Predicting Mortality in Liver Transplant Candidates

Jonathon Byrd, Sivaraman Balakrishnan, Xiaoqian Jiang, and Zachary C. Lipton

Abstract Donated livers are assigned to eligible matches among patients on the transplant list according to a sickest-first policy, which ranks patients by their score via the Model for End-stage Liver Disease (MELD). While the MELD score is indeed predictive of mortality on the transplant list, the score was fit with just three features, for a different task (outcomes from a shunt insertion procedure), and on a potentially un-representative cohort. These facts motivate us to investigate the MELD score, assessing its predictive performance compared to modern ML techniques and the fairness of the allocations vis-a-vis demographics such as gender and race. We demonstrate that assessing the quality of the MELD score is not straightforward: waitlist mortality is only observed for those patients who remain on the list (and don't receive transplants). Interestingly, we find that MELD performs comparably to a linear model fit on the same features and optimized directly to predict same-day mortality. Using a wider set of available covariates, gradient-boosted decision trees achieve .926 AUC (compared to .867 for MELD-Na). However, some of the additional covariates might be problematic, either from a standpoint of procedural fairness, or because they might expose the process to possible gaming due to manipulability by doctors.

J. Byrd (✉) · S. Balakrishnan · Z. C. Lipton
Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: jabyrd@cmu.edu

S. Balakrishnan
e-mail: siva@stat.cmu.edu

Z. C. Lipton
e-mail: zlipton@cmu.edu

X. Jiang
University of Texas Health Science Center at Houston, Houston, TX 77030, USA
e-mail: Xiaoqian.Jiang@uth.tmc.edu

A. Shaban-Nejad et al. (eds.), *Explainable AI in Healthcare and Medicine*, Studies in Computational Intelligence 914,
https://doi.org/10.1007/978-3-030-53352-6_31

1 Introduction

Machine learning algorithms are increasingly being used to drive allocative decisions in applications with potential for high social impact, such as allocating bank loans, ranking job applicants, or matching organs to patients in need of transplants. The use of machine learning systems in consequential domains such as recidivism prediction in criminal justice, or the aforementioned problems, raises concerns of how to audit the quality and equitability of algorithmic decisions. Currently, the ML-based tools used to drive these decisions are classifiers, and perhaps the most mature set of practical tools for analyzing such data-driven decisions are those used for evaluating classifiers. Typically, we observe sample covariates, X, on which we make predictions $\widehat{Y}$, evaluated against observed outcomes, Y. Possibly, we also observe some protected feature Z, used to assess the fairness of classification decisions. Model performance is evaluated by comparing metrics based on true/false positives/negatives such as precision and recall, or comparing metrics based on output score such as calibration by group or ranking metrics like area under the receiving operator characteristic (ROC) curve.

However, this static view of examples and labels paints an incomplete picture of the situation in several real-world tasks. Firstly, we only observe outcomes under the decision made for each candidate. Potential outcomes under alternative policies are unobserved. Secondly, we are concerned with comparing the outcomes of allocating resources among eligible candidates. But aggregation across time periods and locations pools together incomparable candidates. (It is not possible to send a resource obtained in 2013 back through time to a recipient in 2006). To asses the quality of allocative decisions, we must look at the outcomes of each individual decision among the eligible recipients of that resource.

As of November 12, 2019, 12, 965 patients with end-stage liver disease are listed on the Organ Procurement and Transplantation Network (OPTN) waiting list for liver transplantation. In 2018, over 12, 700 patients were added to the list, while only 8, 250 liver transplantations were performed, necessitating a decision policy to select patients to receive available organs. 609 patients died while waiting for an organ, while another 627 were removed from the list due to being too sick for the procedure. In the United States, donor livers are roughly allocated according to a *sickest-first* policy, where compatible transplant candidates are ranked according to disease severity. First, candidates incompatible with the given donor liver are filtered out according to blood type, size, and geography to lower the risks of graft failure. The remaining candidates are ranked according to their Model for End-Stage Liver Disease (MELD) score [18], a simple model calculated from three blood measurements/tests: total bilirubin (g/dL), serum creatinine (g/dL), and international normalized ratio (INR) of prothrombin time. Currently, the MELD-Na score [3] is used, which also incorporates serum sodium.

Although MELD was originally developed to predict post-treatment survival in the transjugular intrahepatic portosystemic shunt (TIPS) procedure [12], it was repurposed by the OPTN to rank liver transplant candidates, after being demonstrated

to be a capable general predictor of mortality in patients with chronic liver disease [9]. The MELD-based allocation system was immediately successful, leading to the first ever reduction in the number of waiting list candidates and a 15% reduction in mortality among those on the waiting list [8]. However, the model is not without its drawbacks. In addition to being fit on a different cohort for a different task, the model thresholds the log-transformed values of its three features at 1.0 to avoid negative values. This is problematic, as a large percentage waiting list candidates possess serum creatinine levels below this threshold, and values below this threshold can reflect very different levels of kidney function [17].

Furthermore, as expected from such a simple model, the correlation between MELD and outcome is not equally strong for all patients. For some patients, MELD may not accurately reflect the severity of their condition. Other patients become ineligible for transplant before their condition deteriorates to the point of producing the higher MELD scores needed to be prioritized for transplantation. For example, in patients with Hepatocellular Carcinoma (HCC), as their tumors become larger and more numerous, post-transplant regrowth in the new liver becomes likely. Such patients are granted MELD exception scores which replace their calculated MELD score when being ranked for organ allocation. Due to the high difference in MELD scores among different regions, most MELD exception scores are based on the median MELD at transplant time of recent transplant recipients in the same region. 18.9% of candidates on the waiting list on December 31, 2017 have active MELD exceptions, 13.3% of which are for HCC. While the final rule prescribes disease severity as the measure with which to prioritize transplant candidates, the MELD system also prioritizes patients who may soon develop high risks for graft failure or disease recurrence. However, this is done in a post-hoc manner via heuristics, which prompts the question: why not estimate both mortality risk and transplant ineligibility risk in a statistically principled way?

In this paper, we evaluate the performance of modern machine learning methods on the task of mortality prediction for waiting list candidates. We train and evaluate models on predicting both same-day and 3-month mortality. We find that gradient boosting ensembles outperform MELD and MELD-Na in terms of area under the ROC curve by a wide margin on both the same-day and 90-day prediction tasks—0.900 (gradient-boosting) vs 0.831 (MELD-Na) for same-day prediction and 0.926 vs 0.867 for 3-month prediction. Removing demographic features including race, gender, education, etc, as well as more subjective features including ascites, encephelopathy, and diagnosis, does not have a large effect on model performance. Both our model and MELD-Na slightly underestimate mortality in female patients as compared to male patients, but we find no similar trends when comparing scores across ethnicities.

2 Related Work

Since the implementation of the MELD-based liver allocation system in 2002, many analyses and validations of the model's performance have been performed, and many modifications to the original MELD formula have been proposed. Merion et al. and Bambha et al. examine prediction using differences in MELD scores updated over time to incorporate information regarding a patient's change in condition over time, rather than just MELD at waiting list registration [1, 13]. Many authors [3, 11] have shown that serum sodium levels are predictive of waiting list mortality, ultimately leading to the adoption of the MELD-Na score for ranking transplant candidates. Sharma et al. refit MELD coefficients using a time-dependent Cox model on liver transplant waiting list patients, and showed that their updated model better ranks patients by waiting list mortality [17]. Leise et al. use a generalized additive form of Cox regression with smoothing splines to propose new cutoff values for MELD features [11]. Myers et al. propose the 5-variable MELD model which incorporates serum sodium and albumin in addition to the three MELD features [14]. 5-variable MELD is also derived using Cox regression, and was found to outperform MELD in prediction of 3-month mortality.

While there have been many attempts to predict post-treatment outcomes in liver transplant patients, comparatively little work has been done in predicting waiting list survival. Cuchetti et al. train a multi-layer perceptron to predict 3-month mortality in a cohort of 188 patients using age, sex, treatment indication, and a set of 10 laboratory values as features [6]. Recently, classification trees trained using mixed-integer optimization techniques were shown to reduce combined waitlist deaths/removals and post-transplant deaths by 17.6% as compared to MELD in simulation [2]. Their model, termed, "Optimized Predictor of Mortality" (OPOM) also outperformed MELD in ranking waiting list patients for 3-month survival. OPOM is comprised of two models, one trained on non-HCC candidates, and the other on HCC candidates. They train on data from the OPTN Standard Transplant Analysis and Research (STAR) dataset, treating every patient check-in as an example, and removing examples for which the patients receive treatment within three months (although they also explore imputing observations for treatment-censored patients).

MELD and other proposed patient-scoring methods have largely been created by fitting Cox regression models to patient features at the time of transplant listing. These models are then validated primarily using area under the ROC curves (ROC AUC) on a separate holdout set of patients. There are several potential issues with this procedure: 1) It is not clear that the proportional hazards assumption should hold among patients with different liver transplant indications. 2) Using standard Cox regression ignores feature updates from check-ins after patients are added to the waiting list (although some papers use time-dependent models [17]). As we would like models to perform well on both new waiting list candidates as well as candidates remaining on the list, it is important to include post-listing observations when evaluating models. When this has been done [2], each check-in update is treated as an example. However, this approach weights our evaluation metric based on the number of measurements of

a patient. 3) Cox models assume uninformative censoring, which clearly does not hold in this application, because treatment is allocated according to patient features. Furthermore, choosing not to consider any patients treated within 3 months for model evaluation, biases metrics to deprioritize patients with more severe conditions, who tend to either quickly receive treatment or quickly die. This is especially problematic when the end goal is to identify patients with the highest mortality risk.

Much work has also been done in predicting graft failure following liver transplants, either as a function of the recipient, or the donor-recipient pair. The SOFT (Survival Outcomes Following Liver Transplantation) score identifies 19 factors as significant predictors of recipient mortality following transplantation, and provides a logistic regression model to estimate recipient mortality [16]. Delen et al. use a support vector machine with a Gaussian kernel, a multi-layer perceptron (MLP), and an M5-based regression tree to select features for a Cox survival model for post-transplant outcomes, demonstrating their method to be superior to traditional feature selection methods on this task [7]. More recently, machine learning methods including random forests and MLP's have been applied to predicting recipient mortality following transplantation [4, 10]. Perez-Ortiz et al. augment a dataset of liver transplant outcomes with recent unlabeled transplants and virtual donor-recipient pairs to combat class imbalance, and use a semi-supervised label propagation method to train support vector classifiers on the augmented data [15]. They then propose an organ allocation policy based on the model.

3 Dataset

Our data is composed of OPTN waiting list histories and Transplant Candidate Registration (TCR) form data from the STAR (Standard Transplant Analysis and Research) file for adult transplant candidates registered on the OPTN liver waiting list. Pre-2016 waiting list histories for candidates added to the waiting list after June 30, 2004 are divided into in-sample training, validation, and test sets using a random respective 50-25-25% split on individual patients. When deploying models to make real-world decisions, we are making decisions regarding future patients using data from past patients. Changes in society, patient care, and healthcare policy may cause to distribution shift among patients seen over time. For this reason, models are also evaluated on an out-of-sample test set of waiting list histories for a randomly selected 50% of patients added to the waiting list between January 1, 2016 and June 30, 2018.

Each day that a patient is on the waiting list (a patient-day) is treated as an example. As we wish to rank patient's chance of survival without treatment, patient-days on which the patient receives treatment or is removed from the waiting list for any reason other than death are excluded for the same-day mortality prediction task. For 3-month mortality prediction, all patient-days for which the patient is removed from the list for any non-death reason within 3 months are excluded, with the exception of patients removed from the list due to condition improvement such that they no longer require transplantation. These patient-days remain in the dataset as nega-

Table 1 Composition of train and test sets for same-day and 3-month prediction tasks

	Patient-days	Patients	Positive examples	Updates
Same-day mortality				
Training set	27.79M	61.0K	7886 (0.03%)	758K
In-sample test set	13.78M	30.5K	3943 (0.03%)	374K
Out-of-sample test set	3.02M	15.2K	991 (0.03%)	142K
3-month mortality				
Training Set	24.66M	61.0K	791K (2.6%)	513K
In-sample test set	12.23M	30.5K	397K (2.6%)	252K
Out-of-sample test set	32.45M	15.2K	90K (2.2%)	85K

tive examples. Table 1 breaks down the numbers of examples and patients in each dataset for the two prediction tasks. Demographic information is presented in Table 2. Figures 1, 2, and 3 examine MELD-Na scores for patients removed from the waiting list for different reasons.

Our models make use of 50 features, 31 of which are known at waiting list registration, while the remaining 19 are updated over time while a candidate remains on the list. Categorical features are encoded as dummy variables, and numerical features are standardized to have zero mean and unit variance in the training set. MELD and MELD-Na values are calculated using their respective formulas rather than read directly from the dataset. This results in a different value being used for 4% of observations, almost all of which have a difference of one point. The match MELD (value used for ranking patients that incorporates MELD exceptions) for inactive patient-days is set as the match MELD pre-inactivity. Missing values for numerical time-series features are forward-filled using the last known value for that feature. Missing values that cannot be forward-filled and missing values for numerical non-updated features are imputed using the feature median from the training set, and a corresponding missing value indicator feature is created for each numerical feature with missing values. When serum sodium values are missing and cannot be forward-filled, MELD-Na is calculated using the original MELD formula. Patient-days missing features to calculate lab MELD are removed from the training set if those values cannot be forward-filled. Numerical features are clipped to within four standard deviations from the mean. After pre-processing, the data has dimensionality 241.

Table 2 Demographic breakdown of train and validation sets. Percentages of patients and patient-days that fall into each category are shown as well as transplant rates, mortality rates, and rates of removal from the list for becoming too sick to transplant for each category. Age and diagnosis categories reflect values at waiting list registration time

Feature	Category	Patients	Patient-days	Transplant	Mortality	Removed: sick
	All Patients	100.0%	100.0%	55.6%	13.0%	10.6%
Age	18–34	6.0%	5.6%	57.5%	8.7%	5.6%
	35–49	20.2%	22.1%	56.0%	11.9%	7.4%
	50–64	60.9%	61.5%	56.0%	13.6%	11.1%
	65+	12.9%	10.8%	52.6%	13.8%	15.4%
Gender	Male	64.7%	62.8%	58.1%	12.3%	10.1%
	Female	35.3%	37.2%	51.0%	14.2%	11.4%
Ethnicity	White	70.6%	70.3%	55.9%	12.9%	10.2%
	Black	9.0%	7.5%	60.9%	11.5%	10.5%
	Hispanic	14.6%	16.5%	50.6%	15.1%	12.1%
	Asian	4.5%	4.5%	56.4%	9.7%	10.9%
	Other	1.3%	1.2%	54.7%	14.0%	11.1%
Diagnosis	Non-Chol. Cirrhosis	76.1%	79.4%	54.7%	13.4%	10.9%
	Chol. Liver Dis-ease/Cirr.	7.9%	8.6%	58.3%	12.0%	8.6%
	Biliary Artesia	0.3%	0.4%	61.1%	10.0%	6.1%
	Acute Hepatic Necrosis	4.7%	3.1%	54.4%	11.8%	10.4%
	Metabolic Diseases	2.2%	1.7%	68.2%	11.2%	6.7%
	Malignant Neoplasms	18.5%	12.6%	68.5%	6.2%	12.1%
	Other	1.9%	1.8%	61.6%	11.2%	7.1%

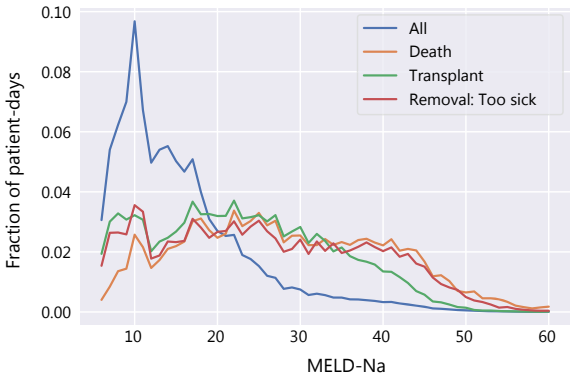

Fig. 1 Fraction of patient-days (y-axis) with different MELD-Na scores (x-axis) in the training and validation sets normalized by group. Lines for removal due to death, transplant, or too sick for transplant show MELD-Na scores at removal time

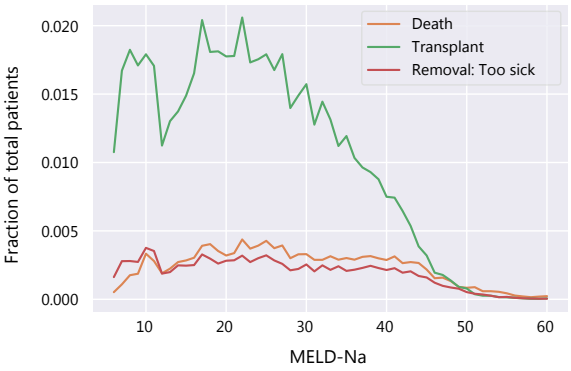

Fig. 2 MELD-Na scores (x-axis) of patients at waiting list removal due to death, transplant, or becoming too sick for transplant. Each point shows the fraction of total patients with corresponding outcome and MELD-Na score. Patients in the training and validation sets are shown

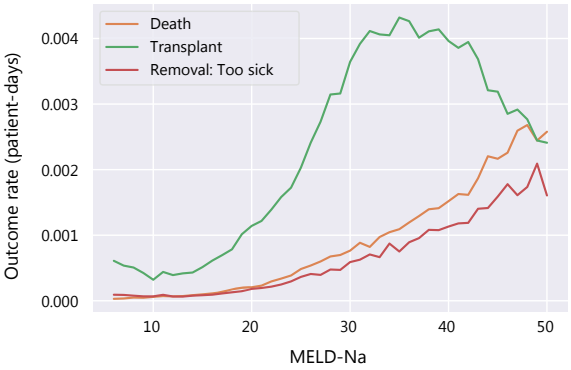

Fig. 3 Outcome rates among patient-days (y-axis) with different MELD-Na scores (x-axis) for removal due to death, transplant, or becoming too sick for transplant

4 Predicting Mortality

When allocating livers, we are much more concerned with giving livers to the patients who most need them, rather than accurately predicting mortality risks. Thus, we evaluate models by comparing their area under the ROC curve (ROC AUC, also termed c-statistic). MELD has traditionally been validated on the task of ranking patients by 3-month mortality. We also consider same-day mortality prediction, which corresponds to learning a hazard function. Under the proportional hazards assumption, both of these rankings would be equivalent. Training and evaluating on same-day mortality allows us to utilize data from patients treated within 3 months of listing, as well as measurements from patients within 3 months of transplantation date. Given that we are interested ranking the sickest compatible patients above others, and the current policy attempts to treat sicker patients sooner, it is important to include patients who receive transplants quickly in our datasets.

Many features to which we have access may not be appropriate to use as input for models used in the organ allocation process. Equitability concerns discourage the use of features such as race or education level, even if they were to be associated

with waiting list outcomes. While knowledge of gender may be useful in interpreting how serum creatinine values reflect kidney function, naively feeding this feature to models invites discrimination on a protected class. We present results for models trained without such features.

Additionally, many features are the result of subjective judgements from physicians. End-stage liver disease symptoms such as ascites and encephelopathy are not judged identically by every physician, and their subjectivity was a primary criticism of the Child-Turcotte-Pugh score [5]. Primary and secondary diagnoses can be especially dependent on the discretion of the physician when multiple indications for liver transplant are present. These features do not only raise concerns due to their variability, but also do to their manipulability. Knowledge of the allocation system may influence physicians when measuring attributes or making decisions that could influence the patient's transplantation ranking. We identify four features with higher measurement subjectivity, and present results from models trained excluding these features.

In total, we train models using four different feature sets. The first contains all available features. The second set excludes the following demographic features: citizenship, education, gender, whether or not patient works for income, ethnicity, blood type, and donor service area (age, height, and weight are kept as features). The third set further excludes the following features which we believe have a higher degree of measurement subjectivity: ascites, encephelopathy, diagnosis, and functional status. The final set includes only features used in MELD-Na: bilirubin, international normalized ratio of prothrombin time (INR), serum creatinine, serum sodium, and dialysis twice within prior week.

We fit logistic regression and gradient-boosting ensembles with decision trees as base classifiers on each feature set and prediction task. Hyperparameters were selected by iterative grid searches on the validation set. Logistic regression models use an L2-penalty of 1. Gradient-boosting ensembles are trained with a learning rate of 0.1. Ensembles of 2000 trees with max tree depth of 4, minimum of 60 examples required to create a split, and a minimum of 7 examples required to create a leaf, are trained for the first three feature sets. On the MELD-Na feature set, we use ensembles of 500 trees with max tree depth of 3, minimum of 6 examples required to create a split, and a minimum of 5 examples required to create a leaf.

Table 3 shows ROC AUC on the test set for the same-day mortality and 3-month mortality prediction tasks respectively. Results are shown for MELD, MELD-Na, and our models across the four feature sets. Results are given for both holdout test sets containing patients from different time periods. We also compare the equitability of scoring models by examining mortality rates of protected classes with similar model scores (Fig. 4).

Table 3 ROC AUC scores for ranking patient-days by same-day and 3-month mortality. The first number in each cell is the performance on the in-sample holdout set which includes patients added to the waiting list between July 8, 2004 and December 31, 2015. The second number is the performance on the out-of-sample holdout set which includes patients added to the waiting list between January 1, 2016 and June 30, 2018. Results are shown for ranking by MELD, MELD-Na, match MELD, and logistic regression and gradient boosting models trained on four different feature sets. Selected features refers to feature set excluding both non-demographic features and four additional features with greater measurement subjectivity. Same-day and 3-month versions of logistic regression and gradient boosting models refer to the target the model was trained to estimate

	All features	Non-demographic features	Selected features	MELD-Na features
Same-day mortality				
MELD	N/A	N/A	N/A	0.825, 0.791
MELD-Na	N/A	N/A	N/A	0.831, 0.793
Match MELD	N/A	N/A	N/A	0.750, 0.729
Logistic Regression (same-day)	0.888, 0.867	0.886, 0.864	0.876, 0.855	0.817, 0.782
Gradient Boosting (same-day)	**0.935**, **0.920**	0.931, 0.918	0.873, 0.857	0.793, 0.735
Logistic Regression (3-month)	0.881, 0.851	0.880, 0.849	0.872, 0.839	0.820, 0.774
Gradient Boosting (3-month)	0.902, 0.873	0.901, 0.873	0.894, 0.864	0.832, 0.796
3-month mortality				
MELD	N/A	N/A	N/A	0.715, 0.674
MELD-Na	N/A	N/A	N/A	0.730, 0.686
Match MELD	N/A	N/A	N/A	0.685, 0.651
Logistic Regression (same-day)	0.786, 0.756	0.786, 0.752	0.778, 0.745	0.700, 0.662
Gradient Boosting (same-day)	0.783, 0.767	0.781, 0.765	0.808, 0.781	0.731, 0.690
Logistic Regression (3-month)	0.820, 0.772	0.818, 0.770	0.809, 0.759	0.734, 0.687
Gradient Boosting (3-month)	**0.834, 0.800**	0.832, 0.798	0.827, 0.789	0.734, 0.696

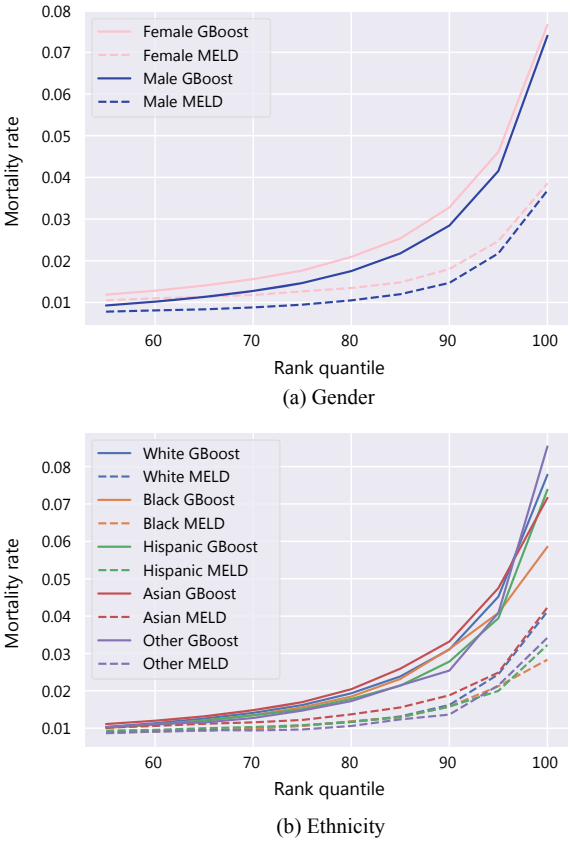

Fig. 4 Mortality rates (y-axis) for rank quantiles (x-axis) by gender (a) and ethnicity (b). The x-axis shows the quantile for patient days ranked by our model (solid lines) or by MELD-Na (dashed lines). The y-axis shows the mortality rates for patient days in the corresponding demographic group. Each bin is 5 percentiles wide, and we only show the top 50% rank days, because lower-ranked patient-days are unlikely to be selected for transplantation. The 'GBoost' model corresponds to the gradient-boosting model trained to predict same-day mortality using the non-demographic feature set

5 Discussion

Firstly, we find that both logistic regression and gradient boosting models trained on a larger feature set outperform MELD and MELD-Na by a wide margin in terms of ROC AUC for both same-day and 3-month mortality prediction tasks (Table 3). These results show that there is potential for a large margin of improvement in ranking patients by mortality risk for the current MELD-based system. However, we do find that when limiting ourselves to the feature set used by MELD-Na, it is hard to perform too much better, and MELD-Na outperforms linear classifiers limited to the MELD-Na feature set. Larger feature sets may provide more of a performance benefit than more complex models. Eliminating demographic features such as gender, race, and location, as well as features with a more measurement subjectivity such as ascites and encephelopathy, does not sacrifice much test performance. These models may still perform well when restricted to more practical feature sets.

Both our model and MELD-Na consistently underestimate mortality for female patients by a small margin (Fig. 4). We do not find many notable trends in scoring

ethnicity groups, but both models seem to slightly under-score Asian patients, except in the highest scoring quantile bin, where Black patients are over-scored by both models. The higher mortality rates among higher-scored patient-days from our model reflect the ability of our model to better select high-risk patient-days as compared to MELD-Na. Overall, the relative scoring between groups is similar between MELD-Na and our model.

Our results show the ability of modern machine learning techniques to more accurately rank patient-days by the existing metric of 90-day mortality, but is this the metric we wish to optimize in practice? Viewing the data in terms of patient-days rather than just patients with a static set of features at registration, or a series of check-in updates, more accurately reflects the task at hand, yet the merit of comparing patient-days occurring years apart is debatable, the distribution of test examples are biased by the allocation policy, and it is unclear what exactly this metric is estimating. While our methods outperform MELD on the waitlist mortality prediction task, we are ultimately interested in improving priority assignment within the matching system - a decision-making, not prediction task.

Furthermore, we wish to ask the following questions: is sickest first even the correct policy with which to allocate livers? Can we do better by directly optimizing waitlist mortality? In future work, we plan to address these questions, investigate the applicability of machine learning methods to estimate post-treatment outcomes for donor-recipient pairs, and explore the possibility of statistically-principled methods for liver allocation to achieve desired outcomes in a practical and equitable manner.

References

1. Bambha, K., Kim, W., Kremers, W., Therneau, T., Kamath, P., Wiesner, R., Thostenson, J., Benson, J., Dickson, E.: Predicting survival among patients listed for liver transplantation: an assessment of serial meld measurements. American Journal of Transplantation (2004)
2. Bertsimas, D., Kung, J., Trichakis, N., Wang, Y., Hirose, R., Vagefi, P.: Development and validation of an optimized prediction of mortality for candidates awaiting liver transplantation. Am. J. Transplantat. **19**, 1109–1118 (2018)
3. Biggins, S., Kim, W., Terrault, N., Saab, S., Balan, V., Schiano, T., Benson, J., Therneau, T., Kremers, W., Wiesner, R., Kamath, P., Klintmalm, G.: Evidence-based incorporation of serum sodium concentration into meld. Gastroenterology **130**, 1652–1660 (2006)
4. Chandra, V.: Graft survival prediction in liver transplantation using artificial neural networks. J. Health Med. Inform. **16**, 72–78 (2016)
5. Child, C.: The liver and portal hypertension. In: Annals of Internal Medicine (1964)
6. Cucchetti, A., Vivarelli, M., Heaton, N., Phillips, S., Piscaglia, F., Bolondi, L., La Barba, G., Foxton, M., Rela, M., O'Grady, J., Pinna, A.: Artificial neural network is superior to meld in predicting mortality of patients with end-stage liver disease. Gut (2007)
7. Delen, D., Oztekin, A., Kong, Z.: A machine learning-based approach to prognostic analysis of thoracic transplantations. Artif. Intell. med. **49**, 33–42 (2010)
8. Freeman, R., Wiesner, R., Edwards, E., Harper, A., Merion, R., Wolfe, R.: Results of the first year of the new liver allocation plan. Liver Transplant. **10**, 7–15 (2004)
9. Kamath, P.S., Wiesner, R.H., Malinchoc, M., Kremers, W.K., Therneau, T.M., Kosberg, C.L., D'Amico, G., Dickson, E.R., Kim, W.R.: A model to predict survival in patients with end-stage liver disease. Hepatology **33**, 464–470 (2001)

10. Lau, L., Kankanige, Y., Rubinstein, B., Jones, R., Christophi, C., Muralidharan, V., Bailey, J.: Machine-learning algorithms predict graft failure following liver transplantation. Transplantation **101**, 125–132 (2016)
11. Leise, M., Kim, W., Kremers, W., Larson, J., Benson, J., Therneau, T.: A revised model for end-stage liver disease optimizes prediction of mortality among patients awaiting liver transplantation. Gastroenterology **140**, 1952–1960 (2011)
12. Malinchoc, M., Gordon, F., Peine, C., Rank, J., Borg, P.: A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. Hepatology **31**, 864–871 (2000)
13. Merion, R., Wolfe, R., Dykstra, D., Leichtman, A., Gillespie, B., Held, P.: Longitudinal assessment of mortality risk among candidates for liver transplantation. Liver Transplant. **9**, 12–20 (2003)
14. Myers, R., Shaheen, A., Faris, P., Aspinall, A., Burak, K.: Revision of meld to include serum albumin improves prediction of mortality on the liver transplant waiting list. PLoS One **8**, e51926 (2013)
15. Pérez-Ortiz, M., Gutiérrez, P.A., Ayllón-Terán, M., Heaton, N., Ciria, R., Briceño, J., Martínez, C.: Synthetic semi-supervised learning in imbalanced domains: constructing a model for donor-recipient matching in liver transplantation. Knowl. Based Syst. **123**, 75–87 (2017)
16. Rana, A., Hardy, M., Halazun, K., Woodland, D., Ratner, L., Samstein, B., Guarrera, J., Brown, R., Emond, J.: Survival outcomes following liver transplantation (soft) score: a novel method to predict patient survival following liver transplantation. Am. J. Transplant. **8**, 2537–2546 (2008)
17. Sharma, P., Schaubel, D., Sima, C., Merion, R., Lok, A.: Re-weighting the model for end-stage liver disease score components. Gastroenterology **135**, 1575–1581 (2008)
18. Wiesner, R., Edwards, E., Freeman, R., Harper, A., Kim, R., Kamath, P., Kremers, W., Lake, J., Howard, T., Merion, R., Wolfe, R., Krom, R., Colombani, P., Cottingham, P., Dunn, S., Fung, J., Hanto, D., McDiarmid, S., Rabkin, J., Teperman, L., Turcotte, J., Wegman, L.: Model for end-stage liver disease (meld) and allocation of donor livers. Gastroenterology **124**, 91–96 (2003)