


OPEN

Machine Learning Prediction Model of Waitlist Outcomes in Patients with Primary Sclerosing Cholangitis

Xun Zhao¹ , MD,¹ Maryam Naghibzadeh, MD,¹ Yingji Sun, MSc,¹ Arya Rahmani, BSc,¹ Leslie Lilly, MD,^{1,2} Nazia Selzner, MD, PhD,^{1,2} Cynthia Tsien, MD, MPH,^{1,2} Elmar Jaeckel, MD,^{1,2} Mary Pressley Vyas,³ Rahul Krishnan, PhD,^{4,5} Gideon Hirschfield, MD, PhD, MB,^{1,2} and Mamatha Bhat, MD, PhD^{1,2}

Background. Liver transplantation is essential for many people with primary sclerosing cholangitis (PSC). People with PSC are less likely to receive a deceased donor liver transplant compared with other causes of chronic liver disease. This disparity may stem from the inaccuracy of the model for end-stage liver disease (MELD) in predicting waitlist mortality or dropout for PSC. The broad applicability of MELD across many causes comes at the expense of accuracy in prediction for certain causes that involve unique comorbidities. We aimed to develop a model that could more accurately predict dynamic changes in waitlist outcomes among patients with PSC while including complex clinical variables. **Methods.** We developed 3 machine learning architectures using data from 4666 patients with PSC in the Scientific Registry of Transplant Recipients (SRTR) and tested our models on our institutional data set of 144 patients at the University Health Network (UHN). We evaluated their time-dependent concordance index (C-index) for mortality prediction and compared it against MELD-sodium and MELD 3.0. **Results.** Random survival forest (RSF), a decision tree-based survival model, outperformed MELD-sodium and MELD 3.0 in both the SRTR and the UHN test data set using the same bloodwork variables and readily available demographic data. It achieved a C-index of 0.868 (SD 0.020) and 0.771 (SD 0.085) on the SRTR and UHN test data, respectively. Training a separate RSF model using the UHN data with PSC-specific achieved a C-index of 0.91. In addition to high MELD score, increased white blood cells, time on the waiting list, platelet count, presence of Autoimmune hepatitis-PSC overlap, aspartate aminotransferase, female sex, age, history of stricture dilation, and extremes of body weight were the top-ranked features predictive of the outcomes. **Conclusions.** Our RSF model offers more accurate waitlist outcome prediction in PSC. The significant performance improvement with the inclusion of PSC-specific variables highlights the importance of disease-specific variables for predicting trajectories of clinically distinct presentations.

(*Transplantation Direct* 2025;11: e1774; doi: 10.1097/TXD.0000000000001774.)

Primary sclerosing cholangitis (PSC) is a chronic and progressive cholestatic liver disease with a median time from diagnosis to death or liver transplantation (LT) of 20 y.^{1,2} There are no approved pharmacological treatments that show

consistent evidence of efficacy, and LT is an essential treatment for many.^{1,3} Organ allocation in the United States and Canada relies on the model for end-stage liver disease-sodium (MELD-Na) or MELD 3.0^{4,5} scores to predict mortality. The

Received 18 December 2024. Revision received .
Accepted 1 January 2025.

¹ Ajmera Transplant Program, University Health Network, Toronto, ON, Canada.

² Division of Gastroenterology and Hepatology, Department of Medicine, University of Toronto, Toronto, ON, Canada.

³ PSC Partners Seeking a Cure Canada, Toronto, ON, Canada.

⁴ Department of Computer Science, University of Toronto, Toronto, ON, Canada.

⁵ Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada.

This study has received funding from the PSC Partners.

The authors declare no conflicts of interest.

X.Z. contributed in writing, conceptualization, and article preparation. M.N. contributed in data curation, writing, and article preparation. Y.S. contributed in statistical analysis and reviewing article for intellectual content. A.R. contributed in article preparation. L.L., N.S., C.T., E.J., and M.P.V. contributed in reviewing article for intellectual content. R.K. and G.H. contributed in conceptualization and reviewing article for intellectual content. M.B. contributed in conceptualization, reviewing of article for intellectual content, article preparation, and funding. X.Z. and M.N. are co-first authors.

Supplemental digital content (SDC) is available for this article. Direct URL citations appear in the printed text, and links to the digital files are provided in the HTML text of this article on the journal's Web site (www.transplantationdirect.com).

Correspondence: Mamatha Bhat, MD, PhD, Multiorgan Transplant Program, University Health Network, MaRS 9-9055, 585 University Ave, Toronto, ON M5G 2N2, Canada. (mamatha.bhat@uhn.ca); Gideon Hirschfield, MD, PhD, MB, Toronto Centre for Liver Disease (TCLD), Toronto General Hospital, University Health Network Eaton Building, 200 Elizabeth St, Toronto, ON M5G 2C4, Canada. (gideon.hirschfield@uhn.ca).

Sponsors did not have any role in study design, data collection, analysis, interpretation, or writing.

Copyright © 2025 The Author(s). *Transplantation Direct*. Published by Wolters Kluwer Health, Inc. This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

ISSN: 2373-8731

DOI: 10.1097/TXD.0000000000001774

MELD score is generalizable to many causes of liver diseases; however, it is not a disease-specific risk prediction model. Bilirubin is underweighted and mortality of cholestatic diseases such as PSC is not well captured.⁶⁻⁸ Therefore, despite having excellent posttransplant overall survival and graft survival,⁹⁻¹¹ people with PSC are currently disadvantaged by the MELD-Na allocation system. They wait longer on the deceased donor LT (DDLT) waitlist and disproportionately rely on living donor LT (LDLT) when compared with other populations.⁷ Recent publications confirm that people with PSC have a higher risk of mortality on the DDLT waitlist when LDLT is not available.^{12,13}

Machine learning (ML) models, unlike traditional statistical regression models, do not assume fixed shapes to hazard functions and can consider changing hazard dynamics to make more accurate predictions. Recent publications demonstrated applications of ML risk prediction in the field of LT.¹⁴⁻¹⁶ We hypothesize that ML models can more accurately predict waitlist outcomes in PSC than the MELD score and that the integration of PSC-specific complications as variables is important to enhance model performance.

MATERIALS AND METHODS

Setting and Study Design

The University Health Network (UHN) in Toronto is the largest LT center in Canada, with >300 people listed for LT annually. The Scientific Registry of Transplant Recipients (SRTR) and its 11 Organ Procurement and Transplantation Network (OPTN/United Network for Organ Sharing) is a US registry with 200 000 waitlisted LT candidates. This study retrospectively retrieved data from consecutive adult patients with PSC listed for LT from November 2012 to June 30, 2021, at the UHN and from February 27, 2002, to December 31, 2020, from the SRTR. The study was ethically approved by the Research Ethics Board of UHN (approval No. REB 21-6104). The study start period corresponded to the adoption of the MELD-based allocation system for the UHN and the OPTN regions. We reported our ML protocol in accordance with the MINIMAR reporting standards¹⁷ for artificial intelligence in healthcare as recommended by the equator network (Enhancing the QUALity and Transparency Of health Research).

Participants

Patients who required multiorgan transplants or retransplantation were excluded. Those listed with HIV, concomitant hepatocellular carcinoma (HCC), and other chronic liver diseases were excluded, given that PSC was not the only contributor to their transplant listing. Those listed with exception MELD scores or status 1 were also excluded, given that their calculated MELD score was not used for organ allocation.

Variables and Outcomes Measured

Variables from the UHN database were collected using the electronic transplant database (Organ Transplant Tracking Record: Transplant Care Platform 6, Organ Transplant Tracking Record Chronic Care Solutions, Omaha, NE). From both the SRTR and the UHN, demographic and clinical variables were collected, namely age, biological sex, race, ABO blood group, body mass index, and presence of ascites. Laboratory variables collected are creatinine, bilirubin, international normalized ratio (INR), sodium, and albumin, which also corresponds to all the variables used in MELD-Na and

MELD 3.0. PSC-specific complications were available in the UHN database. These include episodes of cholangitis, concomitant cirrhosis, pruritus, ascites, stricture dilation, biliary dysplasia, biliary stent, concomitant autoimmune disease including inflammatory bowel disease, colorectal dysplasia on colonoscopy, cholecystectomy, and colectomy. The UHN data set also included expansive laboratory variables such as white blood count, platelet count, serum sodium, creatinine, INR, bilirubin, albumin, aspartate aminotransferase (AST), and alanine aminotransferase.

Waitlist outcomes were defined as follows:

1. Death on the waitlist, which includes patients removed from the waitlist due to being too sick to transplant.
2. Received a LT.
3. Removed from the waitlist due to improvement in their condition or opting out of transplantation.

Statistical Analysis

Comparisons between 2 groups were made by independent 2-sample *t* test or Wilcoxon tests for continuous variables and the Fisher exact test or the chi-square test for categorical variables.

ML Model Architecture

The Cox proportional hazard (CPH) model was used as a baseline model for time-to-event survival analysis. The random survival forest (RSF) model is an ensemble tree architecture adjusted for right-censored survival data.¹⁸ It randomly samples observations that are independent of each other. RSF is nonparametric in nature and does not assume hazard function shape, giving it an advantage in predicting changing survival dynamics. The RSF model makes no assumption of linearity or additivity between variables, allowing it to capture complex interactions between variables. It also readily accommodates censored data, as each decision tree's optimal splitting criteria consider censoring. DeepSurv is a neural network-based architecture adapted for survival analysis.¹⁹ Like RSF, it also makes no assumption on hazard function. Nodes in this neural network architecture receive input, assign a sum of weights, and output an activation function. DeepSurv architecture and layers of nodes can be adjusted for specific data sets and thus perform well with high-dimensional data or large data sets.

Performance Evaluation Metric

The model's performance was evaluated using the time-dependent concordance index (C-index), which compares all possible pairs of participants for any chosen time point (*t*). A pair is considered concordant if the higher predicted risk participant experienced the event at time *t*. Therefore, the time-dependent concordance index evaluates the discrimination of model predictions at any specific time. The ML models C-index were compared with MELD-Na and MELD 3.0.

ML Pipeline

Data Preprocessing

Data preprocessing was completed with guidance from the domain expertise of transplant hepatologists. Only variables that contain <20% missingness in both data sets were retained to minimize need for data imputation. Continuous variables were inputted with population mean and binary variables with 0.5. One hot encoding was applied to categorical variables.

Training, Validation, and Testing

Training and validation were split 70% and 30%. This data-splitting ratio was used throughout the analysis. Data z-normalization to the training distribution was performed on data splitting. Normalizing variables on a common scale allows for more efficient training and can reduce bias introduced in learning due to large differences in the scales of different variables.

Cross-validation and Hyperparameter Tuning

Cross-validation helps to ensure the model's integrity as it avoids relying on a single fixed data split, which may result in a high variance. Five-fold cross-validation was performed during hyperparameter tuning to find the optimal hyperparameters for our models. Five-fold cross-validation divides the data into 5 subsets, or folds, then trains and validates the model multiple times, each time using a different subset. Optimal hyperparameters were subsequently selected and used for training on the training set. Performance on test set was then done using bootstrapping.

Model Selection Process

- Step 1. We compared the performance of model architectures CPH, RSF, and DeepSurv trained and validated on SRTR against MELD-Na and MELD 3.0. The top 2 architectures were retained for further refinement using SRTR data.
- Step 2. Top 2 model architectures were trained on the 11 OPTN regions in the SRTR database and tested on each of the OPTN regions as well as the UHN database. The best performing model was identified.
- Step 3: Best performing model trained on the top 2 OPTN regions was then tested on the UHN data set.

We also separately trained and tested a model using the UHN data set with the inclusion of disease-specific variables.

Model Interpretability

Shapley (SHAP) values²⁰ were assessed to determine the variables with the highest impact on prediction. Population-level SHAP values express the marginal contribution of each variable to the model prediction. For a given variable, a SHAP value is calculated with all the possible subsets of other variables that do not contain the specific variable of interest. The difference in model prediction with and without the variable of interest is retained for all possible subsets, and the average is expressed as a SHAP value. SHAP values can be ranked to assess the most impactful variable to the model. The target user of our models was intended to be clinicians.

Python code for this analysis is made available on GitHub for transparency: https://github.com/iriejy-sun/PSC_waitlist_mortality.

This study was approved by the Research Ethics Board at the UHN (REB study #21-6104.0.1) and was conducted in accordance with the Declaration of Helsinki and Istanbul.

RESULTS

Participant Characteristics

Between February 27, 2002, and December 31, 2020, 9000 adult participants listed for PSC were identified in the SRTR. Applying exclusion criteria, 1186 had previous LTs, 252 had

a previous multiorgan transplant, 263 participants were listed with other concomitant liver disease, 8 were listed with HIV, 185 were listed with HCC, and 9 had HCC prelisting. Another 1266 had received MELD exception points, and 349 were listed under status 1. 692 participants were further removed as they had refused transplant or were lost to follow-up. One hundred twenty-four participants did not have an activation date for LT, and thus we could not establish a time-to-event listing outcome. The study flow diagram is shown in **Figure S1** (SDC, <http://links.lww.com/TXD/A747>).

In total, 4666 participants from the SRTR and 144 participants from the UHN were retained for training and validating our ML models. The demographic data of the cohorts are summarized in Table 1. Participants were mostly of male sex and had a mean age at listing of 48 y. Participants listed at UHN had a higher average MELD score of 20 compared with 18 in the SRTR.

Overall, there were 63 224 follow-up visits in the SRTR database and 4487 follow-up visits in the UHN database, with a mean follow-up of 449 and 348 d, respectively. Up to the time of our review of the database, December 31, 2020, 19.4% of participants on the SRTR group had died on the waitlist or were delisted because of being too sick to transplant. In the UHN group this was 11.8%. The mean time to event of death/delisting was 763 and 927 d, respectively, in the SRTR and UHN. The mean time to transplant was comparable at 284 d in SRTR and 250 d in UHN. However, 60% of all transplanted subjects at UHN received DLT, as opposed to 18% in SRTR. Notably, 8.7% of SRTR subjects and 14.6% of UHN subjects were censored or removed from the waitlist due to their condition improving and no longer requiring a LT.

Step 1. Comparison of 3 ML Survival Analysis Models (CPH, RSF, and DeepSurv) Trained and Tested in the SRTR Against MELD

CPH, RSF, and DeepSurv were compared against MELD-Na and MELD 3.0 to assess mortality. Time-dependent C-index can be inferred at any time frame, but specific time frames of 30, 183, and 365 d are shown in Figure 1 for the purpose of data visualization. CPH and RSF algorithms outperformed MELD-Na and MELD 3.0 with higher concordance indices at 90 d, the time frame for mortality prediction the MELD score was originally built on. CPH and RSF showed the highest time-dependent C-index at 30, 183, and 365 d: 0.872, 0.868, and 0.816, respectively, for CPH and 0.856, 0.857, and 0.812 for RSF. These 2 models were retained for step 2.

Step 2. Identifying the Best Performing ML Model Between CPH and RSF by Training Them on Individual OPTN Regions and Testing Them on Each OPTN Region as well as the UHN Database

To differentiate performance between the top 2 performing models in step 1, we divided the SRTR database into their individual OPTN regions and then repeated the training and validation process. This was done to better capture regional variations in organ availability and patient population by OPTN region. By comparing individual OPTN regions, we can gain insights into how the models perform in various geographical and administrative areas. Subsequently, we evaluated the performance of these OPTN-trained models on both OPTN regions and the UHN data set to assess generalizability of the models to our specific institution. The average

TABLE 1.
Participants characteristics SRTR and UHN

Demographic characteristics	SRTR (N = 4666)	UHN (N = 144)	P
Sex			
Male, n (%)	3155 (67.6)	93 (64.6)	0.159
Female, n (%)	1511 (32.4)	51 (35.4)	
Race			
Non-White, n (%)	958 (20.5)	130 (90.3)	<0.001
White, n (%)	3708 (79.5)	14 (9.7)	
Age at listing, y, mean (SD)	48.1 (13.9)	47.1 (13.4)	0.266
Weight, kg, mean (SD)	77.3 (16.9)	75.2 (17.6)	<0.001
Body mass index, mean (SD)	25.6 (4.8)	25.6 (4.8)	
ABO blood type, n (%)			
A	1777 (38.1)	53 (36.8)	0.926
B	616 (13.2)	16 (11.1)	0.247
AB	182 (3.9)	8 (5.6)	0.545
O	2083 (44.7)	67 (46.5)	0.523
Blood work results at listing			
Serum albumin, g/L, mean (SD)	30.0 (6.9)	33.1 (5.1)	<0.001
Serum creatinine, μmol/L, mean (SD)	107.0 (53.4)	95.4 (57.4)	
Total serum bilirubin, μmol/L, mean (SD)	153.4 (160.4)	125.3 (135.5)	0.023
INR, mean (SD)	1.5 (0.7)	1.4 (0.4)	0.026
Serum sodium, mmol/L, mean (SD)	136.4 (3.9)	137.1 (3.3)	0.040
MELD score (SD)	17.5 (12.9)	20.3 (7.7)	<0.001
Comorbidities			
Presence of ascites			
Yes	289 (6.2)	80 (55.6)	<0.001
No	294 (6.3)	56 (38.9)	
Unknown	4083 (87.5)	8 (5.6)	

INR, international normalized ratio; MELD, model for end-stage liver disease; SRTR, Scientific Registry of Transplant Recipients; UHN, University Health Network.

time-dependent C-indices of this process are shown on 2 heatmaps, 1 for CPH and 1 for RSF (Tables S1 and S2, SDC, <http://links.lww.com/TXD/A747>).

RSF trained on OPTN regions 4 and 8 and generalized well with other OPTN regions while also performing best on our institution’s patient population. The average time-dependent C-index was 0.749 and 0.717 for these regions, respectively.

Step 3. RSF Algorithm was Trained on Combined OPTN Regions 4 and 8 and Tested on the UHN Data Set

Data from OPTN regions 4 and 8 were combined to train a new RSF model (SRTR-RSF model) and validated on our own institution’s data. Training RSF on the combined regions OPTN 4 and 8 yielded a C-index of 0.785, 0.735, and 0.798 at 30, 183, and 365 d when tested on the UHN test set. In Figure 2, we demonstrate how the models compare with each other.

Assessing Impact of PSC Disease-specific Variables

The RSF architecture was subsequently modified to train and test on UHN data only with additional PSC-specific variables (Table S3, SDC, <http://links.lww.com/TXD/A747>). It is important to note at this step that PSC-specific variables could not be used, given that the SRTR does not collect them. We first chose 3 additional UHN data features known to increase morbidity²¹ or mortality in patients with PSC: cholangitis, number of cholangitis episodes, and presence of cirrhosis.²² This modification improved C-index to 0.91 when tested on

UHN data. We then included all 17 additional UHN PSC-specific features available: cholangitis, cholangitis episodes, cirrhosis, associated autoimmune disease, pruritus, esophageal varices, endoscopic retrograde cholangiopancreatography with stricture dilation, biliary dysplasia, biliary stent, inflammatory bowel disease, colorectal dysplasia, cholecystectomy, colectomy, white blood cells (WBCs), platelets, alanine aminotransferase, and AST. This PSC-specific algorithm yielded a C-index of 0.94 when tested on UHN data. In Figure 2, we demonstrate how the models compare with each other.

SHAP Values

To understand the variables that contributed the most to the PSC-specific RSF model prediction, we assessed the SHAP values and listed them (Figure 3). In order of priority, high MELD score, increased WBC, time on a waiting list, platelet count, presence of AIH-PSC overlap, AST, female sex, age, history of stricture dilation, and extremes of body weight were the most impactful variables. Importantly, components of the MELD score, namely bilirubin, creatinine, INR, and sodium, did not individually outrank the top 10 variables for prediction.

Example of Individualized Risk Prediction

Figure S2 (SDC, <http://links.lww.com/TXD/A747>) gives an example of how this ML model can be applied to an individual subject. It provides the example of a 57-y-old women with PSC-cirrhosis and a MELD score of 15 at listing. Under the current

Mean concordance index predicting 3-month mortality on SRTR data, at various times on the waitlist

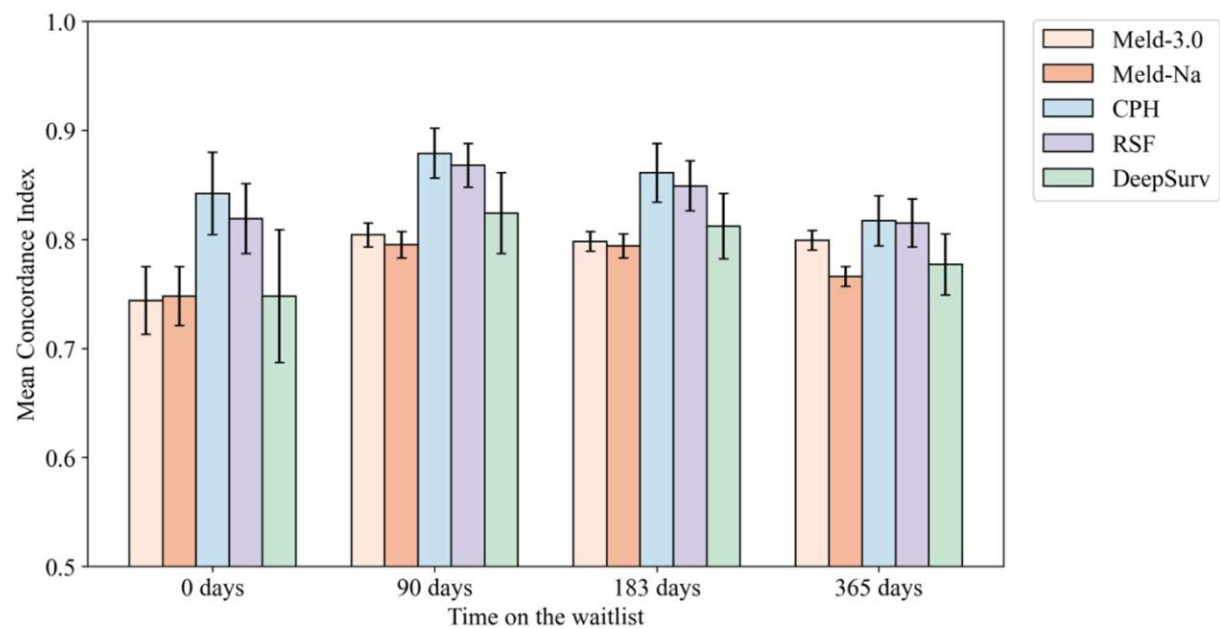


FIGURE 1. Time-dependent C-index performance comparison of mortality prediction for various models, evaluated at different time on the waitlist.

Mean concordance index predicting 3-month mortality on UHN data, at various times on the waitlist

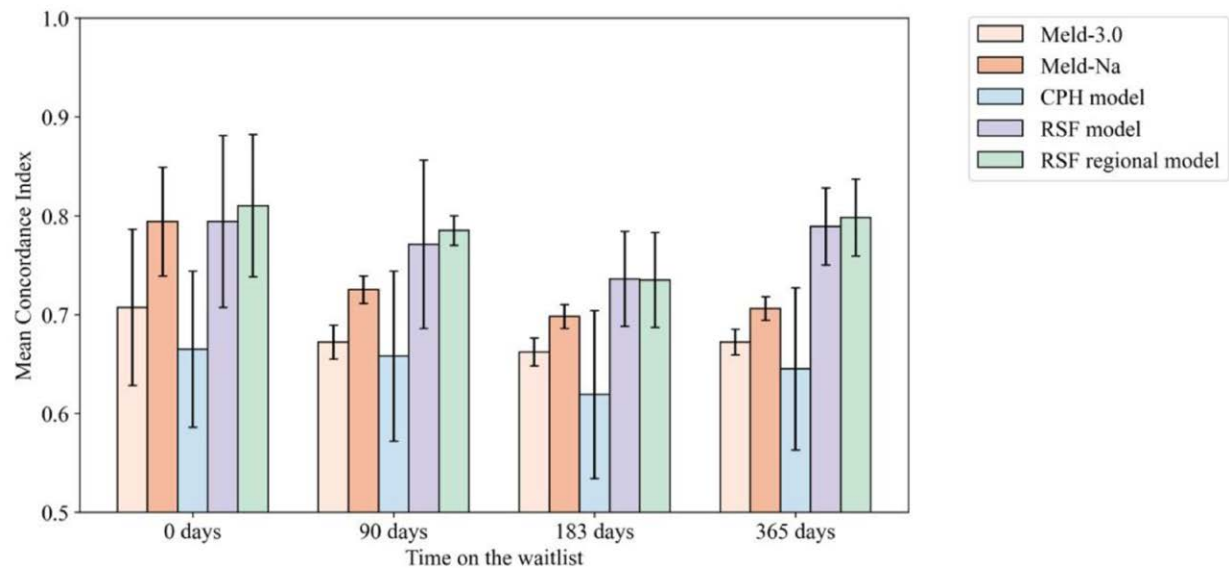


FIGURE 2. Time-dependent C-index performance comparison of mortality prediction for various models on UHN, evaluated at different time on the waitlist.

MELD allocation system, this person was unlikely to receive a DDLT. Our ML-based algorithms more accurately predicted decreased survival than the listed MELD score of 15.

DISCUSSION

The MELD score is a sum of natural logarithms. It was originally trained on a population consisting mostly of hepatitis C, a parenchymal disease,⁶ and not specifically designed to capture the severity of cholestatic diseases such as PSC.² The advantage of the MELD score lies in its overall generalizability over multiple disease causes, and therefore is useful for

organ allocation. However, as a prognostic tool for mortality, the MELD score is less accurate in certain subgroups, such as women and people with cholestatic diseases.^{23,24} Furthermore, as a CPH model, the MELD score assumes a fixed shape to the hazard function to estimate risk of mortality. ML models are more flexible and do not make this assumption. Therefore, they can consider changing risk functions through nonlinear interactions between multiple variables. Leveraging this advantage of ML is particularly interesting in the context of our research purpose, which is to model waitlist mortality in the clinically distinct population of people with PSC.

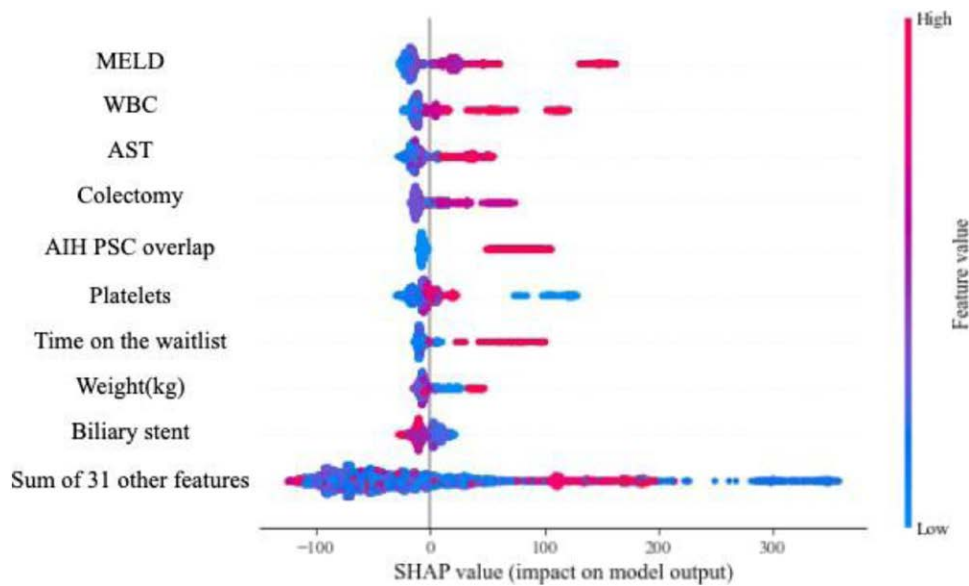


FIGURE 3. SHAP values on PSC-specific model.

In this proof-of-concept project, we have demonstrated that ML models can more accurately predict mortality in subjects listed with PSC. Our RSF model uses the same variables found in the MELD-Na and MELD 3.0 score, supplemented by a few demographic and clinical variables such as age, blood type, body mass index, and ascites. Time-dependent C-index for 90-d mortality was higher in the RSF when compared with MELD-Na and MELD 3.0 whether testing was done on SRTR data or the UHN data set. Interestingly, the C-index of the MELD-Na and MELD 3.0 was lower when tested on the UHN data set (Figure 2) than it was when tested on the SRTR data set (Figure 1). Using our hospital's patient population, which follows the largest PSC cohort in Canada, we demonstrated that MELD and its variants cannot always accurately predict waitlist mortality in PSC as its performance decreased when tested on an independent data set. However, this should not be seen as a major limitation of the MELD-based allocation system. As an organ allocation score, MELD was not designed to be disease specific. It prioritizes generalizability over accuracy when considering a heterogeneous pool of recipient candidates. Nonetheless, our RSF model is a proof of concept that ML and its nonparametric approach to hazard functions can recoup some of this lost accuracy through nonlinear interpretations of relationships between the same variables used as the MELD-Na and MELD 3.0 scores. We also observed that the inclusion of PSC-specific variables into a ML survival model can improve the accuracy of the predictions. Despite the large size of the SRTR database, the RSF model trained and then tested on the SRTR could not achieve an average time-dependent C-index beyond 0.85. This same RSF architecture was trained and tested on the smaller UHN database, with the inclusion of PSC-specific complications, such as cholangitis, and improved this time-dependent C-index to 0.94. Disease-related complications hold important prognostic information and should be explored for better modeling, whether it be using ML or regression models.

The strength of our work is built on the collaboration between ML experts, clinical experts, and patient partners during the conception, model design, and interpretability assessment phase of the study. The input of patient partners provided

valuable insights into the applicability of our work as well as a platform to discuss and reevaluate our findings through patient perspectives. We hope that input from patients will help build trust in future artificial intelligence-related clinical research and facilitate long-term patient engagement in this field of study. Particular attention was given to the interpretability of our model as we showed SHAP values to delineate the variables that contributed most to our algorithm. MELD was the top contributor and remains a good prognostication tool, which is reassuring. However, PSC-specific features, as we hypothesized, significantly enhanced the algorithm's performance. It is also reassuring that these features include previously established risk factors for poor prognosis in PSC, namely advanced age, presence of biliary stricture, and splenomegaly.²⁵⁻²⁷ Other factors such as AST and thrombocytopenia were also validated in established prognostic tools such as the Amsterdam-Oxford Score,²⁸ UK-PSC,²⁹ and PRETo.³⁰ Other high SHAP value variables in our model include high WBC, history of colectomy, concomitant autoimmune hepatitis, and low platelets. Therefore there is biological plausibility that these high SHAP input variables hold prognostic value in our ML model, as opposed to a black box model where the relationship between input variables and outcome is unexplainable. Concomitant PSC and autoimmune hepatitis may have a worse prognosis,³¹ and recent evidence showed that low platelets to WBC ratio may be associated with decompensation in cirrhosis.^{32,33}

Our work has many limitations. First, we did not attempt to build a model that includes all causes of liver disease but restricted to PSC as a proof of concept. Hence, our RSF model, although more accurate than the MELD-Na and MELD 3.0 at predicting waitlist outcomes, cannot be used for allocation prioritization. Second, our external validation data set comprises 144 subjects with PSC, and a larger validation data set would have been beneficial. However, the UHN data set spans 10 y and comes from the largest transplant center in Canada. Third, there were proportionally more LDLT recipients in the UHN data set than in the SRTR data set, which may reflect the difference in organ availability between the OPTN and our institution. Mean time to transplant was

comparable between SRTR and UHN despite this difference in donor type. Importantly however, this difference does not translate in the outcome of interest that our model is predicting. Death occurred in 19.4% of the subjects in SRTR and in 11.8% of subjects in UHN. A class imbalance between the 2 data sets is therefore not drastically different, allowing for a consistent performance evaluation.

We also highlight some limitations of ML models when compared with conventional statistical models. In step 1, RSF outperformed MELD 3.0, however, DeepSurv did not. Importantly, the CPH model was very similar to RSF in terms of performance across all time frames (Figure 1). Therefore, the enhanced performance of RSF can be mostly attributed to the addition of the number of variables compared with the MELD score. We believe that this is an important observation that demonstrates how ML models do not have an inherent advantage compared with conventional time-to-event models. Nonetheless, when comparing performance of CPH and RSF on individual OPTN regions and testing on UHN data, we observe that certain OPTN regions are more performant when validated on the UHN than others. This region-specific gain in performance was only observed using a ML-based RSF (Tables S1 and S2, SDC, <http://links.lww.com/TXD/A747>) rather than CPH, despite using all the same variables. This discrepancy was discussed among our team of ML analysts, patient allies, and clinicians. However, there was no obvious reason within the data or from our clinical knowledge where OPTN regions 4 and 8 would be more similar to our patient population at the UHN. Perhaps the advantage of ML methodologies, through its nonlinear modeling, is its capacity to infer the presence of latent variables not obvious to clinicians. From one perspective, this is a limitation that hinders the external validity of our ML model, but as a proof of concept, this also presents the opportunity to further investigate this latency between indirectly inferred relationships through mathematical models and clinically observed variables.

Our RSF-OPTN regionally trained models, when tested on UHN, outperformed the MELD scores and the CPH at 90 d with significant statistical differences. At other timeframes, there was a trend toward better C-index, but with significant overlap in confidence intervals (Figure 2). However, the MELD-Na score and MELD 3.0 were also conceived for 90-d mortality prediction and used this specific time frame for organ allocation.

Finally, when adding disease-specific variables, our model's performance increase drastically. Given that this model was developed only using data from UHN, we lack a robust externally validated data set and, therefore, must cautiously interpret these results. Nonetheless, by simply adding cholangitis, number of cholangitis episodes, and presence of cirrhosis, the performance of our model increased its overall concordance index to >0.90. These clinical variables are essential to the assessment of patients with PSC, as they are major reasons for referring patients with PSC to LT. However, these variables are not included in the MELD score or any of its updated versions for organ allocation. We recognize that cholangitis is not often seen in patients without PSC, and the presence of cholangitis as a predictor of mortality is not well established.³⁴ Therefore, advocating for the systematic integration of this variable into the current allocation system requires much more robust data.

Overall, our work highlights 2 important lessons for the design of future research on improving LT allocation scoring. First, organ allocation models should explore the added benefit of ML methodology. The advantages of ML for individualized risk prediction can be leveraged for more accurate predictions of waitlist outcomes and perhaps better capture latent variables. We have highlighted that this is feasible through subjects with PSC. Second, data registries for organ transplants, such as the SRTR, can consider expanding their data collection to include more disease-specific variables. Both conventional statistical models and ML models greatly benefit from more granular data to make accurate predictions. We were only able to build a PSC-specific model using UHN data. A more thorough exploration of our research idea would be to build a prediction model that includes disease-specific variables for all causes of liver disease. For example, including cholangitis for PSC, level of alkaline phosphatase for primary biliary cirrhosis, or oxygen saturation of hepatopulmonary syndrome would allow us to compare the performance of an AI allocation model using disease-specific variables across heterogeneous causes of liver disease. This is currently not possible under the SRTR data collection policy. We hope that a future ML model for LT organ allocation would include all causes of liver disease as well as disease-specific variables, expanding generalizability as an allocation tool while maintaining accuracy for individualized risk prediction.

REFERENCES

1. Bowlus CL, Arrivé L, Bergquist A, et al. AASLD practice guidance on primary sclerosing cholangitis and cholangiocarcinoma. *Hepatology*. 2023;77:659–702.
2. Boonstra K, Weersma RK, van Erpecum KJ, et al; EpiPSC/PBC Study Group. Population-based epidemiology, malignancy risk, and outcome of primary sclerosing cholangitis. *Hepatology*. 2013;58:2045–2055.
3. Chazouilleres O, Beuers U, Bergquist A, et al. EASL clinical practice guidelines on sclerosing cholangitis. *J Hepatol*. 2022;77:761–806.
4. Kim WR, Mannalithara A, Heimbach JK, et al. MELD 3.0: the model for end-stage liver disease updated for the modern era. *Gastroenterology*. 2021;161:1887–1895.e4.
5. Kim WR, Biggins SW, Kremers WK, et al. Hyponatremia and mortality among patients on the liver-transplant waiting list. *N Engl J Med*. 2008;359:1018–1026.
6. Malinchoc M, Kamath PS, Gordon FD, et al. A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. *Hepatology*. 2000;31:864–871.
7. Goldberg DS, French B, Thomasson A, et al. Current trends in living donor liver transplantation for primary sclerosing cholangitis. *Transplantation*. 2011;91:1148–1152.
8. Pullen LC. *Liver Allocation for Rare Disease: Does the MELD Score Suffice?* Elsevier, 2020:3271–3272.
9. Satapathy SK, Jones OD, Vanatta JM, et al. Outcomes of liver transplant recipients with autoimmune liver disease using long-term dual immunosuppression regimen without corticosteroid. *Transplant Direct*. 2017;3:e178.
10. Graziadei IW, Wiesner RH, Marotta PJ, et al. Long-term results of patients undergoing liver transplantation for primary sclerosing cholangitis. *Hepatology*. 1999;30:1121–1127.
11. Singal AK, Gurusu P, Hmoud B, et al. Evolving frequency and outcomes of liver transplantation based on etiology of liver disease. *Transplantation*. 2013;95:755–760.
12. Leung KK, Kim A, Hansen BE, et al. The impact of primary liver disease and social determinants in a mixed donor liver transplant program: a single-center analysis. *Liver Transpl*. 2021;27:1733–1746.
13. Onofrio F, Zheng K, Xu C, et al. Living donor liver transplantation can address disparities in transplant access for patients with primary sclerosing cholangitis. *Hepatology Commun*. 2023;7:e0219.
14. Spann A, Yasodhara A, Kang J, et al. Applying machine learning in liver disease and transplantation: a comprehensive review. *Hepatology*. 2020;71:1093–1105.

15. Bhat M, Rabindranath M, Chara BS, et al. Artificial intelligence, machine learning, and deep learning in liver transplantation. *J Hepatol*. 2023;78:1216–1233.
16. Gottlieb N, Azhie A, Sharma D, et al. The promise of machine learning applications in solid organ transplantation. *NPJ Digital Med*. 2022;5:89.
17. Hernandez-Boussard T, Bozkurt S, Ioannidis JP, et al. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. *J Am Med Inform Assoc*. 2020;27:2011–2015.
18. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2:841–860.
19. Katzman JL, Shaham U, Cloninger A, et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol*. 2018;18:1–12.
20. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2:56–67.
21. Cheung AC, Patel H, Meza-Cardona J, et al. Factors that influence health-related quality of life in patients with primary sclerosing cholangitis. *Dig Dis Sci*. 2016;61:1692–1699.
22. van den Brand FF, van der Veen KS, de Boer YS, et al; Dutch Autoimmune Hepatitis Study Group. Increased mortality among patients with vs without cirrhosis and autoimmune hepatitis. *Clin Gastroenterol Hepatol*. 2019;17:940–947.e2.
23. Desai NM, Mange KC, Crawford MD, et al. Predicting outcome after liver transplantation: utility of the model for end-stage liver disease and a newly derived discrimination function1. *Transplantation*. 2004;77:99–106.
24. Hayashi PH, Forman L, Steinberg T, et al. Model for end-stage liver disease score does not predict patient or graft survival in living donor liver transplant recipients. *Liver Transpl*. 2003;9:737–740.
25. Broome U, Olsson R, Lööf L, et al. Natural history and prognostic factors in 305 Swedish patients with primary sclerosing cholangitis. *Gut*. 1996;38:610–615.
26. Farrant JM, Hayllar KM, Wilkinson ML, et al. Natural history and prognostic variables in primary sclerosing cholangitis. *Gastroenterology*. 1991;100:1710–1717.
27. Tischendorf JJ, Hecker H, Krüger M, et al. Characterization, outcome, and prognosis in 273 patients with primary sclerosing cholangitis: a single center study. *Am J Gastroenterol*. 2007;102:107–114.
28. de Vries EM, Wang J, Williamson KD, et al. A novel prognostic model for transplant-free survival in primary sclerosing cholangitis. *Gut*. 2018;67:1864–1869.
29. Goode EC, Clark AB, Mells GF, et al; UK-PSC Consortium. Factors associated with outcomes of patients with primary sclerosing cholangitis and development and validation of a risk scoring system. *Hepatology*. 2019;69:2120–2135.
30. Eaton JE, Vesterhus M, McCauley BM, et al. Primary sclerosing cholangitis risk estimate tool (PREsTo) predicts outcomes of the disease: a derivation and validation study using machine learning. *Hepatology*. 2020;71:214–224.
31. Al-Chalabi T, Portmann B, Bernal W, et al. Autoimmune hepatitis overlap syndromes: an evaluation of treatment response, long-term outcome and survival. *Aliment Pharmacol Ther*. 2008;28:209–220.
32. Zhang J, Qiu Y, He X, et al. Platelet-to-white blood cell ratio: a novel and promising prognostic marker for HBV-associated decompensated cirrhosis. *J Clin Lab Anal*. 2020;34:e23556.
33. Kim JH, Kim S-E, Song D-S, et al. Platelet-to-white blood cell ratio is associated with adverse outcomes in cirrhotic patients with acute deterioration. *J Clin Med*. 2022;11:2463.
34. Goldberg DS, Camp A, Martinez-Camacho A, et al. Risk of waitlist mortality in patients with primary sclerosing cholangitis and bacterial cholangitis. *Liver Transpl*. 2013;19:250–258.