0.01			
HCC-MELD [26]	0.68 ± 0.02	0.65 ± 0.01	0.63 ± 0.01			
MELD-Eq [26]	0.66 ± 0.03	0.62 ± 0.02	0.59 ± 0.01			
Up to the first two observations						
HCC-3D	0.82 ± 0.02	0.80 ± 0.01	0.78 ± 0.01	0.02 ± 0.00	0.04 ± 0.00	0.08 ± 0.00
Deep Hit	0.78 ± 0.00	0.76 ± 0.01	0.75 ± 0.00	0.02 ± 0.00	0.04 ± 0.00	0.08 ± 0.00
RSF	0.73 ± 0.01	0.74 ± 0.01	0.74 ± 0.01	0.06 ± 0.00	0.09 ± 0.00	0.12 ± 0.00
Cox MELD 3.0 Components	0.69 ± 0.02	0.64 ± 0.01	0.63 ± 0.01	0.02 ± 0.00	0.05 ± 0.00	0.11 ± 0.00

Cox MELD 3.0	0.63 ± 0.02	0.60 ± 0.00	0.58 ± 0.01	0.02 ± 0.00	0.05 ± 0.00	0.11 ± 0.00
Multi-HCC [28]	0.71 + 0.02	0.70 ± 0.01	0.67 ± 0.01	0.02 ± 0.00	0.06 ± 0.01	0.12 ± 0.01
Mehta et al [27]	0.69 ± 0.03	0.66 ± 0.01	0.64 ± 0.02			
HALT-HCC [25, 53, 54]	0.70 ± 0.01	0.69 ± 0.01	0.67 ± 0.01			
deMELD [29]	0.65 ± 0.03	0.61 ± 0.01	0.59 ± 0.01			
HCC-MELD [26]	0.67 ± 0.02	0.65 ± 0.01	0.63 ± 0.01			
MELD-Eq [43]	0.63 ± 0.03	0.60 ± 0.01	0.58 ± 0.01			
Up to the first three observations						
HCC-3D	0.80 ± 0.01	0.79 + 0.01	0.78 ± 0.01	0.02 ± 0.00	0.06 ± 0.00	0.09 ± 0.00
Evaluation of HCC-3D on hold-out test set (95% confidence interval in 100 bootstrapped samples)						
First observation only	0.71* (0.68-0.75)	0.70 (0.68-0.73)	0.69 (0.67-0.71)	0.02 (0.01-0.02)	0.05 (0.04-0.05)	0.07 (0.07-0.08)
Up to the first two observations	0.80* (0.78-0.82)	0.78 (0.76-0.80)	0.76 (0.75-0.78)	0.02 (0.02-0.02)	0.05 (0.05-0.05)	0.09 (0.08-0.09)
Up to the first three observations	0.78* (0.75-0.80)	0.77 (0.75-0.79)	0.76 (0.75-0.78)	0.02 (0.02-0.02)	0.06 (0.05-0.06)	0.10 (0.09-0.10)

SD, standard deviation; HCC, hepatocellular carcinoma; HCC-3D, HCC Dynamic-Deep-Drop; RSF, random survival forest; MELD, Model for End Stage Liver Disease; HALT-HCC, Hazard Associated with Liver Transplantation for HCC; deMELD, drop-out equivalent MELD; MELD-Eq, MELD equivalent score

* The 3-month C-index shown for the test set was produced using the post-hoc cubic smoothing spline calibrator, applied to model test set predictions, demonstrating that calibration did not compromise discriminative ability

†. The published formula for calculating the MELD-matched Multi-HCC score was used to calculate C- indices in cross-validation, which differs from the row above where the Multi-HCC model was re-trained and evaluated.

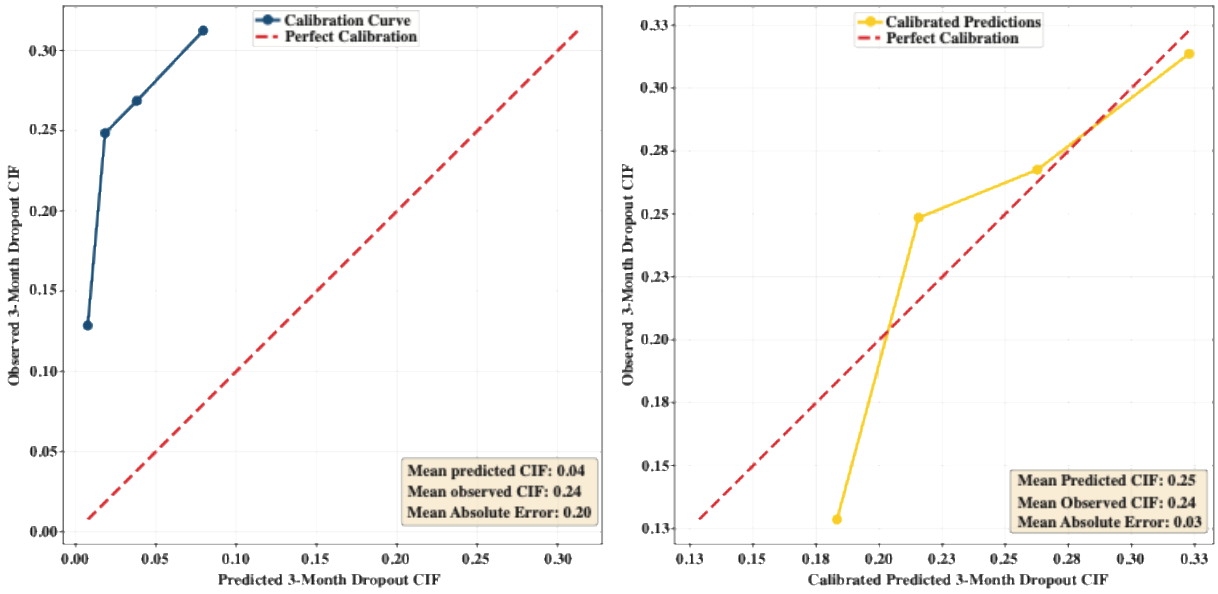


Figure 19. Calibration plots.

Predicted 3-month dropout CIF versus observed (estimated by Aalen-Johansen) in the HCC Dynamic-Deephit model from fold 3 prior to calibration (left) and after application of the post-hoc cubic smoothing spline calibrator (trained on the training set). Both figures show results from the hold-out test set. Differences between predicted and observed 3-month CIFs were assessed using mean absolute error.

Allocation simulation

Patient demographics of those used in simulation are shown in Table 27. To validate the design of the simulation (Figure 20), we compared the results using MMaT-3 as the policy for assigning exception points to HCC patients against real-world outcomes reflected in the SRTR dataset and demonstrated nearly identical outcome counts (Figure 21). These findings were confirmed in sensitivity analysis (Supplementary Appendix: Sensitivity Analysis), which illustrated the stability of the MMaT-3 implementation (Table 28). The identified M_{eq} risk groups were lowest risk (< 27), middle risk (27-29), and highest

risk (≥ 30). Recurrence-free-survival was significantly increased and cumulative incidence of dropout significantly decreased in the low-risk group compared to the middle and high-risk groups (p=0.04 and p<0.001, respectively, Figure 22).

Table 28. Demographics of liver transplant patients (SRTR) used in simulation.

Variable	N = 62,464
Age, y (median [IQR])	57.00 [47.00, 64.00]
Sex = M (%)	38,049 (61)
Status 1A	1,166 (2)
HCC	10,268 (16)
Initial Lab MELD (median [IQR])	18.00 [13.00, 27.00]
Last Lab MELD (median [IQR])	20.00 [13.00, 30.00]
Initial Allocation MELD (median [IQR])	20.00 [14.00, 29.00]
Last Allocation MELD (median [IQR])	24.00 [16.00, 33.00]
Original Cumulative Tumor Diameter, cm (median [IQR])	2.90 [2.20, 4.10]
Original Number of Tumors, n (median [IQR])	1.00 [1.00, 1.00]
Original Tumor Burden (%)	
Within Milan	9,002 (88)
Within UNOS downstage	954 (9)
All-comers	292 (3)
AFP, ng/mL (median [IQR])	6.00 [3.00, 13.00]
Transplant	39,046 (63)
Waitlist Dropout	7,851 (14)
SRTR, Scientific Registry of Transplant Recipients; IQR, interquartile range; HCC, hepatocellular carcinoma; MELD, Model for End-Stage Liver Disease; UNOS, United Network for Organ Sharing; AFP, alpha-fetoprotein	

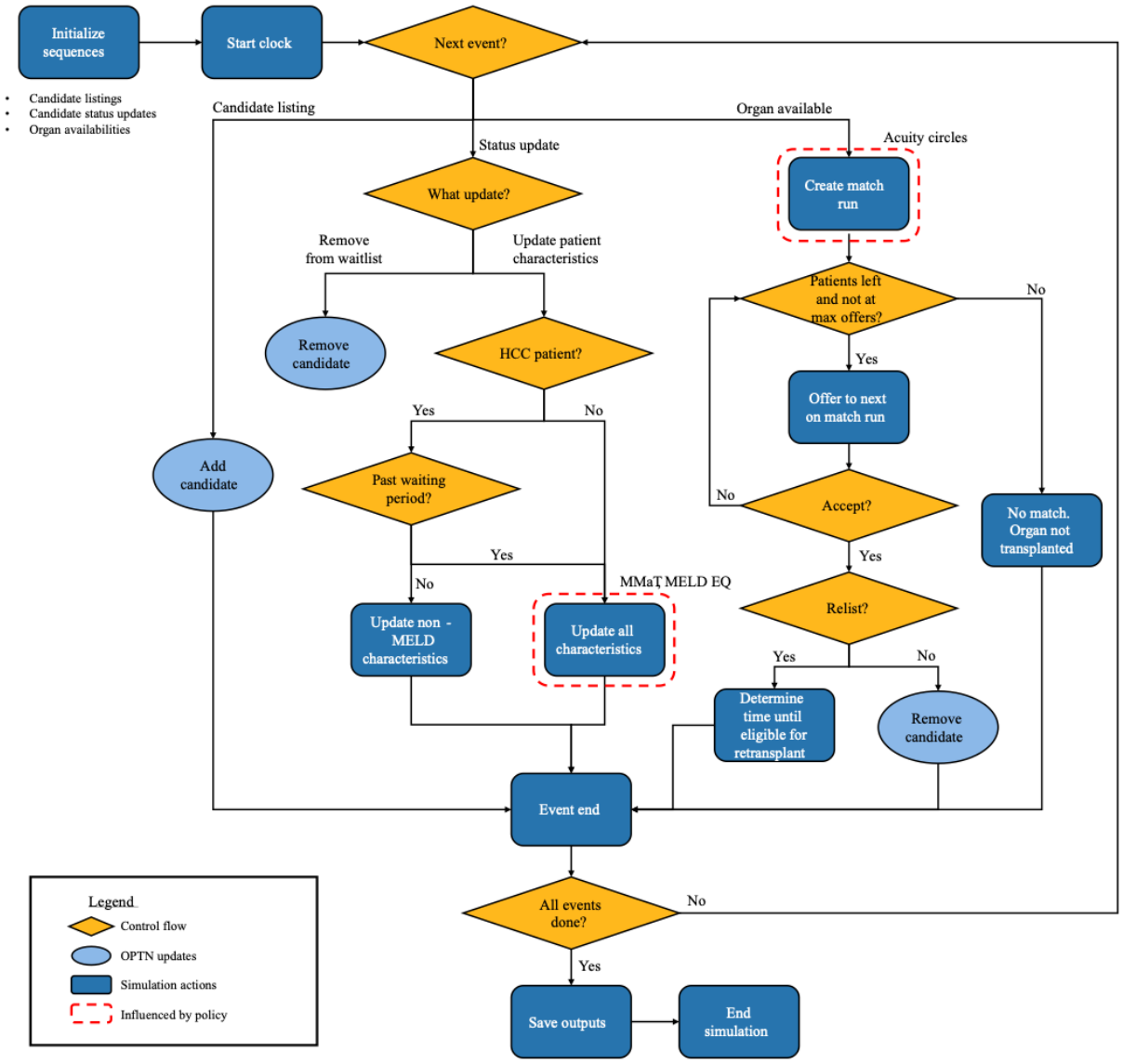


Figure 20. Overview of the simulation design.

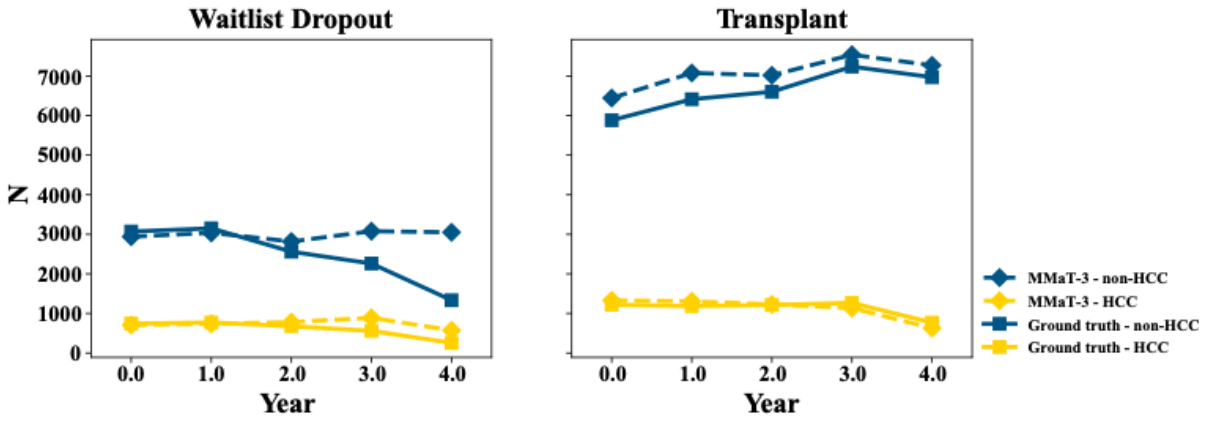


Figure 21. Simulation validation.

Profiles illustrating the yearly (x-axis) trends of number of events between the simulation when using MMAT-3 as the policy for assigning MELD points to HCC patients (dotted lines, diamond markers), and real-world data measured retrospectively (solid lines, square markers). The y-axis indicates the number of candidates listed in the respective year on the x-axis that experienced (could be in the same year or a potentially later year) waitlist dropout (left) or transplant (right). Trends are illustrated for both HCC (orange) patients and non-HCC patients (blue).

Table 29. Simulations results, with counts shown as mean (SD).

Simulation scenario			HCC		Non-HCC	
	Waiting period (months)	Scaling factor	Waitlist dropout	Transplant	Waitlist dropout	Transplant
Ground truth			3,010	5,660	12,377	33,104
Main results						
MMaT-3	6	30	3,674.4 (13.21)	5,645.8 (18.12)	14,919.7 (65.10)	35,362.1 (51.84)
M_{eq} Continuous	6	30	3,646.9 (17.29)	5,684.2 (21.67)	14,912.2 (52.77)	35,353.7 (64.46)
M_{eq} Stratified	6	30	3,638.8 (20.00)	5,679.7 (29.05)	14,934.3 (47.50)	35,337.7 (47.60)
M_{eq} Stratified 3m	--	30	3,387.1 (20.22)	5,992.9 (30.77)	14,973.1 (61.32)	35,290.1 (57.63)

Sensitivity analysis						
MMaT-3	6	20	3,697.5 (16.34)	5,624.4 (20.69)	15,498.7 (79.81)	34,730.8 (74.07)
MMaT-3	6	40	3,653.8 (13.46)	5,671.6 (17.13)	14,604.0 (48.84)	35,684.4 (62.55)
MMaT-3	6	50	3,629.6 (20.22)	5,698.1 (23.87)	14,437.4 (37.40)	35,878.4 (42.28)
M_{eq} Continuous	6	20	3,688.4 (13.91)	5,635.8 (21.32)	15,523.1 (55.06)	34,712.4 (68.23)
M_{eq} Continuous	6	40	3,626.1 (20.89)	5,710.1 (25.35)	14,638.3 (77.03)	35,651.8 (76.19)
M_{eq} Continuous	6	50	3,616.5 (16.82)	5,715.0 (17.80)	14,452.5 (47.56)	35,862.1 (64.40)
M_{eq} Continuous	3	20	2,371.3 (9.18)	7,210.2 (19.56)	15,701.6 (70.51)	34,497.7 (71.86)
M_{eq} Continuous	3	30	2,341.8 (21.44)	7,235.5 (28.13)	15,099.9 (80.49)	35,133.7 (87.44)
M_{eq} Continuous	3	40	2,322.3 (19.45)	7,268.7 (25.32)	14,832.3 (94.00)	35,415.4 (106.27)
M_{eq} Continuous	3	50	2,308.9 (20.88)	7,287.4 (25.05)	14,666.2 (78.22)	35,618.2 (82.25)
Multi-HCC	6	30	5,934.7 (19.32)	2,955.7 (17.80)	14,621.5 (63.24)	35,685.5 (72.44)
Multi-HCC & MMaT-3	6	30	3,581.9 (12.01)	5,752.3 (13.48)	15,047.2 (53.06)	35,207.6 (46.80)

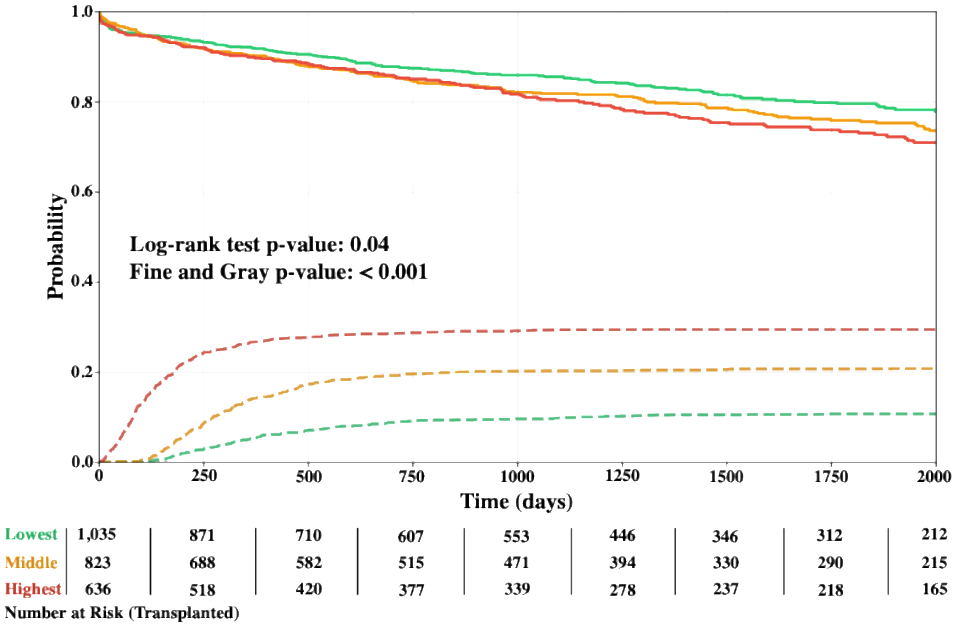


Figure 22. HCC-3D stratified survival by risk score group.

HCC-3D model-derived risk groups in the test set. Groups were formed by taking tertiles of the calculated HCC-3D Meq score. The lowest risk group is shown in green, followed by the middle risk group in orange, and the highest risk group in red. Post-transplant recurrence-free-survival was compared in the three risk groups as shown in the down-sloping Kaplan-meier curves. Differences were assessed using the log-rank test. Cumulative incidence of waitlist dropout was compared between the three risk groups as shown by the up-sloping curves. Differences were assessed using Fine and Gray competing risk regression.

All simulation results are reported as mean ± SD. MMaT-3 resulted in an average of 3,674 ± 13 waitlist dropouts and 5,646 ± 18 transplants for HCC patients and 14,920 ± 65 waitlist dropouts and 35,362 ± 52 transplants for non-HCC patients (Figure 23). With the continuous M_{eq} score applied after a six-month waiting period there were 3,647

$\pm$ 17 waitlist dropouts and 5,684 $\pm$ 22 transplants for HCC patients and 14,912 $\pm$ 53 waitlist dropouts and 35,354 $\pm$ 64 transplants for non-HCC patients. With the stratified M_{eq} score applied there were 3,639 $\pm$ 20 waitlist dropouts and 5,680 $\pm$ 29 transplants for HCC patients and 14,934 $\pm$ 48 waitlist dropouts and 35,290 $\pm$ 58 transplants for non-HCC patients. When allowing high-risk patients to receive exception scores after 3 months, there were 3,387 $\pm$ 20 waitlist dropouts and 5,993 $\pm$ 31 transplants for HCC patients and 14,973 $\pm$ 61 waitlist dropouts and 35,290 $\pm$ 57 transplants for non-HCC patients. When comparing MMaT-3 and the continuous M_{eq} score after six months, there was a significant decrease in mean waitlist dropouts (27; $p<0.001$) and significant increase in mean transplants (38; $p<0.001$) for the HCC patients. On the other hand, the differences in mean waitlist dropouts and transplants in the non-HCC patients were statistically insignificant ($p=0.78$ and $p=0.75$, respectively). When comparing MMaT-3 and the stratified M_{eq} score, there was a significant decrease in mean waitlist dropouts (35; $p<0.001$) and significant increase in mean transplants (34; $p<0.001$) for the HCC patients, with insignificant differences in mean waitlist dropouts and transplants for non-HCC patients ($p=0.57$ and $p=0.29$, respectively). When comparing MMaT-3 and the stratified M_{eq} allowing for high-risk patients to receive exception scores after 3 months, there was a larger decrease in mean waitlist dropouts (287; $p<0.001$) and increase in mean transplants (347; $p<0.001$) for the HCC patients at the expense of decreasing the mean transplants (72; $p=0.009$) for non-HCC patients.

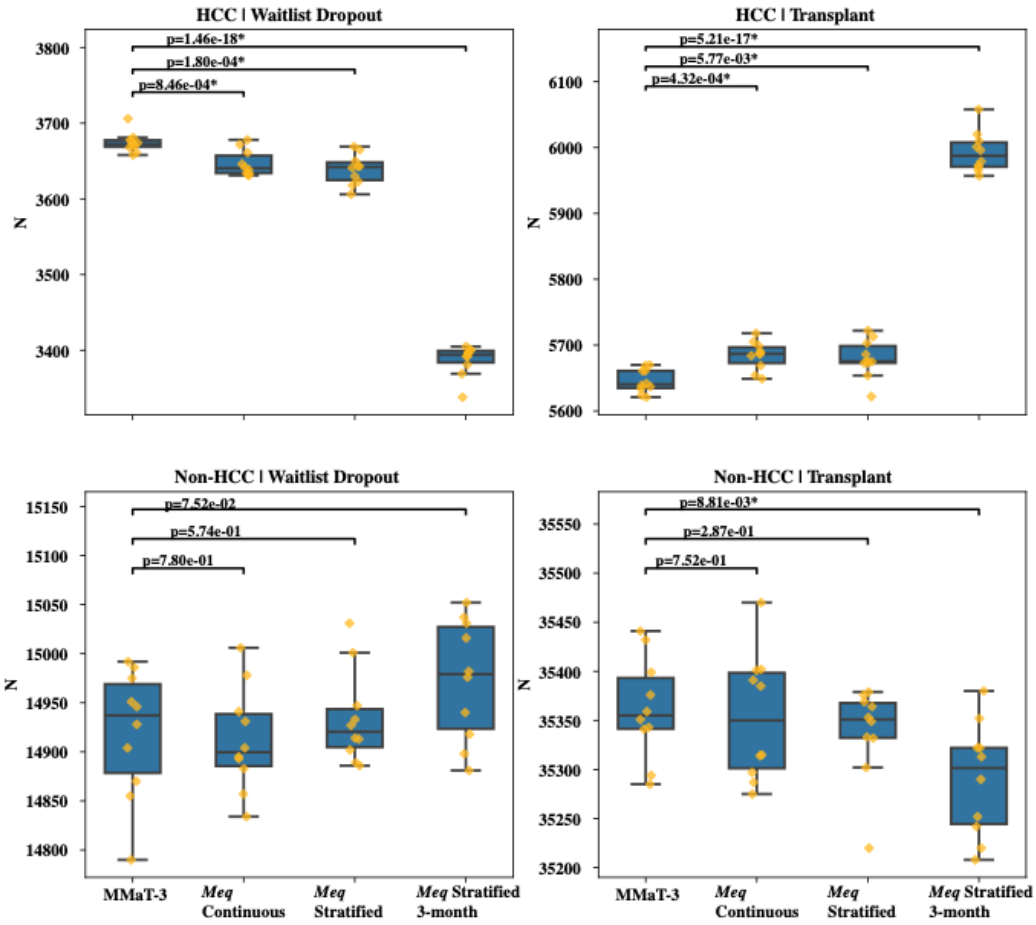


Figure 23. Simulation results.

Outcomes of the simulation when using different policies (x-axis) for assigning MELD points to HCC patients, including MMaT-3, M_{eq} score after a 6-month waiting period (M_{eq} Continuous), stratified M_{eq} with low-risk (< 27), medium-risk (27-29), and high-risk (≥ 30) resulting in modifications to MMaT according to Table 24, and stratified M_{eq} where patients in the highest risk group received exception points at 3 months instead of 6. The top row illustrates the outcomes for the HCC patients, including the total number of waitlist dropouts (left) and transplants (right). The bottom row illustrates the same outcomes for the non-HCC patients, with the total number of waitlist dropouts on the left and transplants on the right. Summarizing box plots are shown with the

individual data points to illustrate the aggregate and individual results from the repeated simulations.

MELD change impact in select HCC patient subgroups

MELD score cross tabulation for patients waitlist dropout is shown in Figure 24 where the diagonal boxes represent those with unchanged allocation and Meq score ranges with 112 patients having Meq score ranges greater than allocation score ranges. These patients would have had increased priority for LT and perhaps would have been transplanted prior to dropout. This is compared to those 41 patients who had allocation score ranges greater than Meq score ranges who may have been underprioritized by the Meq score. Overall, there were 71 more patients who had an increase in priority because of the Meq score compared to a decrease in priority. MELD score cross tabulations for low-risk HCC patients and non-low-risk HCC patients are shown in Supplemental Figure 25 A & B, respectively). The low-risk patients had unchanged priority by the Meq score in 583 (50%) patients, compared to the non-low-risk with 1,068 (35%) patients with unchanged priority and 1,496 (48%) patients with an increase in priority afforded by the Meq score. This demonstrates a preferential increase in organ prioritization for non-low-risk HCC patients by our model.

	HCC-3D *Meq* Score					
Allocation MELD	**< 20**	**20-24**	**25-29**	**30-34**	**≥ 35**	**Total**
< 20	-	-	-	-	-	-
20-24	-	6	46	19	-	71
25-29	-	-	57	47	-	104
30-34	-	1	32	30	-	63
≥ 35	-	-	1	7	43	51
Total	-	7	136	103	43	289

Figure 24. HCC-3D differential prioritization.

MELD equivalent score (X-axis) versus allocation MELD (Y-axis) cross tabulation. The boxes contain numbers of patients. The gold diagonal boxes represent unchanged MELD score ranges between true allocation MELD and MELD equivalent score. Green boxes represent those with increased priority as a result of the MELD equivalent score versus those in red representing decrease priority.

A)

Allocation MELD	HCC-3D *Meq* Score					
	< 20	20-24	25-29	30-34	≥ 35	Total
< 20	-	-	-	-	-	-
20-24	-	9	256	44	-	309
25-29	-	9	543	121	-	673
30-34	-	2	117	31	-	150
≥ 35	-	-	20	5	-	25
Total	-	20	936	201	-	1157

B)

Allocation MELD	HCC-3D *Meq* Score					
	< 20	20-24	25-29	30-34	≥ 35	Total
< 20	-	-	-	-	-	-
20-24	-	22	659	381	-	1062
25-29	-	18	828	451	1	1298
30-34	-	9	314	203	4	530
≥ 35	-	2	107	73	15	197
Total	-	51	1908	1108	20	3087

Figure 25. HCC-3D differential prioritization in distinct subgroups.

HCC-3D Differential prioritization in A) low risk HCC patients (MELD score <15, single tumor < 3 cm, AFP ≤ 20) and B) non-low risk HCC patients. MELD equivalent score (X-axis) versus allocation MELD (Y-axis) cross tabulation. The boxes contain numbers of

patients. The gold diagonal boxes represent unchanged MELD score ranges between true allocation MELD and MELD equivalent score. Green boxes represent those with increased priority as a result of the MELD equivalent score versus those in red representing decrease priority.

Chapter 4 Discussion

In this work, we propose HCC-3D, a novel longitudinal survival network model for HCC patient LT waitlist prognostication. HCC-3D achieved excellent discriminative performance with 3-, 6-, and 12-month c-indices greater than any previous published model for HCC waitlist outcomes. This work offers a fundamentally new approach to the task of waitlist outcome prediction that is inspired by the notion that evolution over time is a key prognostic indicator. However, previous approaches to capture the risk associated with change over time have been too simplistic, relying on landmark timepoints and engineered features that capture a predefined change over time like AFP response to LRT, change in MELD score above some threshold, or even the Modified Response Evaluation Criteria in Solid Tumors (mRECIST) criteria [52]. The problem with such approaches is that they do not allow flexible learning of longitudinal relationships or complex interactions among different feature changes. For example, one patient could have a complete response to LRT with no viable tumor left, but a significant increase in MELD 3.0 or one or more of the component laboratory markers. An effective prognostic model must be able to weigh these different changes simultaneously to adequately predict risk of waitlist mortality as our model does. Our model used similar features as prior work including tumor size and number, AFP, and MELD 3.0 components. The significantly improved performance can thus be attributed

to the longitudinal encoder and MLP structure that allows learning of complex longitudinal relationships and feature interactions.

Another advantage of our approach is the model architecture allows flexible predictions where a prediction of risk can be made when an HCC patient first submits a MELD exception form or is listed for LT. Subsequently, when additional information is available, an updated prediction can be made that will have improved discriminative performance and can increase confidence in the prognostication of that patient. Prior HCC patient waitlist prediction models, including Multi-HCC, primarily focus on obtaining a prediction using data available at listing. However, HCC patients must wait the mandatory six months after exception approval to receive exception points, making an assessment of 6-month prognosis at the time of listing less useful because when the six months has elapsed and patients receive their points, they would have differential organ priority based on a score calculated six months in the past. Our approach allows prediction with up to two time points, which would include information in the two most recent exception forms, which at 6 months would be the second and third exception forms, ensuring up-to-date information is used at the time of exception point allocation. This approach provides a prediction with more up-to-date data when organ allocation priority would be determined, capturing any changes that occurred during the 6-month waiting period.

The Dynamic-DeepHit architecture is designed to accommodate long sequences of data. Our model could easily be adapted to accommodate longer patient sequences though we found that with three months between exception form submissions, most patients have three or fewer forms. Moreover, even at the second form a large number

of patients have no viable tumor reported (58%, Figure 17). We found that inclusion of longer sequences (up to three exception forms) did not significantly improve performance (Table 26), likely due to the inclusion of increasingly large numbers of datapoints representing no viable tumor and the time from last observation to outcome becoming increasingly short for those patients with longer sequences. With a median time to event or last follow up of eight months, inclusion of even the third exception form (nine months from listing) can significantly skew the model's assessment of median time to event in this population. Theoretically, with additional longitudinal datapoints taken at shorter time intervals like one month, model performance could improve even further though such longitudinal granularity for HCC patients is not available in the UNOS or SRTR databases. We thus chose the sequence length two configuration as the final model as it was the most parsimonious while exhibiting optimal performance.

Another major contribution of this work is that our proposed model was robustly evaluated in simulation, to predict what the impact on waitlist dropout and probably of transplant would be on both HCC and non-HCC populations if it were to be implemented. We created an open-source, fully configurable allocation simulation framework based on LSAM and LivSim. To validate the simulation, we compared the results using MMaT-3 as the policy for assigning exception points to HCC patients against real-world outcomes reflected in the SRTR dataset. Figure 21 illustrates the yearly waitlist dropout and transplants that were observed in real life compared to the simulation. Notably, the trends exhibit strong similarity with a tendency for higher rates of transplant in the simulation, likely due to an overestimation of acceptance compared to the real-world. Furthermore, the increasing difference in waitlist dropouts in the final

years of the simulation are attributed to the number of real-world patients with no reported outcomes (due to potential reporting delay), whereas the simulation preserves a more constant rate of dropout reflecting more realistic outcomes with perfect observability. Overall, the discrepancies reflect how the simulations are imperfect representations of reality, but with reasonable similarities warranting their use for analysis, as well as justifiable sources of differences. Improved alignment with reality will require further modifications to the acceptance model to reflect real-world decisions, as well as complete observability or real-world outcomes.

We offer two implementation strategies for HCC-3D that were both simulated and compared to the existing allocation framework (MMaT-3). The continuous M_{eq} score and stratified M_{eq} score connected to modifications to the MMaT system resulted in fewer waitlist dropouts for HCC patients and an increased number of transplants without significantly effecting the non-HCC patients. The raw number of decreases in waitlist dropout and increases in transplant are modest (<100) with maintenance of the 6-month waitlist period, but statistically significant. Based on Figure 24, we know that the M_{eq} score tended to increase priority for patients who experienced waitlist dropout, suggesting that the use of this model is prioritizing those patients at highest risk without affecting organ allocation for non-HCC patients.

The proposed HCC-3D model achieved state of the art discriminative performance for predicting waitlist death or delisting in HCC patients and demonstrated improved waitlist outcomes for HCC patients with no effect on non-HCC patients in a simulated model of liver allocation. This model has the potential to improve liver allocation for HCC patients, specifically in differentiating patients based on risk of

waitlist death or delisting in a more robust way than the current MMaT-3 system. We believe this work can serve to benchmark the necessary steps for proposed models affecting organ allocation including model training, robust validation, calibration, and allocation simulation. Additionally, we have provided a roadmap for integrating a DL model with the existing MELD 3.0 score that can likely be applied to many other domains in clinical medicine.

Chapter 4 Conclusions

We demonstrate state-of-the-art discriminative performance in the prediction of waitlist death or delisting in transplant-eligible HCC patients using a longitudinal deep survival network model (HCC-3D). In simulation, use of our model led to fewer HCC patient waitlist dropouts and more transplants with no significant change to the dropout or transplant rate of non-HCC candidates. Implementation of HCC-3D could improve allocation of donor livers to HCC patients.

Chapter 4 Supplementary Appendix

HCC-3D model architecture

The model utilized a static feature encoder consisting of a two-layer multilayer perceptron (MLP) and a two-layer unidirectional LSTM temporal feature encoder with a temporal attention mechanism to allow differential weighting of individual sequence time steps. Temporal attention was calculated using the transformer-based multi-head self-attention mechanism (6 heads) originally described by Vaswani et al. in the landmark paper “Attention is all you need,” which is commonly used for temporal data in deep learning [55]. Information flows from timestep to timestep via the LSTM and the attended output at the last valid time step is then concatenated to the static feature

encoder output via a two-layer MLP. The combined feature MLP output is then passed to two cause-specific networks for each of two possible event types (waitlist dropout or transplant). Each cause-specific network consists of a three-layer MLP, which outputs raw logits that are converted to joint probabilities by applying the SoftMax activation function across all time bins and event types. This forces the model to allocate probability for each event at each time, ensuring the probability of any event over the full-time horizon sums to one. Prediction time was discretized into one-month bins for a total ten-year prediction horizon (120-time bins total). The cumulative incidence (CIF) of a given event could then be calculated by summing the predicted probability of that event up to the time of interest. All MLPs used rectified linear unit activation functions (ReLU) and dropout was applied to all encoder layers (0.1-0.2).

Loss function design

A weighted combination of negative log-likelihood and ranking losses were used to train the model as displayed in equation 5 below.

I. $$L_{nll_{event}} = -\log P(T = \tau_i, k = k_i | X_i)$$

II. $$L_{nll_{censored}} = -\log[\sum_{\tau=t}^{T_max} \sum_{k=1}^{K} P(T = \tau, k = k | X_i)]$$

III. $$L_{nll} = \left(\frac{1}{N}\right) \sum_{i=1}^{N} L_{nll_i} \quad (5)$$

IV. $$L_{rank} = \frac{1}{|Pa|} \Sigma_{(i,j) \in Pa} \exp(-\frac{CIF_{k(\tau_i|X_i)} - CIF_{k(\tau_j|X_j)}}{\sigma})$$

V. $$L_{total} = \alpha \cdot L_{nll} + \beta \cdot \lambda_{rank}$$

Equation 5 depicts loss functions: I. Negative loglikelihood loss for dropout events. II. Negative loglikelihood loss for censored patients. III. Total negative loglikelihood loss. IV. Ranking loss. V. Total loss. The negative loglikelihood loss for patients with events ($L_{nll_{event}}$) is defined as the negative sum of the log of the probability P of event k_i occurring at time interval τ_i given covariates X_i. For censored patients, $L_{nll_{censored}}$ is defined as the negative log of the survival probability, which is calculated by summing the probability P that event k has occurred by time interval τ, across event types (inner sum) and from the first time interval t to the maximum time interval T_max (outer sum). The total negative loglikelihood loss (L_{nll}) is calculated by taking the average of the calculated losses for censored and non-censored patients.

The ranking loss (L_{rank}) was computed for acceptable pairs of patients Pa where one patient had the event of interest at time interval τ_i and the other had either event or was censored at time τ_j where $\tau_j > \tau_i$. The ranking loss was thus calculated by taking the exponent (exp) of the negative difference in predicted cumulative incidence of event between the patient who truly experienced the event ($CIF_{k(\tau_i|X_i)}$) and other patient in the pair ($CIF_{k(\tau_j|X_j)}$). The exponent of this negative difference divided by σ (a temperature parameter controlling penalty strength, set to 0.1) was summed for each pair and divided by the number of acceptable pairs ($|Pa|$). The total loss was calculated by taking the weighted sum of the negative loglikelihood and ranking losses. The weights of α and β were tuned empirically and set to 0.85 and 0.5 in the final configuration, respectively.

Derivation of the MELD equivalent

Model predictions of 3-month HCC waitlist mortality were used to derive MELD equivalent (M_{eq}) scores that were then interdigitated with the MELD 3.0 scores of non-HCC patients. The Multi-HCC model achieved MELD 3.0 equivalence using the linear predictor, an ability afforded by the linearity and structure of the Cox Proportional Hazards model that both Multi-HCC and MELD 3.0 rely on. In our case, equivalence cannot be achieved this way due to the non-linear nature and architectural differences of our model. MELD score interdigitation was thus accomplished by relying on the underlying survival function inherent to the MELD 3.0 score as detailed in Equation 6 below [56].

$$\text{I.} \quad S(t) = [S_0(t)]^{\exp(\beta \cdot MELD + \alpha)}$$

$$\text{II.} \quad S_{HCC} = 1 - CIF_{dropout}$$

$$\text{III.} \quad E = \frac{\ln(S_{HCC})}{\ln(S_0)} = \frac{\ln(1 - CIF_{dropout})}{\ln(0.946)} = exp(\beta \cdot M_{eq} + \alpha) \tag{6}$$

$$\text{IV.} \quad M_{eq} = \frac{\ln(E) - \alpha}{\beta} = \frac{\ln(E) + 3.56}{0.17698}$$

Equation 6 depicts the derivation of the MELD equivalent M_{eq} score from model predictions of 3-month HCC waitlist mortality. I. Survival function from MELD 3.0 model. II. Survival probability as determined by the model is one minus the predicted 3-month waitlist dropout cumulative incidence. III. Term E is equivalent to the relative hazard exp ($\beta \cdot M_{eq} + \alpha$) and the ratio of the log-survival of the patient to the log survival of the baseline population. The baseline survival constant based on MELD 3.0 is 0.946. IV.

Rearrangement to solve for M_{eq} substitution of intercept α = -3.56 and coefficient β = 0.17698 from MELD 3.0 model.

The survival function from the MELD 3.0 model $S(t)$ is the baseline hazard (90-day survival constant) $S_0(t)$ raised to the exponential function of the MELD 3.0 score regression coefficient β multiplied by the *MELD* 3.0 score plus the intercept α. Survival probability from the HCC-3D model S_{HCC} can be calculated by taking one minus the calibrated predicted 90-day waitlist dropout CIF $CIF_{dropout}$. S_{HCC} is then set equal to $S(t)$ as our goal is to map the two survival functions to one another. Algebraic rearrangement and substitution of the published baseline hazard S_0 = 0.946 from MELD 3.0 yields the relative hazard exponent E, which represents a patient's relative hazard of death compared to baseline. This hazard is then mapped back to the MELD 3.0 scale using the linear relationship of the MELD score and substituting published constants of α = -3.56 and β = 0.17698. The continuous M_{eq} score produced by the model was post-processed to align with the OPTN standards including capping at a minimum of 0 and maximum of 40 and rounding to the nearest whole integer.

Simulation design

Figure 20 summarizes the workflow of the discrete event simulation. Briefly, the main inputs of the simulation are sequences of three types of events: 1) the listing of a candidate; 2) updating of candidate status and/or variables; 3) and an organ becoming available. The listing of a candidate adds them to the Organ Procurement & Transplantation Network (OPTN) with their current patient factors (e.g., MELD, Status 1 designation). A candidate status update takes an existing patient in the OPTN and updates their factors (e.g., updates MELD scores or provides exception points for HCC

patients); furthermore, a patient is removed from the OPTN if they are indicated to experience waitlist dropout. When an organ becomes available, it triggers the allocation process that generates a match run from existing candidates in the OPTN based on the defined allocation policy; offers are then made in order of the match run until the organ is accepted, the maximum number of offers has been reached without acceptance, or the match run is depleted without acceptance. After transplant, candidates are either permanently removed from the OPTN or relisted for re-transplant after a random period drawn from a distribution parameterized the same way as LSAM. The simulation continues until all events occur.

Overall, the design is based on LSAM with key differences, including: 1) re-implementation in Python 3.10; 2) re-formatting of input data streams to accept simpler csv formats; 3) updates to accommodate recent policy changes [57]. Notably, we incorporated the MMaT-3 policy for assigning exception points to HCC patients. Following most recent policy practices, we calculated MMaT based on transplants that occurred within a 150 nautical miles (NM) radius around each donor hospital where a match is being run [58]. The transplants used to perform these calculations were obtained from real-world transplants that occurred within a year prior to the start of simulation. We also implemented allocation policy based on Acuity Circles to generate the match runs for an organ [59]. Consistent with current allocation policy, we first prioritized offers to Status 1 patients within 500 NM of the donor hospital. We then ordered patients by groups of medical urgency defined by MELD score (≥37, 33-36, 29-32, 15-28). Within each group, the first to receive offers were those within 150 NM of the donor hospital. If there were multiple potential candidates, they were ordered first by

MELD score, with further ties broken by time on the waitlist. The next to receive offers were those within 250 NM of the donor hospital, and the last to receive offers were those within 500 NM of the donor hospital. If no offers were made at a given medical urgency group, then the process repeated with the next group of candidates with lower medical urgency. To calculate distances, we leveraged the ZIP codes of the donor and candidate hospitals. We mapped each ZIP code to estimated coordinates (latitude and longitude) using the GeoNames database and then calculated their distance in NM using the Haversine method [60]. We also incorporated specific policies for HCC patients, including waiting periods before they were eligible for exception points, as well as allowance for downstaging. Specifically, we parameterized the duration of the waiting period, and for MMaT-3 we only allowed HCC patients to receive exception points after a six-month waiting period. For the MELD equivalent score, we experimented with different waiting periods (six months and three months). We also required that HCC candidates satisfied the Milan criteria after the waiting period to receive their exception points. If HCC candidates failed to satisfy the Milan criteria after the waiting period, their exceptions points were taken away, and they were no longer eligible for future exception points. While we used the acceptance model defined in LSAM, we modified two features that depended on definitions of Donation Service Areas (DSAs) no longer relevant for current policy. First, we estimated “locality” previously defined as candidates and organs located in the same DSA, as candidates and organs located in within 150 NM. Second, we estimated “time of transport” previously defined using historical average travel times between DSAs using the distance between a candidate and donor scaled by a constant. The median time of transport between DSAs reported by LivSim

was 3.28 hours, and the median distance between ZIP codes of real-world transplants was 142.3 NM. Thus, we elected for a scaling factor on the magnitude of 20 to 50 to preserve approximate magnitude of the variable.

Sensitivity analysis

In the process of validation, we assessed sensitivity to parameters such as the scaling factor applied to the distance to estimate time of transport and reran the simulation with a scaling factor of 20, 30, 40, and 50 (Table 28). We ultimately selected of a factor of 30 that yielded trends (Figure 21) for MMaT-3 that were close to reality, achieving similar rates of HCC transplant between MMaT-3 and reality, with lower rates of non-HCC transplants for MMaT-3. However, we observed that results remained consistent across all scaling factors, with low sensitivity to this parameter. Furthermore, we assessed changes when applying a waiting period of 3 months vs. 6 months for HCC patients before they become eligible for exception points.

Data availability

UNOS: This study used the UNOS database. Use of these data requires compliance with a data use agreement. Access may be obtained upon reasonable request to UNOS.

SRTR: This study used the SRTR database standard analysis files. Use of these data requires compliance with a data use agreement. Access may be obtained upon reasonable request to the SRTR.

Chapter 4 Acknowledgements

This study used data from the Scientific Registry of Transplant Recipients (SRTR). The SRTR data system includes data on all donor, wait-listed candidates, and transplant recipients in the US, submitted by the members of the Organ Procurement and Transplant network (OPTN). The Health Resources and Services Administration (HRSA), U.S. Department of Health and Human Services provides oversight to the activities of the OPTN and SRTR contractors. The data reported here have been supplied by the Hennepin Healthcare Research Institute (HHRI) as the contractor for the SRTR. The interpretation and reporting of these data are the responsibility of the authors and in no way should be seen as an official policy of or interpretation by the SRTR or the U.S. Government. JF was supported by NIH T32 EB016640.

Chapter 4 References

[1] A. Vogel, T. Meyer, G. Sapisochin, R. Salem, and A. Saborowski, "Hepatocellular carcinoma," (in eng), Lancet, vol. 400, no. 10360, pp. 1345-1362, Oct 15 2022, doi: 10.1016/s0140-6736(22)01200-4.

[2] H. Sung et al., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: a cancer journal for clinicians, vol. 71, no. 3, pp. 209-249, 2021.

[3] J. K. Horwitz and V. G. Agopian, "Indication of Liver Transplant for HCC: Current Status and Future Directions," Current Hepatology Reports, pp. 1-8, 2024.

[4] A. B. Massie et al., "MELD exceptions and rates of waiting list outcomes," American Journal of Transplantation, vol. 11, no. 11, pp. 2362-2371, 2011.

[5] K. J. Halazun et al., "Standing the test of time: outcomes of a decade of prioritizing patients with hepatocellular carcinoma, results of the UNOS natural geographic experiment," Hepatology, vol. 60, no. 6, pp. 1957-1962, 2014.

[6] A. J. Kwong et al., "OPTN/SRTR 2021 annual data report: liver," American Journal of Transplantation, vol. 23, no. 2, pp. S178-S263, 2023.

[7] E. W. Beal et al., "Trends in the Mortality of Hepatocellular Carcinoma in the United States," Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract, vol. 21, no. 12, pp. 2033-2038, Dec 2017, doi: 10.1007/s11605-017-3526-7.

[8] J. L. Petrick, S. P. Kelly, S. F. Altekruse, K. A. McGlynn, and P. S. Rosenberg, "Future of Hepatocellular Carcinoma Incidence in the United States Forecast Through 2030," Journal of clinical oncology : official journal of the American Society of Clinical Oncology, vol. 34, no. 15, pp. 1787-94, May 20 2016, doi: 10.1200/JCO.2015.64.7412.

[9] V. Mazzaferro et al., "Liver transplantation for the treatment of small hepatocellular carcinomas in patients with cirrhosis," N Engl J Med, vol. 334, no. 11, pp. 693-9, Mar 14 1996, doi: 10.1056/NEJM199603143341104.

[10] A. Forner, M. Reig, and J. Bruix, "Hepatocellular carcinoma," Lancet, vol. 391, no. 10127, pp. 1301-1314, Mar 31 2018, doi: 10.1016/S0140-6736(18)30010-2.

[11] G. Sapisochin and J. Bruix, "Liver transplantation for hepatocellular carcinoma: outcomes and novel surgical approaches," Nature reviews. Gastroenterology & hepatology, vol. 14, no. 4, pp. 203-217, Apr 2017, doi: 10.1038/nrgastro.2016.193.

[12] M. G. Refolo, C. Messa, V. Guerra, B. I. Carr, and R. D'Alessandro, "Inflammatory Mechanisms of HCC Development," Cancers (Basel), vol. 12, no. 3, Mar 10 2020, doi: 10.3390/cancers12030641.

[13] L. In der Stroth, U. Tharehalli, C. Gunes, and A. Lechel, "Telomeres and Telomerase in the Development of Liver Cancer," Cancers (Basel), vol. 12, no. 8, Jul 24 2020, doi: 10.3390/cancers12082048.

[14] D. Moris et al., "Role of exosomes in treatment of hepatocellular carcinoma," Surg Oncol, vol. 26, no. 3, pp. 219-228, Sep 2017, doi: 10.1016/j.suronc.2017.04.005.

[15] D. Moris et al., "Program death-1 immune checkpoint and tumor microenvironment in malignant liver tumors," Surg Oncol, vol. 26, no. 4, pp. 423-430, Dec 2017, doi: 10.1016/j.suronc.2017.08.005.

[16] G. Wang et al., "Oncogenic driver genes and tumor microenvironment determine the type of liver cancer," Cell Death Dis, vol. 11, no. 5, p. 313, May 4 2020, doi: 10.1038/s41419-020-2509-x.

[17] V. Mazzaferro et al., "Milan criteria in liver transplantation for hepatocellular carcinoma: an evidence-based analysis of 15 years of experience," Liver transplantation, vol. 17, no. S2, pp. S44-S57, 2011.

[18] P. Pouramin, S. E. Allen, J. L. Silburt, and B. L. Gala-Lopez, "Median Meld at Transplant Minus 3 Reduces the Mortality of Non-Hepatocellular Carcinoma Patients on the Liver Transplant Waitlist," Current Oncology, vol. 31, no. 11, pp. 7051-7060, 2024.

[19] S. Bernards et al., "The impact of median model for end-stage liver disease at transplant minus 3 national policy on waitlist outcomes in patients with and without hepatocellular carcinoma," Liver Transplantation, vol. 28, no. 3, pp. 376-385, 2022.

[20] A. J. Kwong et al., "OPTN/SRTR 2022 annual data report: liver," American Journal of Transplantation, vol. 24, no. 2, pp. S176-S265, 2024.

[21] S. Bernards et al., "Awarding additional MELD points to the shortest waitlist candidates improves sex disparity in access to liver transplant in the United States," American Journal of Transplantation, vol. 22, no. 12, pp. 2912-2920, 2022.

[22] R. H. Shah, D. Chyou, and D. S. Goldberg, "Impact of major hepatocellular carcinoma policy changes on liver transplantation for hepatocellular carcinoma in the United States," Liver Transplantation, vol. 28, no. 12, pp. 1857-1864, 2022.

[23] C. Toso et al., "A model for dropout assessment of candidates with or without hepatocellular carcinoma on a common liver transplant waiting list," Hepatology, vol. 56, no. 1, pp. 149-156, 2012.

[24] M. Akabane et al., "Continuous risk score predicts waitlist and post-transplant outcomes in hepatocellular carcinoma despite exception changes," Clinical Gastroenterology and Hepatology, vol. 22, no. 10, pp. 2044-2052. e4, 2024.

[25] M. Akabane et al., "Continuous Risk Score Predicts Waitlist and Post-transplant Outcomes in Hepatocellular Carcinoma Despite Exception Changes," Clinical Gastroenterology and Hepatology, vol. 22, no. 10, pp. 2044-2052.e4, 2024/10/01/ 2024, doi: https://doi.org/10.1016/j.cgh.2024.05.046.

[26] G. P. Guerrini et al., "Value of HCC-MELD Score in Patients With Hepatocellular Carcinoma Undergoing Liver Transplantation," Progress in Transplantation, vol. 28, no. 1, pp. 63-69, 2018, doi: 10.1177/1526924817746686.

[27] N. Mehta, J. L. Dodge, J. P. Roberts, and F. Y. Yao, "A novel waitlist dropout score for hepatocellular carcinoma – identifying a threshold that predicts worse post-transplant survival," Journal of Hepatology, vol. 74, no. 4, pp. 829-837, 2021/04/01/ 2021, doi: https://doi.org/10.1016/j.jhep.2020.10.033.

[28] J. S. Norman et al., "Model of urgency for liver transplantation in hepatocellular carcinoma: a practical model to prioritize patients with hepatocellular carcinoma on the liver transplant waitlist," Gastroenterology, vol. 168, no. 4, pp. 784-794. e4, 2025.

[29] C. Toso et al., "A model for dropout assessment of candidates with or without hepatocellular carcinoma on a common liver transplant waiting list," Hepatology, vol. 56, no. 1, 2012. [Online]. Available: https://journals.lww.com/hep/fulltext/2012/07000/a_model_for_dropout_assessment_of_candidates_with.17.aspx.

[30] K. Han et al., "Preoperative alpha-fetoprotein slope is predictive of hepatocellular carcinoma recurrence after liver transplantation," Canadian Journal of Gastroenterology and Hepatology, vol. 21, no. 1, pp. 39-45, 2007.

[31] B. Magro et al., "Pre-transplant alpha-fetoprotein> 25.5 and its dynamic on waitlist are predictors of HCC recurrence after liver transplantation for patients meeting Milan criteria," Cancers, vol. 13, no. 23, p. 5976, 2021.

[32] H. Yu, Y. Bai, X. Xie, Y. Feng, Y. Yang, and Q. Zhu, "RECIST 1.1 versus mRECIST for assessment of tumour response to molecular targeted therapies and disease outcomes in patients with hepatocellular carcinoma: a systematic review and meta-analysis," BMJ open, vol. 12, no. 6, p. e052294, 2022.

[33] A. Kwong et al., "Machine learning to predict waitlist dropout among liver transplant candidates with hepatocellular carcinoma," Cancer Medicine, vol. 11, no. 6, pp. 1535-1541, 2022, doi: https://doi.org/10.1002/cam4.4538.

[34] S. Nagai et al., "Use of neural network models to predict liver transplantation waitlist mortality," Liver Transplantation, vol. 28, no. 7, pp. 1133-1143, 2022.

[35] S. Grossberg, "Recurrent neural networks," Scholarpedia, vol. 8, no. 2, p. 1888, 2013.

[36] Y. Hua, Z. Zhao, R. Li, X. Chen, Z. Liu, and H. Zhang, "Deep learning with long short-term memory for time series prediction," IEEE Communications Magazine, vol. 57, no. 6, pp. 114-119, 2019.

[37] S. Wiegrebe, P. Kopper, R. Sonabend, B. Bischl, and A. Bender, "Deep learning for survival analysis: a review," Artificial Intelligence Review, vol. 57, no. 3, p. 65, 2024.

[38] C. Lee, J. Yoon, and M. v. d. Schaar, "Dynamic-DeepHit: A Deep Learning Approach for Dynamic Survival Analysis With Competing Risks Based on Longitudinal Data," IEEE Transactions on Biomedical Engineering, vol. 67, no. 1, pp. 122-133, 2020, doi: 10.1109/TBME.2019.2909027.

[39] C. Lee, W. Zame, J. Yoon, and M. Van Der Schaar, "Deephit: A deep learning approach to survival analysis with competing risks," in Proceedings of the AAAI conference on artificial intelligence, 2018, vol. 32, no. 1.

[40] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network," BMC medical research methodology, vol. 18, no. 1, p. 24, 2018.

[41] K. J. Halazun et al., "Dynamic α-fetoprotein response and outcomes after liver transplant for hepatocellular carcinoma," JAMA surgery, vol. 156, no. 6, pp. 559-567, 2021.

[42] Q. Lai et al., "A novel prognostic index in patients with hepatocellular cancer waiting for liver transplantation: Time–Radiological-response–Alpha-fetoprotein–INflammation (TRAIN) score," Annals of surgery, vol. 264, no. 5, pp. 787-796, 2016.

[43] M. R. Marvin, N. Ferguson, R. M. Cannon, C. M. Jones, and G. N. Brock, "MELDEQ: An alternative Model for End-Stage Liver Disease score for patients with hepatocellular carcinoma," Liver Transplantation, vol. 21, no. 5, pp. 612-622, 2015, doi: 10.1002/lt.24098.

[44] P. C. Austin, M. Ibrahim, and H. Putter, "Accounting for competing risks in clinical research," Jama, vol. 331, no. 24, pp. 2125-2126, 2024.

[45] E. R. Perito et al., "The impact of increased allocation priority for children awaiting liver transplant: a liver simulated allocation model (LSAM) analysis," Journal of pediatric gastroenterology and nutrition, vol. 68, no. 4, pp. 472-479, 2019.

[46] C. Blandon, S. J. Karp, M. Shah, R. J. Lynch, and D. S. Goldberg, "Assessing LSAM's ability to account for changes in organ donation and transplant center behavior," Liver Transplantation, vol. 30, no. 11, pp. 1123-1131, 2024.

[47] A. Goel et al., "Liver simulated allocation modeling: were the predictions accurate for share 35?," Transplantation, vol. 102, no. 5, pp. 769-774, 2018.

[48] R. B. Freeman Jr, R. H. Wiesner, J. P. Roberts, S. McDiarmid, D. M. Dykstra, and R. M. Merion, "Improving liver allocation: MELD and PELD," American Journal of Transplantation, vol. 4, pp. 114-131, 2004.

[49] K. Washburn, E. Pomfret, and J. Roberts, "Liver allocation and distribution: possible next steps," Liver Transplantation, vol. 17, no. 9, pp. 1005-1012, 2011.

[50] A. Wey et al., "The effect of acuity circles on deceased donor transplant and offer rates across model for end-stage liver disease scores and exception statuses," Liver transplantation, vol. 28, no. 3, pp. 363-375, 2022.

[51] N. Mehta, J. L. Dodge, R. Hirose, J. P. Roberts, and F. Y. Yao, "Predictors of low risk for dropout from the liver transplant waiting list for hepatocellular carcinoma in long wait time regions: implications for organ allocation," American Journal of Transplantation, vol. 19, no. 8, pp. 2210-2218, 2019.

[52] R. Lencioni and J. M. Llovet, "Modified RECIST (mRECIST) assessment for hepatocellular carcinoma," in Seminars in liver disease, 2010, vol. 30, no. 01: © Thieme Medical Publishers, pp. 052-060.

[53] A. Attia, I. A. Rowe, E. M. Harrison, T. Gordon-Walker, and B. M. Stutchfield, "Implausible algorithm output in UK liver transplantation allocation scheme: importance

of transparency," The Lancet, vol. 401, no. 10380, pp. 911-912, 2023/03/18/ 2023, doi: https://doi.org/10.1016/S0140-6736(23)00114-9.

[54] J. S. Norman, P. J. Li, P. Kotwani, A. M. Shui, F. Yao, and N. Mehta, "AFP-L3 and DCP strongly predict early hepatocellular carcinoma recurrence after liver transplantation," Journal of Hepatology, vol. 79, no. 6, pp. 1469-1477, 2023/12/01/ 2023, doi: https://doi.org/10.1016/j.jhep.2023.08.020.

[55] A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[56] W. R. Kim et al., "MELD 3.0: The Model for End-Stage Liver Disease Updated for the Modern Era," (in eng), Gastroenterology, vol. 161, no. 6, pp. 1887-1895.e4, Dec 2021, doi: 10.1053/j.gastro.2021.08.050.

[57] S. R. o. T. Recipients, "Liver Simulated Allocation Model (LSAM) User Guide," Minneapolis, MN, 2019. [Online]. Available: https://www.srtr.org/media/1361/LSAM-2019-User-Guide.pdf

[58] H. R. a. S. Administration, "FAQ: Changes to Medical Urgency (MMAT) and Sorting at the Donor Hospital ", 1/20/2026 2021. [Online]. Available: https://www.hrsa.gov/sites/default/files/hrsa/optn/1379-faq-mmat-at-donor-hospital-sorting-changes.pdf

[59] H. R. a. S. Administration. "Liver: Policy Issues " U.S. Department of Health and Human Services https://www.hrsa.gov/optn/policies-bylaws/policy-issues/liver-intestine-policy (accessed 1/20/2026.

[60] GeoNames. "GeoNames Geographical Database." unxos gmbh. https://www.geonames.org/ (accessed 1/20/2026).

Chapter 5 Summary and Discussion of Core Studies

This collection of studies explored longitudinal data analysis approaches in the liver transplant (LT) domain. Healthcare data generation has exploded with the use of electronic health records (EHRs) and an increase in ease and availability of laboratory, imaging, and other diagnostic testing. The analysis approaches, however, have remained limited to static snapshots of patient health and status. In LT, the most widely used prognostic model the Model for End-Stage Liver Disease (MELD), is limited to risk assessment at a single timepoint despite the availability of time-series data and growing evidence of its prognostic importance. This dissertation has explored the shift from conventional cross-sectional models to advanced longitudinal frameworks, ranging from mixed-effects clustering to offline reinforcement learning. This work highlights the temporal nuances of LT data while providing tools for the creation of clinically usable longitudinal models. The four core studies of this work provide a broad view of longitudinal data analysis approaches.

Center-level Variations are Associated with Disparate Waitlist Mortality and Probability of Liver Transplant in MELD 35 and Greater Patients

Center-level variation in high-acuity LT candidates was explored using mixed-effects clustering where we identified significant disparities in waitlist outcomes of high-MELD (≥35) patients. Our analysis demonstrated that disparate waitlist outcomes are attributable to transplant center, stemming from institutional practices rather than purely patient factors. Some LT centers may have more experience and comfort in applying aggressive perioperative management strategies to allow for LT in the highest acuity

patients. The variation in waitlist outcomes attributable to center was most pronounced in those patients initially listed at a MELD score < 35 with subsequent decompensation and increase in MELD to ≥ 35, demonstrating the prognostic importance of MELD score change over time.

Several factors that could reduce center effect including UNOS region, transplant volume and median MELD at listing/transplant were assessed. None of these factors were significantly associated with risk-group. However, we also compared causes of death of waitlisted candidates among the three groups and found multi-system organ failure to account for the vast majority of candidate deaths in every group. This finding could suggest that these differences are in part attributable to critical care management expertise and the ability to treat and transplant critically ill patients with multiple organ system failures. In a broader context, we cannot exactly pinpoint what steps could be taken at the “high-risk” centers to reduce the elevated waitlist mortality in high MELD patients, but the data clearly showing that there are center-level disparities in the care of high MELD patients is useful for the transplant community to know in general. Such information can inspire efforts to correct these disparities at the center and national levels even though we cannot entirely expose the specific problems causing the disparity. Additionally, there are patients who are deemed “too sick for transplant” and are de-listed, but the criteria that result in this action are not clearly defined and are influenced by many factors including medical leadership, surgical leadership, hospital finances, and center culture. This work exposes the fact that different centers undoubtedly have differing criteria for illness that precludes transplant and the low-risk group successfully

transplanting the highest acuity patients with good post-transplant outcomes could inspire these centers to re-examine these criteria.

Future studies may examine more granular institutional data to identify specific modifiable factors a center could change to reduce the center-level disparity. Additional clustering approaches including unsupervised approaches may identify previously unknown patient subgroups that contribute to disparate waitlist outcomes.

Liver Transplantation at the Extreme: Analysis of Outcomes and Predictors of Futility in Acute-on-chronic Liver Failure Grade 3 Patients on Maximal Life Support

Acute-on-chronic liver failure (ACLF) grade 3 represents a hyper-acute subgroup of the LT population, characterized by cirrhosis with rapid decompensation and multi-system organ failure. These patients are characterized by extremely high waitlist mortality and poor post-transplant survival without appropriate patient selection. Our work herein identified an even more acute subgroup of patients with ACLF grade 3 that we coined extreme ACLF-3 (eACLF3), characterized by those with ACLF-3, requiring vasopressor support, renal replacement therapy, and mechanical ventilation. Our investigation identified this subgroup of patients at a single center and in a multi-center consortium, and elucidated risk factors for futility after LT including duration of sepsis prior to LT, indication for intubation, donor quality, change in MELD score, and change in respiratory status on the ventilator. This work demonstrated that even in this subgroup of extremely high-risk LT recipients, they could achieve reasonable post-transplant survival with appropriate selection based on our identified risk factors. While the low sample size made model development and validation challenging, this work was still able to highlight a unique subgroup of LT patients that certain centers do have experience transplanting with

success. The experience shown by this work can offer information to providers who find themselves caring for patients that could be classified as eACLF-3 with little formal guidance on when LT is an appropriate option. Patients without the identified risk factors for futility after transplant may provide some degree of confidence in providing LT for those with limited experience treating this population.

Due to the limited sample size and granularity of data required to identify the patient population of interest and risk factors for futility after transplant, complex modeling approaches were ill-suited to this analysis. Likely, regardless of model complexity, high discrimination of risk would be unlikely to be achieved in a generalizable way due to the number of events and in this setting, such was not the goal. Rather, we sought to identify a unique hyper-acute subset of LT recipients that our center has considerable experience in caring for, characterize their outcomes, and identify risk factors for futility after LT. These objectives were achieved using simple regression techniques alone.

Liver Transplant Donor-recipient Matching with Offline Reinforcement Learning

Offline reinforcement learning (RL) was applied to the complex decision-making process of donor-recipient matching in LT. Predictive modeling in LT is typically done for waitlist and post-transplant settings separately. However, in real-world decision making the anticipated waitlist and post-transplant risks are connected and weighed together when match suitability is determined. Additionally, the waitlist episode is never a static set of values as most patients remain on the list for a length of time where patient status changes regularly. The offline RL approach attempts to provide a modeling framework more consistent with the true decision-making process that occurs when a possible donor-recipient match is proposed. Using this methodology, we were able to show that

decisions made by the model could have led to a significant reduction in waitlist dropouts and one-year graft failures without significantly adversely affecting successful transplants. The model's decisions clearly correlated with improved post-transplant survival in those who were transplanted, and such a system could assist providers in making optimal decisions regarding liver donor-recipient matching.

Traditional modeling techniques would be inadequate in terms of predicting an optimal decision based on a balanced risk of waitlist mortality and post-transplant graft failure, which our approach achieves. We relaxed several reality constraints intentionally in our approach to demonstrate the concept's validity. In future work, model performance in different patient subgroups and geographic regions needs to be assessed to determine any potential bias. Additionally, our approach is currently candidate centric where each potential donor-recipient pair and the decision to accept it or not is agnostic to all other candidates and possible donor-recipient pairs at the same time. This is how the current system works where a provider with an offered organ is not expected to consider if the organ could be better placed in a different eligible patient not under his or her care. However, a truly optimal system that maximizes benefit of LT at the population level would consider all possible donor-recipient matches at a given point in time. Such a framework would likely require a "multi-agent" system where many potential donor-recipient matches would be considered simultaneously and the pair resulting in the best predicted outcome would be chosen.

HCC-3D: Improving Liver Transplant Allocation with the Hepatocellular Carcinoma Dynamic-Deep-Drop Model

Patients with hepatocellular carcinoma (HCC) have a risk of waitlist mortality that is independent of their degree of liver dysfunction as measured by the Model for End-Stage Liver Disease (MELD) score. To ensure equitable distribution of donor livers to HCC patients, they are eligible to receive MELD exception points if meeting certain criteria at the median MELD at transplant surrounding the donor hospital minus three points (MMaT-3). While this system has increased the transplant priority of HCC patients in general, it fails to differentiate risk of waitlist dropout among HCC patients who on a given match run (one donor) would have the same MMaT-3 based exception point score regardless of their tumor burden, treatments received, or degree of liver dysfunction. To improve equitable distribution of donor livers to HCC patients, we developed the HCC Dynamic-Deep-Drop Model (HCC-3D), which takes longitudinal changes in tumor size, number, tumor biomarkers, and liver disease biomarkers to discriminate patients based on risk of waitlist dropout with state-of-the-art accuracy.

A major contribution of this work is in bridging the gap between a deep learning model architecture with high performance metrics and use in an existing and established system in clinical medicine, in this case the organ allocation system. To do this, we used our model to create a risk score that is scaled to the established MELD 3.0 score, reporting a novel derivation to link the deep-learning derived survival curve to the MELD 3.0 survival curve. We then provided several simulations using the continuous version of our new risk score as well as the score being used to modify exception point allocation based on MMaT. In the latter case, we included this MMaT-modified simulation because MMaT-3 is an established framework that has significant

advantages enjoyed by the LT community, despite its deficits that we attempted to solve in this paper.

To simulate the use of our model, we created an open-source, fully configurable discrete event simulation based on the Liver Simulated Allocation Model (LSAM). LSAM is controlled by the Scientific Registry of Transplant Recipients (SRTR) with no publicly available source code and limited configurability. Our simulation framework is well-validated in terms of replicating the real-world outcomes at baseline and could be used by future researchers to simulate any other potential allocation policy changes in the US concerning HCC or non-HCC candidates.

Chapter 5 Key Takeaways

The use of traditional regression and clustering techniques with mixed effects is beneficial when sample sizes are limited, the focus is on risk factor identification in a unique patient population, or we want to isolate the effect of a single grouping variable on outcomes. When more complex problems need to be solved like creating a (clinical) decision support tool that balances the risk of waitlist mortality and post-transplant graft failure or death by assessing candidate features and longitudinal trajectories on the waitlist, the use of offline RL was essential. Additionally, to harness the full predictive capability of longitudinal feature changes in HCC LT candidates, a longitudinal survival network allowed achievement of optimal performance, surpassing all previously reported models. As this work demonstrates, excellent performance is only the first step, which must be followed by an implementation plan for how the model could be integrated into existing systems in clinical medicine. The use of simulation can bolster these efforts as well as combining the use of a model with an existing framework that

has aspects that are advantageous as we did with the use of HCC-3D with the MMaT framework.

Lessons Learned

Transplant outcomes are not purely biological, but systemic where a MELD ≥35 can have a different prognostic significance depending on geography and even individual center. The concept of acuity must be filtered through the lens of regional differences in donor availability and local practice patterns, experience, and risk tolerance. National policy and risk-score based prognostication must account for center-level behavior to truly achieve equity and fairness of liver allocation, which will be especially important as more advanced models are introduced.

The proposed longitudinal deep survival model, HCC-3D, demonstrates exceptional predictive performance in HCC waitlist prognostication. However, such advanced models will likely reach a ceiling in performance if variation attributable to non-biological parameters like center variability is not considered. Future studies should continue to examine variability of risk-score based prognostication not attributable to biological parameters to understand how such variability can be modeled and accounted for.

Exceptional model performance alone does not guarantee model adoption or usability. In the case of HCC-3D, the continuous Meq score serves as a purely data-driven assessment of differential waitlist dropout risk in HCC patients that can be calculated in a rolling fashion as new data points become available. However, in the process of doing this study, it was realized that while the existing MMaT-3 framework has clear shortcomings in terms of giving all HCC patients the same priority, it also has

desirable attributes that are accepted and enjoyed by the transplant community. Namely, the MMaT-3 framework is robust to regional wait-time differences because it's based on the median MELD at transplant in the area surrounding the donor hospital so all eligible HCC patients on the match run receive the same MMaT-3 regardless of their geographic location. This prevents patients from intentionally moving locations to a lower wait-time region to receive a transplant faster. We therefore included a heuristic modification to the existing MMaT-3 system using our model where the desirable aspects of MMaT-3 are preserved, while allowing for differential allocation priority based on predicted risk of waitlist dropout. This serves as an emblematic example of how to integrate complex predictive models with high accuracy into an existing and accepted clinical system to maximize the chance of adoption and use of the model. Additionally, the use of a simulated allocation framework provides further evidence that introduction of a new model will not disrupt the existing allocation system in unexpected ways and can further demonstrate improvement in outcomes afforded by adoption of the model. Such simulation is essential for any proposed model that will change the way livers are allocated.

While deep survival networks exhibited high predictive performance for HCC patient waitlist prognostication, the optimal LT donor-recipient match calls for a fundamentally different solution. The donor-recipient match represents a higher level of complexity than standard patient-level risk prediction as donor data and candidate data must be assessed simultaneously with multiple potential outcomes considered separately as opposed to accurate prediction of a single outcome. Regression approaches to donor-recipient matching where graft failure or early post-LT mortality is

predicted from donor-recipient features clearly fail to capture the complexity involved in the real match-making process. However, even more advanced architectures for single outcome prediction will likely be similarly inadequate. The donor-recipient match is an optimization problem with many disparate data points in the donor and recipient that must be considered with multiple undesirable outcomes that must accounted for. Our offline RL model approached this problem at the level of an individual candidate, where decisions were made for this candidate and a potential offered donor as this mirrors the real-life decision-making process. Transplant providers make donor-recipient match decisions based on the candidate under their care with no expectation or responsibility to consider other potential candidates not under their care. However, as further efforts are made to design intelligent donor-recipient match decision support tools, a multi-state system where every potential candidate is considered simultaneously and matches are made based on the largest overall benefit at the population level is attractive and may be possible in the future.

Challenges

The studies described were highly reliant on United Network for Organ Sharing and Scientific Registry of Transplant Recipients databases, which are valuable due to every transplant patient in the US being represented in them. However, these databases have several drawbacks including a lack of granularity that makes a full and robust assessment of patient acuity difficult as well as variables with missing data, ambiguous coding of certain variables, and changes in data collection through the years that are present but not clearly described. The MELD exception form data requires substantive data cleaning and data mining efforts to become easily usable for model

training and evaluation. There is no standard approach for this process, which could lead to issues with reproducibility and verification of documented findings.

There is a large computational-clinical divide where technological advancement in model architecture and ability has outpaced clinical adoption at an incredible speed. An understanding and acknowledgement of this gap is critical to creating machine learning (ML) models that can integrate into existing clinical systems and have successful adoption. There is no shortage of reports of advanced ML models in the LT literature with no evidence of adoption, which I believe is due to model complexity, lack of easy interpretability seen with regression-based approaches, and lack of explanation for how the model could be used in the real-world. Truly novel approaches like offline RL for donor-recipient matching require an initial report and proof-of-concept as we have done, but subsequent work must tackle these barriers to implementation where an approach to implementation in the current allocation system is assessed and the potential impact of doing so is tested using simulation.

Concluding Remarks

Overall, the traditional methods explored in the first two studies highlight potential problems and areas for improvement in terms of geographic variation in waitlist outcomes and the prognostic significance of longitudinal risk factors. The HCC-3D model bridges the gap between existing outcome-centric patient-level prediction frameworks, utilizing a more complex longitudinal deep learning architecture to maximize predictive accuracy while functioning within the existing allocation system. Finally, the use of offline RL explores donor-recipient matching, a clinical problem space

that is likely not completely addressed by simple outcomes prediction and better suited to RL-based optimization.

The culmination of this work is a step away from static “point-in-time” decision-making and highlighting how longitudinal data can be better leveraged in different clinical applications. Taken together, this work could improve the liver allocation system by integrating state-of-the art predictive analytics (longitudinal deep survival networks) with a donor-recipient matching framework (offline RL) and testing prior to implementation using discrete event simulation. We have thus provided a blueprint for the eventual development of an allocation system that is more equitable through heightened predictive accuracy, more reflective of an individual’s trajectory on the waitlist, and that strives to achieve the maximum benefit afforded by transplant given each possible donor-recipient pair.